WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 1, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n1.06
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Point Cloud Semantic Segmentation Based on Rotation Invariance and Feature Aggregation

Xiujuan Liang *, Shiye Zhang

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China

*Corresponding Author: 212304020034@home.hpu.edu.cn

ABSTRACT

As a key task in 3D scene understanding, point cloud semantic segmentation has broad application prospects in fields such as autonomous driving and robot navigation. Existing point cloud segmentation methods suffer from insufficient local feature extraction and a lack of effective integration of global contextual information, leading to inaccurate recogni-tion and incomplete segmentation of categories with similar surface textures and geometric structures. In view of this, this paper proposes an improved point cloud segmentation method for RandLA-Net : (1) Local polar coordinate posi-tion encoding module is introduced to eliminate the impact of Z-axis rotation on feature learning; (2) Global information acquisition module composed of attention mechanisms is constructed to enhance the network's contextual perception ability; (3) Hybrid pooling mechanism is integrated to improve the extraction of local features. The proposed method is evaluated on the self-built HPU dataset and public datasets S3DIS and Toronto-3D. The results show that the improved network achieves mean intersection over union (mIoU) values of 90.7%, 71.2%, and 76.4% respectively, demonstrating improvements compared to other algorithms. The model exhibits excellent generalization and segmentation perfor-mance in different types of point cloud scenes.

KEYWORDS

3D point cloud; Point cloud semantic segmentation; Attention mechanism; Global information; Feature learning

1. INTRODUCTION

With the continuous development of sensor technologies such as LiDAR and stereo cameras, point cloud data has gradually become an important carrier of three-dimensional spatial information due to its high precision, high density, and rich semantic information [1]. It has shown immense potential in numerous fields. For example, it is used for environmental perception and path planning in autonomous driving [2], for precise object recognition and interaction in robotics [3], and for constructing 3D worlds in virtual reality [4]. Point cloud semantic segmentation can fully leverage the value of point cloud data in semantic parsing, aiming to assign corresponding semantic category labels to each point in the point cloud, thereby achieving fine-grained scene classification and recognition, and providing fundamental semantic support for scene understanding [5].

In recent years, the rapid development of deep learning technology has revolutionized point cloud semantic segmentation, with neural networks becoming the recognized mainstream solution in this field. This advancement can be attributed to the ability of these networks to learn complex features from data, enabling more accurate segmentation and classification of 3D objects. Deep learning-based point cloud segmentation methods are broadly categorized into three types: voxel-based methods [6–

9], multi-view based methods [10–13], and point-based methods. Voxel-based methods discretize irregular 3D point clouds into uniform cubic, cylindrical, or spherical grids. This approach allows for the preservation of essential structural information inherent in 3D data. By transforming irregular data into a structured format, voxel-based methods leverage traditional deep learning techniques, which often perform better with regular inputs. However, this benefit comes with significant drawbacks. The process requires handling and storing a vast number of voxels, which can lead to heightened memory consumption and computational overhead. The increased complexity of processing these large grids can limit the scalability of the method, particularly when dealing with highly detailed or extensive point cloud datasets. Multi-view based methods, also known as 2D projection methods, engage in projecting 3D objects from various angles to create multiple 2D views. Features are extracted from each view before being fused into a comprehensive global descriptor that supports accurate object recognition. Although this method enables the leveraging of established 2D convolutional networks, it poses a risk of losing critical geometric information during the projection phase. This loss can adversely affect the overall recognition accuracy, particularly in complex scenarios where the 3D spatial relationships are pivotal to understanding the object's shape and structure. Point-based methods represent a more direct approach, processing raw point clouds without converting them into structured formats like voxels or images. This direct handling helps avoid the information loss that may occur during conversion processes, allowing for the use of inherent point cloud characteristics. While PointNet and PointNet++ can directly process point cloud data, they have insufficient local geometric modeling capabilities [14, 15]. Given that convolution operations can efficiently capture local spatial features in regular data, researchers have explored applying traditional convolutional mechanisms to irregular, unordered point cloud data. Works such as PointCNN [16], PointConv [17], and KPConv define a "convolution-like" operation directly on raw point clouds to aggregate neighborhood information and generate features with local context awareness [18], significantly enhancing the expressive power of point cloud features. However, they still face the challenge of insufficiently capturing complex scene geometric structures. Subsequent research such as PointWeb [19], ShellNet [20], and RandLA-Net extract local features of point clouds through local adaptive aggregation [21], concentric spherical shells, and local feature aggregation, proving highly efficient in local feature aggregation, but still lacking in global context modeling.

To further deal with these problems, this paper proposes a point cloud segmentation method based on rotation invariance and feature aggregation. This method primarily consists of a Local Polar Representation (LPR) module, a Spatial and Strengthen Channel Attention Mechanism (SSAM) module, and a Mixed Pooling (MP) module. By effectively extracting global and local features of point clouds, it aims to better accomplish 3D point cloud segmentation tasks. The main contributions of this paper are as follows: 1) Integrating the LPR and MP modules, LPR precisely encodes the local geometric information of point clouds by calculating relative positions, distances, etc., and transforms it from the Cartesian coordinate system to the polar coordinate system to better capture geometric relationships under rotation. Mixed Pooling effectively enhances the efficiency of feature extraction while ensuring feature representation. 2) Proposing the SSAM module, which suppresses background noise in point clouds and models and integrates global information through a spatial attention mechanism and an enhanced channel attention mechanism. 3) Conducting experiments on real-world datasets, the results demonstrate that this method exhibits good generalization ability and segmentation performance in various types of point cloud scenarios.

2. PROPOSED METHODS

2.1. Folding Model

The baseline model is comprised of an encoder layer and a decoder layer. The encoder layer primarily consists of random sampling and local feature integration modules, while the decoder layer mainly performs upsampling operations. The point cloud semantic segmentation method proposed in this

paper primarily focuses on improving the local feature integration module within the encoder layer of the baseline model. The overall architecture of the model is illustrated in Figure 1. The entire model learns complex local semantic features by progressively increasing the receptive field size.

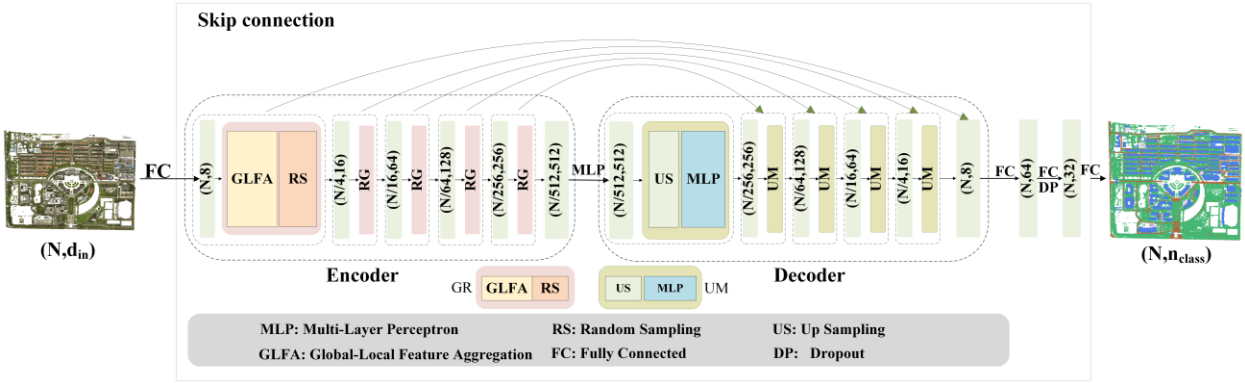


Figure 1. The structure of the model

Specifically, the model's input is a point cloud of size $N*d_{in}$, where N represents the total number of points and d_{in} is the feature dimension of each input point. Initially, the dimension of the raw point cloud data is elevated from (N,d_{in}) to $(N,8)$ through a fully connected layer. This dimension-enhanced semantic information then serves as the input to the global-local feature integration module. After the feature aggregation module progressively extracts both global information and local detailed features from the point cloud layer by layer, random sampling is applied to compress the point cloud. The decoder layer employs efficient upsampling operations to restore the number of points. A Multi-Layer Perceptron is utilized to adjust the point cloud feature dimensions [22], and skip connections are implemented to concatenate with intermediate feature maps from the encoder layer, thereby achieving cross-level fusion of feature information. Finally, Finally, after the feature dimensions are adjusted by two fully connected layers, a dropout layer is applied for regularization to enhance the model's generalization capability. Subsequently, another FC layer performs classification, ultimately outputting the predicted semantic class labels for the point cloud, with its final feature dimension being (N,n_{class}), where n_{class} is the number of predefined semantic categories.

2.2. Global-Local Feature Aggregation Module

The Global-Local Feature Aggregation Module is located within the encoder layer, following the random sampling operation. The module's architecture is illustrated in Figure 2. This module integrates sub-modules such as Local Polar Coordinate Position Encoding (LPR), a Spatial and Channel-Enhanced Global Feature Integration Module, and Hybrid Pooling, all connected via residual structures. Specifically, the LPR module is employed to learn the local geometric structure of the point cloud, achieving Z-axis rotation-invariant representation for the same object and significantly reducing geometric information loss caused by random sampling. Subsequently, the features re-encoded by the LPR module are concatenated with the original spatial features, and global contextual information is acquired through the Global Feature Integration Module. The Hybrid Pooling (MP) module then adaptively preserves salient local features, enhancing the model's ability to perceive local details. To further enhance the model's feature expression capability, residual modules are stacked to enlarge the receptive field for feature extraction, thereby strengthening the model's capacity to capture features at different scales. Finally, the enhanced global and local features are fused to generate the ultimate feature representation.

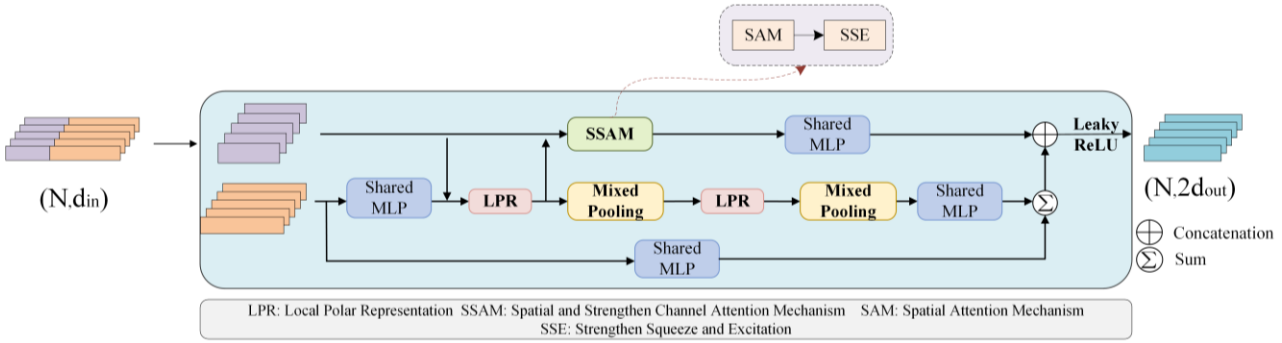


Figure 2. Global-Local Feature Aggregation Structure

2.2.1. Local Polar Coordinate Position Encoding Module

In point cloud segmentation tasks across various categories of scenes, objects of the same class often exhibit significant directional variability due to natural or anthropogenic factors. This directional variability poses challenges for traditional feature representations based on Cartesian coordinates, as they fail to provide rotation-invariant features. Specifically, these features do not remain consistent when an object is rotated in space, particularly within the horizontal plane. The lack of rotation invariance can lead to a marked decrease in the robustness of the model during feature extraction, subsequently impacting segmentation accuracy. To effectively address this challenge, this study introduces a Local Polar Coordinate Position Encoding module, which transforms the relative positional information of point clouds into a polar coordinate representation [23]. This transformation facilitates the generation of rotation-invariant features, significantly enhancing the model's ability to perceive objects from different orientations. The structure of this module is illustrated in Figure 3.

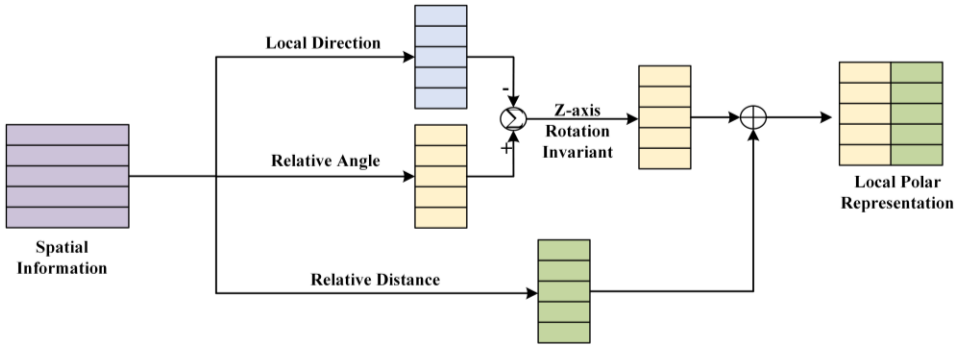


Figure 3. Local Polar Coordinate Position Encoding Structure

This module takes the spatial geometric information of all points as input. Before calculating the local direction, relative angles, and relative distances, it requires the aggregation of neighboring points $\{p_i^1, \ldots p_i^k, \ldots p_i^K\}$ for each central point p_i using the K-NN (K nearest neighbors) algorithm. The calculations for the Euclidean distance (dis_i^k), azimuth angle (φ_i^k), and elevation angle (θ_i^k) of each neighboring point p_i in polar coordinates are as follows:

$$dis_i^k = \sqrt{x_i^{k^2} + y_i^{k^2} + z_i^{k^2}} \tag{1}$$

$$\varphi_i^k = \arctan\left(\frac{y_i^k}{x_i^k}\right) \tag{2}$$

$$\theta_i^k = \arctan\left(\frac{z_i^k}{\sqrt{x_i^{k^2} + y_i^{k^2}}}\right) \tag{3}$$

In the equations, (x_i^k, y_i^k, z_i^k) denotes the coordinates of p_i in the Cartesian coordinate system.

To compute the azimuth angle α_i and elevation angle β_i for point p_i relative to the average position of its neighborhood, updates are performed through equations (4) and (5), yielding the azimuth angle difference $\varphi_i^{k'}$ and elevation angle difference $\theta_i^{k'}$.

$$\varphi_i^{k'} = \varphi_i^k - \alpha_i \tag{4}$$

$$\theta_i^{k'} = \theta_i^k - \beta_i \tag{5}$$

Finally, the angle's magnitude is adjusted based on the relative distance in the polar coordinate system, facilitating the expansion of spatial features. This process refines the relationship between point k and its neighboring spatial features.

2.2.2. Global Feature Integration Module

Local feature extraction often struggles to sufficiently capture the long-range dependencies between points, leading to poor segmentation results in complex scenes. To address this issue, we propose a Global Feature Integration module composed of a spatial attention mechanism and an enhanced channel attention mechanism. This module aims to extract global information and improve the model's segmentation performance in complex scenarios.

When noise is present in the point cloud data, it can interfere with the extraction of effective features. The introduction of a Spatial Attention Mechanism (SAM) can enhance the response of key regions while suppressing background interference. Its structure is illustrated in Figure 4.

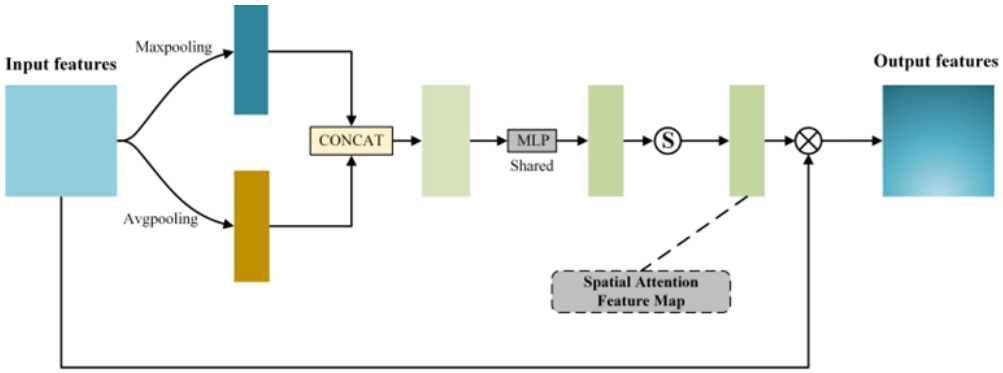


Figure 4. Spatial Attention Structure

The spatial information of the original point cloud and the results obtained from the segmentation stage are combined as input $F^{N\times D}$, where N denotes the number of point clouds and D represents the spatial dimensionality. The spatial dimensionality is utilized to maximize the spatial information at each point through a pooling operation, obtaining features $F_{Max}^{N\times 1}$ and $F_{Avg}^{N\times 1}$. These two sets of features will be concatenated, and the concatenated features are then passed into a Multi-Layer Perceptron for further processing. Finally, the output is processed through a Sigmoid activation function to obtain the spatial attention weights $F_{SA}^{N\times 1}$, expressed as follows:

$$F_{SA}^{N\times 1} = \sigma\left(M\left(C\left(F_{Max}^{N\times 1}, F_{Avg}^{N\times 1}\right)\right)\right) \tag{6}$$

In the equation, C represents a concatenation operation, M indicates a shared-weight MLP layer, and σ denotes the Sigmoid activation function.

By using equation (7), the initial point features $F^{N\times D}$ are combined with the spatial attention weights $F_{SA}^{N\times 1}$, resulting in the final spatial feature enhancement of the elements. The output $F_{SAout}^{N\times D}$ is expressed as $F^{N\times D} \otimes F_{SA}^{N\times 1}$, where $\otimes$ represents the element-wise multiplication.

$$F_{SAout}^{N\times D} = F_{SA}^{N\times 1} \otimes F^{N\times D} \tag{7}$$

The different channel features of point clouds exhibit variability in their expressive capabilities. By combining the channel attention module with a squeeze and excitation mechanism, we obtain an improved enhanced channel attention module. This module can effectively suppress noise in spatial key regions while simultaneously strengthening the important channel features. The structure of the Enhanced Channel Attention (Strengthen Squeeze and Excitation, SSE) is illustrated in Figure 5.

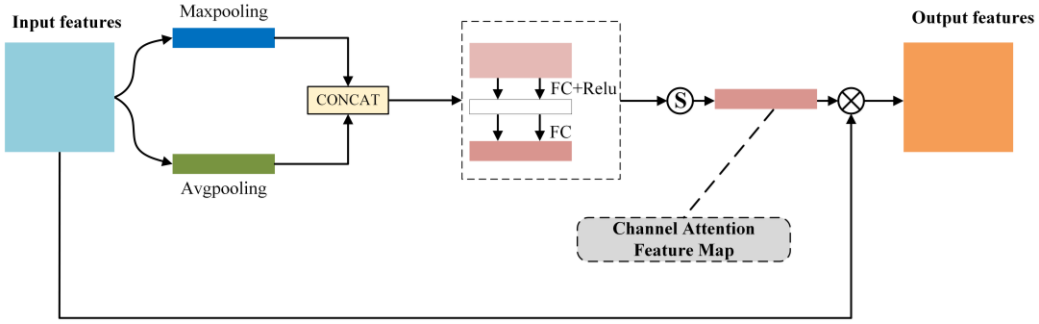


Figure 5. Enhanced Channel Attention Structure

Specifically, let $F \in R^{N\times D}$ be the input, where N denotes the number of point clouds, and D reflects the dimensionality of the spatial features. The channel attention module aims to learn the relationships between these channel dimensions, allowing for adjustments based on their dependencies.

In the enhanced channel attention module, we can obtain the max pooling channel features $F_{Max}^{1\times D}$ and the average pooling channel features $F_{Avg}^{1\times D}$. Subsequently, the max pooling features and the average pooling features are concatenated. The concatenated features are then subjected to a shared two-layer MLP for nonlinear transformation. The first layer compresses the channel dimensionality and uses the ReLU activation function, while the second layer restores the channel dimensionality.

After processing through the Sigmoid activation function, the features of each channel are mapped to the range of (0, 1), resulting in the channel attention weights, which contain the weight information for each channel, as shown in equation (8). This channel attention feature map is then expanded to match the size of the intermediate features $F \in R^{N\times D}$. Finally, by performing element-wise multiplication, we obtain the final output features $F_{CAout}^{N\times D}$ that incorporate the intermediate features, which can be expressed as follows:

$$F_{CA}^{1\times D} = \sigma\left(W_2\delta\left(W_1\left(C\left(F_{Max}^{1\times D}, F_{Avg}^{1\times D}\right)\right)\right)\right) \tag{8}$$

$$F_{CAout}^{N\times D} = F_{CA}^{1\times D} \otimes F^{N\times D} \tag{9}$$

In the equation, C denotes the concatenation operation, W_1 represents the channel compression operation, δ signifies the ReLU activation function, W_2 indicates the channel restoration operation, σ refers to the Sigmoid activation function, $\otimes$ represents the element-wise multiplication

2.2.3. Hybrid Pooling

In point cloud processing tasks, attention pooling is utilized to aggregate local features, thereby highlighting key information. However, the weights assigned in this process can be easily affected by noise and other non-critical features, resulting in a decrease in the accuracy and effectiveness of feature extraction. To address this issue, this paper proposes a hybrid pooling method that combines attention pooling with max pooling to merge point features into a single feature value vector. Max pooling is employed to extract key point features, while attention pooling is used to differentiate important neighboring point features. The structure of the hybrid pooling approach is illustrated in Figure 6.

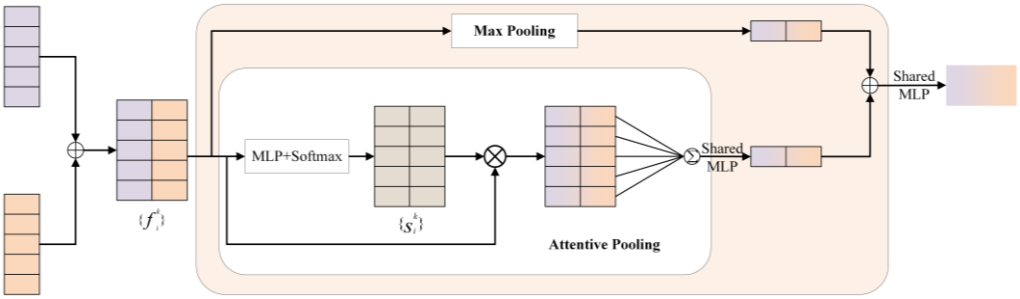


Figure 6. Hybrid Pooling Structure

The features obtained from the spatial information through the local polar coordinate positional encoding module, combined with the point cloud's k-nearest neighbor features, serve as the input for the hybrid pooling. The feature vectors obtained from the max pooling branch and the attention pooling branch are concatenated, and then processed through an MLP layer to adjust the channel dimensions, resulting in the aggregated features. The calculation for max pooling is depicted in equation (10), where p_i^k denotes the k-nearest neighboring points for each point, and $f\left(p_i^k\right)$ represents the feature vectors of all neighboring points.

After passing through the shared MLP layer and applying the Softmax activation, the attention pooling yields an attention weight score. The input features are then weighted and summed using these attention scores to obtain the feature vector from attention pooling, as shown in equation (11), where $g(\bullet)$ represents the learned attention score.

Finally, the results from max pooling and attention pooling are concatenated and processed through a shared MLP layer to produce the final aggregated features, as expressed in equation (12), where C denotes the concatenation operation.

$$F_{\max} = MaxPool\left\{f\left(p_i^k\right)\right\} \tag{10}$$

$$F_{att} = \sum\left\{f\left(p_i^k\right) * g\left(f\left(p_i^k\right), W\right)\right\} \tag{11}$$

$$F_{fus} = C\left(F_{\max}, F_{att}\right) \tag{12}$$

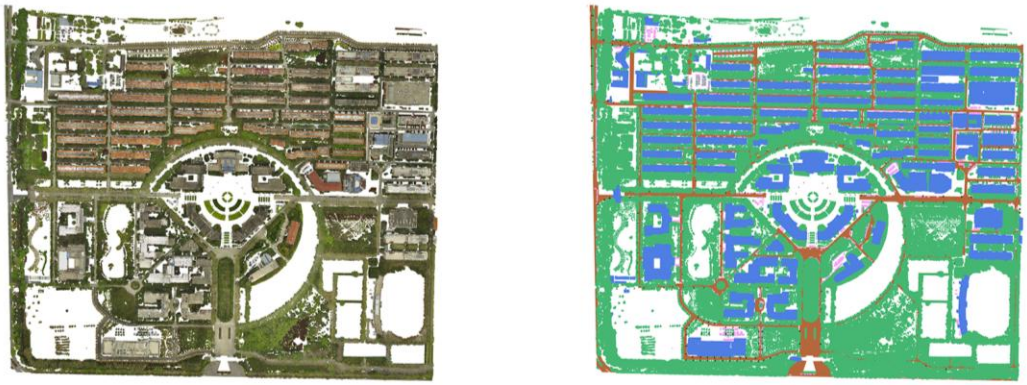
3. ANALYSIS AND DISCUSSION

3.1. Dataset and Environment Setup

This paper evaluates the effectiveness and robustness of the model using the self-built HPU dataset, the S3DIS dataset, and the Toronto-3D dataset. An introduction to the datasets is provided in Table 1, with the self-built HPU dataset illustrated in Figure 7.

Table 1. Dataset Introduction

Attribute Information	Self-built HPU	S3DIS [24]	Toronto-3D [25]
Number of Points	Approximately 280 million	Approximately 690 million	Over 78 million
Point Cloud	3D coordinates, color	3D coordinates, color	3D coordinates, color, intensity, GPS time, scanning angle
Scene Tpye	Campus scene	Indoor scene	Outdoor urban road scene
Acquisition Method	LiDAR	Laser scanner	Mobile LiDAR
Semantic Elements	Classified into 5 categories: buildings, vehicles, street lights, roads, and vegetation based on current public outdoor point cloud datasets and the characteristics of the scene	Classified into 13 categories: ceiling, floor, wall, beam, column, window, door, table, chair, sofa, bookshelf, board, and others	Classified into 8 categories: road, road markings, natural trees, buildings, utility lines, telephone poles, vehicles, and fences
Dataset Split	Divided into training, validation, and test sets in a 6:2:2 ratio	6 regions used for six-fold cross-validation	4 regions, with Region 2 as the test set



(a) image (b) label

Figure 7. HPU Dataset

The experiments were conducted on an Ubuntu 18.04 operating system, utilizing an NVIDIA RTX A6000 GPU with 48GB of memory, and the model was built using the TensorFlow-GPU 2.6 framework. The loss function employed is the weighted cross-entropy loss, and the optimizer used is Adam. The training consists of 100 iterations, with a batch input size of 4.

For the HPU dataset, the initial learning rate is set to 0.001, with the number of points in the input scene being 65,536. For the S3DIS dataset, the initial learning rate is set to 0.01, with the number of points in the input scene being 40,960. In the Toronto-3D dataset, the initial learning rate is also set to 0.01, with the number of points in the input scene being 65,536.

3.2. Experimental Results and Analysis

Representative models were selected to evaluate the proposed HPU dataset, and robustness tests were conducted on the S3DIS and Toronto-3D datasets. The experimental results are represented using

Overall Accuracy (OA), mean Accuracy (mAcc), Mean Intersection Over Union (mIoU), and Intersection Over Union (IoU).

Tables 2, 3, and 4 present the experimental results of the proposed model compared to representative models on the HPU dataset, S3DIS dataset, and Toronto-3D dataset, respectively. From Table 2, it can be observed that RandLA-Net outperforms several other models in terms of the mIoU metric, with significantly better segmentation results for buildings. Compared to the RandLA-Net model, the proposed model shows improvements of 4%, 3.7%, and 5% in OA, mAcc, and mIoU, respectively, achieving the best segmentation results for vehicles, roads, and vegetation.From

Table 3, it is evident that the proposed model improves the mIoU metric by 1.2% compared to RandLA-Net, with enhancements in both OA and mAcc metrics. Moreover, compared to other publicly available point cloud segmentation models, the proposed model achieves the best segmentation results for categories such as beams, windows, sofas, bookshelves, and boards.Table 4 indicates that the proposed model also improves the mIoU metric by 2.6% relative to RandLA-Net, with significantly better segmentation results for natural trees and buildings. In comparison with the RandLA-Net model, the proposed model achieves the best segmentation results for categories like roads, buildings, and vehicles, and demonstrates commendable segmentation performance for natural trees, utility lines, telephone poles, and fences.

Table 2. Evaluation results on the HPU dataset (Unit: %)

Method	OA	mAcc	mIoU	IoU				
				car	light	road	veg	build
PointNet++ [15]	95.8	72.9	66.9	61.5	8.0	80.0	97.9	87.2
DLA-Net [26]	96.7	93.6	84.7	95.3	51.8	86.5	97.7	92.4
BAAF-Net [27]	94.8	94.5	85.6	94.9	67.7	79.1	93.5	92.8
RandLA-Net [21]	94.8	93.1	85.7	95.5	65.7	79.3	89.2	98.9
Ours	98.8	96.8	90.7	95.7	66.0	95.5	97.7	98.8

Table 3. Evaluation results on the S3DIS dataset (Unit: %)

Method	OA	mAcc	mIoU	Iou												
				ceil.	floor	wall	beam	col.	wind.	door	table	chair	sofa	book.	board	clut.
PointNet [14]	78.6	66.2	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
SPGraph [28]	86.4	73.0	62.1	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
PointWeb [19]	87.3	76.2	66.7	93.5	94.2	80.8	52.4	41.3	64.9	68.1	71.4	67.1	50.3	62.7	62.2	58.5
ShellNet [18]	87.1	—	66.8	90.2	93.6	79.9	60.4	44.1	64.9	52.9	71.6	84.7	53.8	64.6	48.6	59.4
DLA-Net [26]	86.0	79.9	65.8	91.3	96.8	75.6	59.0	48.7	56.8	60.1	71.8	81.0	53.8	63.8	41.2	55.5
RandLA-Net [21]	88.0	82.0	70.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
Ours	88.2	82.1	71.2	93.2	96.0	80.5	64.4	48.0	65.2	69.1	71.2	80.3	66.6	64.8	66.3	59.8

Table 4. Evaluation results on the Toronto3D dataset (Unit: %)

Method	OA	mIoU	Iou							
			road	rd mrk.	natural	building	util. line	pole	car	fence
PointNet++ [15]	92.6	59.5	92.9	0.0	86.1	82.2	60.9	62.8	76.4	14.4
DGCNN [29]	94.2	61.7	93.9	0.0	91.3	80.4	62.4	62.3	88.3	15.8
MS-PCNN [30]	90.0	65.9	93.8	3.8	93.5	82.6	67.8	71.9	91.1	22.5
DLA-Net [26]	93.3	76.1	90.9	32.4	96.0	92.3	87.1	75.8	89.9	44.2
BAAF-Net [27]	94.2	70.8	89.1	6.3	96.3	93.2	86.1	82.1	86.6	29.9
RandLA-Net [21]	94.9	73.8	93.6	17.1	96.1	92.2	86.4	79.2	87.9	37.8
Ours	93.8	76.4	91.6	29.4	96.8	93.3	86.7	79.8	90.8	42.9

The experiment visualized a portion of scene data from the three datasets, as illustrated in Figures 8, 9, and 10. In each figure, from left to right, the images represent the original point cloud, the ground truth labels, the segmentation results of the baseline model, and the segmentation results of the proposed model, respectively. From Figures 8 and 10, it can be observed that the baseline model achieves relatively good segmentation results for large-scale categories such as buildings. However, due to its lack of global context information capture and the inability to extract detailed local features, it tends to misclassify when segmenting different types of objects with similar local features, such as roads and low vegetation in Figure. In Figure, when processing categories such as walls, beams, roofs, doors, windows, and boards that are on the same plane, incomplete and inaccurate segmentation results are observed for categories like boards, walls, and bookshelves. Furthermore, in Figure 10, while dealing with fine linear structures such as road markings and utility lines, there are instances of disconnection, fragmentation, and even loss of segmentation.

The proposed model effectively models global context information, enhancing its ability to recognize classes with long-range dependencies, thereby accurately segmenting most categories. This improvement is evident in Figures 8 and 9, where the model can accurately segment ground objects with similar structures. The segmentation results for large-scale objects such as walls and complex objects like bookshelves are more complete, effectively alleviating misclassification issues. By modeling the Z-axis rotational invariance of the relative positions of the points, the segmentation accuracy for directional objects, such as sofas and chairs, is improved. In the mixed scene of sparse and dense point clouds shown in Figure 10, the proposed method results in more continuous and complete segmentation of road markings, demonstrating a notable improvement.

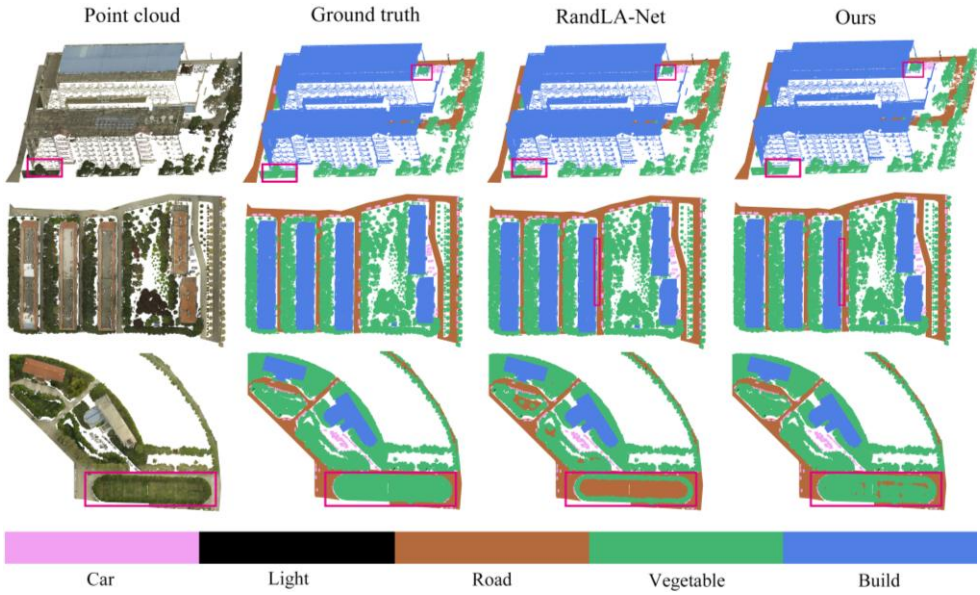


Figure 8. Semantic segmentation results of the HPU dataset

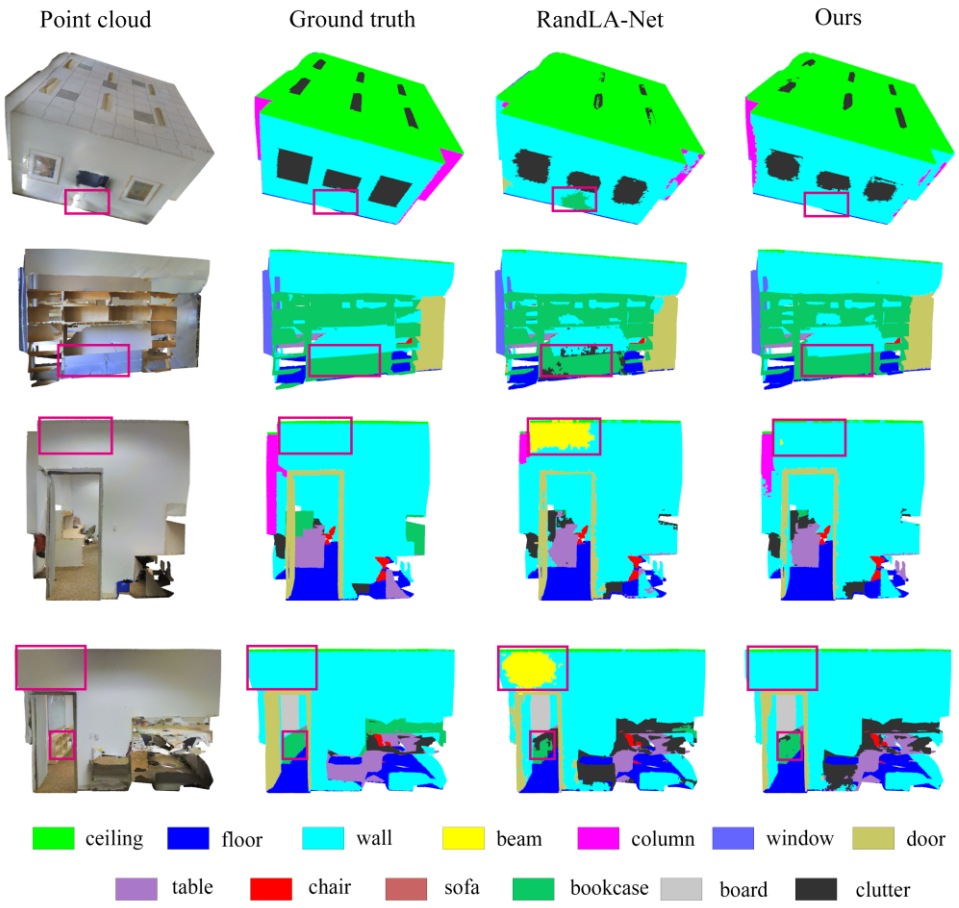


Figure 9. Semantic segmentation results of the S3DIS dataset (Area5)

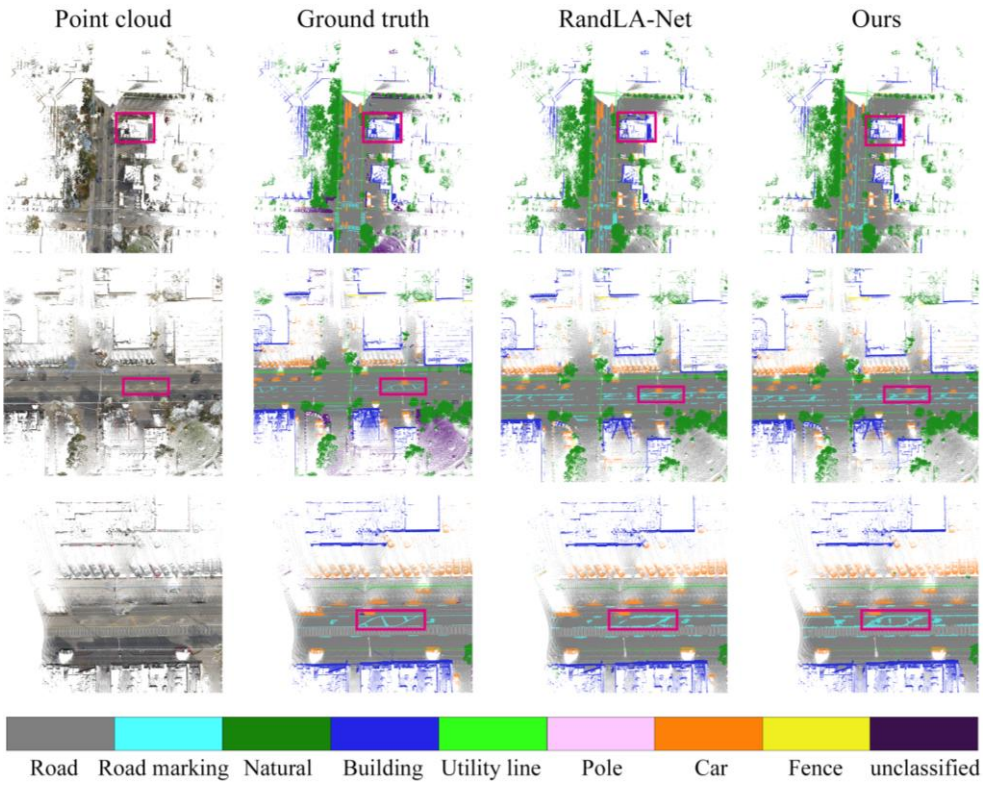


Figure 10. Semantic segmentation results of the Toronto3D dataset

3.3. Ablation Experiments

To explore the contributions of different modules in the proposed model, ablation experiments were conducted on the Position Encoding Module (LPR), the Mixed Pooling Module (MP), and the Global Feature Integration Module (SSAM). The following variations were tested: 1) Baseline model results. 2) Replacement of the original position encoding module with the local polar coordinates position encoding module. 3) Replacement of the original attention pooling module with the mixed pooling module. 4) Addition of the global feature integration module. 5) Replacement of the position encoding module and addition of the global feature integration module. 6) Replacement of the position encoding module and attention pooling module, and addition of the global feature integration module. The above experiments were conducted on the HPU, S3DIS (Area 5), and Toronto-3D point cloud datasets. The experimental results are presented in Table 5.

Table 5. Ablation study results on different datasets(mIoU)

Method	Dataset		
	HPU	S3DIS (Area5)	Toronto-3D
Baseline	85.7	62.5	73.8
LPR	89.3	62.6	72.1
MP	87.8	63.2	73.9
SSAM	89.9	63.3	73.0
LPR+SSAM	90.4	63.4	74.3
LPR+SSAM+MP	90.7	63.5	76.4

From Table 5, it can be seen that by introducing the local polar coordinates position encoding module to model the local spatial features of the point cloud, the model can learn characteristics that are invariant to Z-axis rotation. This enables a more comprehensive learning of the geometric structure of the same ground object, achieving mean Intersection Over Union (mIoU) scores of 89.3%, 62.6%, and 72.1% on the three datasets, respectively. Building upon this, after adding the global feature integration module, the model's mIoU performance improves across the three datasets, indicating that the model effectively integrates the global context information of the point cloud and enhances its capacity for global feature representation of the scene. Lastly, to enhance the model's ability to perceive details, the performance reached its optimal level after incorporating the mixed pooling module.

The global feature integration module is formed by combining the spatial attention module and the enhanced channel attention module. If the two modules use different connection methods, there may be slight differences in the experimental results. To investigate this, experiments were conducted on the HPU dataset with different connection methods. The two modules were connected in both series and parallel configurations. CM1 and CM2 represent the series connection, where they are concatenated in different orders, while CM3 represents the parallel connection, where the SAM and SSE modules are connected simultaneously. We conducted experiments to evaluate their impact on semantic segmentation performance, with the specific connection methods illustrated in Figure 11.

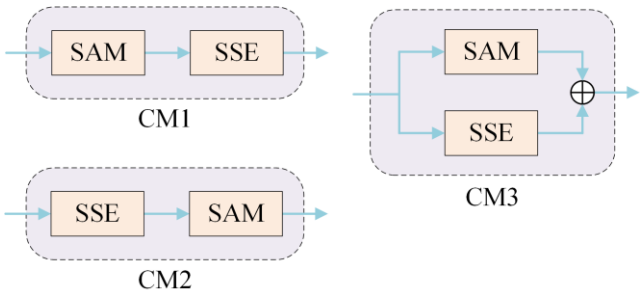


Figure 11. Different Connection Methods

Table 6. Ablation Results of Different Connection Methods

Different Connection	mIoU/%
CM3	89.92
CM2	90.30
CM1	90.74

From Table 6, it can be observed that the series connection method CM1 achieves the highest mean Intersection Over Union (mIoU), followed by CM2, while CM3 exhibits the lowest accuracy. This experimental result indicates that, in this specific semantic segmentation model, the strong dependence on the spatial position information of the point cloud allows the series connection method CM1 to yield better segmentation results. In practical applications, the choice of an appropriate connection method should be weighed and selected based on the specific model structure and task requirements.

4. DISCUSSION

The proposed point cloud semantic segmentation model, based on rotation invariance and feature aggregation, effectively enhances multi-scale feature representation and improves the model's capacity to perceive complex geometric structures through the integration of local polar coordinate encoding, global feature fusion, and hybrid pooling mechanisms. Experimental evaluations demonstrate superior performance across diverse datasets, particularly in addressing the segmentation accuracy and completeness for geometrically similar categories. Nevertheless, certain challenges remain to be addressed.The integration of spatial and enhanced channel attention mechanisms strengthens the model's ability to capture global contextual information, segmentation performance still declines under highly complex scenarios such as sparse local point clouds, noise interference, and fine-grained category distinctions. This indicates that the current approach to balancing local geometric detail extraction and global semantic integration requires further refinement. Future work may explore the incorporation of richer geometric priors or multi-scale multimodal information to improve robustness and generalization.The rotation-invariant characteristic of the model successfully mitigates the impact of spatial rotations on feature extraction, its stability and applicability under complex rotational transformations and multi-angle 3D variations require more comprehensive empirical validation. Subsequent studies should investigate more generalized geometric transformation invariance theories to bolster adaptability in dynamic real-world environments.

Computational efficiency and resource demands remain practical considerations. Although hybrid pooling contributes to improved feature extraction efficiency, the reliance on attention mechanisms and deep multi-layer fusion structures entails substantial computational overhead, potentially limiting deployment in resource-constrained or real-time scenarios. Therefore, research into model compression and inference acceleration is equally critical.

In summary, this study presents a promising framework that effectively integrates local and global information to advance segmentation accuracy. Future research directions should focus on enhancing the model's adaptability, operational efficiency, and explainability to better address the challenges posed by increasingly complex three-dimensional vision tasks.

5. CONCLUSIONS

Based on the above research,the following conclusions are derived: 1) Introduces a deep learning-based point cloud semantic segmentation method that leverages rotation-invariant feature encoding and multi-mechanism feature aggregation to achieve robust local geometric understanding and global contextual integration. By incorporating a local polar coordinate position encoding module, spatial

and channel-enhanced attention mechanisms, and a hybrid pooling strategy, the proposed model substantially improves segmentation accuracy and generalization. 2) Extensive experiments on a self-constructed HPU dataset as well as the public S3DIS and Toronto-3D datasets demonstrate that the proposed method outperforms existing state-of-the-art algorithms, delivering consistent improvements in multiple metrics across diverse semantic categories and scene types.The rotation-invariant encoding effectively addresses feature distortion caused by spatial rotations, the global feature integration module enhances long-range dependency modeling, and the hybrid pooling module preserves key local details. Ablation studies confirm the independent and synergistic contributions of these components, validating the soundness of the design. 3) Despite these advances, challenges remain in addressing extremely complex and dynamic scenarios, particularly concerning fine-grained feature extraction and computational efficiency. Future work will focus on integrating multimodal data and optimizing network architectures to enhance detailed feature capture, while pursuing lightweight designs suitable for practical deployment.

REFERENCES

[1] Ding, Z.; Sun, Y.; Xu, S.; Pan, Y.; Peng, Y.; Mao, Z. Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing. Robotics 2023, 12, 100, doi:10.3390/robotics12040100.

[2] Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. IEEE Access 2020, 8, 58443–58469, doi:10.1109/ACCESS.2020.2983149.

[3] Ning, D.; Huang, S. L-PCM: Localization and Point Cloud Registration-Based Method for Pose Calibration of Mobile Robots. Information 2024, 15, 269, doi:10.3390/info15050269.

[4] Lyu, B.; Wang, Y. Immersive Visualization of 3D Subsurface Ground Model Developed from Sparse Boreholes Using Virtual Reality (VR). Underground Space 2024, 17, 188–206, doi:10.1016/j.undsp.2023.11.004.

[5] Khairnar, S.; Thepade, S.D.; Kolekar, S.; Gite, S.; Pradhan, B.; Alamri, A.; Patil, B.; Dahake, S.; Gaikwad, R.; Chaudhari, A. Enhancing Semantic Segmentation for Autonomous Vehicle Scene Understanding in Indian Con-text Using Modified CANet Model. Methodsx 2025, 14, 103131, doi:10.1016/j.mex.2024.103131.

[6] Wang, W.; Tan, X.; Li, L.; Liu, Y.; Chang, Q. 3D-NLM: Voxel-Based Non-Local Means for 3D Point Cloud Noise De-tection and Smoothing. Comput. Graphics 2025, 132, 104348, doi:10.1016/j.cag.2025.104348.

[7] Singh, D.P.; Yadav, M. Deep Learning-Based Semantic Segmentation of Three-Dimensional Point Cloud: A Comprehensive Review. Int. J. Remote Sens. 2024, 45, 532–586, doi:10.1080/01431161.2023.2297177.

[8] Guo, Z.; Zhang, Y.; Zhu, L.; Wang, H.; Jiang, G. TSC-PCAC: Voxel Transformer and Sparse Convolution-Based Point Cloud Attribute Compression for 3D Broadcasting. IEEE Trans. Broadcast. 2025, 71, 154–166, doi:10.1109/TBC.2024.3464417.

[9] Li, Y.; Li, Q.; Gao, C.; Gao, S.; Wu, H.; Liu, R. PFENet: Towards Precise Feature Extraction from Sparse Point Cloud for 3D Object Detection. Neural Networks 2025, 185, 107144, doi:10.1016/j.neunet.2025.107144.

[10] Zeng, Z.; Xu, Y.; Xie, Z.; Tang, W.; Wan, J.; Wu, W. Large-Scale Point Cloud Semantic Segmentation via Local Per-ception and Global Descriptor Vector. Expert Syst. Appl. 2024, 246, 123269, doi:10.1016/j.eswa.2024.123269.

[11] Liu, Q.; Yuan, H.; Su, H.; Liu, H.; Wang, Y.; Yang, H.; Hou, J. PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-View Projection. IEEE Trans. Circuits Syst. Video Technol. 2021, 31, 4645–4660, doi:10.1109/TCSVT.2021.3100282.

[12] Wang, Z.; Yin, M.; Dong, J.; Zheng, H.; Ou, D.; Xie, L.; Yin, G. Multi-View Point Clouds Registration Method Based on Overlap-Area Features and Local Distance Constraints for the Optical Measurement of Blade Profiles. IEEE/ASME Trans. Mechatronics 2022, 27, 2729–2739, doi:10.1109/TMECH.2021.3119435.

[13] Yu, H.-T.; Song, M. MM-Point: Multi-View Information-Enhanced Multi-Modal Self-Supervised 3D Point Cloud Understanding. Proc. AAAI Conf. Artif. Intell. 2024, 38, 6773–6781, doi:10.1609/aaai.v38i7.28501.

[14] Charles, R.Q.; Hao, S.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, July 2017; pp. 77–85.

[15] Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space 2017.

[16] Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on X-Transformed Points. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2018; Vol. 31.

[17] Wu, W.; Qi, Z.; Fuxin, L. PointConv: Deep Convolutional Networks on 3D Point Clouds.; 2019; pp. 9621–9630.

[18] Thomas, H.; Qi, C.R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds 2019.

[19] Zhao, H.; Jiang, L.; Fu, C.-W.; Jia, J. PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2019; pp. 5560–5568.

[20] Zhang, Z.; Hua, B.-S.; Yeung, S.-K. ShellNet: Efficient Point Cloud Convolutional Neural Networks Using Concentric Shells Statistics 2019.

[21] 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 2020.

[22] Ding, X.; Chen, H.; Zhang, X.; Han, J.; Ding, G. RepMLPNet: Hierarchical Vision MLP with Re-Parameterized Locality 2022.

[23] Fan, S.; Dong, Q.; Zhu, F.; Lv, Y.; Ye, P.; Wang, F.-Y. SCF-Net: Learning Spatial Contextual Features for Large-Scale Point Cloud Segmentation.; 2021; pp. 14504–14513.

[24] Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3D Semantic Parsing of Large-Scale Indoor Spaces. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2016; pp. 1534–1543.

[25] 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.; 2020; pp. 202–203.

[26] Su, Y.; Liu, W.; Yuan, Z.; Cheng, M.; Zhang, Z.; Shen, X.; Wang, C. DLA-Net: Learning Dual Local Attention Features for Semantic Segmentation of Large-Scale Building Facade Point Clouds. Pattern Recognit. 2022, 123, 108372, doi:10.1016/j.patcog.2021.108372.

[27] Qiu, S.; Anwar, S.; Barnes, N. Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion 2021.

[28] Landrieu, L.; Simonovsky, M. Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs 2018.

[29] Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. Acm T. Graphic. 2019, 38, 1–12, doi:10.1145/3326362.

[30] Ma, L.; Li, Y.; Li, J.; Tan, W.; Yu, Y.; Chapman, M.A. Multi-Scale Point-Wise Convolutional Neural Networks for 3D Object Segmentation from LiDAR Point Clouds in Large-Scale Environments. IEEE Trans. Intell. Transp. Syst. 2021, 22, 821–836, doi:10.1109/TITS.2019.2961060.