International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 4 ǁ April 2024 ǁ PP. 300-310

Survey of 3D Segmentation Research Based on Deep Learning

Yaoda Zhu [1,*]
*[*1]College of Communication and Art Design, University of Shanghai for Science and TechnologyShanghai 200093, China; 213342990@st.usst.edu.cn (Y.Z.)*
Corresponding Author: Yaoda Zhu

Abstract
With the rapid development of deep learning technology, three-dimensional point cloud semantic segmentation has been widely applied in indoor scenes, robotic arm grasping, autonomous driving, and other fields. As one of the key technologies in three-dimensional scene understanding, three-dimensional point cloud segmentation has attracted extensive attention from researchers and has significant research significance and broad application prospects. Based on this, this paper provides a detailed review of deep learning-based three-dimensional point cloud segmentation methods and the latest research status. From the perspective of deep learning, three-dimensional point cloud semantic segmentation methods can be divided into direct and indirect semantic segmentation methods. This paper subdivides and analyzes the research contents of each method and summarizes their basic ideas and advantages and disadvantages. At the same time, this paper summarizes and analyzes the current mainstream public datasets and evaluation metrics. Finally, it outlines the future development direction of three-dimensional point cloud semantic segmentation technology.

Keywords: *computer vision, deep learning, three-dimensional point cloud, semantic segmentation*

Date of Submission: 10-04-2024 Date of acceptance: 23-04-2024

I. INTRODUCTION

In recent years, with the continuous improvement of point cloud acquisition devices such as LiDAR[1-4]and RGB-D depth cameras[5,6], the cost of obtaining data has gradually decreased, making the application of three-dimensional point clouds increasingly comprehensive. In the field of computer vision, research on three-dimensional point clouds has attracted widespread attention. The main task of three-dimensional point cloud semantic segmentation is the process of perceiving, analyzing, and ultimately predicting the semantic categories for each three-dimensional point cloud data. Three-dimensional point cloud semantic segmentation technology is mainly applied in various fields such as robotic arm grasping in the mechanical industry, obstacle avoidance for robots in indoor scenes[7-10], vehicle road perception and segmentation recognition in the field of autonomous driving[3,4,11-13], and urban building environment perception in satellite remote sensing fields[14]. Compared to two- dimensional image data, three-dimensional point clouds not only contain the real geometric and shape features of objects but also reduce the influence of changes in lighting intensity and occlusion of viewpoints that exist in two- dimensional images. However, compared to two-dimensional image data, current three-dimensional point cloud data exhibits characteristics such as differences in point cloud distribution density, unorderedness of points, invariance after rotation or translation, and irregularity, posing greater challenges in semantic segmentation. Among them, differences in point cloud density may lead to some areas being densely populated while others are sparse, posing challenges for subsequent feature extraction and semantic segmentation. Additionally, the unorderedness of point clouds and their invariance after rotation or translation result in different point cloud representations of the same object, requiring algorithms to understand the invariance of objects after rotation or translation and segment them correctly. Furthermore, the irregularity of three-dimensional point clouds increases the complexity of processing and analysis, requiring algorithms to adapt to the irregularities between different point clouds.

Currently, three-dimensional point cloud semantic segmentation mainly employs two methods[15,16]: traditional machine learning and deep learning. Traditional machine learning performs well when dealing with small-scale data. However, three-dimensional point cloud data has characteristics such as high feature dimensionality and dense point clouds. Traditional machine learning methods for three-dimensional point cloud semantic segmentation suffer from several drawbacks, including manual feature extraction, loss of local information, poor model generalization ability, and reliance on domain knowledge. These limitations restrict their application effectiveness and performance in complex scenarios.In contrast, deep learning demonstrates powerful computational and training capabilities in three-dimensional point cloud semantic segmentation. It can fully learn feature representations, is sensitive to local information, exhibits good generalization ability, and reduces manual intervention. These advantages enable deep learning to better cope with the complex point cloud data structure and semantic information, providing a more efficient and accurate solution for three-dimensional scene understanding. Therefore, this paper will analyze deep learning-based methods for three-dimensional point cloudsemantic segmentation.

This paper extends and improves upon existing point cloud segmentation reviews. It organizes recent advanced research methods on point cloud semantic segmentation and categorizes them into two main types based on how they handle three-dimensional point cloud data: indirect semantic segmentation methods and direct semantic segmentation methods. Additionally, it provides an analysis of the latest public datasets and commonly used evaluation metrics. Finally, it offers an in-depth outlook on the future research directions in the field of three-dimensional point cloud semantic segmentation for the reference of relevant researchers.

II. DATASETS AND EVALUATION METRICS

In deep learning algorithms, selecting appropriate training datasets is crucial because the quality and diversity of the datasets directly impact the model's generalization ability and practical application effectiveness. To ensure maximized model performance and generalization, it's essential to choose datasets that are representative, diverse, and well-annotated so that the network can fully learn and adapt to different scenarios. Establishing effective and diverse datasets is of significant importance for evaluating the performance of different segmentation methods, the adaptability and robustness of evaluation methods, as well as advancing theoretical research and promoting the emergence of new methods.

2.1 DATASETS

Table 1 lists six commonly used datasets, including S3DIS[17], Semantic3D[18], SemanticKITTI[19], and ShapeNetpart[20]. These datasets have become important resources widely used by researchers in the field of three-dimensional point cloud semantic segmentation, providing significant support for research in this area.

Tabel 1: Datasets Comparison

Datasets	PublicYear	Collectionequipment	Point cloud scene	Application scenarios
Semantic3D[18]↓	2017	Laser Scanner	Semantic segmentation ofoutdoor scenes	Semantic segmentation
S3DIS[17]↓	2016	Depth camera	Semantic segmentation ofindoor scenes	Semantic segmentation
Shapenet[20]↑	2016	Artificialsynthesis	Object model component segmentation	Component segmentation
SemanticKITTI[19]↑	2019	Camera/LiDAR	Semantic segmentation ofoutdoor scenes	Semantic segmentation
STPLS3D[21]↑	2022	Camera/LiDAR	Semantic segmentation ofoutdoor scenes	Semantic segmentation

S3DIS[17] dataset is a large-scale three-dimensional point cloud dataset for indoor scenes, created by the Stanford University Computer Vision Lab, as shown in Figure 1. This dataset, used for indoor scenes, includes spatial geometric coordinate information and color information, and is widely applied in various tasks and applications related to indoor environments in the mechanical industry field. The S3DIS dataset collects six large indoor scenes, including three-dimensional point cloud data and corresponding semantic labels for places such as offices, conference rooms, and libraries, including categories such as floors, walls, and furniture. The S3DIS dataset has become one of the widely used benchmark datasets for indoor three-dimensional point cloud semanticsegmentation.

The Semantic3D[18] dataset is created by Technische Universität München (TUM) and is a large -scale three-dimensional point cloud dataset designed specifically for studying outdoor environmental scenes. This dataset collects rich outdoor three-dimensional point cloud data covering various terrains and scenes such as urbanstreets, parks, and buildings. Each point cloud contains spatial geometric coordinates and color information, along with corresponding semantic labels such as ground, buildings, and trees. The Semantic3D dataset has become one of the widely used benchmark datasets for three-dimensional point cloud semantic segmentation in outdoor environments.

The ShapeNetPart[20] dataset contains 55 classes totaling 51,300 3D models and is a richly annotated dataset. ShapeNetPart is a subset of the ShapeNet dataset, jointly developed by Stanford University in the USA and the Toyota Technological Institute at Chicago. It includes 16 object categories with 16,881 shapes, 31,693 grids, and a total of 50 component categories.

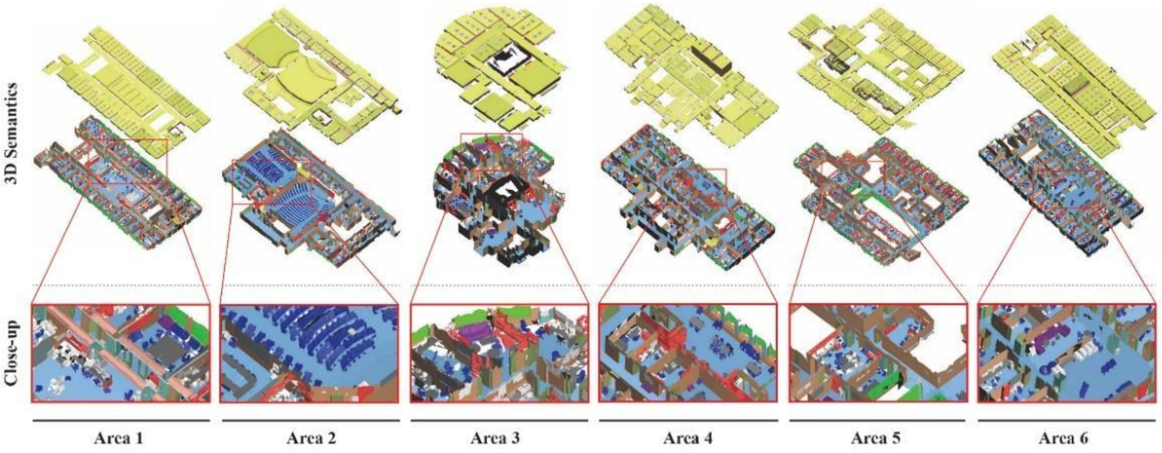


Figure 1: Schematic diagram of S3DIS[17] dataset

The SemanticKITTI[18] dataset is a large-scale three-dimensional point cloud dataset created by the Karlsruhe Institute of Technology (KIT) for autonomous driving scenarios. As an extension of the KITTI dataset, SemanticKITTI provides more detailed semantic annotation information. It contains a large amount of LiDAR point cloud data from driving cars, covering various terrains and scenes in cities and suburbs. Each point cloud contains rich geometric and color information, along with detailed semantic labels such as vehicles, pedestrians, roads, and buildings. SemanticKITTI has become one of the widely used benchmark datasets in autonomous driving technology and practical applications.

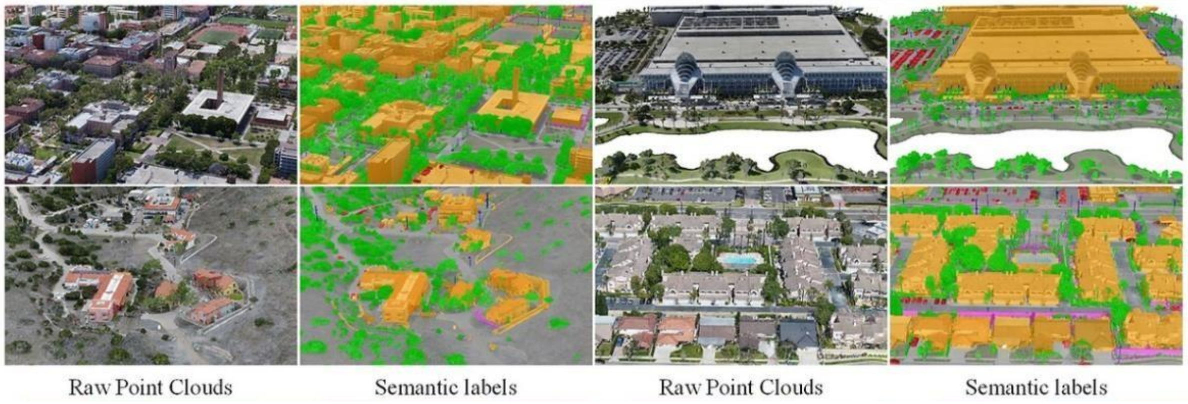


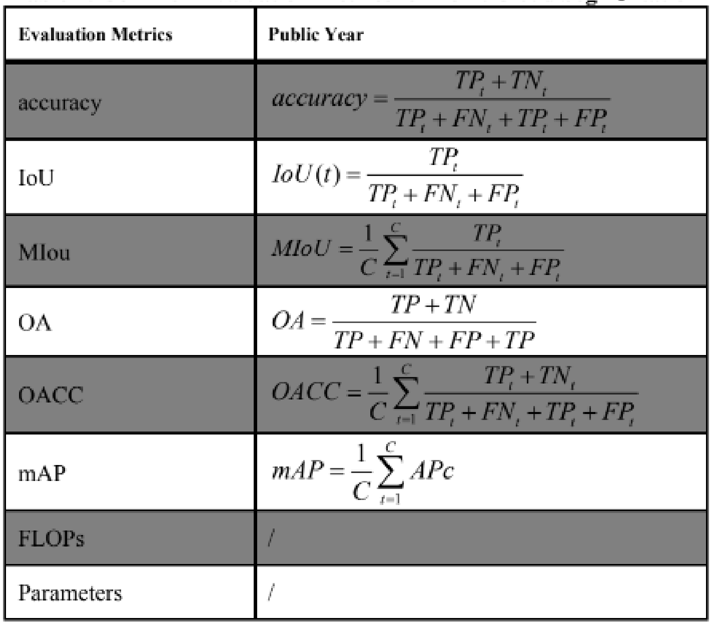
Figure 2: Schematic diagram of STPLS3D[21] dataset

The STPLS3D[21] dataset is a large-scale three-dimensional point cloud dataset designed for outdoor urban scenes, as shown in Figure 2. This dataset contains vast urban scenes with over 16 square kilometers of urban data and up to 18 fine-grained semantic categories, including ground, walls, vegetation, vehicles, and more. The application scope of the STPLS3D dataset is broader and can be used in research and applications such as urban building perception, traffic management, urban building monitoring, and more.

2.2 EVALUATING INDICATOR

Three-dimensional point cloud semantic segmentation evaluation metrics are used to compare the performance of different segmentation algorithms. These metrics typically include Overall Accuracy (OA), Mean Intersection over Union (mIoU), Mean Average Precision (mAP), the number of parameters, and Floating Point Operations per Second (FLOPs). The specific formulas are shown in Table 2. Where: N represents the total number of samples, C represents the number of segmentation categories, TP (True Positive points) represents the portion where the true label and the predicted label overlap, which is the number of points correctly predicted as true by the model. FN (False Negative points) represents the number of true value points not predicted by the model. FP (False Positive points) represents the number of points misjudged by the model as true, i.e., the portionnot overlapping with the true label.

Table 2: Common Evaluation Metrics for Point Cloud Segmentation



Evaluation Metrics	Public Year
accuracy	$accuracy = \frac{TP_t + TN_t}{TP_t + FN_t + TP_t + FP_t}$
IoU	$IoU(t) = \frac{TP_t}{TP_t + FN_t + FP_t}$
MIou	$MIoU = \frac{1}{C}\sum_{t=1}^{C}\frac{TP_t}{TP_t + FN_t + FP_t}$
OA	$OA = \frac{TP + TN}{TP + FN + FP + TP}$
OACC	$OACC = \frac{1}{C}\sum_{t=1}^{C}\frac{TP_t + TN_t}{TP_t + FN_t + TP_t + FP_t}$
mAP	$mAP = \frac{1}{C}\sum_{t=1}^{C} APc$
FLOPs	/
Parameters	/

In the task of three-dimensional point cloud semantic segmentation, the mean Intersection over Union (MIoU) directly reflects the accuracy of the three-dimensional point cloud segmentation. The Intersection over Union (IoU) represents the accuracy of a specific category prediction in three-dimensional point cloud segmentation, which is the ratio of the intersection to the union between the predicted segmentation region and the ground truth label for that category. Overall Accuracy (OA) refers to the average prediction accuracy of the network model across all semantic segmentation categories, which is the proportion of correctly predicted point clouds to the total number of point clouds. Overall Class Accuracy (OACC) refers to the average prediction accuracy for each semantic segmentation category, which is the average of the proportion of correctly predicted point clouds in each category to the total number of point clouds in that category.

III. RESEARCH METHODS FOR 3D POINT CLOUD SEMANTIC SEGMENTATION

With the development of deep learning technology, significant progress has been made in the field of point cloud semantic segmentation. In recent years, researchers have proposed numerous deep learning-based segmentation models to address the challenges of point cloud data. Compared to traditional algorithms, these models have shown significant improvements in performance, reaching higher standards. In this paper, based on the processing methods of point cloud data, we categorize deep learning-based three-dimensional point cloud semantic segmentation methods into two types: indirect and direct semantic segmentation methods.

3.1 INDIRECT SEMANTIC SEGMENTATION METHOD

3.1.1 Research Methods Based On Multi-View Approaches

In the field of point cloud semantic segmentation, the multi-view approach typically involves projecting three-dimensional point clouds into two-dimensional images. Traditional 2D segmentation methods are then employed to segment the data, and the results are subsequently projected back into the 3D point cloud, achieving the segmentation of point clouds. Thanks to the successful application of Convolutional Neural Network (CNN) on 2D images, Su, Hang[7] et al. proposed a multi-view feature extraction point cloud semantic segmentation network, MV-CNN. This network transforms three-dimensional objects into several two-dimensional images from different viewpoints. It aggregates and completes the semantic segmentation task based on the feature information at each position. Zeid, K.A[22]. et al. applied the self-supervised framework data2vec to point cloud segmentation, proposing the Point2Vec network. This network addresses the problems of missing positional information and severe occlusions. However, the projection process may lead to information loss, and it cannot fully utilize the sparsity of the three-dimensional point cloud, thus unable to learn geometric information.To address the issue of mapping between the two-dimensional pixels learned in large scenes and the 3D point cloud, requiring grid reconstruction to recover occlusions, Damien Robert[23] et al. proposed an end-to-end trainable aggregation network based on multi-view, named DeepViewAgg. This model merges the features in 2D images captured from arbitrary positions by utilizing the features of the three-dimensional point cloud.

Bird's Eye View[6] (BEV) has been widely used in the field of autonomous driving. However, in the process of extracting and fusing two-dimensional image data captured by multiple cameras and finally projecting them onto the top-view grid, there are issues related to geometric feature mapping errors. Direct dense mapping using attention mechanisms would lead to wastage of significant computational resources. Florent Bartoccioni[24] et al. proposed a semantic segmentation network called LaRa (Latents and Rays) based on the Cross-Attention mechanism. LaRa utilizes the Cross-Attention mechanism to aggregate features from multiple two-dimensional image sensors into a compact set. It also employs the Self-Attention mechanism to learn semantic information from the feature representation set.Currently, most point cloud semantic segmentation methods adopt the K- Nearest Neighbors (KNN) algorithm[25]. However, its computational complexity also affects the network's computational efficiency. Chuanyu Luo[5] et al. proposed an end-to-end architecture for the multi-view per-point network, MVP-Net (Multiple View Pointwise). MVP-Net can directly perform inference using large-scale outdoor point clouds without the need for K-Nearest Neighbors (KNN) algorithms or other complex preprocessing operations, reducing wastage of computational resources. However, issues such as image occlusion and low image quality may affect the accuracy of semantic segmentation, requiring the introduction of additional post-processingto address such problems.

Although using the multi-view approach avoids the irregularity of point clouds, it still has the following drawbacks: Rendering the original point cloud into two-dimensional images results in a significant waste of computational resources, making real-time performance unattainable. In current autonomous driving applications, using the multi-view method to process data from multiple viewpoints requires a considerable amount of additional computational resources, unable to meet the system's real-time performance requirements. Utilizing the multi-view approach cannot fully exploit the sparsity of the three-dimensional point cloud, resulting in the loss of depth information and overall geometric features of the point cloud. Different view selections can significantly affect the network's recognition performance. When merging occluded or blurry image information from different viewpoints, the network needs to introduce additional post-processing to improve the accuracy of the network segmentation results.

3.1.2 Research On Voxel-Based Methods

Voxel-based methods for point cloud semantic segmentation typically involve preprocessing point clouds into dense or sparse three-dimensional voxels. This approach allows the application of traditional 2D image processing techniques. After learning the features, these are devoxelized and returned to the point cloud to achieve three-dimensional semantic segmentation. Voxel-based methods reduce the impact of holes and missing parts in point clouds during the voxelization process, particularly when dealing with poor quality or heavily occluded point clouds. Additionally, voxel-based methods offer the advantage of interpretability.

Iro Armeni[26] et al., based on the 3D Fully Convolutional Neural Network (3D-FCNN), applied the Fully Connected Conditional Random Field (FC-CRF) to maintain spatial consistency and proposed the SegCloud network for 3D point cloud semantic segmentation. The network achieves point cloud semantic segmentation through trilinear interpolation.Truc Le[27] et al. proposed a method using the embedded voxel grid, PointGrid, which avoids feature loss by uniformly quantizing and hierarchically extracting global information at different scales. However, increasing the resolution during the voxelization process leads to a significant increase in computational complexity.In response to the issue of low voxelization resolution due to limited computational resources, Haotian Tang[28] et al. introduced a lightweight sparse point voxel convolution module called SPVConv and built the SPVCNN network upon it. This innovation addresses the problem of low resolution brought by sparse convolution and overcomes the challenge of applying three-dimensional voxel convolution to large-scale scenes.

Yuenan Hou[29] et al. supplemented sparse supervised signal features using voxel distillation and proposed the PVKD network. The network samples low-frequency categories and distant objects by perceiving sampling difficulty, but it performs poorly in local feature processing.Xinge Zhu[30] et al. introduced a novel point cloud segmentation framework, Cylindrical3D. It achieves semantic segmentation by segmenting scenes cylindrically and using a symmetric three-dimensional convolution network, addressing the irregularity and density variation of outdoor large scenes.

While using voxel-based methods can address the issues of unordered point clouds and differences in point cloud distribution, they also have the following drawbacks: Currently, commonly used voxel-based methods encounter high runtime memory consumption, high computational costs, and large storage space issues when dealing with smaller voxel sizes. Many detailed features of small-sized object point clouds cannot be adequately described during the conversion to discrete voxel representations, resulting in the loss of local features and the inability to retain the original point cloud information. Selecting a voxel size that is too small in the network will lead to a waste of computational resources due to a large number of empty voxels. Conversely, choosing a voxel size that is too large will result in the loss of local features and poorer segmentation results. Furthermore, the usagescenarios become too narrow.

3.2 DIRECT SEMANTIC SEGMENTATION METHODS

In the task of three-dimensional point cloud semantic segmentation, methods based on multi-view and voxelization face some challenges, stemming from the limitations of the algorithms themselves and the unstructured and unordered nature of point cloud data. These methods have certain limitations in segmentation accuracy and practical application range, posing challenges to the task of three-dimensional point cloud semantic segmentation. Direct semantic segmentation methods extract feature information directly from point cloud data, without the need for voxelization or multi-view transformation, retaining the inherent information of the original points for point-level semantic prediction. Therefore, direct processing of irregular point clouds is currently the mainstream method for three-dimensional point cloud semantic segmentation.

3.2.1 Research On Methods Based On Multilayer Perceptron

In recent years, researchers have abandoned the complex preprocessing steps of the above two methods to fully utilize the geometric spatial characteristics of point cloud data, while also considering lower computational complexity and memory requirements. Consequently, they have adopted direct point cloud-based methods for end-to-end learning of the point cloud. Addressing the shortcomings of the multi-view and voxel-based methods, Charles R. Qi[31] et al. first proposed the PointNet network, using unordered raw point clouds as input, thereby avoiding complex preprocessing. They then mapped features to a high-dimensional space to learn the features of each point. However, the PointNet network only considers global feature information and ignores local structural information between points. This led to a problem of not using the spatial structure information of points and resulted in the loss of local features.

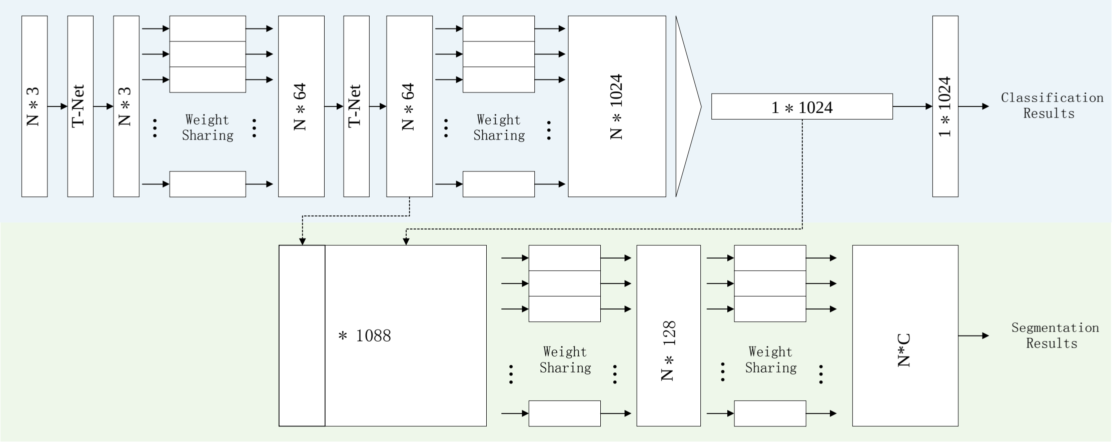


Figure 3: PointNet[31] Network Architecture

To address the issue of local feature loss in the PointNet network, Charles R. Qi[32]et al. proposed the PointNet++ network. Based on the original network, it drew inspiration from the concept of the multi-layer perception field proposed by Krizhevsky[33] et al. They used the farthest point sampling (FPS) algorithm to iteratively extract features from the local neighborhood of the point cloud. The 3DMAX-Net network proposed by Ma, Yanxin[34] et al. similarly adopted a multi-scale approach, effectively capturing local and global contextual information in the 3D point cloud. The network proposed a fusion of multi-scale convolutional layers and a cross- attention module, which can adaptively focus on different spatial resolution information areas according to the point cloud of different scenarios, achieving accurate semantic segmentation of complex 3D scenes. EiyueWang[35] et al. proposed a segmentation network, SGPN, that uses raw point clouds as data input. The network predicts the point cloud through a separate similarity matrix module, thus providing accurate prediction results and grouping suggestions for each point. However, the similarity matrix in the SGPN network grows quadratically with the number of points. While this method saves memory space compared to voxel-based methods, it is not suitable for large-scale point cloud data. To address the various issues with SGPN, QingyongHu[1] et al. proposed an efficient and lightweight RandLA-Net network structure, depending on different application scenarios. The network proposed a novel downsampling strategy, Random Sampling, and a local feature aggregation module, avoiding problems such as local feature loss and low computational efficiency while maintaining real-time performance. However, this method does not show obvious effects on boundary segmentation. Addressing the above issues, TiangeXiang[36] et al. proposed a method for aggregating assumed curves in point clouds. This method is based on the classic segmentation network framework ResNet proposed by Tan, MX[37], which adds a curve grouping module and embeds it into the CurveNet network. By guiding and grouping the connection points in the point cloud in sequence and then aggregating them back to enhance their point-by-point features, the segmentation effect is improved. Currently, most methods used for feature aggregation widely adopt the max-pooling function, which results in the loss of granular information.

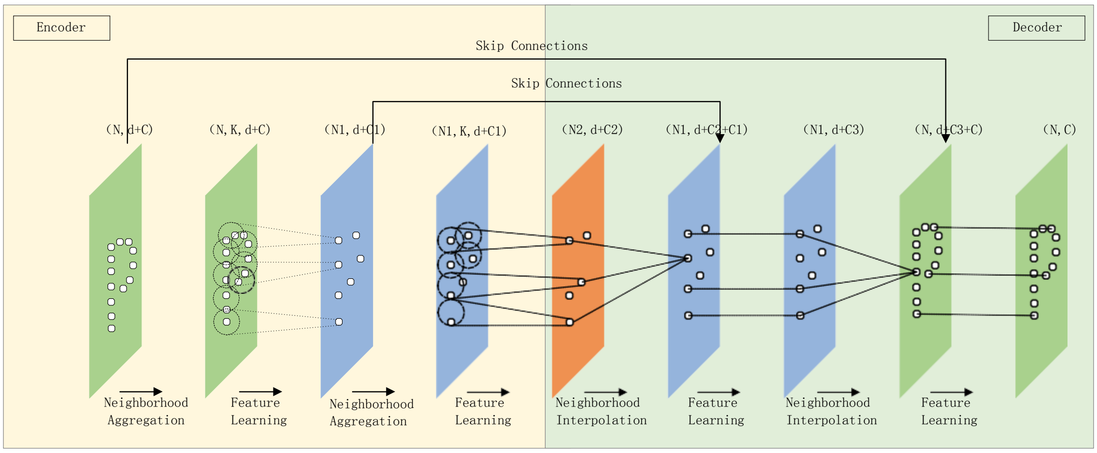


Figure 4: PointNet++[32] Network Architecture

3.2.2 Research On Point Convolution-Based Methods

In addition to using the point-by-point MLP approach to handle point cloud data, some researchers are also exploring the use of point convolution methods for processing point-based data. Due to the unordered and irregular nature of 3D point cloud data, traditional 2D image grid convolutions are not suitable. The convolution- based feature extraction method designs specific 3D convolution kernels and then utilizes convolutional operations to extract features from the point cloud. Wenxuan Wu[38] et al. proposed a convolution operator, PointConv, which effectively calculates the weight function on 3D point cloud data. Additionally, they extended it to a deconvolution operator, PointDeconv, combined with a linear interpolation algorithm, resulting in better segmentation results. However, PointConv does not consider rigid transformations such as rotation and translation during the convolution process. To address this issue, Yongcheng Liu[39] et al. proposed the Relation-ShapeCNN, a network that infers implicit 3D shape information from the relationships between points, achieving permutation invariance and robustness to rigid transformations. Hugues Thomas[40] et al. introduced a kernel point convolution (KPConv) network for point cloud processing. Unlike traditional grid convolutions, KPConv is composed of convolutional kernels with weights, where each kernel point convolution has a corresponding distance influence. However, the KPConv network increases the complexity of the entire calculation and can cause overfitting for simple classification and segmentation tasks. To address the problems with KPConv, Kangcheng Liu[14] et al. designed a novel noise and outlier filtering method to promote subsequent advanced tasks. They proposed a deep convolutional neural network, FG-Net, which utilizes correlation feature mining and geometric perception modeling based on deformable convolutions, making full use of local feature relationships. They also proposed inverse density sampling operations and feature-based residual learning strategies to save computational costs and memory consumption, respectively. Wu, Wenxuan[41] et al. introduced the PAConv network, which processes point-based data by effectively calculating the weight function on 3D point cloud data using convolutional operators. By designing specific 3D convolution kernels and using convolution operations to extract features from the point cloud, it improves the accuracy of point cloud segmentation.

3.2.3 Research On Attention Mechanism-Based Methods

Inspired by the significant success of the Transformer proposed by Vaswani, A.[42] et al. in the field of Natural Language Processing (NLP), many researchers have recently applied the Transformer to direct point-based semantic segmentation of point clouds. Guo, M. H.[43] et al. proposed a Point Cloud Transformer (PCT), suitable for irregular and unstructured point cloud learning networks. The adopted offset attention and normalization mechanism contribute to the network. Zhao, H. S.[44] et al. introduced the Point Transformer network, which utilizes a subtracive vector attention mechanism. In each layer of the downsampling process, a max-pooling layer is used to aggregate information, but a large amount of non-maximum point information is lost during the layer-by-layer downsampling process. To address the issues related to feature loss caused by the max- pooling layer, KevinTirta Wijaya[45] et al. proposed a new point cloud feature learning network named Point Stack. The network adopts the concept of multi-resolution feature learning and learnable pooling layers, avoiding the complete ignorance of non-maximum point features after using max-pooling for feature aggregation.

Simultaneously, the network uses multi-resolution feature learning, ensuring that the final point cloud features contain both high semantic and high-resolution information. Renrui Zhang[46] et al. proposed a dual-scale point cloud segmentation network, DSPoint, based on high-frequency fusion. Global features are extracted using the voxel method and local features are extracted through point convolution. The network pays more attention to high- frequency information, but loses some low-frequency information. Two-dimensional images contain color and texture features that can complement the three-dimensional point cloud features. Youquan Liu[4] et al. proposed a laser radar point cloud semantic segmentation network named UniSeg, which uses features from both two- dimensional images and three-dimensional point clouds, and integrates them using a mechanism called the Learnable Cross-Modal Association (LMA). The UniSeg network effectively utilizes the semantic information of two-dimensional images to achieve more accurate segmentation results.

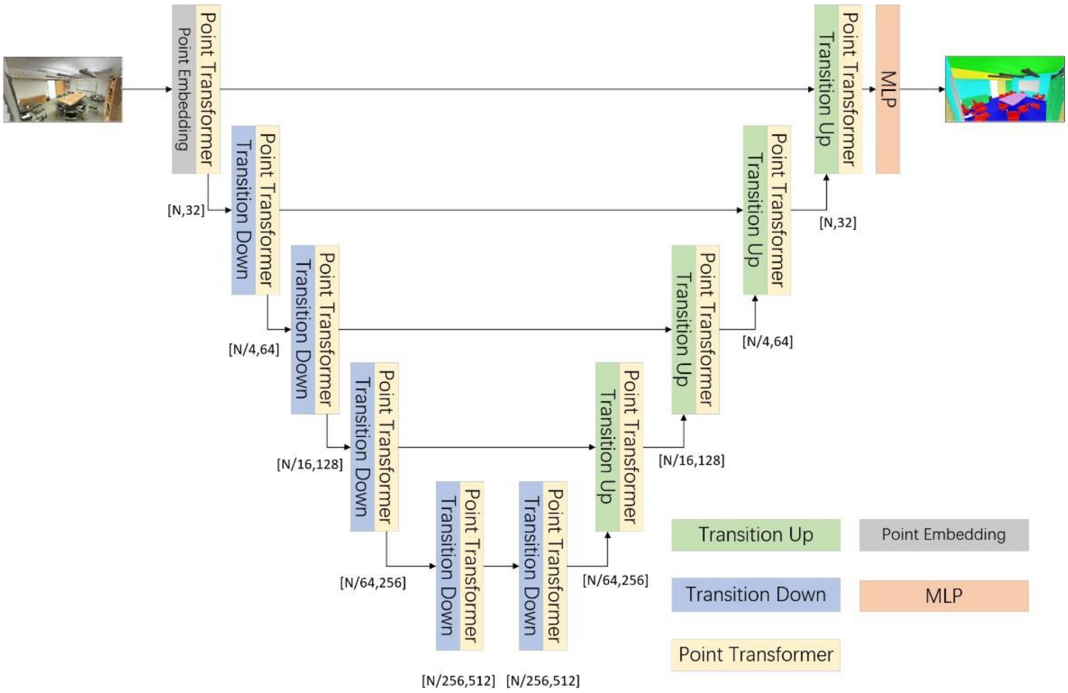


Figure 5: Point Transformer[44] Network Architecture

Influenced by the significant impact of the Swin Transformer series sliding window networks in the image domain [47,48], Lai, Xin et al. proposed the Stratified Transformer network, which introduces a grid-based local attention mechanism to operate Transformer modules in a series of sliding windows, solving the memory consumption issue. Park, Jinyoung[49] et al. introduced a self-positioning Transformer network, SPoTransformer, to reduce computational complexity and better acquire local neighborhood features. Zhou J et al. proposed the Point Cloud Size-Aware Transformer, which can provide different effective receptive fields for objects of different sizes. Zhuoxu Huang[50] et al. utilized the Local Context Propagation (LCP) module to propose the LCPFormer network, which achieves semantic segmentation by weighting the overlapping points in adjacent local regions to share point features. Kunyu Peng[3] et al. proposed a Multi-Attention Semantic Segmentation (MASS) network model. The network framework converts the input three-dimensional point cloud into cylindrical features and occupancy features around the vehicle. Through a key-point-driven graph attention mechanism, attention from the spatial input vector embedding of the LSTM (Long Short-Term Memory), and based on the cylindrical features, segmentation masks are obtained to achieve semantic segmentation. Transformers, with their ability to capture long-range dependencies, have become the preferred choice for most direct point-based point cloud semantic segmentation methods.

IV. FUTURE IMPROVEMENTS

4.1 FUTURE IMPROVEMENT STRATEGIES

The existing methods have made significant progress in improving the accuracy of semantic segmentation. However, there are still some limitations, including but not limited to the insufficient dataset, algorithm complexity, and real-time processing. Therefore, future research on 3D point cloud semantic segmentation algorithms will focus on multiple feature fusion, multimodal fusion, real-time processing and lightweighting, as well as self-supervised learning and reinforcement learning. Multiple feature fusion and multimodal fusion will utilize information from different types of data to enhance the robustness and accuracy of the algorithm. Research on real-time processing and lightweighting will optimize algorithm structure and model design to meet the requirements for speed and resource consumption. Self-supervised learning and reinforcement learning will provide new avenues for unsupervised learning and intelligent decision-making, thus promoting further application and development of 3D point cloud semantic segmentation technology in fields such as autonomous driving and intelligent robotics.

4.2 EVALUATION METRICS

i. Research on Three-Dimensional Point Cloud Semantic Segmentation Algorithms with Multiple Feature Fusion

In future research, more attention will be paid to effectively utilizing and integrating various types of data. Among these, multi-feature data fusion is an important direction, including utilizing the spatial geometric features, color information of point cloud data, as well as other data obtained from different sensors, such as image information, normal vector information, etc. By comprehensively using these different types of data, the robustness and accuracy of semantic segmentation algorithms can be improved, further expanding their applicability and effectiveness in practical applications.

ii. Research on Three-Dimensional Point Cloud Semantic Segmentation Algorithms with Multimodal Fusion In future research, more attention will be given to multimodal data fusion, which can be carried out at multiple levels, including feature fusion, information fusion, and model fusion. In terms of feature fusion, the feature information obtained from different sensors can be combined to form a more comprehensive and rich feature representation. From the perspective of information fusion, the semantic information of different types of data can be cross-verified and complemented to improve the consistency and accuracy of the semantic segmentation results. From the viewpoint of model fusion, an end-to-end multimodal semantic segmentation model can be designed to integrate information from different data sources into a unified framework, achieving more comprehensive and accurate segmentation results. Research on multimodal fusion of three-dimensional point cloud semantic segmentation algorithms helps to address the challenges encountered in practical applications, promoting the widespread application of semantic segmentation technology in fields such as autonomous driving and intelligent robotics.

iii. Research on Real-time and Lightweight Techniques

Currently, although the proposed semantic segmentation network models have made significant progress in accuracy, the increase in complexity and processing speed remains a major challenge. Especially in applications such as autonomous driving, pedestrian detection, and environmental perception, the demand for real-time semantic segmentation is increasing. Additionally, as the application scenarios of three-dimensional point clouds in industries such as mechanical engineering continue to expand, higher requirements and challenges are posed for the real-time and efficiency of point cloud semantic segmentation algorithms. Therefore, future research directions will focus on algorithmic improvements in real-time and lightweight techniques to meet the demands of different application scenarios.

Specific methods to address this challenge include the study of real-time and lightweight algorithms. To enhance real-time processing, the algorithm's response speed can be improved by optimizing network structure, designing lightweight models, or using hardware acceleration. For example, lightweight network structures can be adopted, reducing the number of network layers or parameters, and further reducing model complexity through techniques such as pruning and quantization. Furthermore, for lightweight techniques, exploration of the characteristics of point cloud data can lead to the design of more efficient algorithms or model structures to enhance the algorithm's operational efficiency in resource-constrained environments.

iv. Research on Self-Supervised Learning and Reinforcement Learning

Future research will also focus on self-supervised learning and reinforcement learning. These two learning methods can provide new ideas and technical support for three-dimensional point cloud semantic segmentation tasks

Self-supervised learning is an unsupervised learning method that trains models by utilizing the characteristics of the data itself, without the need for manually labeled labels. In the field of three-dimensional point cloud semantic segmentation, self-supervised learning can utilize features such as spatial structure and color information of point cloud data, designing self-supervised tasks to train models. For example, autoencoders or generative adversarial networks can be designed to learn representations of point cloud data, and then use the learned representations for semantic segmentation tasks. The introduction of self-supervised learning can effectively solve the problem of high data annotation costs while improving the model's generalization ability androbustness.

Reinforcement learning is a method of learning through interaction with the environment, with the goal of enabling agents to learn to take actions in the environment to maximize cumulative rewards. In three-dimensional point cloud semantic segmentation, the segmentation task can be modeled as a reinforcement learning problem, where the agent learns the optimal segmentation strategy by observing point cloud data and performing semantic segmentation operations. By introducing reinforcement learning, semantic segmentation algorithms can become more flexible and intelligent, dynamically adjusting segmentation strategies in different scenarios, thereby improving segmentation accuracy and efficiency.

V. CONCLUSION

This paper provides a comprehensive review and summary of the deep learning-based semantic segmentation methods for three-dimensional (3D) point clouds. Although the field of deep learning-based 3D point cloud semantic segmentation is relatively new, this paper clearly demonstrates the rapid growth and effectiveness within this field. The paper gives a detailed introduction to 3D point cloud semantic segmentation, from the perspective of deep learning, categorizing the methods into direct and indirect types, and providing a detailed analysis and summary of various methods, including their basic principles and advantages and disadvantages. Furthermore, this paper also provides a comprehensive overview of the current mainstream public datasets and evaluation metrics, to better understand the research status. Finally, this paper outlines the future development direction of 3D point cloud semantic segmentation technology, providing valuable reference and inspiration for researchers in this field. By reading this paper, it is hoped that scholars and researchers will further delve into the field of 3D point cloud semantic segmentation, thus promoting more prominent progress in research and practice in areas such as intelligent manufacturing, autonomous driving, and intelligent robotics. It is anticipated that various innovative research ideas will continue to emerge in the coming years.

REFERENCES

[1]. Qingyong H, Bo Y, Linhai X, et al. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds[J]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Proceedings, 2020: 11105-11114.

[2]. Kundu A, Xiaoqi Y, Fathi A, et al. Virtual Multi-view Fusion for 3D Semantic Segmentation[J]. Computer Vision - ECCV 2020.

[3]. 16th European Conference. Proceedings. Lecture Notes in Computer Science (LNCS 12369), 2020: 518-535.

[4]. Peng K Y, Fei J C, Yang K L, et al. MASS: Multi-Attentional Semantic Segmentation of LiDAR Data for Dense Top-View Understanding[J]. Ieee Transactions on Intelligent Transportation Systems, 2022, 23(9): 15824-15840.

[5]. Liu Y, Chen R, Li X, et al. UniSeg: A Unified Multi-Modal LiDAR Segmentation Network and the OpenPCSeg Codebase[J].

[6]. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 21605-21616.

[7]. Luo C, Li X, Cheng N, et al. MVP-Net: Multiple View Pointwise Semantic Segmentation of Large-Scale Point Clouds[J], 2022.

[8]. Ma Y, Wang T, Bai X, et al. Vision-Centric BEV Perception: A Survey[J], 2022.

[9]. Su H, Maji S, Kalogerakis E, et al. Multi-view Convolutional Neural Networks for 3D Shape Recognition[C]. IEEE International Conference on Computer Vision, 2015: 945-953.

[10]. Lai X, Liu J H, Jiang L, et al. Stratified Transformer for 3D Point Cloud Segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 8490-8499.

[11]. Park C, Jeong Y, Cho M S, et al. Fast Point Transformer[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 16928-16937.

[12]. Zhang C, Wan H C, Shen X Y, et al. PVT: Point-voxel transformer for point cloud learning[J]. International Journal of Intelligent Systems.

[13]. Qin Z, Yu H, Wang C J, et al. Geometric Transformer for Fast and Robust Point Cloud Registration[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 11133-11142.

[14]. Yan X, Gao J T, Zheng C D, et al. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds[C]. 17th European Conference on Computer Vision (ECCV), 2022: 677-695.

[15]. Anh-Thuan Tran H-S L, Suk-Hwan Lee, Ki-Ryong Kwon ·. PointCT: Point Central Transformer Network for Weakly -supervised Point Cloud Semantic Segmentation[J]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[16]. Liu K, Gao Z, Lin F, et al. Fg-net: A fast and accurate framework for large-scale lidar point cloud understanding[J], 2022, 53(1): 553-564.

[17]. Xiao A, Zhang X, Shao L, et al. A Survey of Label-Efficient Deep Learning for 3D Point Clouds arXiv[J]. arXiv, 2023.

[18]. Xiao A R, Huang J X, Guan D Y, et al. Unsupervised Point Cloud Representation Learning With Deep Neural Networks: A Survey[J]. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11321-11339.

[19]. Armeni I, Sener O, Zamir A R, et al. 3D Semantic Parsing of Large-Scale Indoor Spaces[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 1534-1543.

[20]. Hackel T, Savinov N, Ladicky L, et al. Semantic3D.net: A new Large-scale Point Cloud Classification Benchmark[J], 2017.

[21]. Behley J, Garbade M, Milioto A, et al. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences[J],

[22]. 2019.

[23]. Chang A X, Funkhouser T, Guibas L, et al. ShapeNet: an information-rich 3D model repository arXiv[J]. arXiv, 2015: 11 pp.- 11 pp.

[24]. Meida Chen Q H, Thomas Hugues. STPLS3D: A Large-Scale Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset.[J], 2022.

[25]. Zeid K A, Schult J, Hermans A, et al. Point2Vec for Self-Supervised Representation Learning on Point Clouds arXiv[J]. arXiv, 2023.

[26]. Robert D, Vallet B, Landrieu L. Learning Multi-View Aggregation In the Wild for Large-Scale 3D Semantic Segmentation[J], 2022.

[27]. Florent Bartoccioni ÉZ, Andrei Bursuc, Patrick Pérez, Matthieu Cord, Karteek Alahari LaRa: Latents and Rays for Multi -Camera Bird's-Eye-View Semantic Segmentation[J], 2022.

[28]. Keller J M, Gray M R, Givens J A, Jr. A fuzzyK-nearest neighbor algorithm[J]. IEEE Transactions on Systems, Man and Cybernetics, 1985, SMC-15(4): 580-585.

[29]. Tchapmi L P, Choy C B, Armeni I, et al. SEGCloud: Semantic Segmentation of 3D Point Clouds[C]. International Conference on 3D Vision (3DV), 2017: 537-547.

[30]. Le T, Duan Y, Ieee. PointGrid: A Deep Network for 3D Shape Understanding[C]. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 9204-9214.

[31]. Tang H, Liu Z, Zhao S, et al. Searching efficient 3d architectures with sparse point-voxel convolution[C]. European conference on computer vision, 2020: 685-702.

[32]. Hou Y N, Zhu X G, Ma Y X, et al. Point-to-Voxel Knowledge Distillation for LiDAR Semantic Segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 8469-8478.

[33]. Zhu X G, Zhou H, Wang T, et al. Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR-Based Perception[J]. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6807-6822.

[34]. Qi C R, Su H, Mo K C, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 77-85.

[35]. Qi C R, Yi L, Su H, et al. PointNet++ deep hierarchical feature learning on point sets in a metric space[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5105-5114.

[36]. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[37]. Ma Y, Guo Y, Lei Y, et al. 3DMAX-Net: A multi-scale spatial contextual network for 3D point cloud semantic segmentation[C].

[38]. 2018 24th International Conference on Pattern Recognition (ICPR), 2018: 1560-1566.

[39]. Wang W Y, Yu R, Huang Q G, et al. SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation[C].

[40]. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 2569-2578.

[41]. Xiang T G, Zhang C Y, Song Y, et al. Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis[C]. 18th IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 895-904.

[42]. Muzahid A a M, Wan W G, Sohel F, et al. CurveNet: Curvature-Based Multitask Learning Deep Networks for 3D Object Recognition[J]. Ieee-Caa Journal of Automatica Sinica, 2021, 8(6): 1177-1187.

[43]. Wu W X, Qi O G, Li F X, et al. PointConv: Deep Convolutional Networks on 3D Point Clouds[C]. 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 9613-9622.

[44]. Liu Y, Fan B, Xiang S, et al. Relation-Shape Convolutional Neural Network for Point Cloud Analysis[J], 2019.

[45]. Thomas H, Qi C R, Deschaud J E, et al. KPConv: Flexible and Deformable Convolution for Point Clouds[C]. IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6420-6429.

[46]. Xu M, Ding R, Zhao H, et al. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds[C].

[47]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3173-3182.[42]. [42] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J], 2017.

[48]. Guo M H, Cai J X, Liu Z N, et al. PCT: Point cloud transformer[J]. Computational Visual Media, 2021, 7(2): 187-199.

[49]. Zhao H S, Jiang L, Jia J Y, et al. Point Transformer[C]. 18th IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 16239-16248.

[50]. Kevin Tirta Wijaya D-H P, Seung-Hyun Kong. Advanced Feature Learning on Point Clouds using Multi-resolution Features and Learnable Pooling[J], 2022.

[51]. Zhang R, Zeng Z, Guo Z, et al. DSPoint: Dual-scale Point Cloud Recognition with High-frequency Fusion[J], 2021: arXiv: 2111.10332.

[52]. Liu Z, Hu H, Lin Y, et al. Swin Transformer V2: Scaling Up Capacity and Resolution[J]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 11999-12009.

[53]. Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[C]. 18th IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 9992-10002.

[54]. Park J, Lee S, Kim S, et al. Self-positioning Point-based Transformer for Point Cloud Understanding[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21814-21823.

[55]. Huang Z, Zhao Z, Li B, et al. LCPFormer: Towards Effective 3D Point Cloud Analysis via Local Context Propagation in Transformers[J]. Arxiv, 2023.