IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 20, 2019, accepted December 4, 2019, date of publication December 10, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958671

A Review of Deep Learning-Based Semantic Segmentation for Point Cloud

JIAYING ZHANG, XIAOLI ZHAO, ZHENG CHEN, AND ZHEJUN LU
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Corresponding author: Xiaoli Zhao (evawhy@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772328.

ABSTRACT In recent years, the popularity of depth sensors and 3D scanners has led to a rapid development of 3D point clouds. Semantic segmentation of point cloud, as a key step in understanding 3D scenes, has attracted extensive attention of researchers. Recent advances in this topic are dominantly led by deep learning-based methods. In this paper, we provide a survey covering various aspects ranging from indirect segmentation to direct segmentation. Firstly, we review methods of indirect segmentation based on multi-views and voxel grids, as well as direct segmentation methods from different perspectives including point ordering, multi-scale, feature fusion and fusion of graph convolutional neural network (GCNN). Then, the common datasets for point cloud segmentation are exposed to help researchers choose which one is the most suitable for their tasks. Following that, we devote a part of the paper to analyze the quantitative results of these methods. Finally, the development trend of point cloud semantic segmentation technology is prospected.

INDEX TERMS 3D point clouds, deep learning, feature fusion, graph convolutional neural network, semantic segmentation.

I. INTRODUCTION

Semantic segmentation, as one of the most important research technologies for computer vision, was first put forward in the 1970s, and aims at classifying every pixel or point in the scene into several regions with specific semantic categories [1]. Nowadays, semantic segmentation is the basis of three-dimensional scene understanding and achieves several gratifying performances in the fields of mapping geographic information, navigation and positioning, computer vision, pattern recognition [2], etc., which has important research significance and broad application prospect.

Semantic segmentation based on two-dimensional images has made great progress in recent years. However, due to the limitations of two-dimensional data in occlusion and other aspects, the performance on segmentation is unsatisfactory. Therefore, researchers gradually turn their attention to three-dimensional data, like 3D voxel grids or 3D point clouds. Compared with traditional measurement technology, non-contact technology that is widely used for collecting point cloud data has the superiority of rapidity, penetration, real-time, dynamic, high density and high efficiency. Besides, three-dimensional data such as point clouds not only makes up for the issues of illumination and posture encountered in two-dimensional images, but also provides rich spatial information for complex scenes. Therefore, point cloud has become the research emphasis of three-dimensional data, and makes a lot of contributions to indoor navigation [3], unmanned driving [4], analysis of urban morphology, protection of digital cultural heritage and other aspects, and also changes people's lifestyle dramatically.

In this paper, we provide a comprehensive deep learning-based point cloud semantic segmentation methods. The goal of our review is to summary various kinds of approaches related to this topic, ranging from indirect ways to direct ways. Apart from reviewing the existing point cloud semantic segmentation based on deep learning, we introduce several primary datasets for point cloud (S3DIS, ScanNet, Semantic3D to name a few). Finally, we make an analysis of the results of point cloud semantic segmentation and point out its future development direction.

A. CHALLENGES OF POINT CLOUD SEMANTIC SEGMENTATION

Point cloud composed of a series of points is a point set with significant geometric data representation structure. Compared with two-dimensional data, there are many

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

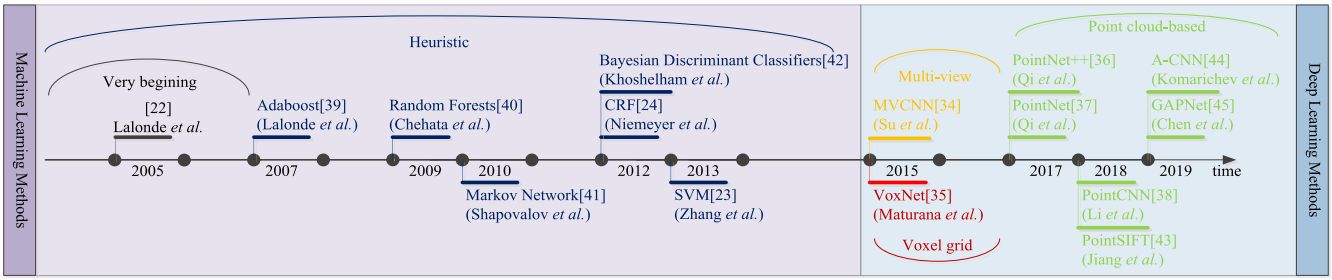

FIGURE 1. A brief chronology of point cloud semantic segmentation.

advantages with point cloud, but its characteristics of sparsity, randomness and non-structure make the semantic segmentation based on point cloud full of challenges. Nevertheless, many researchers are interested in the semantic segmentation of point cloud, because of its effective application prospect.

With the emergence of new technologies such as deep learning and convolutional neural networks in recent years, point cloud semantic segmentation has enjoyed a continuous development [5]–[8]. Although deep learning has been widely used in the processing of two-dimensional images, it is still difficult to perform convolution operations on irregular and disordered 3D point clouds directly. In order to make convolutional neural networks suitable for the point clouds, researchers convert such data into a regular structure (i.e., multi-view, voxel grids, point cloud) and then input it into the network for processing to achieve the segmentation.

This paper concentrates on concluding the existing approaches based on indirect segmentation and direct segmentation. Such classifications solve the problem that convolution operations are hardly applied to irregular 3D point clouds directly. A more in-depth classification of our paper is summarized in section II.

B. RELATED PREVIOUS WORKS

The development of 3D data capturing devices, for instance, LiDAR and Microsoft Kinect, makes the acquisition of point cloud data become ever more convenient. Traditionally, point cloud segmentation algorithms mainly include: methods based on attribute clustering [9], [10], methods based on model fitting [11], [12], methods based on region growth [13]–[15], methods based on graph-cut [16]–[18] and methods based on edge [19]–[21]. However, those approaches adopt handcrafted features from geometric constraints and several are limited by the assumed prior knowledge. Also, parameter adjustment is difficult and the segmentation results are uncontrollable.

Point cloud semantic segmentation has been put forward to understand the 3D scenes sufficiently. Different from other computer vision tasks, point cloud semantic segmentation goes through a short history and can be traced back to the pioneer works in [22]. Originally, researchers adopted the methods based on machine learning, such as Maximum Likelihood classifiers based on Gaussian Mixture Models, Support Vector Machines [23], Conditional Random Fields [24], Markov Random Fields [25], etc. to realize the task.

Along with the popularity of deep learning, deep neural networks have greatly promoted the progress of computer vision technology. And more and more models using deep neural networks (Convolutional neural networks [26]–[28], Recurrent neural networks [29]–[31], Deep belief networks [32], [33], etc.) have been springing up to extract distinguished features to realize the point cloud semantic segmentation. However, due to the irregularity and non-structure of point clouds, the application of deep network of 3D data still faces enormous challenges. At first, to circumvent this barrier, researchers transform point cloud into a regular structure (i.e., multi-views and voxel grids) suitable for convolutional neural networks to process. But these methods [34], [35] can cause problems such as information loss and computational complexity. Recently, PointNet [36], which directly works on point cloud, not only accelerates the speed of computation but improves the performance of the segmentation. Nowadays, there are many methods based on PointNet [37], [38] having been proposed. While, different methods operating on raw point cloud may use diverse knowledge and models during training. Especially, some point cloud semantic segmentation models further add the image processing algorithm to increase the accuracy of semantic segmentation. A brief chronology is shown in Figure 1.

Different from previous point cloud semantic segmentation technologies, in this paper, we divide those deep learning-based methods into two categories: indirect and direct point cloud segmentation methods. Indirect methods converting the point cloud into the regular structure are based on multi-views and voxel grids, and direct methods work directly on point cloud, which are composed of four categories: point ordering based methods, multi-scale based methods, feature fusion based methods and fusion of graph convolutional neural network (GCNN) based methods. For a review, Figure 2 is a summary of relevant methods that are involved in this paper.

C. OUR CONTRIBUTIONS

Our contributions in this paper are concluded as follows:

1) A comprehensive review of point cloud semantic segmentation models based on deep learning. We classify and summarize the existing semantic segmentation

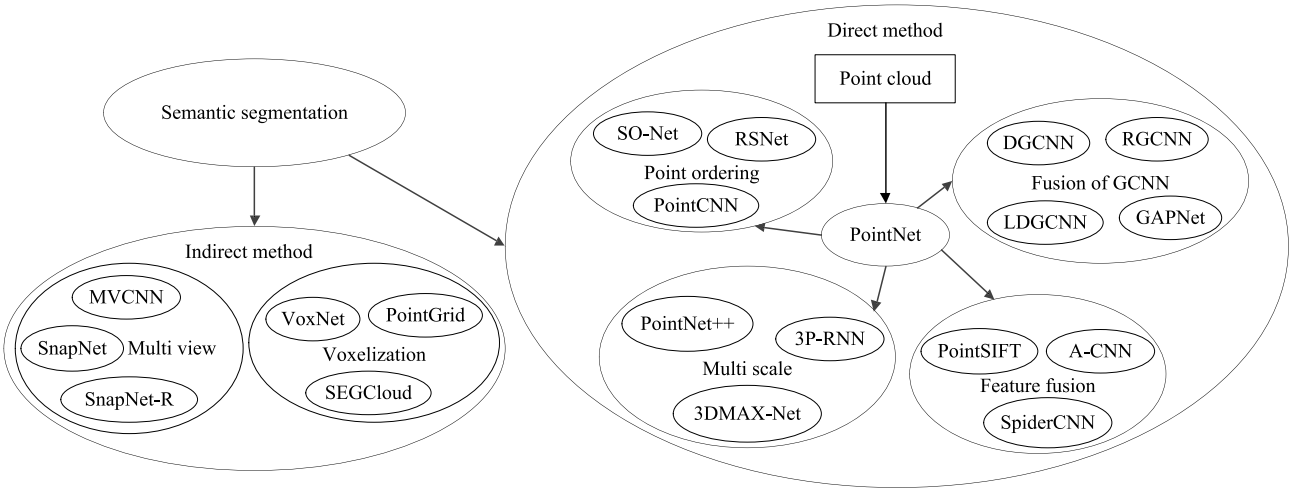

FIGURE 2. Visual representation of point cloud semantic segmentation method.

models according to different representations of 3D point clouds. The purpose is to make people have a deep understanding of the semantic segmentation model of point cloud based on deep learning.

2) Various datasets related to point cloud segmentation. We describe in the collection and composition of several common datasets used for point cloud, as well as the number of training and testing sets in deep network construction.
3) Analysis of the experimental results. According to some evaluation indicators, we summarize the results of point cloud semantic segmentation using different models on different datasets.
4) Discussion regarding the disadvantages of existing methods and future directions. We thoroughly analyze several problems for model design, which need to improve for future research.

The remainder of this paper is organized as following. Section II introduces the semantic segmentation models of point cloud based on deep learning. Section III describes common datasets widely used for point cloud. Next, section IV performs a quantitative evaluation on different indicators. Section V presents a brief discussion of the described models and predicts research directions in this field. Finally, section VI concludes the paper.

II. POINT CLOUD SEMANTIC SEGMENTATION MODEL BASED ON DEEP LEARNING

With the advent of deep learning techniques, point cloud semantic segmentation achieves a tremendous improvement. In recent years, a large number of models using these methods have been proposed to process the point cloud. Compared with traditional algorithms, models based on deep learning techniques have superior performance and reach a higher benchmark.

Based on the irregularity of point cloud, we classify point cloud semantic segmentation models based on deep learning into two categories: indirect ways and direct ways. In the rest of this section, we will introduce the specific models in these two categories comprehensively.

A. INDIRECT WAYS FOR POINT CLOUD SEMANTIC SEGMENTATION

We will briefly analyze the deep learning methods based on the transformation of point cloud for semantic segmentation in this section. So far, there are two kinds of 3D representations (i.e., multi-views and voxel grids), which convert the point cloud into a regular structure to realize the segmentation. However, those models have several drawbacks that need to be strengthened. Unfortunately, not many articles have used these transformation approaches to realize the segmentation and we will make a brief introduction.

1) MULTI-VIEW BASED METHODS [34], [45], [47]

Deep networks have been popular to process 2D data with a regular structure. Owing to the irregularity of point cloud, 2D networks cannot be directly extended to 3D applications on point clouds. Hence, a simple way is tantamount to transform 3D data into 2D views and then apply existing knowledge to extract features for point cloud processing.

a: MVCNN

Guided by 2D images, Su *et al.* [34] propose a multi-view convolutional neural network (MVCNN) based on images in 2015, which promote the development of 3D data processing. On the one hand, it successfully applies CNN to unstructured data like point cloud. On the other hand, it effectively completes the tasks such as classification and segmentation of point cloud. The main idea of this method is to project 3D point cloud into some 2D images from multiple perspectives, and employ CNN to extract features for each view using the

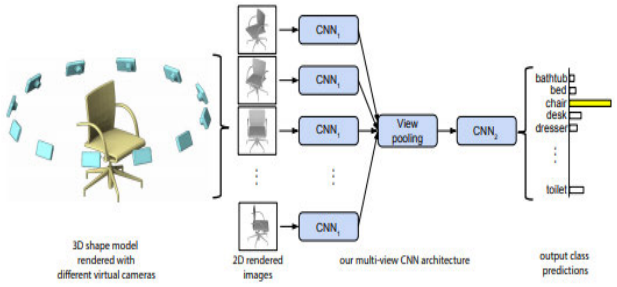

FIGURE 3. The framework of multi-view convolutional neural network (MVCNN [34]) Figure extracted from [34].

methods of image processing, and then aggregate features that extracted from different perspectives through the view pooling layer. Finally, aggregated features are input into CNN for processing, thus receiving the results of classification and segmentation. From the description, MVCNN is suitable for the segmentation of individual objects, rather than complex scenes, because it ignores the spatial relationship between objects. Figure 3 is the illustration of this network framework.

b: SnapNet

Projection, which means 3D point clouds are transformed into several 2D images from multiple perspectives, results in the problem of losing information, in order to address that, SnapNet [46] selects some snapshots of the point cloud to generate pairs of RGB and depth images. Then, using fully convolutional networks labels each pair of 2D images pixel by pixel. Finally, this model projects the marked points into 3D space to achieve the task. Although it adds depth information to assist the realization of semantic segmentation, there are also some problems that affect the accuracy of segmentation.

c: SnapNet-R

SnapNet addresses the problem of information loss, but it encounters problems in the process of image generation. Therefore, SnapNet-R [47] is put forward on the basis of SnapNet. It directly processes multiple views to obtain dense 3D point markers to further enhance the result of segmentation. The process of generating a marked point cloud can be divided into the following two steps: 2D labeling of RGB-D images obtained from stereo images and 3D labeling using SnapNet. The model has an algorithm that makes it easy to realize, however, its segmentation accuracy on object boundaries is still required to be strengthened.

From those, we can make a conclusion that compared with the traditional methods based on artificial features, the method of point cloud segmentation based on multiple views has achieved excellent results, but the projection of 3D point clouds will lead to the loss of a large number of important geometric spatial information, which finally affects the accuracy of point cloud segmentation, and it is also seriously influenced by the angle of projection.

2) VOLUMETRIC METHODS [35], [56], [57]

Voxelization of point cloud [48], [49] refers to transforming unstructured point clouds and making it into the regular volumetric occupancy grid, then learning its features by using neural networks to achieve the semantic segmentation of point cloud.

a: VoxNet

VoxNet [35] that using volumetric methods is to convert unstructured geometric data to a regular 3D grid over which standard CNN operations can be applied, and then use a 3D CNN to predict a class label directly from the occupancy grid. The method solves the problem of point clouds' non-structure, but it also has shortcomings like low efficiency of voxel grid arrangement caused by the sparsity of point clouds, large memory occupied during the computing process, long time for training and the problem of information loss, etc. With respect to the aforementioned problems, researchers have made many improvements to address them.

b: SEGCloud

Considering the sparsity of point cloud, B. Graham designs a sparse convolution network [50] and applies it to the 3D segmentation task [51]. Li et al. attempt to sample sparse 3D data and then input it into network to process, which reduces the computational cost [52]. In order to overcome the problem of spatial resolution of voxel grids, literature [53]–[55] introduce the methods of spatial partition, such as K-d tree [54] or octree [55]. However, the drawback of such methods is that it only depends on the voxel boundary and does not pay attention to the geometric structure of the local region. SEGCloud [56] subdivides the large point cloud into voxel grids by using a 3D fully convolutional neural network (3D-FCNN), and then exploits a trilinear interpolation layer to interpolate class score to 3D points. Finally, conditional random field (CRF) is used to combine original 3D point features with interpolation scores for post-processing to get fine-grained class distributions. This model effectively combines machine learning with deep learning to achieve specific tasks, and performs well in the field of semantic segmentation. Figure 4 shows the overall framework of this semantic segmentation network.

c: PointGrid

PointGrid [57] is a 3D convolutional network integrated by points and grids. In the mixed model, it employs 3D CNN to learn the grid cells with fixed points to obtain the details of local geometry. The model uses the same transformation method as VoxNet, but it can better express the change of scale, avoid information loss, and occupy a small memory. In addition, Hua *et al.* [58] also propose a three-dimensional convolution operator based on the unified grid kernel for point cloud semantic segmentation and target recognition. Compared to the state-of-the-art methods, PointGrid is simpler and faster in training and testing.

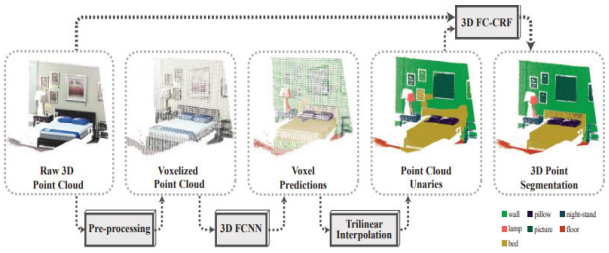

FIGURE 4. The overall architecture of SEGCloud [56] for semantic segmentation Figure extracted from [56].

The above-mentioned methods have solved the problem of non-structured point cloud and improved the disadvantages of voxel grids in various degrees, nevertheless, there are no specific approaches to deal with the quantitative artifacts that caused in the conversion process, and the calculation cost increased after transformation.

Indirect ways, which transform point cloud into regular views and voxel grids to reach the task of semantic segmentation, have solved the problem that CNN cannot be applied to point cloud and achieved excellent segmentation result. Unfortunately, it also has problems such as loss of information, complex computation and large memory occupation to be improved.

B. DIRECTLY WAYS FOR POINT CLOUD SEMANTIC SEGMENTATION

A number of shortcomings of the point cloud semantic segmentation model based on the indirect approach (that is, the transformation of the point cloud) are listed in the previous section. Therefore, the model based on raw point clouds is gradually proposed to make full use of the characteristics of point cloud data and reduces the computational complexity of the network. PointNet [36], proposed by Qi et al. is a pioneering network architecture that directly applies deep learning on the unstructured point cloud to deal with the classification and segmentation of point cloud, and its model is shown in Figure 5. This framework mainly addresses the problem of sparsity, permutation invariance and transformation invariance of point clouds. Considering the sparsity of point clouds, researchers who design the PointNet do not convert point clouds into multi-view or voxel grids, but process the points directly. For the permutation invariance, multi-layer perceptron (MLP) is employed to extract features for each point independently, and then the information of all points is aggregated to obtain global features by using the maximum pooling layer. Besides, in order to solve the problem of transformation invariance, this framework also adds the transformation network [59], which constructs the transformation matrix to spatially align the input point clouds and features. Although PointNet has a beneficial effect on point cloud classification and segmentation, it fails to take the relationship between points and local neighborhood information into account. Therefore, when dealing with point clouds in large scenes, it leads to the loss of critical information and reaches a bad segmentation result.

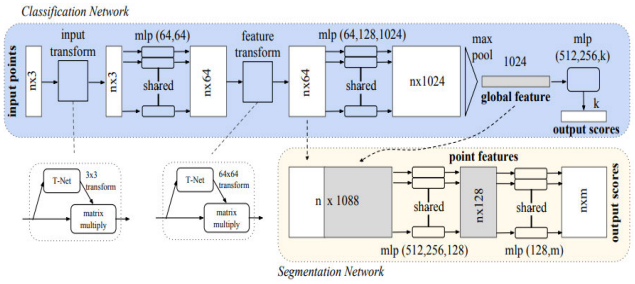

FIGURE 5. The framework of PointNet [36] for point cloud classification and segmentation Figure extracted from [36].

Aiming at improving the results of point cloud semantic segmentation, researchers begin to take actions to improve the algorithm on the basis of PointNet. The following section primarily summarizes the semantic segmentation methods based on raw point cloud in recent years from these four categories: methods of point ordering [38], [60], [61] methods of multi-scale [37], [62], [63] methods of feature fusion [43], [43], [67] and methods of fusing GCNN [45], [69], [70], [72].

1) METHODS OF POINT ORDERING [38], [60], [61]

The difficulty of point cloud semantic segmentation based on neural networks primarily lies in the irregularity and disorder of point cloud. Nowadays, there are a lot of network models designed to deal with such problems and those are paying off in point cloud semantic segmentation.

a: PointCNN

With respect to the disorder of point clouds, Li *et al.* [38] propose PointCNN, which performs quite well both in some complex datasets and challenging tasks. The key to the PointCNN model is the X-Conv operator. X transform is a group of weight X learned from the input points, and it can be used to re-weight and arrange the associated features of each point. Since X transform is learned from the input points, its weight may change with the order of the input points. Because the proposed model avoids the change of features with the order of input points, it almost remains unchangeable to the X-transformed features. The advantage of PointCNN is that convolution applying on the X-transformed features can greatly improve the utilization of convolution kernels and enhance the ability of convolution operations for extracting features on unordered data. The deficiency is that the X-transformation learnt from this network is not perfect and influences the point cloud segmentation. As is shown in Figure 6, (a) illustrates the application of hierarchical convolution on a regular grid and a point cloud; (b) is the structure of PointCNN for segmentation which is constructed by using the X-Conv operator.

b: RSNet

Huang *et al.* [60] propose a novel framework for 3D segmentation (RSNet), and Figure 6 is the overall structure of the network model. As is shown in the figure, it mainly

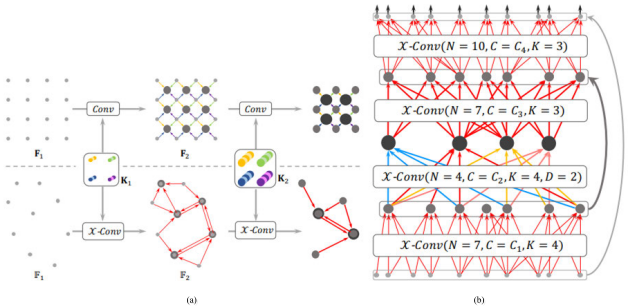

FIGURE 6. (a) Hierarchical convolution and (b) the framework of PointCNN [38] for semantic segmentation Figure extracted from [38].

consists of the slice pooling layer, the RNN layer and the slice de-pooling layer. The slice pooling layer is used to project irregular point features into feature vectors with regular order that apply to RNN. The RNN can be implemented to simulate the relevance between feature vectors. In addition, the slice de-pooling layer assigns features that in the sequence to the points to achieve the task of point cloud segmentation. The characteristic of this model is to extract the feature vectors of XYZ respectively, and output ordered feature sequence for post-processing. According to the description, this model relieves the influence of point cloud irregularity and achieves a better performance in point cloud semantic segmentation.

c: SO-Net

SO-Net [61] is a model possessing the characteristic of permutation invariance, which is used to simulate the spatial distribution of point clouds by constructing self-organizing mapping to fix the position of points and realize the efficient segmentation of point cloud. Moreover, to improve the network performance in various tasks, it proposes a point cloud auto-encoder as pre-training. However, because of the huge amount of point cloud data included in the large scene and the great complexity of the scene, there are a lot of limitations when the network is used to process point clouds. Figure 8 is a network structure of point cloud classification and segmentation formed by a self-organizing mapping (SOM).

Models that proposed based on point ordering, compare with the previous methods, can solve the disorder of point cloud and accelerate the speed of processing, and make larger contributions to the point cloud semantic segmentation. However, those methods may run into problems in other aspects, for example, the encoder is not powerful enough to capture fine-grained structures.

2) METHODS BASED ON MULTI-SCALE [37], [62], [63]

With the development of deep learning, researchers often use convolutional neural networks to extract the features of objects. Meanwhile, the receptive field becomes more and more crucial for the task of segmentation. If the receptive field is too small, only local features can be obtained. Otherwise, if the receptive field is too large, it contains much invalid information affecting the result. Therefore, researchers have been designing various multi-scale model architectures to get features to solve the problem.

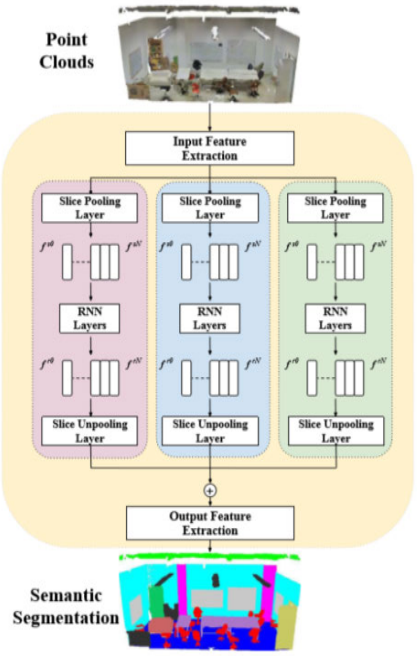

FIGURE 7. The overall framework of RSNet [60] for point cloud semantic segmentation Figure extracted from [60].

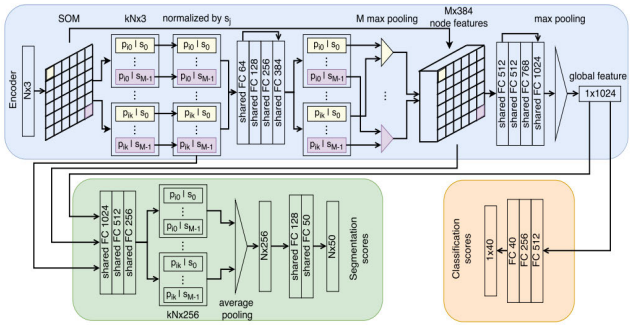

FIGURE 8. The point cloud semantic segmentation network of SO-Net [61] formed by SOM Figure extracted from [61].

a: PointNet++

Qi *et al.* [37], for the sake of improving the results of point cloud segmentation, introduce an upgraded version, which is called PointNet++. PointNet++ is made up of sampling layer, grouping layer and PointNet layer. The model firstly selects several points from input points as the centroid of the local areas by using FPS, then adds a local region grouping module based on the original network to construct local regions. Finally, PointNet is recursively used to extract local features. The framework of this network is shown in Figure 9. Although the model effectively solves the problem of extracting local features and enhances the results of segmentation, it still independently processes the points in the point set

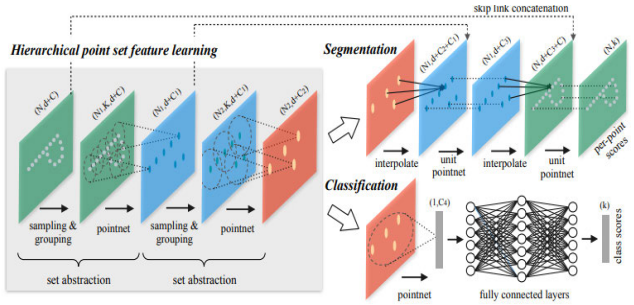

FIGURE 9. The architecture of PointNet++ [37] for point cloud classification and segmentation Figure extracted from [37].

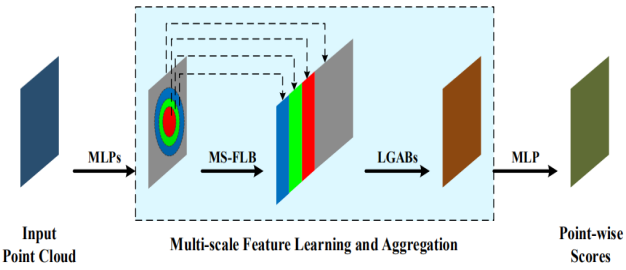

FIGURE 10. The network of 3D-MAXNet [62] for point cloud semantic segmentation (MS-FLB: Multi-scale feature learning block, LGAB: Local and global feature aggregation block) Figure extracted from [62].

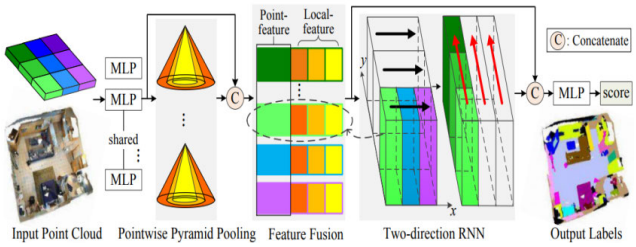

FIGURE 11. The framework of 3P-RNN [63] for point cloud semantic segmentation Figure extracted from [63].

and does not take into consideration the relationship between points, such as distance and direction.

b: 3DMAX-Net

3DMAX-Net [62] adopts the idea of multi-scale. Its structure is very simple and consists of two core parts (MS-FLB and LGAB). In this model, it firstly fuses the features learned at multiple scales, and then aggregates the local features and global features that merge to improve the accuracy of segmentation. MLP ultimately computes the score of each point to realize the task. Figure 10 is the network of 3DMAX-Net for semantic segmentation, in which MS-FLB is a multi-scale feature learning block, and LGAB is a block that aggregates the local features and the global features. The model can aggregate features learnt for different scales and reach a better performance.

c: 3P-RNN

Most methods are used to fuse the feature maps from the front and back layers, and it fails to fully obtain the spatial information, thus the result is not particularly well. To improve the performance of point cloud semantic segmentation comprehensively, Ye *et al.* [63] propose a pointwise pyramid pooling module, which can be utilized to aggregate the features of local neighborhoods at different scales. Meanwhile, the hierarchical two-direction recurrent neural networks (RNNs) are used to learn spatial context information to achieve the fusion of semantic features with multiple levels. Figure 11 shows the structure of this network. Besides, different from others, this method takes into account the spatial information and shows high accuracy in segmentation both on challenging indoor and outdoor 3D datasets.

Methods based on multi-scale are motivated by the knowledge of two-dimensional image processing, they can adjust the receptive field to extract the feature according to the scale of objects. Thus, no matter the size of the target object, it can be accurately segmented and capture fine features. Meanwhile, there are still some drawbacks that need researchers to be improved.

3) METHODS OF FEATURE FUSION [43], [43], [67]

Feature fusion [64]–[66], an important technology for semantic segmentation, combines the global features with local features acquired from the network to improve the performance of point cloud semantic segmentation. The model of PointNet, however, only extracts the global features of point cloud to achieve the semantic segmentation, without considering the characteristics of its local regions. In view of this deficiency, researchers have made many improvements.

a: PointNet

From the above analysis, we know that both PointNet and PointNet++ directly use the raw point clouds to extract feature, in order to achieve the understanding of 3D scenes. However, the description of shape features in point clouds also plays a crucial role in improving the results of point cloud segmentation. Inspired by the Scale-invariant Feature Transform (SIFT) used in 2D images, Jiang *et al.* [43] design the PointSIFT module, which can be embedded in the underlying network. This module encodes the information of eight main directions into a direction coding unit, and then stacks several direction coding units to get different features. In Figure 12, (a) is the structure of the PointSIFT module, and (b) is the overall network structure by embedding the PointSIFT module into the network to reach the semantic segmentation of point cloud. The model effectively introduces the knowledge of 2D image processing into 3D point cloud and obtains the local feature of the scene to reach the segmentation task.

b: A-CNN

A-CNN [43] which applies to the newly designed annular convolution in a hierarchical neural network is to achieve the semantic segmentation of large scenes. The function of that annular convolution is to extract the geometric features of the local neighborhood around each point. In addition, inspired

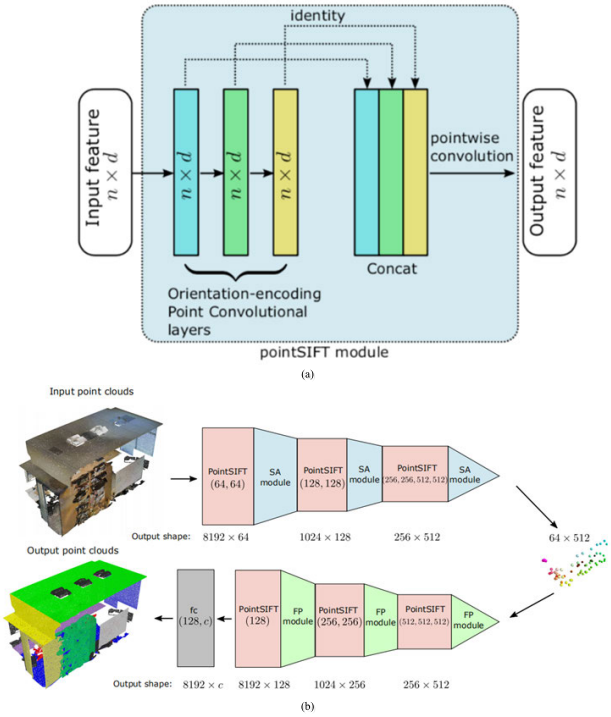

FIGURE 12. The mode of PointSIFT [43] ((a) the architecture of PointSIFT model and (b) the whole architecture of PointSIFT for point cloud segmentation) Figure extracted from [43].

by dilated convolution, the annular convolution proposed in the model also adopts the form of dilated rings to better capture the details of the object. And in the following step of processing the point cloud, the method of feature fusion is used to combine the global features with the local features to improve the result of segmentation.

c: SpiderCNN

Many methods based on feature fusion have also been proposed to process large-scale point cloud scenes, such as SpiderCNN [67], in addition to the models described above. This model consists of a unit called SpiderConv, which extends convolution operations on regular grids to embeddable irregular sets of points by parameterizing a series of convolution filters. Furthermore, it can effectively extract geometric features from point clouds in the scene.

According to the models that proposed by using feature fusion, we make a conclusion that the method of deep learning can get local and global features of different scenes, then fuse them to improve the result of segmentation. This method solves the problems caused by indirect methods and has significant advantages over the traditional method based on artificial features.

4) METHODS OF FUSING GCNN [45], [69], [70], [72]

Graph is a type of structured data composed of a series of nodes and edges. Nowadays, graph convolutional neural network [68] (GCNN) is commonly used in the field of computer vision, which operates directly on the graph structure and can capture the dependencies of graph by transferring the information between the nodes.

a: DGCNN

Wang *et al.* [69] first applied GCNN to the process of point cloud and combined it with PointNet to realize the semantic segmentation of point cloud. DGCNN is inspired by graph CNN, however, the most significant difference is that the graph constructed is dynamic and updates after each layer of the network. In [69], an edge convolution operation is mainly designed to extract the feature of center points. Meanwhile, it can obtain the edge vector of the center points and the K nearest neighbor (KNN) points. Not only that, the architecture of this network is almost similar to PointNet, DGCNN only replaces the multi-layer perceptions that stacked with edge convolution. This algorithm searches the neighborhoods in Euclidean space, as well as clusters analogous features in the feature space, so it has a significant effect in the task of point cloud classification and segmentation.

b: LDGCNN

In the model DGCNN, introducing a space-transformed network increases the complexity of the network, and the parameters for training in the network also increase accordingly. On the basis of DGCNN, Zhang *et al.* [70] adopted the network structure of DenseNet [71] to modify the original model and proposed LDGCNN to deal with the aforementioned problems. The basic idea of this model is to connect the hierarchical features extracted from different dynamic graphs, and replace the transformation network with MLP. This method effectively avoids the problem of gradient disappearance, reduces the size of the network, and achieves a superior semantic segmentation result on the representative dataset of the point cloud. Figure 13 shows the network structure of LDGCNN for point cloud classification and segmentation.

c: RGCNN

Another way to achieve the point cloud segmentation by using GCNN is RGCNN [72]. This network is composed of three regular graph convolutional layers, each of which contains graph construction, graph convolution and feature filtering. The purpose of adaptively capturing the dynamic graph structure is achieved by designing the graph Laplacian matrix describing the inter-layer feature relationships. At the same time, the matrix will be updated continuously according to the relevant features that learned. The model fusing GCNN not only addresses the problem of permutation invariance of point clouds, but also has strong robustness to noise and density in point clouds.

d: GAPNet

The problem that how to properly use visual information to process resources, and obtain more suitable results for human perception has become an active research topic. To overcome that, researchers proposed an attention mechanism which has

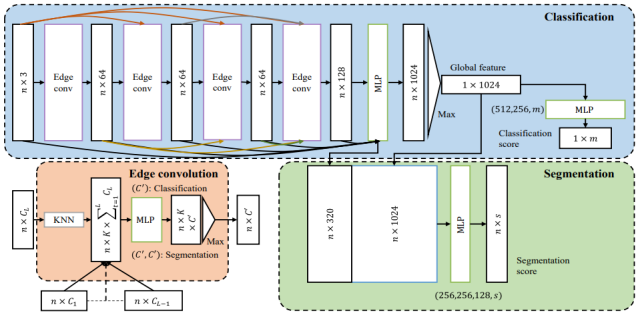

FIGURE 13. The network structure diagram of LDGCNN [70] for point cloud semantic segmentation Figure extracted from [70].

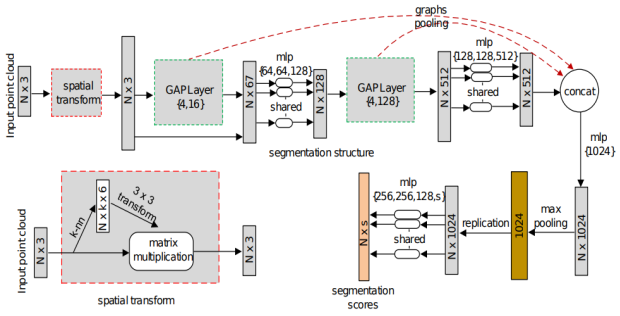

FIGURE 14. The network framework of GAPNet [45] for point cloud semantic segmentation Figure extracted from [45].

two essential aspects: selecting a specific part of the visual area as input; focusing on and allocating limited resources for processing the information to critical areas. Nowadays, this technology is gradually mature and it is combined with GCNN for the segmentation of point cloud. GAPNet [45] is a new neural network of point cloud, which embeds a graphical attention mechanism into multi-layer perceptions that stacked to learn the local geometric information. The structure of this network is shown in Figure 14 and it is similar to PointNet. The crucial distinction is that the GAPLayer is introduced to learn the attention characteristics of each point by highlighting the different attention weight in the neighborhood. Moreover, in order to provide sufficient features for the model, the multi-head mechanism is added to aggregate the features acquired from different GAPLayers. In the picture, the numbers below GAPLayers represent the number of heads and coding feature channels respectively. Finally, to achieve the effective segmentation of point cloud, GAPNet applies the stacked MLP layer in the attention feature and the local signature to fully extract the information of local geometry. The model firstly adds the attention mechanism in the human visual system to the point cloud segmentation and segments the scene well.

GCNN is now widely used in the point cloud to accomplish the semantic segmentation, and achieves several wonderful segmentation results. Methods based on GCNN, compared with other methods, not only examine the relationship between points, but also get the boundary feature. Therefore, it has numerous advantages in point cloud segmentation, but there are many aspects of other tasks that need to be improved.

So far, researchers have proposed many kinds of models based on raw point cloud for understanding 3D scenes, especially the emergence of deep learning. Although this segmentation method processing the point cloud directly has reached a better result for the segmentation of point clouds, it needs a large amount of data for training and the requirement for GPU's computing ability is higher.

III. RELATED DATASETS OF POINT CLOUD SEMANTIC SEGMENTATION

To verify the effect of the algorithms that are proposed on the semantic segmentation of point cloud, a valid dataset is especially critical. Owing to the rise of deep learning, we often construct a deep neural network and train the network to accomplish some tasks. Then, if the network model is deeper and more complex, huge training data is required to make the model effective, so the dataset plays an indispensable role during the model training. The dataset should not only be effective, but also contains rich and varied data. Only in that way, the model can be well trained, thus guaranteeing the results for subsequent processing. However, creating a large and efficient dataset demands a lot of manpower, material and financial resources. Some research institutions have provided several reliable open datasets, such as: ShapeNet, S3DIS, ScanNet, etc., to promote the study of point cloud semantic segmentation. In the next section, we will briefly describe the datasets which are commonly used for the segmentation of point cloud.

- PartNet [73]: PartNet is a large 3D dataset annotated with fine-grained, instance-level and part of 3D hierarchical information. It contains 573,585 partial instances of 24 different object categories and approximately 26,671 3D models. This dataset can be applied to many tasks such as shape synthesis, dynamic modeling of the 3D scene and feasibility analysis.
- UWA Dataset [74]: The dataset which possesses 50 different scenes scanned by Minolta scanner is mainly used in the target recognition and segmentation of chaotic scene based on a 3D model, and each scene contains 4 to 5 objects that randomly placed. In addition, it can be obtained from Point Cloud Library (PCL) or 3D key point detection benchmarks.
- ShapeNet Part [75]: It is a large scale of 3D shape dataset with rich annotations and often used for 3D object part segmentation. This dataset contains 16881 shapes from 16 categories, with a total of 50 parts that labeled, where each object typically has 2 to 5 markers.
- S3DIS [76]: The Stanford Large-scale 3D Indoor Spaces Dataset (S3DIS) that obtained by the Matterport scanner is a widely used dataset, which contains 272 3D room scenes in 6 regions for semantic segmentation, where each point in the scene is represented by a semantic label in one of the 13 categories (chair, table, wall, etc.). As shown in Figure 15, there are some scenes in this dataset and different color annotations in those pictures indicate different categories.

FIGURE 15. Point cloud scene and semantic segmentation diagram in S3DIS [76] dataset Figure extracted from [76].

FIGURE 16. Annotated indoor scene map in ScanNet [77] dataset Figure extracted from [77].

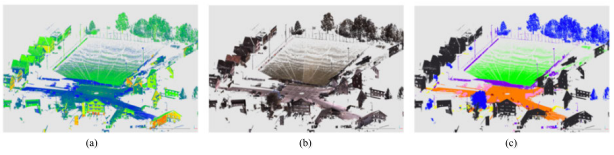

FIGURE 17. Pictures related to Semantic3D [78] dataset ((a) Point cloud scene, (b) intensity diagram and (c) semantic segmentation diagram) Figure extracted from [78].

• ScanNet [77]: ScanNet that is a RGB-D video dataset possesses 2.5 million views scanned from 1,513 3D indoor scenes with a total of 21 semantic categories. Furthermore, the dataset contains the information of XYZ and label, but lacks the information of color. It is common to divide this dataset into two categories: 1201 scenes are used for training and 312 scenes are used for testing. Figure 16 is the result of several indoor scenes marked in this dataset.

• Semantic3D [78]: This is currently the largest available Lidar dataset that consists of eight semantic categories covers a wide range of urban outdoor scenes: churches, streets, railway, squares, villages, football fields and castles with more than 3 billion points. Each point in the scene has RGB and the value of intensity. Then, fifteen scenarios in this dataset are used for training, the remaining fifteen scenarios are for testing. In Figure 17, we respectively show a point cloud scene, its intensity diagram and the result of semantic segmentation in Semantic3D.

• vKITTI [79]: The vKITTI dataset is a large outdoor dataset that is simulated from the KITTI dataset having real-world scenes, which contains 13 semantic categories in the urban scene. Through projection, researchers can map the two-dimensional semantic label into the three-dimensional space to get the annotated point cloud, and the semantic segmentation results of outdoor scenes are presented in Figure 18.

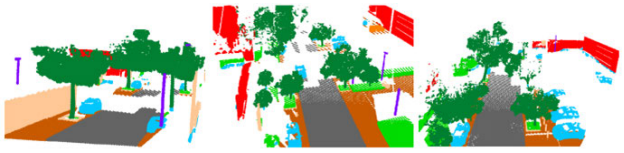

FIGURE 18. Semantic segmentation results of outdoor scenes in vKITTI [79] dataset Figure extracted from [79].

TABLE 1. Common datasets of point cloud segmentation.

Datasets	Number of categories	Number of training sets	Number of verification sets	Number of testing sets
PartNet [73]	24	----	----	----
UWA Dataset [74]	55	----	----	----
ShapeNet Part [75]	16	12137	1870	2874
S3DIS [76]	13	224	----	48
ScanNet [77]	21	1201	----	312
Semantic3D [78]	8	15	----	15
KITTI(Zhang) [81]	10	140	----	112
KITTI(Ros) [82]	11	170	----	46
vKITTI [79]	13	----	----	----

• KITTI Raw [80]: This dataset collected by Velodyne Lidar is a sparse and colorless point cloud, which is commonly applied in the domain of mobile robots and autonomous driving. KITTI Raw cannot be used for supervised training, because its semantic label lacks authenticity. However, its density is consistent with that of vKITTI, and it can be used for generalized verification of experiments. In addition, researchers also implement their requirements by manually labeling datasets, in order to implement this dataset to semantic segmentation successfully. Zhang *et al.* [81] annotate 252 images into 10 object categories, of which 140 images are used for training and 112 are for testing. While, Ros *et al.* [82] mark 216 images of 11 categories (170 training images and 46 testing images).

Table 1 shows the common datasets for point cloud segmentation and provides the number of categories and the number of training/validation/testing datasets. Furthermore, this paper pays more attention to the results of point cloud semantic segmentation on S3DIS, ShapeNet, ScanNet and other datasets.

IV. ANALYSIS OF EXPERIMENTAL RESULTS

In the previous section, this paper starting from the representation of 3D data mainly reviews the methods of semantic segmentation based on raw point cloud, which does not take any quantitative results into account. To make a significant contribution to the semantic segmentation, the performance of the designed model must be quantitatively evaluated. Besides, in order to fully reflect the fairness of the assessment and the effectiveness of the model, it is necessary to use a variety of standards from different aspects and some well-known evaluation indicators for evaluation. In this section, we analyze the existing methods according to the result data. Firstly, the performance of the existing semantic segmentation

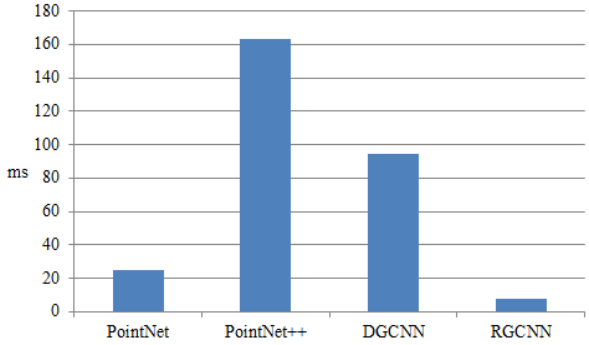

FIGURE 19. Forward time of different network models.

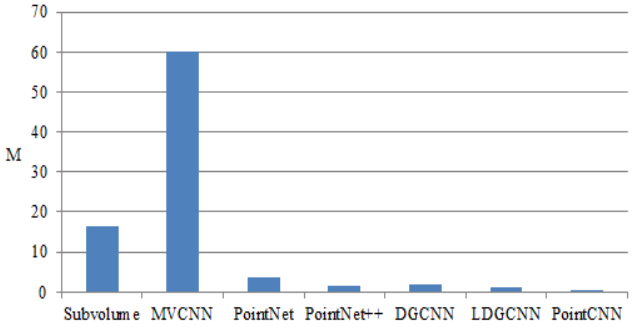

FIGURE 20. Network parameters of different semantic segmentation models.

models is measured by the evaluation indicators of execution time, space complexity and accuracy. Furthermore, the above metrics which are obtained from the segmentation results using different models on the most representative datasets (S3DIS, ScanNet, ShapeNet, etc.) are collected. Finally, we make a summary of the segmentation results and draw conclusions.

A. EXECUTION TIME

Execution time is an important and valuable measurement index that is used to evaluate the performance of the model, and especially with the development of deep learning and convolutional neural networks, it becomes more and more important. The processing performance of the network can be effectively judged by the training time of the model, but this indicator is heavily dependent on the hardware, and sometimes it makes no sense for comparison. However, in order to help researchers to explore in-depth, the hardware of the execution system needs to be set out in detail. If it is properly utilized, the performance of the model can be effectively detected, and the training time of the model with different segmentation methods can be fairly compared under the same condition of hardware, and the training speed can also be tested. Owning to the different equipment used during training, it is hard to analyze them comprehensively. Figure 19 is the forward time estimated under four different models based on the computing power of an NVIDIA GPU with the same type.

B. SPACE COMPLEXITY

Space complexity is a measurement index, which refers to the storage space temporarily occupied by an algorithm during the running process of the procedure. This indicator, when used to test the performance of the deep learning model, means the quantity of parameters that the model needs. We always hope that the parameters of the model are fewer when building a model to accomplish different tasks. In an algorithm, its time complexity and space complexity often interact. If we pursue a better time complexity blindly, the performance of spatial complexity must be degraded, which means it may result in occupying more storage space; and vice versa. Therefore, when we design the model, both the time and space complexity should be taken into consideration, after that the optimal model can be designed. Figure 20 summarizes the number of parameters of different point cloud semantic segmentation models (other models are not introduced here in some detail because the original paper does not compare the quantity of parameters).

C. ACCURACY

Nowadays, many evaluation criteria are proposed to evaluate the accuracy of segmentation for different semantic segmentation models. Among them, the mean Intersection over Union (mIoU) and overall accuracy (OA) are the two most important indicators for evaluating the result of point cloud semantic segmentation.

- mIoU: mIoU is an effective indicator for checking the accuracy of segmentation. The IoU mainly calculates the ratio between the intersection and the union of two sets, which in the segmentation refers to the overlap ratio between the real area and the predicted area. While the mIoU is the calculation of the IoU based on each category and then takes the average. IoU can be calculated by the following equation:

$$IoU = \frac{TP}{T + P - TP}$$

TP: the number of true positives; T: the number of ground true positive samples; P: the number of predicted positives belonging to that class.

- OA: It is one of the simplest metrics, simply calculating the probability that the semantic annotation result of each random sample is consistent with the annotation type of real data.

Table 2, Table 3, Table 4, Table 5 and Table 6 illustrate the performance of point cloud semantic segmentation based on deep learning in many representative datasets. Since the original paper, which the model selected in this paper, does not conduct a comprehensive experiment on the accuracy and mIoU of the following five datasets, and some articles' codes are not open source, we only summarize the experimental results obtained from the original paper in those tables. Then, we conduct an in-depth comparative analysis of experimental results based on the same dataset, according to the different properties between datasets.

TABLE 2. Segmentation results of different models on ShapeNet part dataset.

	Yi[83]	KD-Net[54]	SEGCloud[56]	PointNet[36]	PN++(MSG+DP)[37]	O-CNN+CRF[84]	SSCNN[85]	
mIoU(%)	81.4	82.3	79.4	83.7	85.1	85.9	84.7	
	DGCNN[69]	RGCNN[72]	RSNet[60]	SO-Net[61]	SpiderCNN[67]	LDGCNN[70]	GAPNet[45]	SGPN[86]
mIoU(%)	85.1	84.3	84.9	84.6	85.3	85.1	84.7	85.8

TABLE 3. Segmentation results of different models on S3DIS dataset.

	SEGCloud[56]	PointNet[36]	MS+CU[87]	G+RCU[87]	DGCNN[69]	RSNet[60]	SGPN[86]
mIoU(%)	48.92	47.71	47.8	49.7	56.1	53.83	50.37
OA(%)	----	78.62	79.2	81.1	84.1	----	80.78
	3DMAX-Net[62]	SPGraph[88]	3P-RNN[63]	PointCNN[38]	PointSIFT[43]	ASIS[89]	A-CNN[43]
mIoU(%)	47.5	62.1	56.3	62.74	70.23	59.3	----
OA(%)	79.5	85.5	86.9	88.1	88.72	86.2	87.3

TABLE 4. Segmentation results of different models on ScanNet dataset.

	SEGCloud[56]	PointNet[36]	PN++[37] (SSG)	PN++[37] (MSG+DP)	PN++[37] (MRG+DP)	RSNet[60]	PointCNN[38]	PointSIFT[43]
mIoU(%)	----	14.69	----	34.26	----	39.35	----	----
OA(%)	73.0	73.9	83.3	84.5	83.4	----	85.1	86.2

TABLE 5. Segmentation results of different models on Semantic3D dataset.

	SEGCloud[56]	TMLC-MSR[90]	DeePr3SS[91]	SnapNet[45]	SPGraph[88]
mIoU(%)	61.3	54.2	58.5	59.1	73.2
OA(%)	88.1	86.2	88.9	88.6	94.0

TABLE 6. Segmentation results of different models on vKITTI dataset.

	PointNet[36]	3P-RNN[63]	G+RCU[87]
mIoU(%)	34.4	41.6	36.2
OA(%)	79.7	87.8	80.6

In Table 2, we collect the segmentation results of different models on ShapeNet Part dataset. This dataset is commonly applied to object part segmentation and contains many shapes. Furthermore, due to its specification, lots of papers just evaluate the value of mIoU. According to the results, we see that O-CNN+CRF outperform other models, because it combines deep learning methods with traditional methods, and processes the point cloud directly without transformation.

In the above two tables, we list the experimental results of two indoor scene datasets (S3DIS and Scannet) on different models. Semantic segmentation of indoor scenes is challenging, because indoor scenes contain more objects and some of them attaching to each other (like board and wall) are difficult to segment. Thus, when we segment such scenes, we must consider the relationship between adjacent objects and the overall properties of the scene. Common global features include color features, texture features and shape features, while those features are not suitable to image aliasing and occlusion. We need to take into account the features extracted from the local area of the image to achieve better semantic segmentation results. From two tables, PointSIFT fusing local features and global features shows strong performance and even better than state-of-the-art. We can conclude

that methods of feature fusion are popular with indoor complex scene datasets.

For outdoor scenes, similar to indoor scenes, there are many things in the scene, but those are complex and variable. In addition, the forms of point cloud collected by different methods also make the implementation of point cloud semantic segmentation difficult. Therefore, researchers seldom experiment on outdoor datasets and the experimental results are not rich. The following two tables are the results obtained on the two outdoor datasets Semantic3D and vKITTI. From the statistics in those tables, we know that methods based on multi-scale are outstanding for point cloud semantic segmentation, because they not only take into account low-level semantic information but also high-level semantic information, meanwhile, solving the problem of information loss.

The analysis of execution time, space complexity and accuracy of point cloud semantic segmentation shows that different methods have their own advantages and disadvantages. Although the improved model based on PointNet performs well in the segmentation, its network model is complex, resulting in long execution time. For the space complexity, compared with the semantic segmentation methods based on multi-view and voxel, the semantic segmentation model using the raw point clouds has fewer parameters and reaches a better result. Moreover, with the continuous improvement of algorithms and network models, the accuracy of point cloud segmentation on different datasets is also getting increasingly higher.

V. DISCUSSION

The understanding and analysis of 3D scenes is the key to the research of unmanned driving, smart cities, smart medical treatment and other fields [92]–[94]. Semantic segmentation, as the basis of 3D scene understanding, is the core of future research. With the continuous development of deep learning, this technology has been extensively used in 3D point clouds. And the semantic segmentation based on 3D point clouds takes significant advantages of its rich data. But, there are also some challenges to the semantic segmentation of point cloud now. These challenges range from problems based on the point clouds itself to challenging issues resulting from the task of semantic segmentation. This paper analyzes and summarizes the 3D point cloud semantic segmentation technology in recent years and the following aspects need some further research:

(1) Given the disorder and irregularity of point clouds, when using the neural network to achieve the point cloud semantic segmentation, researchers need to transform it at first, such as the performance of voxel and multi-view. However, this kind of method that transforms the data into a regular structure inevitably increases the calculation of the algorithm, and leads to the loss of some valid information at the same time. To improve the accuracy of point cloud semantic segmentation, constructing a semantic segmentation model based on the raw point cloud by using the technology of deep learning is an important research direction in the future.

(2) Although the semantic segmentation of point cloud based on deep learning has achieved an excellent result, it requires a large amount of data when training the model. However, the collection of datasets consumes not only a huge number of human resources, but also strong funds. Therefore, collecting a dataset with abundant and efficient data is the primary condition for semantic segmentation.

(3) There are rich and complicated things contained in the outdoor scene, thus it is necessary to consider various factors when using neural networks for segmentation. Therefore, most of the network models that proposed at present are used to solve the problem of semantic segmentation of indoor scenes, rarely involving the segmentation of outdoor scenes. So, future research on the semantic segmentation of point cloud should pay more attention to constructing a network model suitable for processing outdoor scenes.

(4) Instance segmentation is one of the challenging tasks in computer vision, which is a combination of target detection and semantic segmentation. Nowadays, the model of point cloud semantic segmentation that proposed is mostly used to segment the same kind of objects, and rarely to separate different individuals in the same category. In the future, with the development of unmanned driving, environmental awareness and other fields, how to design a deep learning model to achieve instance segmentation of point cloud has broad prospects for development.

(5) To the best of our knowledge, the current semantic segmentation methods of point cloud are only based on 3D data. While, the semantic segmentation of two-dimensional images has been quite mature and is easy to be realized. Therefore, the fusion of 2D image and 3D point cloud is the general trend in the future, which can improve the effect of semantic segmentation.

VI. CONCLUSION

This paper, focus on deep learning technology, presents a comprehensive survey of existing point cloud semantic segmentation methods. Firstly, we review the deep learning-based point cloud semantic segmentation models from two perspectives: indirect way and direct way. Moreover, several special models among two categories are introduced carefully. We then describe popular datasets for point clouds, and analyze the results of different models. In the end, we discuss the challenges of existing point cloud semantic segmentation using deep learning methods, and provide insight for future research directions.

In conclusion, point cloud semantic segmentation has been approached with many superior performances thanks to the development of deep learning techniques, but it still remains problems for improvement. We expect our paper to present a detailed summary to understand state-of-the-arts and, insights for future research in point cloud semantic segmentation.

REFERENCES

[1] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun and Y. Tang, "Methods and datasets on semantic segmentation: A review," *Neurocomputing*, vol. 304, pp. 82–103, Aug. 2018.

[2] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*. [Online]. Available: https://arxiv.org/abs/1512.03012

[3] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Jun. 2017, pp. 3357–3364, doi: 10.1109/ICRA.2017.7989381.

[4] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 918–927.

[5] A. Nguyen and B. Le, "3D point cloud segmentation: A survey," in *Proc. 6th IEEE Conf. Robot., Autom. Mechatronics (RAM)*, Nov. 2013, pp. 225–230.

[6] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, Sept. 2018.

[7] S. Arshad, M. Shahzad, Q. Riaz, and M. M. Fraz, "DPRNet: Deep 3D point based residual network for semantic segmentation and classification of 3D point clouds," *IEEE Access*, vol. 7, pp. 68892–68904, 2019.

[8] Y. Ishikawa, R. Hachiuma, N. Lenaga, W. Kuno, Y. Sugiura and H. Saito, "Semantic segmentation of 3D point cloud to virtually manipulate real living space," in *Proc. 12th Asia–Pacific Workshop Mixed Augmented Reality (APMAR)*, Mar. 2019, pp. 1–7.

[9] B. Liu, S. He, D. He, Y. Zhang, and M. Guizani, "A spark-based parallel fuzzy c-means segmentation algorithm for agricultural image big data," *IEEE Access*, vol. 7, pp. 42169–42180, Mar. 2017.

[10] Q. Zhan, L. Yu, and Y. Liang, "A point cloud segmentation method based on vector estimation and color clustering," in *Proc. 2nd Int. Conf. Inf. Sci. Eng.*, Dec. 2010, pp. 3463–3466.

[11] R. Schnabel, R. Wahl, and R. Klein, "Efficient RANSAC for point-cloud shape detection," *Comput. Graph. Forum*, vol. 26, pp. 1–12, May 2007.

[12] B. Yang and Z. Dong, "A shape-based segmentation method for mobile laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 81, pp. 19–30, Jul. 2013.

[13] E. Che and M. J. Olsen, "Fast ground filtering for TLS data via scanline density analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 129, pp. 226–240, Jul. 2018.

[14] A.-V. Vo, L. Truong-Hong, D. F. Laefer, and M. Bertolotto, "Octree-based region growing for point cloud segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 88–100, Jun. 2015.

[15] L. Huang, W. Li, Q. Yang and Y. Chen, "Segmentation algorithm of three-dimensional point cloud data based on region growing," *Appl. Mech. Mater.*, vol. 741, pp. 382–385, Mar. 2015.

[16] A. Golovinskiy and T. Funkhouser, "Min-cut based segmentation of point clouds," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2009, pp. 39–46.

[17] S. Ural and J. Shan, "Min-cut based segmentation of airborne lidar point clouds," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XXXIX-B3, pp. 167–172, Sep. 2012.

[18] J. Yan, J. Shan, and W. Jiang, "A global optimization approach to roof segmentation from airborne lidar point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 94, pp. 183–193, Aug. 2014.

[19] T. Rabbani, F. Van Den Heuvel, and G. Vosselmann, "Segmentation of point clouds using smoothness constraint," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 36, no. 5, pp. 248–253, 2006.

[20] M. A. Wani and H. R. Arabnia, "Parallel edge-region-based segmentation algorithm targeted at reconfigurable multiring network," *J. Supercomput.*, vol. 25, pp. 43–62, May 2003.

[21] E. Castillo, J. Liang, and H. Zhao, "Point cloud segmentation and denoising via constrained nonlinear least squares normal estimates," in *Innovations for Shape Analysis*. Berlin, Germany: Springer, Nov. 2012, pp. 283–299.

[22] J.-F. Lalonde, R. Unnikrishnan, N. Vandapel, and M. Hebert, "Scale selection for classification of point-sampled 3D surfaces," in *Proc. 5th Int. Conf. 3-D Digit. Imag. Modeling (3DIM)*, Jun. 2005, pp. 285–292.

[23] J. Zhang, X. Lin, and X. Ning, "SVM-based classification of segmented airborne lidar point clouds in urban areas," *Remote Sens.*, vol. 5, no. 8, pp. 3749–3775, Jul. 2013.

[24] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Conditional random fields for lidar point cloud classification in complex urban areas," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vols. I–3, pp. 263–268, Sep. 2012.

[25] Y. Lu and C. Rasmussen, "Simplified Markov random fields for efficient semantic labeling of 3D point clouds," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 2690–2697.

[26] M. Hatt, C. Parmar, J. Qi, and I. E. Naqa, "Machine (deep) learning methods for image processing and radiomics," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 104–108, Mar. 2019.

[27] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018.

[28] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1457–1470, May 2017.

[29] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[30] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Sep. 2017.

[31] Z. Zhiqiang, L. Zhihui, C. De, Z. Huaxzhang, Z. Kun, and Y. Yi, "Two-stream multi-rate recurrent neural network for video-based pedestrian re-identification," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3179–3186, 2017.

[32] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[33] Z. Li, X. Cai, Y. Liu, and B. Zhu, "A novel Gaussian-Bernoulli based convolutional deep belief networks for image feature extraction," *Neural Process. Lett.*, vol. 49, pp. 305–319, Feb. 2018.

[34] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.

[35] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Dec. 2015, pp. 922–928.

[36] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vision. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[37] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst. 30 (NIPS)*, 2017, pp. 5105–5114.

[38] Y. Li, R. Bu, M. Sun, W. Wu, X. Di and B. Chen, "PointCNN: Convolution On X-transformed points," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*. Dec. 2018, pp. 828–838.

[39] S. K. Lodha, D. M. Fitzpatrick, and D. P. Helmbold, "Aerial lidar data classification using AdaBoost," in *Proc. 6th Int. Conf. 3-D Digit. Imag. Modeling (3DIM)*, Aug. 2007, pp. 435–442.

[40] N. Chehata, L. Guo, and C. Mallet, "Airborne lidar feature selection for urban classification using random forests," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 207–212, Sep. 2009.

[41] R. Shapovalov, E. Velizhev, and O. Barinova, "Nonassociative Markov networks for 3d point cloud classification," in *Proc. Photogramm. Comput. Vis. Image Anal. (PCV)*, vol. 3, Jan. 2010, pp. 103–108.

[42] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, Feb. 2012.

[43] M. Jiang, Y. Wu, T. Zhao, Z. Zhao, and C. Lu, "PointSIFT: A SIFT-like network module for 3D point cloud semantic segmentation," 2018, *arXiv:1807.00652*. [Online]. Available: https://arxiv.org/abs/1807.00652

[44] A. Komarichev, Z. Zhong, and J. Hua, "A-CNN: Annularly convolutional neural networks on point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7421–7430.

[45] C. Chen, L. Z. Fragonara, and A. Tsourdos, "GAPNet: Graph attention based point neural network for exploiting local feature of point cloud," 2019, *arXiv:1905.08705*. [Online]. Available: https://arxiv.org/abs/1905.08705

[46] A. Boulch, B. Le Saux, and N. Audebert, "Unstructured point cloud semantic labeling using deep segmentation networks," in *Proc. Eurograph. Workshop 3D Object Retr.*, vol. 2, 2017, pp. 17–24.

[47] J. Guerry, A. Boulch, B. Le Saux, J. Moras, A. Plyer, and D. Filliat, "SnapNet-R: Consistent 3D multi-view semantic labeling for robotics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 669–678.

[48] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.
[49] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.
[50] B. Graham, "Spatially-sparse convolutional neural networks," *Comput. Sci.*, vol. 34, pp. 864–866, Sep. 2014.
[51] F. Verdoja, D. Thomas, and A. Sugimoto, "Fast 3D point cloud segmentation using supervoxels with geometry and color for 3D scene understanding," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Aug. 2017, pp. 1285–1290.
[52] Y. Li, S. Pirk, H. Su, C. R. Oi, and L. J. Guibas, "FPNN: Field probing neural networks for 3D data," in *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, Oct. 2016, pp. 307–315.
[53] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3577–3586.
[54] R. Klokov and V. Lempitsky, "Escape from cells: Deep Kd-networks for the recognition of 3D point cloud models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 863–872.
[55] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs," in *Proc. IEEE Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2088–2096.
[56] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "SEGCloud: Semantic segmentation of 3D point clouds," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 537–547, doi: 10.1109/3DV.2017.00067.
[57] T. Le and Y. Duan, "PointGrid: A deep network for 3D shape understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 9204–9214.
[58] B.-S. Hua, M.-K. Tran, and S.-K. Yeung, "Pointwise convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 984–993.
[59] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2015, pp. 2017–2025.
[60] Q. Huang, W. Wang, and U. Neumann, "Recurrent slice networks for 3D segmentation of point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2626–2635.
[61] J. Li, B. M. Chen, and G. H. Lee, "SO-net: Self-organizing network for point cloud analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 9397–9406.
[62] Y. Ma, Y. Guo, Y. Lei, M. Lu, and J. Zhang, "3DMAX-net: A multi-scale spatial contextual network for 3D point cloud semantic segmentation," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1560–1566, doi: 10.1109/ICPR.2018.8546281.
[63] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3D recurrent neural networks with context fusion for point cloud semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 403–417.
[64] R. Zhang, G. Li, M. Li, and L. Wang, "Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 143, no. 9, pp. 85–96, Sep. 2018.
[65] A. Liu, Y. Yang, Q. Sun, and Q. Xu, "A deep fully convolution neural network for semantic segmentation based on adaptive feature fusion," in *Proc. 5th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2019, pp. 16–20, doi: 10.1109/ICISCE.2018.00013.
[66] Y. Li, G. Tong, X. Li, L. Zhang, and H. Peng, "MVF-CNN: Fusion of multilevel features for largr-scale point cloud classification," *IEEE Access*, vol. 7, pp. 46522–46537, Apr. 2019.
[67] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao, "SpiderCNN: Deep learning on point sets with parameterized convolutional filters," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 87–102.
[68] Y. Zhang and M. Rabbat, "A graph-CNN for 3D point cloud classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6279–6283, doi: 10.1109/ICASSP.2018.8462291.
[69] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," 2018, *arXiv:1801.07829*. [Online]. Available: https://arxiv.org/abs/1801.07829
[70] K Zhang, M. Hao, J. Wang, C. W. de Silva, and C. Fu, "Linked dynamic graph CNN: Learning on point cloud via linking hierarchical features," 2019, *arXiv:1904.10014*. [Online]. Available: https://arxiv.org/abs/1904.10014
[71] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[72] G. Te, W. Hu, A. Zheng, and Z. Guo, "RGCNN: Regularized graph CNN for point cloud segmentation," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 746–754.
[73] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A Large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 909–918.
[74] A. S. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1584–1601, Oct. 2006.
[75] L. Yi *et al.*, "Large-scale 3D shape reconstruction and segmentation from ShapeNet core55," 2017, *arXiv:1710.06104*. [Online]. Available: https://arxiv.org/abs/1710.06104
[76] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3D semantic parsing of large-scale indoor spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1534–1543.
[77] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5828–5839.
[78] T. Hackel, N. Savinov, L. Ladicky, J. D. Wegner, K. Schindler and M. Pollefeys, "Semantic3D. net: A new large-scale point cloud classification benchmark," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 91–98, Apr. 2017.
[79] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4340–4349.
[80] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, Aug. 2013.
[81] R. Zhang, S. A. Candra, K. Vetter, and A. Zakhor, "Sensor fusion for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jul. 2015, pp. 1850–1857.
[82] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. M. Lopez, "Vision-based offline-online perception paradigm for autonomous driving," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Feb. 2015, pp. 231–238.
[83] L. Yi, V. G. Kim, D. Ceylan, I. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer and L. Guibas, "A scalable active framework for region annotation in 3D shape collections," *Acm Trans. Graph.*, vol. 35, no. 210, pp. 1–12, Nov. 2016.
[84] P. Wang, Y. Liu, Y. Guo, C. Sun, and X. Tong, "O-CNN: Octree-based convolutional neural networks for 3D shape analysis," *Acm Trans. Graph.*, vol. 36, no. 72, pp. 1–11, Jul. 2017.
[85] L. Yi, H. Su, X. Guo, and L. J. Guibas, "SyncSpecCNN: Synchronized spectral CNN for 3D shape segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2282–2290.
[86] W. Wang, R. Yu, Q. Huang, and U. Neumann, "SGPN: Similarity group proposal network for 3D point cloud instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 2569–2578.
[87] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe, "Exploring spatial context for 3D semantic segmentation of point clouds," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 716–724.
[88] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4558–4567.
[89] X. Wang, S. Liu, X. Shen, C. Shen, and J. Jia, "Associatively segmenting instances and semantics in point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4096–4105.
[90] T. Hackel, J. D. Wegner, and K. Schindler, "Fast semantic segmentation of 3D point clouds with strongly varying density," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3 pp. 177–184, Jul. 2016.
[91] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3D semantic segmentation," in *Proc. Int. Conf. Comput. Anal. Images Patterns (CAIP)*, vol. 10424, Jul. 2017, pp. 95–107.
[92] M. Chen, Q. Zou, C. Wang, and L. Liu, "EdgeNet: Deep metric learning for 3D shapes," *Comput. Aided Geometric Des.*, vol. 72, pp. 19–33, Jun. 2019.
[93] W. Shi, W. Ahmed, N. Li, W. Fan, H. Xiang, and M. Wang, "Semantic geometric modelling of unstructured indoor point cloud," *ISPRS Int. J. Geo-Inf.*, vol. 8, pp. 1–20, Jan. 2019, doi: 10.3390/ijgi8010009.
[94] J. Chen, Z. Kira, and Y. K. Cho, "Deep learning approach to point cloud scene understanding for automatched scan to 3D reconstruction," *J. Comput. Civil Eng.*, vol. 33, pp. 1–22, May. 2019, doi: 10.1061/(ASCE)CP.1943-5487.0000842.

JIAYING ZHANG is currently pursuing the master's degree with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. Her research interests include computer vision, digital image processing, and 3D point cloud processing.

XIAOLI ZHAO is currently an Associate Professor with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. Her research interests include video and image processing, pattern recognition, computer vision, and intelligent computing.

ZHENG CHEN is currently pursuing the master's degree with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. Her research interests include computer vision, video processing, and pattern recognition.

ZHEJUN LU is currently pursuing the master's degree with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. His research interests include location and navigation, data fusion, and sensor signal processing.