electronics

MDPI

Article

MFFRand: Semantic Segmentation of Point Clouds Based on Multi-Scale Feature Fusion and Multi-Loss Supervision

Zhiqing Miao [1], Shaojing Song [2,*], Pan Tang [1], Jian Chen [2], Jinyan Hu [2] and Yumei Gong [2]

[1] School of Resources and Environmental Engineering, Shanghai Polytechnic University, Shanghai 201209, China
[2] School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China
* Correspondence: sjsong@sspu.edu.cn

Abstract: With the application of the random sampling method in the down-sampling of point clouds data, the processing speed of point clouds has been greatly improved. However, the utilization of semantic information is still insufficient. To address this problem, we propose a point cloud semantic segmentation network called MFFRand (Multi-Scale Feature Fusion Based on RandLA-Net). Based on RandLA-Net, a multi-scale feature fusion module is developed, which is stacked by encoder-decoders with different depths. The feature maps extracted by the multi-scale feature fusion module are continuously concatenated and fused. Furthermore, for the network to be trained better, the multi-loss supervision module is proposed, which could strengthen the control of the training process of the local structure by adding sub-losses in the end of different decoder structures. Moreover, the trained MFFRand network could be connected to the inference network by different decoder terminals separately, which could achieve the inference of different depths of the network. Compared to RandLA-Net, MFFRand has improved mIoU on both S3DIS and Semantic3D datasets, reaching 71.1% and 74.8%, respectively. Extensive experimental results on the point cloud dataset demonstrate the effectiveness of our method.

Keywords: point clouds; semantic segmentation; feature fusion; multi-loss supervision

check for updates

Citation: Miao, Z.; Song, S.; Tang, P.; Chen, J.; Hu, J.; Gong, Y. MFFRand: Semantic Segmentation of Point Clouds Based on Multi-Scale Feature Fusion and Multi-Loss Supervision. *Electronics* **2022**, *11*, 3626. https://doi.org/10.3390/electronics11213626

Academic Editors: Jungong Han and Guiguang Ding

Received: 11 October 2022
Accepted: 3 November 2022
Published: 7 November 2022

1. Introduction

The visual sensors are simple in structure and inexpensive, but they are greatly affected by illumination and can hardly work in dark scenes. Compared with the 2D images collected by visual sensors, 3D point clouds obtained by LiDAR have irreplaceable advantages, such as higher measurement accuracy, faster response speed and stronger anti-interference capability, which not only avoid the problems of illumination and object pose in the process of image acquisition, but also obtain the spatial and distance information of the targets. Achieving efficient and accurate semantic segmentation of point clouds is an important research content in autonomous driving.

A series of processing methods have been proposed in the field of point cloud semantic segmentation in recent years, including the projection-based methods [1–5], voxel-based methods [6–10] and point-based methods [11–18]. The projection-based methods are derived from the processing of 2D images; researchers convert irregular point clouds data into conventional multi-views and input them into the deep network architecture for semantic segmentation. Despite mature 2D image processing methods being used after projecting, there is still much spatial geometric information lost in the transformation of point cloud data from 3D to 2D projection. Therefore, there are some inherent limitations for projection-based methods. For voxel-based approaches, the collected point cloud data is transformed into voxels. The voxelization allows point clouds to be processed by using the 3D neural network model, and achieves good results. However, compared with 2D images, the voxelization of point cloud leads to greater computational complexity, so the practicability of this method is relatively low in real scenarios. The point-based methods start from the 3D original data, and the spatial characteristics and relative position information of the point clouds are directly processed through the network. Therefore, there are usually good effects on point clouds' segmentation with point-based methods. However, due to the design of these network structures, there is still plenty of room for improvement of the segmentation effect.

Most existing networks are still inadequate for point cloud feature extraction and the utilization of high and low levels semantic information. Inspired by the feature processing [19,20] and deep supervision [21,22] in 2D image segmentation, the Multi-Scale Feature Fusion Based on RandLA-Net (MFFRand) is proposed in this paper, which makes better use of the feature information extracted from the network in point clouds.

The main contributions of this paper can be summarized as the following three aspects:

- A multi-scale feature fusion module is proposed, which different levels of encoder-decoders interconnect to achieve effective feature fusion between high and low levels of semantic information.
- A multi-loss supervision module is proposed, of which multiple sub-losses connected to different levels of encoder-decoder for supervision of the network training. By this way, the local structures could be trained more sufficiently to achieve better feature fusion, so as to further optimize the final segmentation results.
- The MFFRand allows a one-time training for deep networks to obtain the outputs of network with different depths. This is so that the optimal network depth could be selected according to the results inferred by MFFRand.

2. Related Work

LiDAR is an important tool for environmental perception in autonomous driving. Many experts and scholars have conducted a lot of researches on how to use point cloud data of LiDAR to achieve efficient and accurate semantic segmentation, which mainly include projection-based, voxel-based and point-based methods.

2.1. Projection-Based Methods

Earlier researchers applied deep learning on point cloud data by projecting point clouds into 2D images with multiple views. Su et al. [1] first proposed the MVCNN network based on 3D shape descriptors, which constructed 2D images from different perspectives of 3D objects, and the improved VGG-M model was used to achieve the final semantic segmentation effect. On the basis of MVCNN, Feng et al. [3] considered the correlation of different views and grouped the visual descriptors extracted by CNN under different perspectives, which made full use of the feature relationships in multiple views. To compensate for the loss of geometric information in the process of projecting 3D objects to 2D images, Qi et al. [2] introduced multi-resolution filtering in 3D to deeply explore the feature information of 3D objects and further enhance the segmentation capability. Meanwhile, to address the problem of long-tailed distribution of LiDAR data across space, the points were quantified in the polar coordinate in PolarNet [23]. As a way to better comprehend driving scenes, Peng et al. [24] proposed a multi-attention mechanism for dense top-view semantic segmentation based on sparse LiDAR data. Meanwhile, Lyu et al. [25] considered how to project the features of 3D point clouds onto 2D more efficiently; the 2D mesh projection was obtained by the Kamada-Kawai (KK) algorithm with integer programming in this method, and they used a hierarchical approximation strategy to improve computational efficiency. The selection of viewpoints plays a crucial role in projection-based methods. Li et al. [26] considered selecting viewpoints in a data-driven manner and designed an end-to-end network to learn local multi-view descriptors of point clouds, which is robust. By introducing the attention mechanism, the model can also mine feature information more effectively, especially context information from multiple views [27–29]. Although the projection-based approaches have solved the problem of unstructured the point cloud data, they inevitably lead to the loss of internal geometric structure information.

2.2. Voxel-Based Methods

VoxNet [6] is the first deep learning network to process point clouds with voxelization. For VoxNet, the 3DCNN was used in processing voxels to address the problem of disorder and unstructured point clouds. The prediction of 3DCNN was limited to the rough output at the voxel level; thus, the voxel size became a limiting factor to the overall accuracy. To address this problem, the trilinear interpolation was used in the SegCloud [7] network to send rough predictions back to the original 3D point cloud and inferred the semantic labels of the point cloud through FC-CRF. Compared with 2D images, the voxel-based approaches are computationally expensive due to the additional dimension; to reduce computational cost, the octree [8,30] and Kd-tree structure [9] were both applied to the semantic segmentation of the point clouds. Sparsity and varying the point density are also common problems; solving such problems is commonly performed using the cylindrical model [31,32]. The 3D models' inherent properties are usually overlooked by mainstream point cloud detection frameworks; Han et al. [33] proposed a 3D point cloud segmentation model based on occupancy perception, in which a geometric appearance constraint is used to divide the point cloud into supervoxels, and occupancy signals are used to guide the network to achieve semantic segmentation. In addition, a data volume decomposition was used by Chew et al. [34] to analyze point clouds and generate useful features for semantic segmentation. To simplify computations and supplement the details of smaller instances, an attention-focusing feature fusion and an adaptive feature selection module were proposed by Cheng et al. [35]. Nevertheless, the practicality of voxel-based methods is still poor due to the high spatial complexity of voxelization. However, with the increasing computation efficiency, there is still much room for development in voxel-based methods.

2.3. Point-Based Methods

To reduce the loss of information and computational complexity, some scholars started to study the point cloud processing based on the original point cloud. PointNet [11] is the pioneer of point-based methods. It addressed the disorder of the point cloud with symmetric functions and T-Net was introduced to handle the rotation invariance. PointNet++ [12] is a hierarchical structure of PointNet. Although this structure realized the perception of the local visual field of the point cloud, it does not yet address the problem of the relationship between points. The Adaptive Feature Adjustment (AFA) was proposed in PointWeb [36] to simulate the interaction between points, and it could realize the information exchange and local feature learning of points. The point-based convolution is difficult to apply on point clouds due to the disorder of point clouds; to address this problem, a transformation matrix x-transformation was predicted in PointCNN [37], which made the prediction after point cloud transformation independent of the order of the input points. Inspired by image convolution, the spatial convolution of the point clouds was realized in KPConv [18] by constructing a Kernel point that could be extended to variable convolution. Different from KPConv, the fuzzy mechanism was introduced into the discrete convolution kernel of 3D point clouds in SegGCN [38], which divided the neighbor occupied space in a deterministic manner, and it could reduce the influence of boundary effects suffered during the spatial discretization. The sampling methods in various schemes were usually reasons of inefficiency in the processing of large-scale point clouds; therefore, a random sampling strategy was used in RandLA-Net [15] network to improve the point cloud processing speed. Lu et al. [39] proposed to utilize different aggregation strategies between the same category and different categories. The ability to perceive detailed features can be effectively improved by the aggregation or enhancement strategies for local features [40,41]. In addition, in order to efficiently learn the features of various types of targets from large-scale point clouds, Fan et al. [42] proposed the SCF-Net, which used the dual-distance attention mechanism and global contextual features to improve the performance of semantic segmentation. The point-based methods directly start from the raw data of 3D point clouds, which can obtain more effective and relevant information.

However, existing methods have limited capability to capture local details of different scales, since they have not utilized the given information fully. The method proposed in this paper achieves feature extraction at different scales by adding cross-layer connections to the extracted backbone network, and a multi-loss supervision for the feature extraction process is designed. Therefore, our method is able to leverage the advantages of point-based methods and mine the local feature information at different scales further, which in turn improves the performance of the network.

3. Methodology

Figure 1a is a multi-scale feature extraction network composed of five down-sampling and five up-sampling module in RandLA-Net. The overall structure of 6-layers of MFFRand is demonstrated in Figure 1b. Compared with RandLA-Net, MFFRand model adopts the combination of multiple encoder–decoder structures for multi-scale feature fusion in the process of 3D point clouds. In Figure 1b, the orange part is the multi-scale feature fusion module, and the blue lines are the multi-loss supervision module.

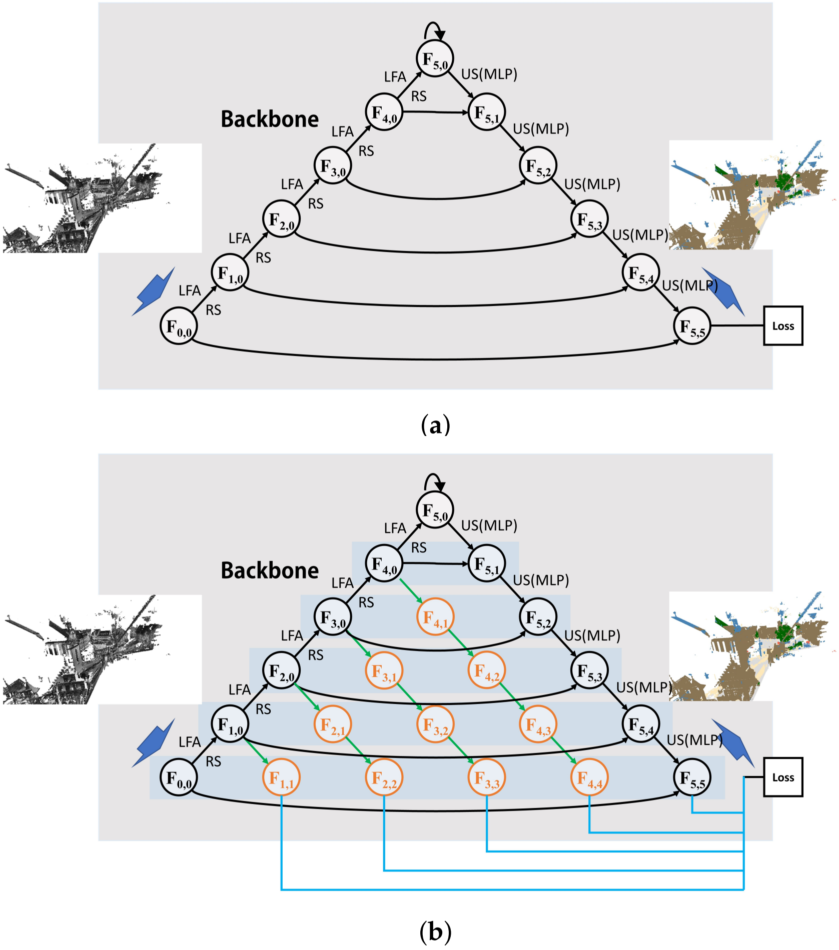


Figure 1. (**a**) RandLA-Net follows the standard encoder–decoder structures with skip connections; this architecture is constructed by low-resolution with semantically strong features and high-resolution with semantically weak features. Where $F_{i,j}$: the feature map after i encoding and j decoding (the initial values of i and j are both 0); LFA: Local feature aggregation, RS: Random sampling, US(MLP): Up-sampling by shared multi-layer perception. (**b**) Based on RandLA-Net, MFFRand concatenate multiple encoder-decoder structures with different depths. The same backbone is used for encoders of different levels of encoder–decoder structures. In the decoding stage, MFFRand starts decoding from different depths and performs feature fusion. The orange part indicates feature fusion, and the blue lines indicate that the training process is supervised by sub-losses.

3.1. Multi-Scale Feature Fusion Module

In many studies, the fusion of different scale features is an important method to improve the segmentation performance. The low-level features are usually with high-resolution and contain more local and detail information. However, due to fewer convolution operations, low-level features have less semantics and more noise. The high-level features are usually with low-resolution and contain stronger semantic information, but their perception for details is weak. How to fuse the semantic features of different levels efficiently is the key to improve the performance of the segmentation network. To meet the requirements of more accurate segmentation in autonomous driving, a feature fusion module based on multiple encoder–decoder structures is proposed in this paper.

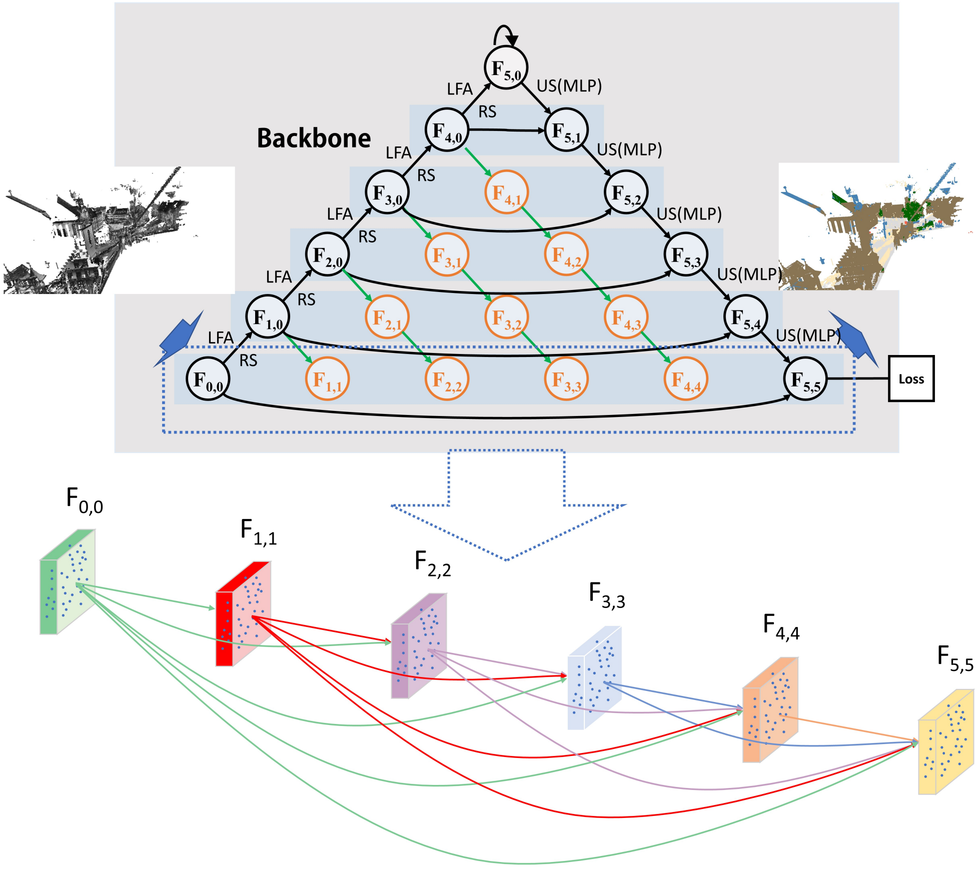


Figure 2. Multiple encoder–decoders are concatenated together to form the main structure of feature fusion. Features in the same blue rectangular block are fully fused by skip connection.

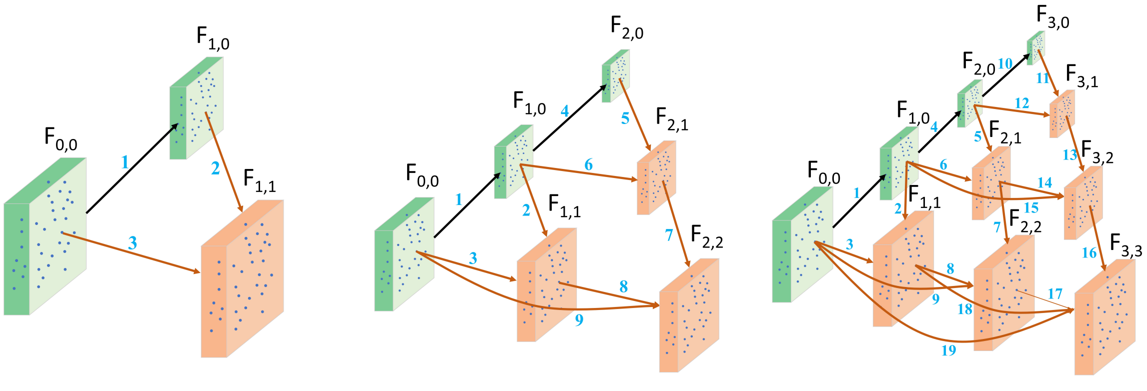
After being processed by encoder–decoder, the common networks usually have the same number of points, feature dimensions and feature sizes. However, this is due to multiple down-sampling and up-sampling, which leads to a large semantic gap between the feature maps of encoder and decoder. To deal with this problem, multiple encoder–decoder structures are added to the original RandLA-Net, and the structure is shown in the top of Figure 2. This module is composed of five encoder-decoders with different levels, and the same backbone is used as the encoder in each encoder–decoder structure. In addition to up-sampling at the decoder phase, feature concatenation is performed in each layer (blue rectangular block). Taking the bottom feature layer $F_{0,0}$, $F_{1,1}$, $\cdots$, $F_{5,5}$ as an example, the concatenation of this layer is demonstrated in the bottom of Figure 2. The feature maps of different encoder–decoders are fused by the skip connection, so as to form a dense feature fusion network. At the same time, the parameters of network could be updated effectively through the back propagation gradient flow and ensure the network could be trained properly.

In terms of presentation, for different levels of encoder–decoder structures, $F_{i,j}$ denotes the feature map obtained by i encoding and j decoding. In the same layer (horizontal blue rectangular block), there are the same number of points and feature dimensions in feature maps. For example, $F_{0,0}$, $F_{1,1}$, $\cdots$, and $F_{5,5}$ are in the same layer of the network. Although the feature maps belong to different levels of encoder–decoders, the number of points and the dimension of feature in this layer remain the same after multiple encoding and decoding operations.

There is a relationship between different encoder–decoder structures. For the encoder–decoder of the 2-layer in Figure 3a, the calculation order is $F_{0,0}$-$F_{1,0}$-$F_{1,1}$, and for the encoder–decoder of 3-layer in Figure 3b, the calculation order is $F_{0,0}$-$F_{1,0}$-$F_{1,1}$-$F_{2,0}$-$F_{2,1}$-$F_{2,2}$. Similarly, the encoding and decoding operations are carried out on the encoder–decoder of layer 4 (Figure 3c), layer 5 and layer 6 successively. In the process of down-sampling, the random sampling strategy is used to improve the processing speed, and the LFA module is used to enhance the receptive field of network. In the process of up-sampling, each up-sampling is accompanied by the fusion with the information extracted from the previous encoder–decoder. In this way, the semantic gap is continuously bridged. The calculation of the feature fusion is given by Equation (1):

$$F_{i,j} = \begin{cases} R(L(F_{i,j})), & j = 0 \\ C(\Delta(\Delta(F_{i-j+t,t})_{t=0}^{j-1}, U(F_{i,j-1}))), & j > 0 \end{cases} \tag{1}$$

where R denotes random sampling, L denotes LFA, C represents convolution, Δ represents feature concatenation, and U denotes up-sampling. When $j = 0$, the encoding operation is carried out firstly, then the down-sampling and LFA is performed in the backbone for 5 times. When $j > 0$, each feature map is obtained through the convolution of $j + 1$ inputs. The input is composed of the first j feature maps ($F_{i-j,0}$, $F_{i-j+1,1}$, $\cdots$, $F_{i-1,j-1}$) in the same layer and the output by the up-sampling with MLP of the feature map ($F_{i,j-1}$) in the previous layer.



(**a**) (**b**) (**c**)

Figure 3. (**a**) 2-layer encoder-decoder structure, where the numbers in the figure represent the order of feature propagation in this structure. (**b**) 3-layer encoder-decoder structure, whose feature propagation is a further encoding, decoding and lateral propagation based on the completion of 2-layer encoder–decoder. (**c**) 4-layer encoder-decoder structure.

3.2. Multi-Loss Supervision Module

Since the network consists of multiple encoder–decoder structures, there are more complex computational steps in the feature fusion process, and the overall training process is more difficult to be controlled. Inspired by DSN [43] and Multi-Scale Structure-Aware network [22], a multi-loss supervision module is introduced to control the local training process. On the basis of the multi-scale feature fusion module, the training process is controlled by adding sub-losses at the end of each decoder separately, which make the local structure trained fully and achieve better feature fusion.

In the five levels of encoder–decoders, the cross-entropy loss function with category weight is added, respectively.

The weight of each category is shown as Equation (2):

$$w_i = \frac{1}{\left(\frac{num_i}{\sum_{t=1}^{n=8} num_t} + 0.02\right)}, \quad i = 1, 2, 3, \cdots, 8 \tag{2}$$

where num_i is the number of points belonging to ith category. There are eight categories in the Semantic3D [44] dataset. The decimal 0.02 in the denominator is to avoid the case where the denominator is 0. When the number of points in a category is less, there will be a higher weight in this category.

Depending on the number of point clouds contained in different categories, the cross-entropy loss output of each category accounts for different proportions, and the calculation of cross-entropy loss is given by Equation (3):

$$H(p,q) = -\sum_{t=1}^{c} (p(x_t) \times \log(q(x_t))) \times w_i \tag{3}$$

where $H(p,q)$ is the cross-entropy loss function with weights per point, $p(x)$ is the real distribution of the target, $q(x)$ is the predicted distribution of the target, c is the number of categories, and the output value of each point will be correspondingly multiplied by that weight as the predicted loss of the target.

The sub-losses output of each encoder–decoder is given by Equation (4):

$$sub\_loss = \frac{\sum_{i=1}^{n} H(p,q)_i}{n} \tag{4}$$

where n is the number of input points, and $H(p,q)_i$ is the loss of ith point.

The loss function of the entire network is given by Equation (5):

$$loss = \frac{1}{m} \sum_{i=1}^{m} sub\_loss_i \tag{5}$$

where m is the number of sub-losses.

The detailed structure of the network with multi-loss supervision is shown in Figure 4.

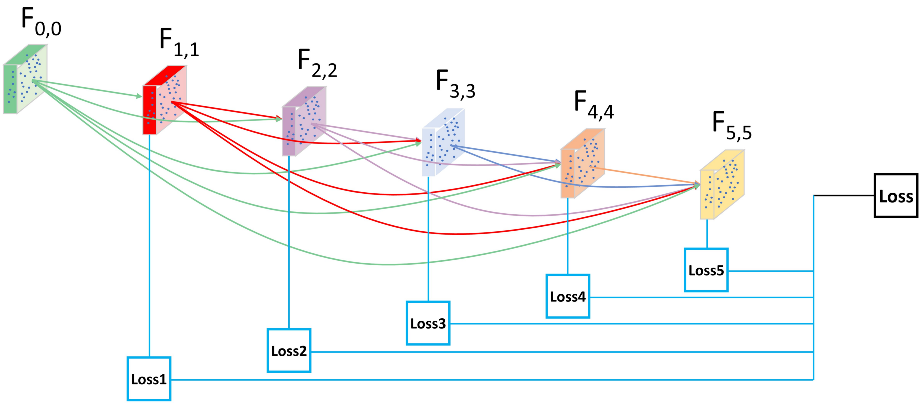


Figure 4. Multi-loss supervision module. The sub-losses are attached to the end of each decoder, which could control the training of the local structure of the network. The loss function of 6-layers MFFRand consists of five sub-losses.

4. Experiments

To evaluate the performance of MFFRand, a series of relevant experiments are conducted on MFFRand. The experiments are divided into four parts: evaluation of mIoU (mean Intersection-over-Union) and OA (Overall Accuracy), time consumption, ablation study and the effectiveness of network segmentation at different depths. All the experiments are conducted on Intel(R) Xeon(R) Gold 5118 @2.30 GHz CPU and NVIDIA GeForce

GTX 1080Ti GPU. Meanwhile, in order to ensure the reliability of experimental results, all results are averaged over three experiments.

4.1. Evaluation of mIoU and OA

S3DIS [45] is a large-scale indoor dataset which consists of 6 large areas; each area contains several scenes, such as offices, hallway, conference rooms, and so on. The whole dataset has around 273 million points annotated with 13 semantic labels. In our experiments, we use a six-fold cross-validation to evaluate the proposed network.

Semantic3D [44] dataset was developed by a research group at ETH Zurich, Switzerland. Semantic3D provides a large labeled 3D point cloud with a total of over 4 billion points, which contains 8 types of objects, including buildings and cars, etc. The dataset is split into 15 training scenes, 15 testing scenes and 4 reduced testing scenes. The reduced-8 is the reduced version of the testing set provided by Semantic3D for a convenient online evaluation, which includes 4 of the 15 testing scenes, and the point clouds are down-sampled at 0.01 m. In our experiments, two of the training scenes are used for validation during training, and reduced-8 is used as the testing set.

In this part, the MFFRand is tested on the S3DIS dataset (6-fold cross-validation) and Semantic3D dataset (reduced-8). Recent representative works are also evaluated on the same dataset. OA and mIoU are used as evaluation metrics.

Table 1 presents the quantitative segmentation results of different networks on the S3DIS dataset with 6-fold cross-validation. MFFRand has achieved a mIoU metric of 71.1%, which is better than other methods, as shown in the table. Additionally, the best segmentation results are obtained on categories such as the wall, beam, column, table, and clutter. The experimental results demonstrate that our method is based on multi-scale feature fusion, and the multi-loss supervision performs effectively in 3D point cloud semantic segmentation and is superior to other methods.

Table 1. Quantitative results of different networks on the S3DIS (6-fold cross-validation).

	mIoU (%)	**OA (%)**	**Ceil.**	**Floor**	**Wall**	**Beam**	**Col.**	**Wind.**	**Door**	**Table**	**Chair**	**Sofa**	**Book.**	**Board**	**Clut.**
DPFA [46]	61.7	89.2	94.6	98.0	79.2	40.7	36.6	52.2	70.8	65.9	74.7	27.7	49.8	51.6	60.6
PointCNN [37]	65.4	88.1	94.8	97.3	75.8	63.3	51.7	58.4	57.2	71.6	69.1	39.1	61.2	52.2	58.6
PointWeb [36]	66.7	87.3	93.5	94.2	80.8	52.4	41.3	64.9	68.1	71.4	67.0	50.3	62.7	62.2	58.5
ShellNet [47]	66.8	87.1	90.2	93.6	79.9	60.4	44.1	64.9	52.9	71.6	84.7	53.8	64.6	48.6	59.4
Liu et al. [48]	68.3	88.6	93.5	96.0	81.5	42.6	46.3	61.0	74.1	67.4	82.7	63.5	59.5	56.6	62.9
MuGNet [49]	69.8	88.5	92.0	95.7	82.5	64.4	60.1	60.7	69.7	82.6	70.3	64.4	52.1	52.8	60.6
RandLA-Net [15]	70.0	88.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
KPConv [18]	70.6	-	93.6	92.4	83.1	63.9	54.3	66.1	76.6	57.8	64.0	69.3	74.9	61.3	60.3
MFFRand (ours)	71.1	88.9	93.9	94.4	83.6	64.8	54.7	62.7	68.2	71.7	79.3	65.5	64.2	58.8	63.1

Table 2 presents the quantitative segmentation results of different networks on the Semantic3D dataset (reduced-8). In terms of mIoU and OA, MFFRand is significantly better than the compared methods. In addition, MFFRand has achieved very good results in four categories, including man-made, natural, etc. The visualization of the prediction results on the testing set is shown in Figure 5.

Table 2. Quantitative results of different networks on the Semantic3D (reduced-8).

	mIoU (%)	**OA (%)**	**Man-Made**	**Natural**	**High Veg**	**Low Veg**	**Building**	**Hard Scape**	**Scanning Art**	**Cars**
SEGCloud [7]	61.3	88.1	83.9	66.0	86.0	40.5	91.1	30.9	27.5	64.3
MSDeepVoxNet [50]	65.3	88.4	83.0	67.2	83.8	36.7	92.4	31.3	50.0	78.2
PointConv [51]	69.2	91.8	92.2	79.2	73.1	62.7	92.0	28.7	43.1	82.3
ShellNet [47]	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
PointGCR [52]	69.5	92.1	93.8	80.0	64.4	66.4	93.2	39.2	34.3	85.3
GACNet [53]	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
RandLA-Net [15]	72.7	92.7	96.4	83.9	85.1	40.1	94.8	46.4	61.4	73.8
KPConv [18]	72.8	92.8	92.9	88.6	82.4	42.4	93.2	38.5	64.5	80.1
MFFRand (ours)	74.8	93.7	97.2	91.9	84.2	47.8	94.0	38.8	66.6	77.8

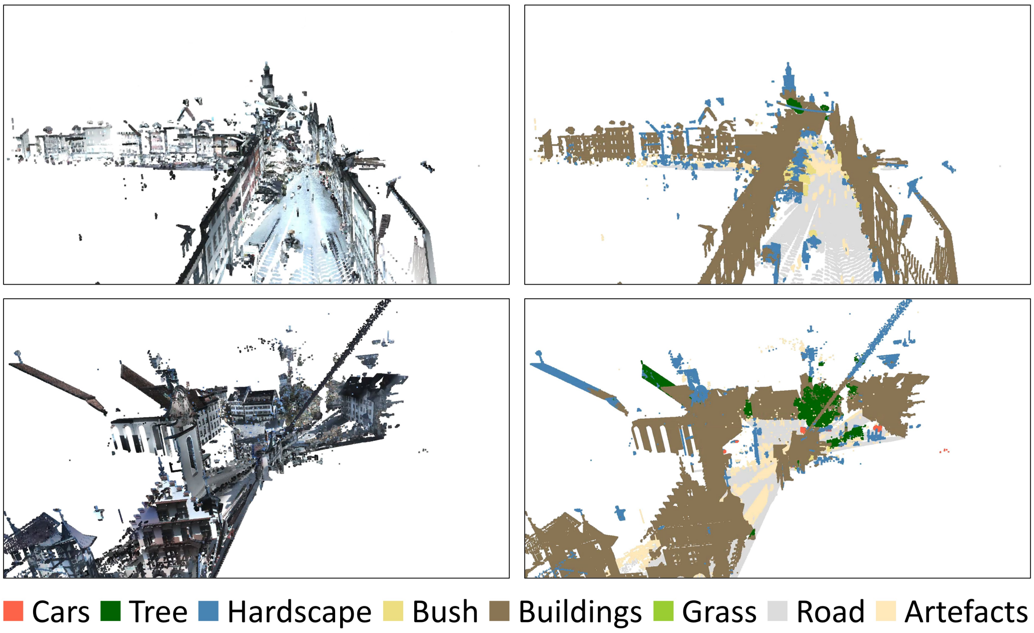


Figure 5. Visualization of partial prediction results of Semantic3D (reduced-8). **Left**: Raw RGB images of the input point clouds. **Right**: Visualization of inference results of testing set by MFFRand. **Top**: Scenario of MarketplaceFeldkirch_station4. **Bottom**: Scenario of StGallenCathedral_station6.

4.2. Evaluation of Time Consumption

SemanticKITTI [54] dataset was developed by a research group at the University of Bonn, Germany in 2019, which is a semantic segmentation dataset for large 3D point clouds of outdoor streets based on radar sequence scene understanding. The dataset contains 21 sequences, including 19 categories such as cars, bicycles, motorcycles and so on. Sequences 00–07 and 09 of SemanticKITTI dataset are used as the training set; sequence 08 is used as a validation set and sequences 11–21 are used as the testing set.

Different from the complete street scenes in the Semantic3D dataset, SemanticKITTI is a large-scale road sequence scene point clouds that is collected by vehicle-mounted LiDAR. In SemanticKITTI, we could evaluate the network by the number of that frames processed per second, so the SemanticKITTI is used to evaluate the efficiency of MFFRand in this paper. The experiments are conducted on sequence 19, which has a total of 4981 frames of data. The SPVNAS [55] and RandLA-Net [15] are also evaluated as contrast experiments on the sequence 19. Processing time (seconds) is used as evaluation metrics. The quantitative results of the experiments are shown in Table 3. According to the results of experiments, the MFFRand has almost no change in time consumption, and still maintains a great advantage.

Table 3. The inference time of different approaches (Tested on sequence 19 of the SemanticKITTI dataset).

	Total Time (seconds)
SPVNAS	617.33 (8.07 frame/s)
RandLA-Net	169.5 (29.39 frame/s)
MFFRand (ours)	170.4 (29.23 frame/s)

4.3. Ablation Study

To further validate the contribution of each module in MFFRand, the following ablation study was conducted on the Semantic3D dataset:

- Remove multi-loss supervision module. The better training of local structure in MFFRand could be achieved by the multi-loss supervision module. With the removal of the multi-loss supervision module, the training of the entire network is controlled by the loss function of the deepest encoder-decoder structure.

- Remove multi-feature fusion module (& multi-loss supervision module). This module enables the feature of encoder-decoders could be fused more sufficiently, bridging the semantic gap between the high and low layers effectively. With the removal of the multi-feature fusion module, the encoder and the decoder sub-networks are directly connected to each other through the skip connections.

The ablation study of two modules is conducted as shown in Table 4. The experimental results demonstrate that the application of the multi-scale feature fusion module could help improve the segmentation effect significantly, and the mIoU increase by 1.1 on the basis of RandLA-Net. Meanwhile, the multi-loss supervision module also has a gain effect on the semantic segmentation of the network, with another 1.0 mIoU improvement on the basis of the multi-scale feature fusion module.

Table 4. Results of ablated network on the Semantic3D (reduced-8).

	mIoU (%)
MFFRand	74.8
Remove multi-loss supervision module	73.8
Remove multi-scale feature fusion module & multi-loss supervision module (RandLA-Net)	72.7

4.4. Evaluation of Output at Different Network Depths

To verify the semantic segmentation performance of MFFRand at different depths, we conduct experiments on MFFRand and RandLA-Net, respectively. The Table 5 shows the results of the experiment and performance comparison. Compared with RandLA-Net, there is an absolute advantage in inference results of 3-6 layers network of MFFRand.

In addition, in order to obtain the inference results of RandLA-Net (3–6 layers), the RandLA-Net with different depths needs to be trained separately and then the inference is conducted. Therefore, for RandLA-Net, four times of training is performed. Since the multi-scale feature fusion module in MFFRand consists of multiple encoder–decoder structures, the MFFRand only needs to be trained once, and then the inference network is connected to the end of each decoder in trained MFFRand separately. By this way, the four inference results could be obtained on MFFRand with different depths, as shown in Figure 6. Therefore, this is an advantage of MFFRand in network training, which can obtain segmentation results of different layers of network with a one-time training process.

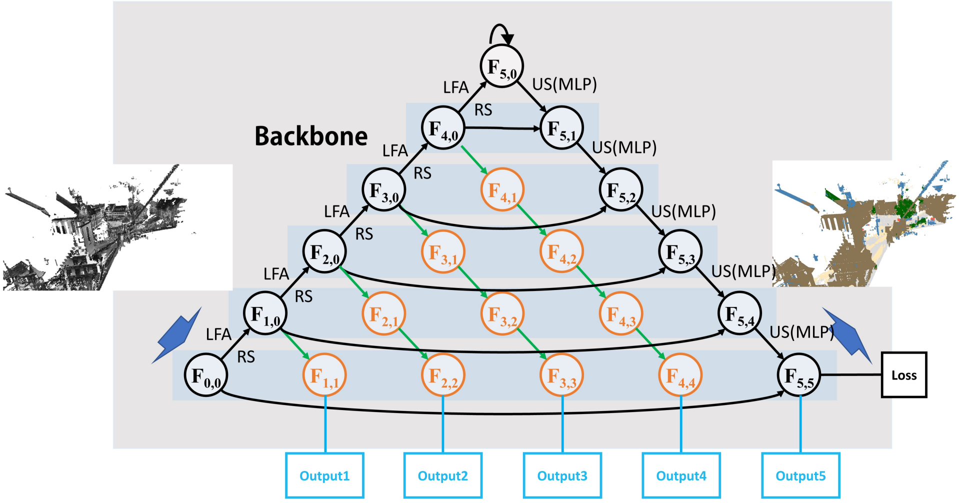


Figure 6. The segmentation results inferenced by MFFRand with different depths; there are outputs in every layer.

Table 5. Influence of the number of network layers on inference results.

Number of Network Layers	mIoU of RandLA-Net (%)	mIoU of MFFRand (%)
3	53.03	53.73
4	67.63	67.63
5	72.70	74.07
6	72.70	74.80

5. Conclusions

MFFRand is proposed in this paper to handle the problem that high and low levels feature information could not be utilized sufficiently. The proposed block mainly consists of two modules, including the multi-scale feature fusion module and the multi-loss supervision module. The multi-scale feature fusion module is introduced to effectively fuse the semantic information between feature maps of different levels, which consists of a dense feature fusion network by concatenating several different levels of encoder–decoder structures. For multi-loss supervision module, the sub-losses are connected, respectively, to the end of each decoder to control the training of the local structure. Quantitative experiments and ablation studies are conducted on the S3DIS and Semantic3D dataset, which demonstrate the effectiveness of MFFRand in 3D point cloud semantic segmentation and the contribution of each module in MFFRand.

The method proposed in this paper improves the performance of semantic segmentation, but there is no breakthrough in efficiency. In the future, we will further explore general and effective methods for improving performance and investigate which cross-layer connections and losses in MFFRand are critical. Moreover, we will extend our method to end-to-end 3D panoptic segmentation and 3D object detection on large-scale point clouds.

Author Contributions: Conceptualization, Z.M. and S.S.; methodology, Z.M. and P.T.; software, Z.M.; validation, Z.M., P.T., J.C., J.H. and Y.G.; formal analysis, Z.M. and J.C.; investigation, Z.M. and J.H.; resources, P.T.; data curation, Z.M.; writing—original draft preparation, Z.M. and S.S.; writing—review and editing, Z.M., S.S. and P.T.; visualization, P.T.; supervision, S.S.; project administration, Z.M.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: The generous support of the Discipline Construction of Computer Science and Technology of Shanghai Polytechnic University under Grant A01GY22G004-01 and the Collaborative Innovation Platform of Electronic Information Master under Grant No. A10GY21F015 of Shanghai Polytechnic University are gratefully acknowledged.

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

References

1. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953.
2. Qi, C.R.; Su, H.; Nießner, M.; Dai, A.; Yan, M.; Guibas, L.J. Volumetric and Multi-view CNNs for Object Classification on 3D Data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5648–5656.
3. Feng, Y.; Zhang, Z.; Zhao, X.; Ji, R.; Gao, Y. GVCNN: Group-view convolutional neural networks for 3D shape recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 264–272.
4. Wu, B.; Wan, A.; Yue, X.; Keutzer, K. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation, Brisbane, Australia, 21–26 May 2018; pp. 1887–1893.
5. Wu, B.; Zhou, X.; Zhao, S.; Yue, X.; Keutzer, K. SqueezeSegV2: Improved Model Structure and Unsupervised Domain Adaptation for Road-Object Segmentation from a LiDAR Point Cloud. In Proceedings of the 2019 International Conference on Robotics and Automation , Montreal, QC, Canada, 20–24 May 2019; pp. 4376–4382.
6. Maturana, D.; Scherer, S. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–1 October 2015; pp. 922–928.

7. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. SEGCloud: Semantic Segmentation of 3D Point Clouds. In Proceedings of the 2017 International Conference on 3D vision, Qingdao, China, 10–12 October 2017; pp. 537–547.
8. Riegler, G.; Osman Ulusoy, A.; Geiger, A. OctNet: Learning Deep 3D Representations at High Resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 27 January–1 February 2017; pp. 3577–3586.
9. Zeng, W.; Gevers, T. 3DContextNet: K-d Tree Guided Hierarchical Learning of Point Clouds Using Local and Global Contextual Cues. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
10. Meng, H.Y.; Gao, L.; Lai, Y.K.; Manocha, D. VV-Net: Voxel VAE Net with Group Convolutions for Point Cloud Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 8500–8508.
11. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 27 January–1 February 2017; pp. 652–660.
12. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv* **2017**, arXiv:1706.02413.
13. Jiang, M.; Wu, Y.; Zhao, T.; Zhao, Z.; Lu, C. PointSIFT: A SIFT-like network module for 3D point cloud semantic segmentation. *arXiv* **2018**, arXiv:1807.00652.
14. Li, J.; Chen, B.M.; Lee, G.H. SO-Net: Self-Organizing Network for Point Cloud Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9397–9406.
15. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11108–11117.
16. Liang, Z.; Yang, M.; Deng, L.; Wang, C.; Wang, B. Hierarchical Depthwise Graph Convolutional Neural Network for 3D Semantic Segmentation of Point Clouds. In Proceedings of the 2019 International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 8152–8158.
17. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4558–4567.
18. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6411–6420.
19. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
20. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 27 January–1 February 2017; pp. 2117–2125.
21. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In Proceedings of the Artificial Intelligence and Statistics, PMLR, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
22. Ke, L.; Chang, M.C.; Qi, H.; Lyu, S. Multi-scale structure-aware network for human pose estimation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 713–728.
23. Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; Foroosh, H. PolarNet: An Improved Grid Representation for Online LiDAR Point Clouds Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9601–9610.
24. Peng, K.; Fei, J.; Yang, K.; Roitberg, A.; Zhang, J.; Bieder, F.; Heidenreich, P.; Stiller, C.; Stiefelhagen, R. MASS: Multi-Attentional Semantic Segmentation of LiDAR Data for Dense Top-View Understanding. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15824–15840. [CrossRef]
25. Lyu, Y.; Huang, X.; Zhang, Z. Learning to Segment 3D Point Clouds in 2D Image Space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12255–12264.
26. Li, L.; Zhu, S.; Fu, H.; Tan, P.; Tai, C.L. End-to-End Learning Local Multi-View Descriptors for 3D Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1919–1928.
27. Peng, B.; Yu, Z.; Lei, J.; Song, J. Attention-guided fusion network of point cloud and multiple views for 3D shape recognition. In Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing, Virtual Conference, 1–4 December 2020; pp. 185–188.
28. Nie, W.; Zhao, Y.; Song, D.; Gao, Y. DAN: Deep-Attention Network for 3D Shape Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 4371–4383. [CrossRef] [PubMed]
29. Zhang, J.; Zhou, D.; Zhao, Y.; Nie, W.; Su, Y. MV-LFN: Multi-view based local information fusion network for 3D shape recognition. *Vis. Inform.* **2021**, *5*, 114–119. [CrossRef]
30. Que, Z.; Lu, G.; Xu, D. VoxelContext-Net: An Octree based Framework for Point Cloud Compression. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 6042–6051.

31. Zhu, Z.; Li, X.; Xu, J.; Yuan, J.; Tao, J. Unstructured road segmentation based on road boundary enhancement point-cylinder network using LiDAR sensor. *Remote Sens.* **2021**, *13*, 495. [CrossRef]
32. Zhou, H.; Zhu, X.; Song, X.; Ma, Y.; Wang, Z.; Li, H.; Lin, D. Cylinder3D: An Effective 3D Framework for Driving-scene LiDAR Semantic Segmentation. *arXiv* **2020**, arXiv:2008.01550.
33. Han, L.; Zheng, T.; Xu, L.; Fang, L. OccuSeg: Occupancy-Aware 3D Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2940–2949.
34. Chew, A.W.Z.; Ji, A.; Zhang, L. Large-scale 3D point-cloud semantic segmentation of urban and rural scenes using data volume decomposition coupled with pipeline parallelism. *Autom. Constr.* **2022**, *133*, 103995. [CrossRef]
35. Cheng, R.; Razani, R.; Taghavi, E.; Li, E.; Liu, B. (AF)2-S3Net: Attentive Feature Fusion with Adaptive Feature Selection for Sparse Semantic Segmentation Network. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 12542–12551.
36. Zhao, H.; Jiang, L.; Fu, C.W.; Jia, J. PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 January 2019; pp. 5565–5573.
37. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on X-transformed points. In *Advances in Neural Information Processing Systems 31*; Neural Information Processing Systems Foundation, Inc.: Montreal, QC, Canada, 2018.
38. Lei, H.; Akhtar, N.; Mian, A. SegGCN: Efficient 3D Point Cloud Segmentation with Fuzzy Spherical Kernel. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11611–11620.
39. Lu, T.; Wang, L.; Wu, G. CGA-Net: Category Guided Aggregation for Point Cloud Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 11688–11697.
40. Zeng, Z.; Xu, Y.; Xie, Z.; Tang, W.; Wan, J.; Wu, W. LEARD-Net: Semantic segmentation for large-scale point cloud scene. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102953. [CrossRef]
41. Qiu, S.; Anwar, S.; Barnes, N. Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1757–1767.
42. Fan, S.; Dong, Q.; Zhu, F.; Lv, Y.; Ye, P.; Wang, F.Y. SCF-Net: Learning Spatial Contextual Features for Large-Scale Point Cloud Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14504–14513.
43. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 27 January–1 February 2017; pp. 4700–4708.
44. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M. Semantic3d.net: A new large-scale point cloud classification benchmark. *arXiv* **2017**, arXiv:1704.03847.
45. Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3D Semantic Parsing of Large-Scale Indoor Spaces. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1534–1543.
46. Chen, J.; Kakillioglu, B.; Velipasalar, S. Background-Aware 3-D Point Cloud Segmentation With Dynamic Point Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5703112. [CrossRef]
47. Zhang, Z.; Hua, B.S.; Yeung, S.K. ShellNet: Efficient Point Cloud Convolutional Neural Networks Using Concentric Shells Statistics. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 1607–1616.
48. Liu, C.; Zeng, D.; Akbar, A.; Wu, H.; Jia, S.; Xu, Z.; Yue, H. Context-Aware Network for Semantic Segmentation toward Large-Scale Point Clouds in Urban Environments. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5703915. [CrossRef]
49. Xie, L.; Furuhata, T.; Shimada, K. Multi-Resolution Graph Neural Network for Large-Scale Pointcloud Segmentation. *arXiv* **2020**, arXiv:2009.08924.
50. Roynard, X.; Deschaud, J.E.; Goulette, F. Classification of point cloud scenes with multiscale voxel deep network. *arXiv* **2018**, arXiv:1804.03583.
51. Wu, W.; Qi, Z.; Fuxin, L. PointConv: Deep Convolutional Networks on 3D Point Clouds. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 January 2019; pp. 9613–9622.
52. Ma, Y.; Guo, Y.; Liu, H.; Lei, Y.; Wen, G. Global Context Reasoning for Semantic Segmentation of 3D Point Clouds. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2920–2929.
53. Wang, L.; Huang, Y.; Hou, Y.; Zhang, S.; Shan, J. Graph Attention Convolution for Point Cloud Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 January 2019; pp. 10288–10297.
54. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 9297–9307.
55. Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; Han, S. Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution. In Proceedings of the European Conference on Computer Vision, Edinburgh, UK, 23–28 August 2020; pp. 685–702.