IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 June 2023, accepted 4 July 2023, date of publication 7 July 2023, date of current version 12 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293161

RESEARCH ARTICLE

A Local-Global Feature Fusing Method for Point Clouds Semantic Segmentation

YUANWEI BI, LUJIAN ZHANG, YAOWEN LIU, YANSEN HUANG, AND HAO LIU
School of Computer and Control Engineering, Yantai University, Yantai 264000, China

Corresponding author: Lujian Zhang (lujzhang@s.ytu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62272405, in part by the Youth Innovation Science and Technology Support Program of Shandong Provincial under Grant 2021KJ080, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2022MF238.

ABSTRACT In recent years, the abundance of information in 3D data has made the semantic segmentation of 3D point clouds a topic of great interest. However, current methods often rely solely on the original three-dimensional coordinates of the point cloud as input geometric features, leading to poor generalization performance. Additionally, occlusion of the point cloud data can negatively impact segmentation accuracy when only local information is considered. To address these issues, this paper proposes a network named LGFF-Net. To fully utilize the original information of point clouds, we designed a Local Feature Aggregation (LFA) module that treats geometric and semantic information equally and preserves the original properties while cross-augmenting them. On the other hand, we proposed a simple and effective Global Feature Extraction (GFE) module to extract global features. Finally, we hierarchically fuse local and global features using a U-shaped segmentation structure. Compared to state-of-the-art networks, our method achieves competitive results on several benchmark datasets, including Semantic Topographic Point Labeling-Synthetic 3D, Toronto_3D, Stanford Large 3-D Indoor Space, and ScanNet. We also conduct multiple ablation experiments to validate the efficacy of LGFF-Net.

INDEX TERMS Feature cross enhancement, global feature extraction, point clouds, semantic segmentation.

I. INTRODUCTION

As data representation in 3D vision, point clouds have a richer spatial structure and more complex geometric information than 2D images. Additionally, the rapid development of 3D sensors [1], [2], [3] has made it easier to acquire point cloud data. As a result, point cloud processing has been widely used in fields such as autonomous driving [4], [5], [6], virtual and augmented reality [7], [8], and robotics [9], [10], [11] in recent years. This paper focuses on the semantic segmentation of point clouds, which is a subtask of point cloud processing that aims to assign a label to each point.

Various deep learning methods have been developed for the semantic segmentation of point clouds. Voxel-based methods [12], [13], [14] can handle large-scale point clouds and benefit from the downsampling effect of voxelization, while projection-based methods [15], [16] can leverage mature 2D algorithms. However, both voxel-based and projection-based methods may destroy structural information during transformation. To address this problem, PointNet [17], the pioneer of point-based methods, introduced the idea of directly consuming point clouds and proposed using Multiple Layer Perceptron (MLP) and global aggregation operations to learn point cloud features, as shown in Figure 1(a). However, PointNet does not consider the local region features of the point cloud.

A series of follow-up works [18], [19], [20], [21], [22], [23], [24], [25], [26], [27] based on PointNet [17] have focused on designing local feature aggregation modules to extract features as local descriptors by combining information from neighboring points with the center of mass. RandLA-Net [20] is one of the representative works, and its local feature aggregation module is shown roughly in Figure 1(b). However, RandLA-Net only encodes the position information of the point cloud, ignoring the processing of semantic information. It is important to note that point clouds contain

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

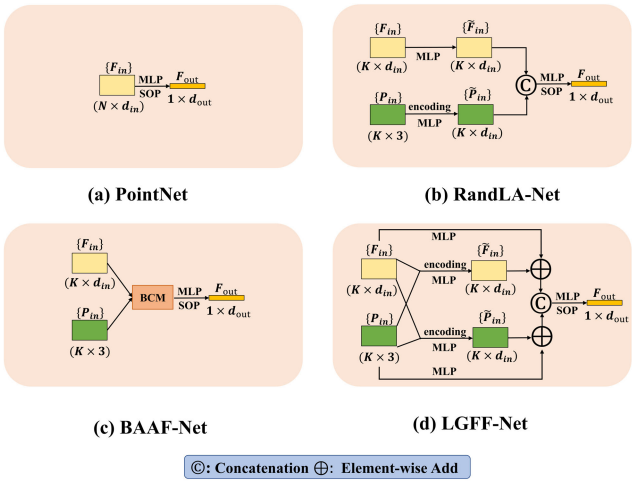


FIGURE 1. Overview of the ways of feature aggregation of PointNet [17], RandLA-Net [20], BAAF-Net [25], and our LGFF-Net. Notice, the listed PointNet considered is the global feature, and others considered are the local region feature. SOP denotes symmetric operations, like MAX.

both geometric and semantic information, and the model needs to treat both types of features equally. Otherwise, the generalization ability of the model may be reduced. To solve this problem, Qiu et al. [25] proposed a Bilateral Context Module (BCM), and its network structure is shown roughly in Figure 1(c). However, while the BCM enables a good mixture of geometric and semantic features, it loses the original properties of both types of features. The advantages of rich positional information that geometric features can represent are no longer apparent, and the richness of semantic features is also reduced.

The proposed Local Feature Aggregation (LFA) module is illustrated in Figure 1(d). LFA is designed by analyzing the geometric and semantic features of the point cloud, with the aim of enriching both types of features. By comprehensively considering geometric and semantic features and using a Res-Connection [28] joint, LFA allows them to influence each other without losing their original properties. LFA combines Attention Pooling and Max Pooling to capture comprehensive local and prominent features, ensuring that the model can focus on more diverse features.

Another problem with most of the methods mentioned above is that they only consider the aggregation of local region features, which can result in the relationship between local regions being ignored and insufficient extraction of long-range dependence relationships between points. Furthermore, it is worth noting that occlusion within the point cloud itself can blur the local structure and hinder the model's ability to extract accurate semantic features. Therefore, deep learning models need to consider the impact of global features on point cloud semantic segmentation to effectively understand 3D scenes in complex environments with occlusions.

To meet the challenge of global feature extraction, the Global Feature Extraction(GFE) module was designed, inspired by PointNet [17]. In the GFE module, the initial point cloud information is fed into an MLP Block for dimension processing. The processed information is then joined with the original point cloud data and the global feature is returned by Max Pooling. By integrating local and global features, the GFE module can better capture the complex structures and semantic features of 3D objects.

The proposed Local-Global Feature Fusing for Point Clouds Semantic Segmentation Network (LGFF-Net) consists of the LFA and GFE modules. To validate the effectiveness of our method, we provide experimental results on four annotated datasets: Semantic Terrain Points Labeling-Synthetic 3D (STPLS3D) [29], Toronto_3D [30], Stanford large-scale 3-D Indoor Spaces (S3DIS) [31], and ScanNet [32]. We also conduct ablation experiments to evaluate the effect of each module. Compared to state-of-the-art methods, our experimental results show that our method achieves competitive performance. In summary, our contributions are as follows.

1) The LFA module is employed to aggregate local features. By fully considering the geometric and semantic features of point clouds, LFA combines them through feature cross-enhancement. This approach ensures that the two types of features can influence each other while maintaining their unique properties.
2) The GFE module is responsible for global feature extraction. This module enhances features using an MLP and employs Max Pooling to extract prominent global features. The GFE module has a straightforward structure and requires minimal computation. Experimental results demonstrate its effectiveness in improving the performance of 3D point cloud semantic segmentation.
3) We propose a novel network called LGFF-Net to enhance feature processing in point cloud semantic segmentation. LGFF-Net combines global and local point cloud features to jointly perform semantic segmentation. By leveraging the strengths of both types of features, LGFF-Net aims to improve the overall performance of 3D point cloud semantic segmentation.

Section II of this paper introduces recent work related to ours. Section III provides a detailed description of our work. Section IV presents our experimental results and analysis.

II. RELATED WORK

Traditional methods [33], [34], [35] for 3D scene understanding are valued for their low computational burden and ease of operation. However, their reliance on manual segmentation of regions and feature labeling limits their ability to achieve optimal results. Deep learning-based techniques have shown remarkable performance in image processing tasks, but the sparsity and irregularity of 3D data make it difficult to apply these methods directly to point clouds. In recent years, an increasing number of researchers have attempted to apply

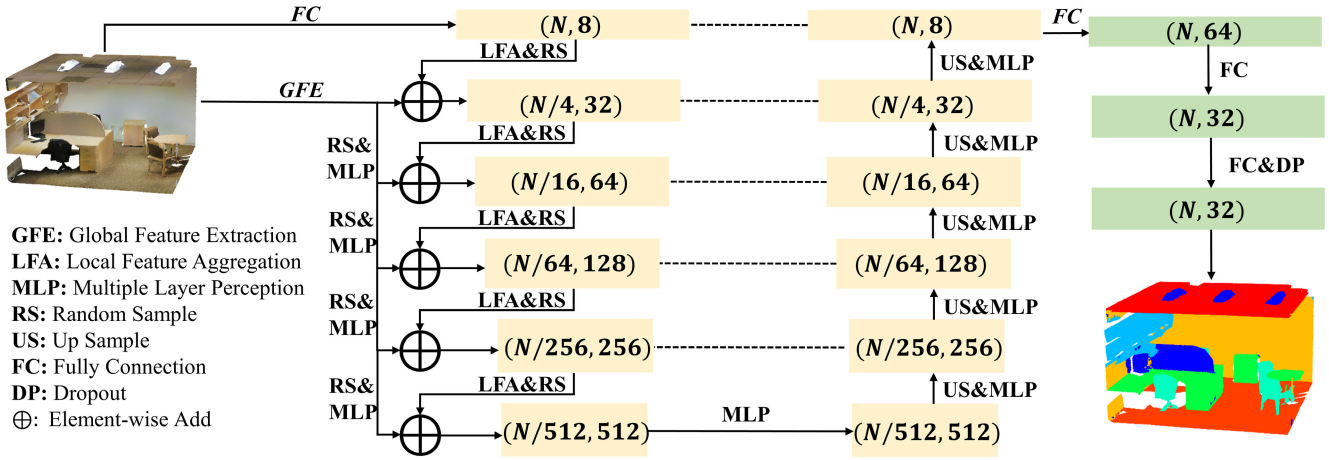


FIGURE 2. Network architecture of LGFF-Net.

deep learning techniques to 3D scene understanding. Existing methods can be categorized into three groups based on the input data representation: projection-based, voxel-based, and point-based methods.

A. PROJECTION-BASED AND VOXEL-BASED METHODS

Projection-based methods [15], [16], [36], [37], [38] convert 3D objects into 2D projections using specialized algorithms. These projections are then analyzed and processed using 2D image algorithms to extract features, which are subsequently fused to obtain the overall characteristics of the 3D object. While these methods have demonstrated promising results in point cloud semantic segmentation, the conversion of 3D data into 2D images can result in the loss of geometric information. Additionally, variations in viewing angles and distances can affect the performance of point cloud semantic segmentation.

Voxel-based methods [12], [13], [39] represent point clouds as voxels, which are then processed using convolutional neural networks to extract general features of 3D objects. While this method can significantly improve performance, it can also result in many empty voxels, which can consume considerable computing resources and time.

B. POINT-BASED METHODS

The limitations of projection-based and voxel-based methods have prompted researchers to explore the direct use of point cloud data for 3D scene understanding. PointNet [17] was a pioneering method in this category, using MLP to extract global features from point clouds and incorporating symmetric functions to address the disorder of point clouds. However, PointNet did not take into account the local structure and context of point clouds. To address these issues, a series of works [40], [41], [42], [43], [44] were initiated, inspired by PointNet. Qi et al. [18] proposed PointNet++, which built upon PointNet by introducing a hierarchical structure through farthest point sampling to learn local area features. Other methods, such as KP-Conv [41] and PA-Conv [42], implemented adaptive convolution operations on 3D point clouds by designing variable convolution kernels.

Although the methods mentioned above are extensions of PointNet [17], they all use MLP to process the overall information of point clouds without distinguishing between geometric and semantic information. However, point clouds comprise an extensive collection of ordered points that typically encompass both geometric and semantic information. Therefore, to effectively use geometric and semantic information for semantic segmentation, it is essential to carefully handle the relationship between these two types of information.

Recent research [20], [21], [23], [26] has started focused on separately encoding and processing geometric and semantic features to achieve more comprehensive representations. For example, RandLA-Net [20] is a method that encodes geometric information and then combines it with semantic features in parallel to obtain an overall feature representation of the point cloud. Other approaches, such as CSA [23] and BAAF-Net [25], propose treating geometric and semantic features equally. CSA introduced cross self-attention to enable interaction between the attention scores and weights of geometric and semantic features. BAAF-Net proposed a bilateral structure to combine geometric and semantic features and perform multiple encodings and connections for feature representation.

However, in most methods that treat geometric and semantic features separately, using the same encoding process may inadvertently diminish the distinction between the two types of features. The unique advantages of location information provided by geometric features may no longer be prominent, and the richness of the overall feature representation provided by semantic features may also be reduced. This can potentially limit the network's ability to effectively utilize both types of features. In general, geometric and semantic features should mutually influence and enhance each other, but it is also crucial to carefully consider and account for their inherent differences.

C. EXTRACTION OF GLOBAL FEATURE

PointNet [17], as a pioneering method for using point-based representations, initially focused on obtaining global representations of point cloud features by processing the global features of point clouds using MLP. However, it was not effective at representing local information. Subsequent research began to emphasize the processing of local area information while overlooking the role of global features.

Recently, the Non-Local Block [45] has emerged as a powerful tool for obtaining global features, and a series of studies [43], [46], [47], [48] have followed its lead. For example, Du et al. [43] proposed a local-global graph convolution method that aggregates local features and then feeds them into a global spatial attention module. Nie et al. [46] designed a scale pyramid architecture to explore how different scales should interact and merge.

The studies mentioned above demonstrate that further mining of global features can provide sufficient contextual information for scene prediction. Inspired by these ideas, our research focuses on the combination of both local and global information to improve the performance of semantic segmentation. However, the Non-Local Block [45] requires significant computational resources and high-end equipment, making it challenging to apply in practice. Therefore, finding a concise global feature extraction method with low computational cost remains a worthwhile research direction.

Taking into account the methods mentioned above, we propose the LFA module to enable interaction between geometric and semantic features while fully considering their differences and connections and preserving their respective characteristics. Additionally, our GFE module, designed with a simple and effective structure, can efficiently obtain global information with low computational cost.

III. METHODS

Our research introduces a novel point cloud semantic segmentation network called LGFF-Net. This network comprises two modules: the LFA module and the GFE module. These modules effectively integrate local and global features into our network architecture. In this paper, we will discuss the overall network architecture, the LFA module for aggregating local features, and the GFE module for extracting global features.

A. OVERALL ARCHITECTURE

Figure 2 illustrates the overall architecture of LGFF-Net for semantic segmentation. The network adopts a U-shaped structure with skip connections, a widely used approach. However, LGFF-Net deviates from the traditional U-shaped network due to incorporating global information. To enhance efficiency, each layer performs local feature aggregation with the LFA module and efficient random sampling. The GFE module extracts global features and adds them to each layer by changing the feature dimension through MLP operation, preserving intricate global geometric information to a significant extent. We will provide more detailed explanations of these components below.

1) NETWORK INPUT

LGFF-Net takes a large point cloud of size $N \times d_{in}$ as input, where N represents the number of points and d_{in} represents the feature dimension of each input point. For datasets such as S3DIS [31], STLP3D [29], and Toronto_3D [30] with RGB, each point is represented by its 3D coordinates and color information. For datasets such as ScanNet [32] and Toronto_3D without RGB, each point is represented only by its 3D coordinates.

2) ENCODING LAYER

The network structure hierarchically encodes the input feature through five encoding layers in series. Each layer comprises a proposed LFA module and a random sampling [20] operation. The global features obtained by GFE are added to each encoding layer through downsampling and MLP. The input of other layers combines the global features extracted by GFE with the output of the previous layer.

3) DECODING LAYER

The network structure includes five decoding layers designed symmetrically with the encoding layer. Each decoding layer consists of an upsampling operation and an MLP, and the feature map is restored to its original resolution using nearest neighbor interpolation. The intermediate feature maps of the encoding layers are combined with the upsampled feature maps via skip connections.

4) FINAL SEGMENTATION RESULT

Finally, the network uses a dropout layer with a ratio of 0.5 and three fully-connected layers to obtain the semantic segmentation result, which assigns a semantic label to each poinat.

B. LOCAL FEATURE AGGREGATION

In general, a point cloud $P = \{p_i\}_{i=1}^{N}$ contains N points. Each point p_i has inherent three-dimensional coordinates $d_i = \{x_i, y_i, z_i\}$ in space, representing its geometric information. Additionally, each point has a feature f_i obtained in the feature space, representing its semantic information.

Most existing methods fail to distinguish between geometric and semantic features or overlook their disparities when encoding them, resulting in subpar overall performance. To address this issue, we thoroughly consider the appropriate utilization of both geometric and semantic features. Our approach enhances the geometric and semantic characteristics of point clouds to facilitate the learning of comprehensive feature maps for precise semantic segmentation. In designing the LFA module, we fully account for the differentiation and correlation between geometric and semantic features. The FCE unit illustrated in Figure 3 gives due importance to the interaction between geometric and semantic features, with the original attributes of each feature occupying a central role.

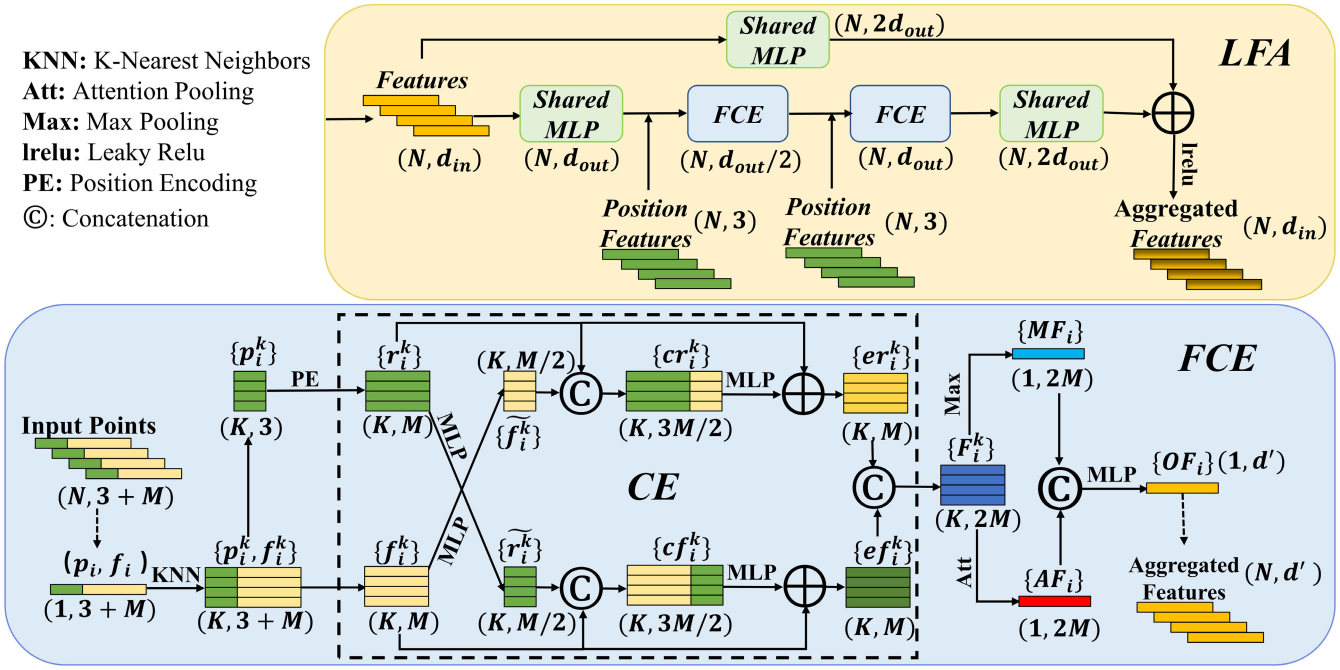


FIGURE 3. Proposed LFA. Geometric and semantic features are mutually augmented, and after aggregation by Max Pooling and Attention Pooling, the MLP is finally performed.

Furthermore, the addition of a residual connection ensures that the geometric and semantic features retain their original characteristics even after their interaction.

1) FEATURE CROSS ENHANCEMENT

Our method constructs local regions by using the K Nearest-neighbors algorithm (KNN) to obtain surrounding neighboring points. The neighboring point set $P_i = \{p_i^j\}_{j=1}^{K}$ contains K points of p_i and reflects its features to a certain degree. We follow the local position encoding strategy of RandLA-Net [20], as shown in equation 1. This strategy has been fully proven to effectively retain and utilize local geometric information.

$$r_i^k = MLP(d_i \bigoplus d_i^k \bigoplus (d_i - d_i^k) \bigoplus ||d_i - d_i^k||). \quad (1)$$

In the Cross Encoding (CE) unit, the Feature Cross Enhancement (FCE) black dotted box in Figure 3 encodes the geometric feature of adjacent points r_i^k and the semantic feature f_i^k to enhance each other without affecting the nature of the original feature. The feature dimensions of r_i^k and f_i^k are reduced to half of their original size to obtain $\tilde{r_i^k}$ and $\tilde{f_i^k}$. Then, the following operations are performed:

$$\begin{cases} er_i = MLP(Concat(\tilde{f_i^k}, r_i^k)) + r_i^k \\ ef_i = MLP(Concat(\tilde{r_i^k}, f_i^k)) + f_i^k. \end{cases} \quad (2)$$

In the above equation, er_i^k and ef_i^k represent the geometric and semantic features after CE. The features are further generalized through MLP, and the integrity of the original features is emphasized through the residual connection. These operations allow r_i^k and f_i^k to interact with each other while maintaining their original feature properties.

Point-wise feature representation is crucial for semantic segmentation [25]. Existing methods typically use Max Pooling or Mean Pooling to aggregate local features, but this can result in significant information loss. In our work, we use both Max Pooling and Attention Pooling to capture salient local features and the entire local region. The mixed feature of p_i is obtained by enhanced feature concatenation and is represented as $F_i^k = \{er_i^k \bigoplus ef_i^k\}$. After Max Pooling, the local max feature MF_i is obtained, as follows:

$$MF_i = Max(F_i^k). \quad (3)$$

Inspired by RandLA-Net [20] and SE-Net [49], the more essential features the higher attention scores. To do this, we first calculate the attention score for each point as follows:

$$s_i^k = SoftMax(MLP(F_i^k)). \quad (4)$$

The attention scores s_i^k are multiplied by F_i^k as attention-weighted features. We focus on the entire local region, so we use a symmetric function to sum the attention-weighted features of each point to get AF_i: The attention scores s_i^k are multiplied by F_i^k to obtain attention-weighted features. To focus on the entire local region, we use a symmetric function to sum the attention-weighted features of each point and obtain AF_i:

$$AF_i = \sum_{k=1}^{K} F_i^k \cdot s_i^k. \quad (5)$$

The final output feature OF_i is obtained by combining the local max feature MF_i and the attention-weighted feature AF_i as follows:

$$OF_i = MLP(Concat(MF_i, AF_i)). \tag{6}$$

2) LOCAL FEATURE AGGREGATION

During the downsampling process of point clouds, some points are discarded, resulting in the loss of the features carried by these points. To address this issue, we increase the receptive field of the retained points to make it more likely for them to retain more features. As shown in Figure 3, inspired by the successful ResNet [28] and RandLA-Net [20], we combine the FCE unit with residual connections to form the LFA module.

$$\begin{cases} p'_j = FCE(p_j), K_j = \{p_j^k\}_{k=1}^{K} \\ p'_m = FCE(p_m), K_m = \{p_m^k\}_{k=1}^{K} \\ p'_i = FCE(p_i), K_i = \{p'_j, p'_m, p_i^k\}_{k=1}^{K-2}. \end{cases} \tag{7}$$

For a given point p_i, there are neighboring points p_j and p_m. Let K_j and K_m be the sets of neighboring points of p_j and p_m, respectively. After aggregation by the first FCE unit, got feature aggregation points p'_j and p'_m. The neighboring points of p_i are represented by the set K_i, which includes p'_j, p'_m, and other points. Finally, after processing by the second FCE unit, obtained the feature aggregation point p'_j.

This process can obtain information from up to K^2 adjacent points. After stacking N units, the number of adjacent points can reach K^N. In theory, stacking more units can result in more adjacent points being obtained and more features being extracted. However, more units also consume more computing resources. Additionally, having too many neighboring points can increase the number of irrelevant features, which may affect the segmentation result. In the LFA module, we stack two FCE units to achieve a good balance between effectiveness and computational efficiency. We provide ablation experiments on the number of LFA stacks in Section IV-C4.

C. GLOBAL FEATURE EXTRACTION

It is important to note that the global feature of point clouds has always been a significant factor in the performance of semantic segmentation of point clouds. Therefore, an effective and straightforward method for extracting global features is crucial. Although PointNet [17] does not consider the local region or the connection between contexts, its effectiveness and simplicity in extracting global features are noteworthy. Inspired by PointNet, we designed the Global Feature Extraction (GFE) structure shown in Figure 4.

$$\begin{cases} F_g = MLP(F_p) \\ F'_g = MB(F_g) \\ \tilde{F}_g = MLP(Max(F_g + F'_g)). \end{cases} \tag{8}$$

In the GFE structure, F_p represents the original information of the point cloud and serves as the input. After dimensionality enhancement with an MLP, obtain the initial global feature

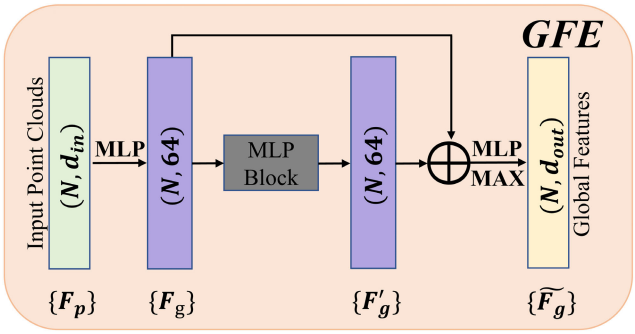


FIGURE 4. Proposed GFE. In the MLP Block, the dimension of the point clouds is changed to 128 and 1024 in that order.

F_g. The MLP Block (MB) is a unit composed of multiple MLPs that processes the feature dimensions of F_g to 64, 128, 1024, and then to 64, to obtain the intermediate global feature F'_g. After applying a residual connection and max pooling, obtain the final global feature $\tilde{F}_g$.

We used the general structure of PointNet [17] but made some modifications, including changing the number of MLPs and the dimensions of features in all aspects. We also removed the T-Net module, which has been shown to have little impact on semantic segmentation. Through these changes, we fully utilized the critical process of PointNet for global feature extraction to further improve the extraction efficiency of GFE and fully utilize the efficiency of MLP. Additionally, using residual connections after multiple MLPs helps to avoid vanishing gradients.

IV. RESULTS

Our LGFF-Net is implemented using TensorFlow on a server equipped with a Tesla V100 GPU. We use the Adam optimizer [50] with an initial learning rate of 0.01, which decreases by 5% after each epoch. Following RandLA-Net [20], the number of training epochs is set to 100, and the number of K in the LFA module is 16. As proposed in the original papers for the Semantic3D and SemanticKITTI datasets [14], [51], we use overall accuracy (OA) and mean intersection over union (mIoU) as our primary evaluation metrics, defined as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

$$mIoU = \frac{\sum_{i=1}^{n} IoU_i}{n} \tag{10}$$

$$OA = \frac{TP}{N}, \tag{11}$$

where TP is the number of true positive samples, FP is the number of false positive samples, FN is the number of inaccurate negative samples, n is the number of semantic labels, and N is the total number of samples.

A. DATASET

While our research primarily focuses on enhancing the intersection of geometric and semantic information, we have also

TABLE 1. Number of labeled points for each class of STPLS3D (in million).

Set	Section	Ground	Building	Tree	Car	Light pole	Fence	Total
Training	OCCC	16.25	14.89	6.69	0.96	0.07	0.06	38.91
	RA	18.70	17.31	8.15	1.10	0.09	0.39	45.74
	USC	41.72	59.52	36.14	1.58	0.27	1.21	140.43
Testing	WMSC	48.35	60.65	38.23	1.62	0.28	1.29	150.41
Total		125.02	152.36	89.20	5.26	0.72	2.94	375.50

TABLE 2. Number of labeled points for each class of Toronto_3D (in thousand).

Set	Section	Unclassified	Road	Road mrk.	Natural	Building	Util. line	Pole	Car	Fence	Total
Training	L001	391	11178	433	1408	6037	210	263	1564	83	21567
	L003	1760	20587	786	1908	11672	332	408	1969	300	39722
	L004	582	3738	281	1310	525	37	71	200	4	6748
Testing	L002	360	6353	301	1942	866	84	155	199	24	10284
Total		3093	41856	1801	6568	19100	663	897	3932	411	78321

TABLE 3. Number of labeled points for each class of S3DIS (in million).

Section	ceil	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clut.	Total
Area1	8.32	6.45	11.46	2.13	1.30	1.48	2.24	1.62	1.12	0.17	1.77	0.80	5.17	44.03
Area2	10.00	9.48	12.54	0.40	0.37	0.18	2.53	0.72	3.62	0.15	1.89	0.47	4.93	47.29
Area3	3.69	3.01	5.13	0.35	0.35	0.33	0.87	0.55	0.47	0.19	1.28	0.29	2.15	18.66
Area4	7.68	7.01	13.44	0.10	0.92	1.09	2.80	1.20	1.31	0.27	2.65	0.21	5.03	43.68
Area5	15.38	13.00	22.95	0.02	1.38	2.76	2.38	2.95	1.47	0.21	8.13	0.93	7.01	78.59
Area6	7.64	6.26	10.57	1.74	1.21	1.05	2.25	2.22	1.41	0.17	1.60	0.69	4.55	41.35
Total	52.71	45.21	76.08	4.74	5.53	6.89	13.07	9.27	9.40	1.16	17.32	3.39	28.84	273.61

conducted experiments to evaluate the performance of point clouds that contain only geometric information. Specifically, we evaluated our method using four datasets: STPLS3D [29], Toronto_3D [30], S3DIS [31], and ScanNet [32], which consist of both indoor and outdoor scenes and include datasets with only geometric information as well as those with both geometric and semantic information. Our model was trained using a batch size of 3 for Toronto_3D and a batch size of 6 for STPL3D, S3DIS, and ScanNet.

It is important to note that our method focuses on the mutual enhancement of geometric and semantic features to improve the performance of semantic segmentation. However, the ScanNet dataset and the Toronto_3D without RGB dataset contain only geometric information and lack semantic information. Due to the limitations of these datasets and our work, we use the geometric information of the point cloud to construct semantic information in this study. Specifically, we use the geometric information of the point cloud as the semantic information.

1) STPLS3D

The STPLS3D [29] dataset, provided by the University of Southern California, is a large-scale photogrammetry 3D point cloud dataset that comprises both real and synthetic scenes. The real scenes encompass an area of approximately 1.27 km^2 and include locations such as the University of Southern California Park Campus (USC), Wrigley Marine Science Center (WMSC) on Catalina Island, Orange County Convention Center (OCCC), and a residential area (RA). The synthetic point clouds span roughly 16 km^2 of cityscapes and contain up to 18 fine-grained semantic and 14 instance classes. In our study, we utilized the point clouds from the real scenes of the STPLS3D dataset. Adhering to the guidelines set forth in the original paper, we classified the point clouds into six categories and used OCCC, RA, and USC for training while reserving WMSC for testing. Table 1 provides further details about the STPLS3D dataset.

2) TORONTO_3D

Toronto_3D [30] is a large-scale outdoor urban point cloud dataset collected along Avenue Road in Toronto, Canada, specifically designed for semantic segmentation. The dataset encompasses approximately 1 km of point clouds, comprising roughly 78.3 million points, and has been divided into four roads and eight labels. Each point contains 3D coordinates, RGB information, intensity, GPS time, scan angle rating, and one label. In our work, adhered to the guidelines provided in the original paper of the Toronto_3D dataset, where L001, L003, and L004 were used for training while L002 was reserved for testing, as shown in Table 2.

3) S3DIS

The S3DIS [31] dataset, also known as Stanford Large-Scale 3D Indoor Spaces, is a large-scale indoor 3D point cloud dataset provided by Stanford University. The dataset comprises 271 rooms, 11 scenes, and approximately 273.61 million points, divided into six teaching and office areas and 13 labels, as shown in Table 3. Each point in the dataset

TABLE 4. Quantitative evaluation results on STPLS3D dataset.

Method	OA (%)	mIoU (%)	Ground	Building	Tree	Car	Light pole	Fence
PointTransformer [24]	54.3	36.3	40.0	20.9	62.6	36.1	49.3	8.8
SCF-Net [22]	75.8	45.9	68.8	37.3	65.5	51.5	31.2	**21.3**
MinkowskiNet [52]	70.4	46.5	64.2	30.0	61.3	46.0	**65.3**	12.4
KPConv [41]	70.7	45.2	60.9	32.1	69.1	53.8	52.1	3.4
PointRas [53]	-	47.4	-	-	-	-	-	-
RandLA-Net [20]	60.2	42.3	46.1	24.2	**72.5**	53.4	44.8	13.0
LGFF-Net(Ours)	**78.8**	**49.1**	**70.7**	**55.0**	57.3	**59.1**	38.8	13.6

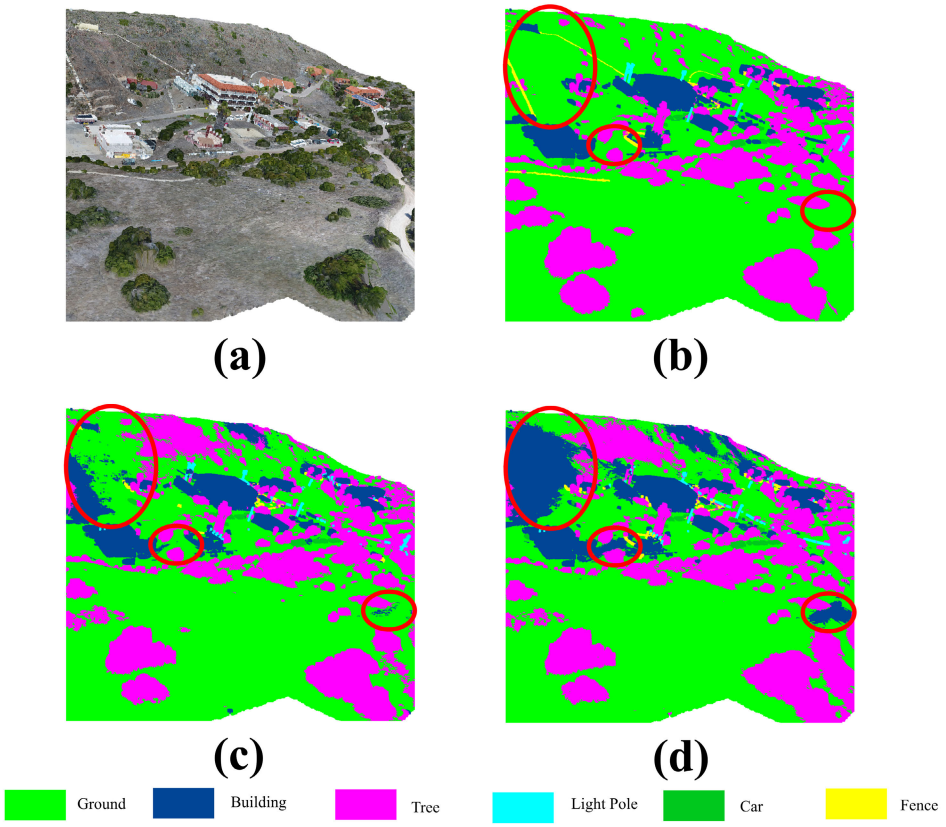


FIGURE 5. Visual comparison diagram of STPLS3D dataset, (a) is full RGB, (b) is ground truth, (c) results with our LGFF-Net, and (d) results with RandLA-Net.

includes coordinate information, RGB information, and one label. Area 5 of S3DIS dataset has the largest number of points and an unbalanced category distribution, making it a challenging area for evaluation. We evaluated the performance of LGFF-Net on Area 5 of S3DIS dataset and designed a series of ablation experiments on the Area.

4) ScanNet

The ScanNet [32] dataset comprises 1513 indoor scenes that have been scanned and reconstructed. The dataset is divided into 1201 training scenes and 312 test scenes and includes 20 categories. Each point in the dataset contains *XYZ* coordinates and a label. In our evaluation, we reported the per-voxel accuracy using the method employed in Point2Node [54] to ensure a fair comparison.

B. EVALUATION ON DIFFERENT DATASETS

In this section, we present the experimental results of LGFF-Net on the datasets described in Section IV-A and provide an analysis of the results.

1) EVALUATION ON STPLS3D

Table 4 displays the quantitative evaluation results of LGFF-Net on STPLS3D [29], with the best results highlighted in bold. The methods listed include PointTransformer [24], SCF-Net [22], MinkowskiNet [52], KPConv [41], and RandLA-Net [20], all of which are mentioned in the original STPLS3D paper, as well as PointRas proposed by Zheng et al. In comparison, LGFF-Net achieved 78.8% in OA and 49.1% in mIoU. LGFF-Net demonstrated state-of-the-art performance overall, with satisfactory results

TABLE 5. Quantitative evaluation results on Toronto_3D dataset.

RGB	Method	OA (%)	mIoU (%)	Road	Road mrk.	Natural	Building	Util. line	Pole	Car	Fence
N	PointNet++ [18]	84.9	41.8	89.3	0.0	69.0	54.1	43.7	23.3	52.0	3.0
	DGCNN [19]	94.2	61.8	93.9	0.0	91.3	80.4	62.4	62.3	88.3	15.8
	KPConv [41]	95.4	69.1	94.6	0.1	96.1	91.5	87.7	81.6	85.7	15.7
	MS-PCNN [55]	90.0	65.9	93.8	3.8	93.5	82.6	67.8	72.0	91.1	22.5
	TGNet [56]	94.1	61.3	93.5	0.0	90.8	81.6	65.3	63.0	88.7	7.9
	MS-TGNet [30]	95.7	70.5	94.4	17.2	95.7	88.8	76.0	74.0	**94.2**	23.6
	RandLA-Net [20]	93.0	77.7	94.6	42.6	96.9	93.0	86.5	78.1	92.9	37.1
	LGFF-Net(Ours)	95.2	71.1	94.6	0.0	95.1	91.6	83.9	73.0	88.4	42.3
Y	Rim et al. [57]	83.6	71.0	92.8	27.4	89.9	**95.2**	85.6	74.5	44.4	**58.3**
	ResDLPS-Net [58]	96.5	80.3	95.8	59.8	96.1	90.9	86.8	79.9	89.4	43.3
	BAAF-Net [25]	94.2	81.2	96.8	67.3	96.8	92.2	86.8	82.3	93.1	34.0
	RG-GCN [59]	96.5	74.5	**98.2**	**79.4**	91.8	86.1	72.4	69.9	82.1	16.0
	MFA [60]	97.0	79.9	96.8	70.0	96.1	92.3	86.3	80.4	91.5	29.4
	NeiEA-Net [61]	97.0	80.9	97.1	66.9	**97.3**	93.0	87.3	**83.4**	93.4	43.1
	Wang et al. [62]	95.3	73.9	95.7	25.9	94.0	86.3	81.5	71.8	78.1	58.1
	RandLA-Net [20]	94.4	**81.8**	96.7	64.2	96.9	**94.2**	**88.1**	77.8	93.4	42.9
	LGFF-Net(Ours)	**97.2**	81.4	96.9	65.5	96.1	92.7	86.0	78.8	93.6	41.4

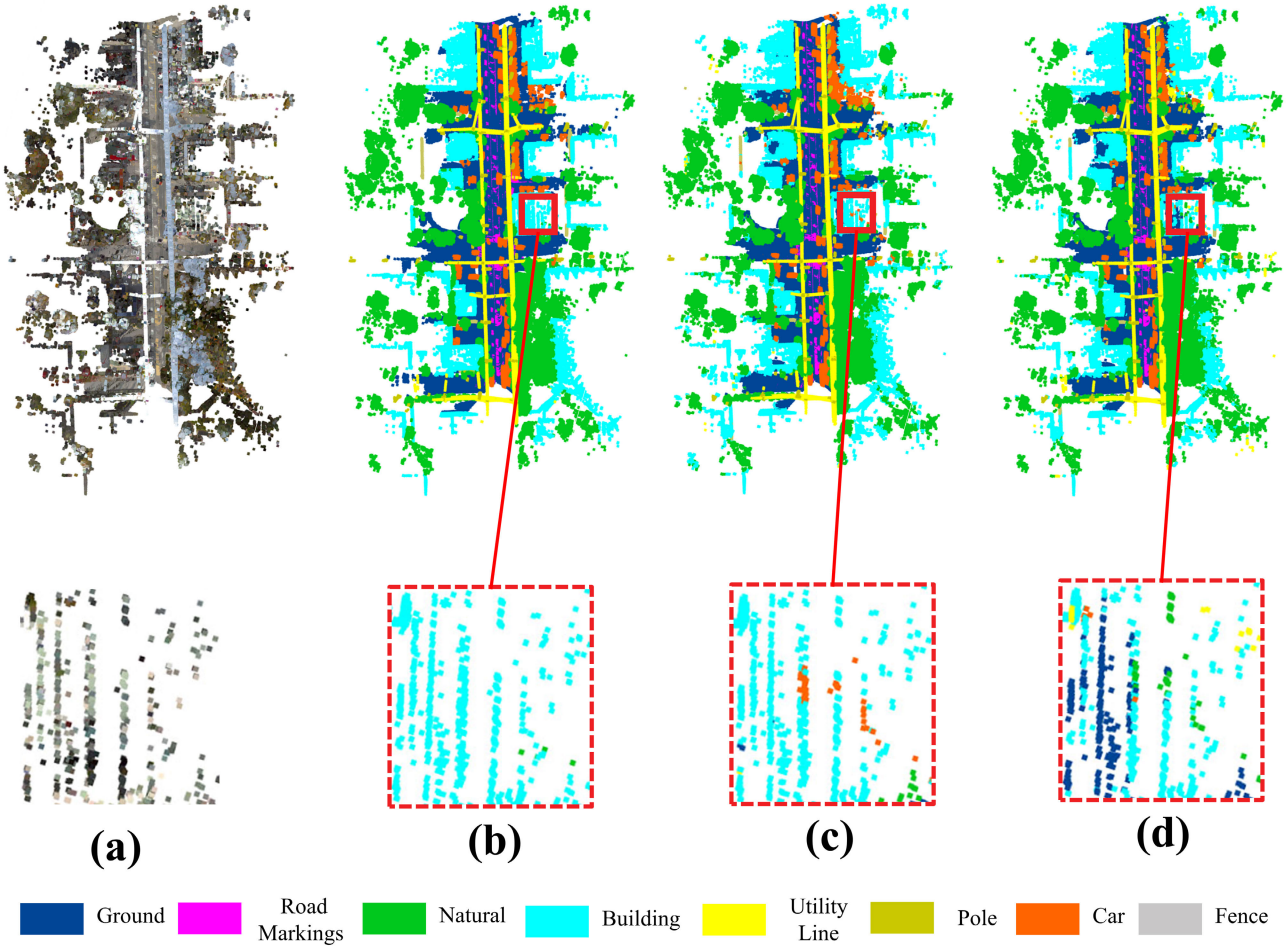


FIGURE 6. Visual comparison diagram of Toronto_3D dataset with RGB, (a) is full RGB, (b) is ground truth, (c) results with our LGFF-Net, and (d) results with RandLA-Net.

in over half of the categories. Notably, LGFF-Net outperformed RandLA-Net by a significant margin, with an improvement of 18.6% in OA and 6.8% in mIoU.

It is worth noting that LGFF-Net performed better for categories such as *Ground*, *Building*, and *Car*, which are closely related in terms of geometric and semantic information.

TABLE 6. Quantitative results on the S3DIS dataset Area5.

Method	OA (%)	mIoU (%)	ceil	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clut.
PointNet++ [18]	-	50.0	90.8	96.5	74.1	0.0	5.8	43.6	25.4	69.2	76.9	21.5	55.6	49.3	41.9
PointGCR [63]	-	54.4	90.7	96.1	74.9	**0.1**	16.1	50.2	32.3	69.0	78.1	41.3	60.7	53.8	43.8
PointCNN [40]	86.0	57.3	92.3	98.2	79.4	0.0	17.6	22.8	62.1	74.4	80.6	31.7	66.7	62.1	56.7
PointWeb [64]	87.0	60.3	92.0	98.5	79.4	0.0	21.1	59.7	34.8	76.3	88.3	46.9	69.3	64.9	52.5
Point2Node [54]	**88.8**	63.0	93.9	98.3	**83.3**	0.0	35.7	55.3	58.8	79.5	84.7	44.1	71.1	58.7	55.2
GA-Net [65]	87.6	63.7	92.9	97.8	81.3	0.0	27.8	60.3	41.7	78.3	86.7	**71.4**	69.9	65.8	53.9
SC-CNN [44]	-	63.1	93.8	**98.7**	80.0	0.0	17.4	55.6	50.6	76.4	88.0	66.8	**71.3**	64.1	56.8
DGCNN [19]	87.0	56.5	92.7	93.6	77.5	0.0	**36.3**	52.5	**63.7**	77.6	62.8	48.7	33.6	45.8	50.2
SPH3D-GCN [66]	86.6	58.6	92.2	97.2	79.9	0.0	32.0	52.2	41.6	85.3	76.9	67.2	36.5	50.7	50.0
LGGCM [43]	88.8	63.3	**94.8**	98.3	81.5	0.0	35.9	63.3	43.5	80.2	**88.4**	68.8	55.7	64.6	47.8
PCT [67]	-	61.3	92.5	98.4	80.6	0.0	19.4	61.6	48.0	76.6	85.2	46.2	67.7	67.9	52.3
PointRas [53]	88.5	62.6	92.8	97.3	73.4	0.0	18.7	**68.4**	50.3	77.3	85.9	63.8	68.0	60.8	**57.3**
MPVCNN++ [68]	88.8	60.2	-	-	-	-	-	-	-	-	-	-	-	-	-
DPFA-Net [69]	88.0	55.2	93.0	98.6	80.2	0.0	14.7	55.8	42.8	72.3	73.5	27.3	55.9	53.0	50.5
Fan et al. [70]	-	61.9	92.8	97.9	80.8	0.0	20.0	56.9	43.0	73.5	82.8	50.8	65.8	64.3	50.5
KPConv [41]	-	63.0	93.4	98.3	78.9	0.0	18.7	52.2	56.6	**87.0**	76.8	71.3	71.0	62.1	52.9
RandLA-Net [20]	86.7	61.6	91.2	95.6	79.5	0.0	20.6	59.9	43.4	76.5	82.8	60.8	70.4	67.9	52.0
LGFF-Net(Ours)	87.0	**63.9**	91.3	96.6	81.4	0.0	32.4	62.5	52.5	76.0	84.2	66.7	68.0	**69.0**	50.6

However, discrete categories such as *Light pole* and *Tree*, which are widely distributed, is a significant challenge for the proposed LFA module. MinkowskiNet [52], with its generalized sparse convolutions, performed better on discrete categories such as *Light pole*. While RandLA-Net [20], with attention pooling, performed better on widely distributed *Trees*. Among the methods listed, PointTransformer's [24] performance was unsatisfactory, possibly due to its large number of parameters that require training in the network but the limited number of training samples provided by STPL3D [29].

Figure 5 displays the quantitative visualization results of LGFF-Net and RandLA-Net [20] on the STPLS3D [29] dataset, which are consistent with the results presented in Table 4. Overall, LGFF-Net has demonstrated significant progress in comparison to other leading methods.

2) EVALUATION ON TORONTO_3D

Table 5 presents the quantitative evaluation results on Toronto_3D [30], with the best results highlighted in bold. On the Toronto_3D dataset with RGB, LGFF-Net achieved 97.2% in OA and 81.4% in mIoU, demonstrating the best performance in OA and the second-best mIoU compared to RandLA-Net [20]. On the Toronto_3D dataset without RGB, LGFF-Net obtained 95.2% in OA and 71.1% in mIoU. Although our method did not achieve the best mIoU, it demonstrated significant progress in OA for the Toronto_3D dataset, both with and without RGB.

On the Toronto_3D dataset with RGB, LGFF-Net achieves superior overall performance compared to BAAF-Net [25], with better OA and mIoU. While the mIoU is slightly lower than that of RandLA-Net [20], the OA is significantly improved. Moreover, LGFF-Net achieves competitive results in all categories, on par with RandLA-Net and BAAF-Net.

On the other hand, LGFF-Net does not achieve the best results in OA and mIoU for the Toronto_3D dataset without RGB. This is primarily due to LGFF-Net's focus on combining geometric and semantic features. Although it did not achieve the best results on the Toronto_3D dataset without RGB, LGFF-Net still demonstrated satisfactory overall performance.

As illustrated in Figure 6, while the addition of LFA can improve the accuracy of local area segmentation to a certain extent, it is still unable to completely avoid incorrectly segmented points. However, the segmentation accuracy in the local area has improved significantly compared to the baseline RandLA-Net [20].

3) EVALUATION ON S3DIS

Table 6 presents the quantitative evaluation results of S3DIS [31], with the best results highlighted in bold. Notably, LGFF-Net achieves 87.0% in OA and 63.9% in mIoU, with its superior performance in mIoU attributed to the proposed GFE and LFA modules. However, the limited number of room points in each area of S3DIS hinders GFE from extracting comprehensive global features. In contrast, LFA's constraints on geometric and semantic features enable LGFF-Net to perform better. This observation is also supported by the ablation experiments mentioned in Section 8. KPConv [41] stands out as an exceptional method due to its ability to achieve the best performance with variable convolution kernels on objects of varying sizes, such as *chair* and *bookcase*. Furthermore, LGGCM [43], with LSA-Conv, achieves both short-term and long-term point-to-point dependencies and performs well overall.

Figure 7 presents a visualization of the results of LGFF-Net and RandLA-Net [20] on S3DIS Area5 [31]. LGFF-Net achieves smoother segmentation in some scenes, such as *door* and *board*, than RandLA-Net, with segmentation results even closely resembling ground truth in some scenes. While the random point sampling employed by RandLA-Net is very efficient, its subsequent use of only attention pooling may be the primary reason for its limitations in performance. Similarly, LGFF-Net utilizes random sampling but incorporates

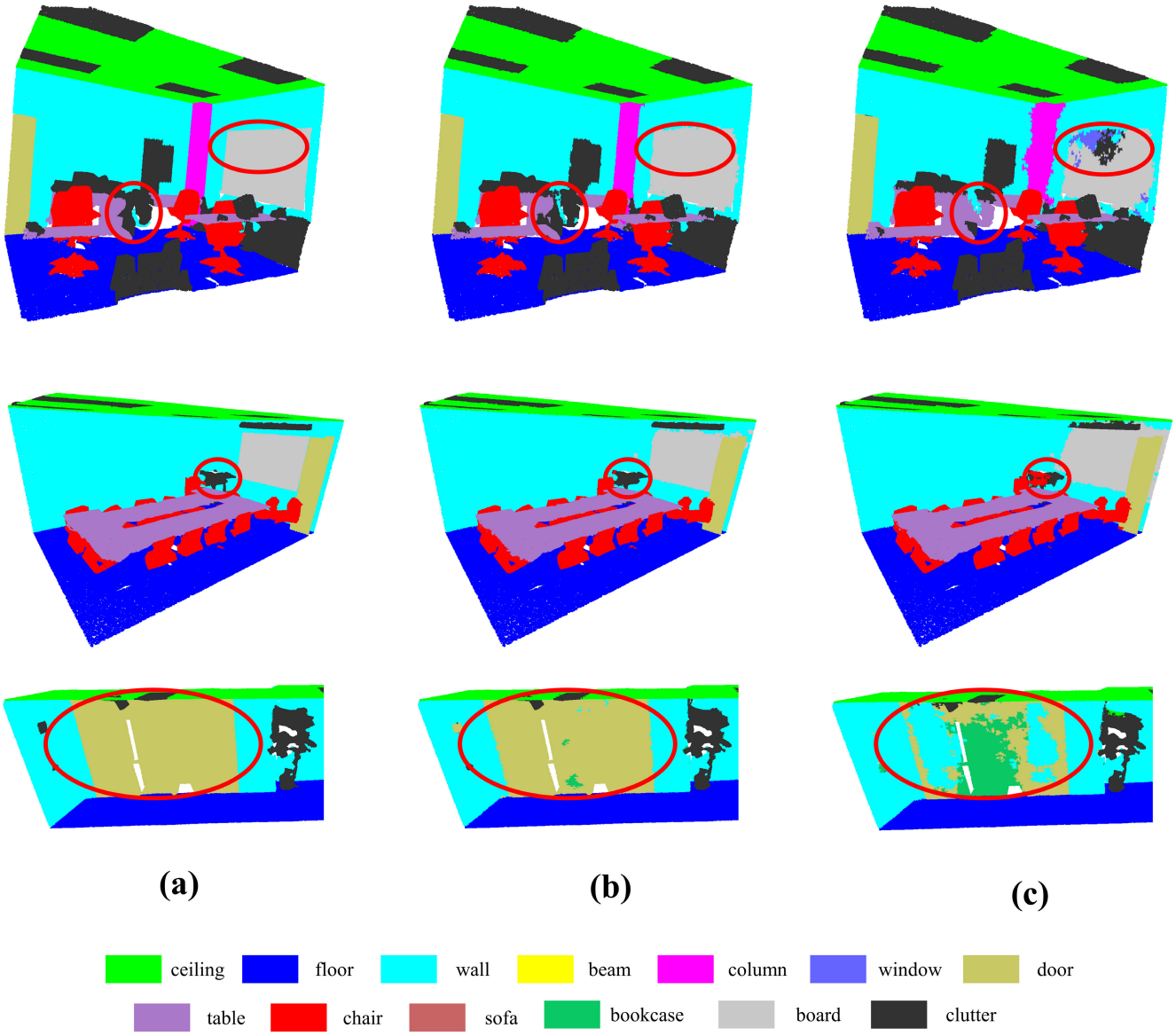


FIGURE 7. Visual comparison diagram of S3DIS Area5, (a) is ground truth, (b) results with our LGFF-Net, and (c) results with RandLA-Net.

global information to ensure the diversity of information to improve performance. The LFA module's constraints on semantics and locations further enhance the segmentation accuracy.

4) EVALUATION ON ScanNet

Table 7 presents the quantitative evaluation results of ScanNet [32], with the best results highlighted in bold. The performance of our method is unsatisfactory, primarily due to two reasons. Firstly, ScanNet only provides geometric information, while our proposed LFA focuses on making geometric and semantic features interact. Secondly, for indoor point clouds, the GFE module struggles to extract comprehensive global features.

TABLE 7. Quantitative results on the ScanNet dataset.

Method	Per-voxel accuracy(%)	mIoU (%)
PointNet++ [18]	84.5	38.3
PointCNN [40]	85.1	45.8
PointGCR [63]	85.3	42.7
GA-Net [65]	86.6	-
LG-Net [71]	**87.1**	**52.3**
RandLA-Net [20]	86.1	-
LGFF-Net (Ours)	85.3	46.6

C. ABLATION STUDIES

To further verify the contribution of each module in our network and to examine the details of each part of the network, we designed multiple sets of ablation experiments on S3DIS Area 5. The results are presented below.

TABLE 8. Ablation study of the proposed modules, "M" means million.

Model	LFA	GFE	Parameters (M)	mIoU (%)
A			4.99	61.6
B		✓	1.65	62.6
C	✓		2.01	63.1
D	✓	✓	2.40	63.9

TABLE 9. Ablation study of the feature dimension of U-shape segmentation network, "M" means million.

Model	Dimension	Parameters (M)	mIoU (%)
A	(16, 64, 128, 256, 512)	8.18	62.5
B	(16, 32, 64, 128, 256)	2.40	63.9
C	(8, 16, 32, 64, 128)	0.72	62.1

1) ABLATION OF PROPOSED MODULE

In our study, we refer to RandLA-Net [20] as Model A. The baseline with the GFE module is represented as Model B. We replaced the local feature aggregation unit in the baseline with LFA and denoted it as Model C. Our proposed LGFF-Net, which includes both GFE and LFA, is labelled as Model D. Models B, C, and D have increased modules but decreased parameters because we reduced the parameters in the U-shape segmentation network. We used mIoU as the evaluation metric and compared the number of parameters to demonstrate the effectiveness of our method.

As shown in Table 8, Model B outperforms Model A, demonstrating the importance of exploring global features for semantic segmentation. Model C also outperforms Model A, possibly due to the cross-enhanced aggregation operation. Our proposed LGFF-Net achieves the best results, attributed to the combination of GFE and LFA, significantly improving semantic segmentation performance.

Furthermore, we reduced the feature dimension of the encoding part of the U-shaped segmentation network while still achieving reliable results. It did not make sense that simply reduce the number of parameters, the goal of ours was to find a suitable feature dimension and balance between accuracy and efficiency. To this end, we designed an ablation experiment of the feature dimension of the U-shaped segmentation network. See section IV-C2 for more details.

2) ABLATION OF THE FEATURE DIMENSION OF U-SHAPE SEGMENTATION NETWORK

To balance model performance and efficiency, we conducted an ablation experiment on the feature dimension of the U-shaped segmentation network. The results are shown in Table 9. We found that reducing the feature dimension model B that based on Baseline achieved a balance between performance and efficiency. B's performance was slightly better than A's and had significantly fewer parameters, demonstrating the effectiveness of reducing the feature dimension. However, further reducing the feature dimension based on B in C resulted in insufficient feature representation and unsatisfactory performance. Therefore, we chose B's feature dimension for our proposed LGFF-Net model.

TABLE 10. Ablation studies of the aggregation way in LFA.

Model	Cross Encoding	Aggregation way	mIoU (%)
A1	N	Mean	54.8
B1	N	Max	60.2
C1	N	Attention	62.6
D1	N	Mean+Attention	62.2
E1	N	Max+Attention	63.3
A2	Y	Mean	56.7
B2	Y	Max	62.0
C2	Y	Attention	61.0
D2	Y	Mean+Attention	60.5
E2	Y	Max+Attention	63.9

TABLE 11. Ablation study of the number of LFA, "M" means million.

Model	Number	Parameters (M)	mIoU (%)
A	1	1.50	59.5
B	2	2.40	63.9
C	3	2.95	63.0

3) ABLATION OF THE AGGREGATION WAY AND CROSS ENCODING IN LFA

This section presents a quantitative analysis of the impact of different aggregation methods and Cross-Encoding (CE) on the LFA module. We evaluated models A1-E1 without CE and using distinct types of aggregation, and models A2-E2 with CE and using distinct types of aggregation.

Table 10 shows that using Max Pooling or Attention Pooling alone resulted in better performance than using Mean Pooling. This may be because Mean Pooling reduces the differences in local region features, making it a suboptimal choice. By using Max Pooling and Attention Pooling, the network can focus on prominent features in the local region while considering the entire local region.

We observed that CE generally improved the model's performance. However, in the model with CE, Max Pooling performed better than Attention Pooling, while the opposite was true in the model without CE. Specifically, in the model using Max Pooling, performance with CE was better than without, whereas in the model using Attention Pooling, performance without CE was better. This difference may be due to prominent features becoming more apparent after cross-encoding. Compared to Attention Pooling, Max Pooling has a stronger selection intention for prominent features, while Attention Pooling's inherent averaging inhibits the selection of such features.

4) ABLATION OF THE NUMBER OF LFA

The selection of the number of LFA modules is related to the selection of the number of adjacent points. The more LFA modules stacked, the more adjacent points are considered. As shown in Table 11, a single LFA module does not select

TABLE 12. Number of FLOPs and parameters of different methods, "M" and "G" mean million and gigabytes.

Method	FLOPs (G)	Parameters (M)	mIoU (%)
PointNet [17]	4.87	3.57	41.1
PointNet++ [18]	1.57	1.88	55.6
GA-Net [65]	3.14	5.01	63.7
LGGCM [43]	1.87	4.26	63.3
LG-Net [71]	3.26	5.16	**64.9**
KPConv [41]	6.50	14.93	63.0
RandLA-Net [20]	3.13	4.99	61.6
LGFF-Net (Ours)	1.53	2.40	63.9

enough adjacent points to represent the features of the entire local region, leading to poor performance. In contrast, stacking three LFA modules results in too many unwanted adjacent points affecting feature representation. Therefore, we chose two LFA modules as the best option, effectively capturing the contextual information of the point cloud and balancing computational complexity. Two LFA modules selected a suitable number of adjacent points to represent the features of the entire local region, leading to better performance. Stacking a third LFA module not only increases computational cost but also degrades performance.

D. EFFICIENCY ANALYSIS

Our proposed method is compared with representative approaches in terms of the number of parameters and computational complexity. We selected floating-point operations (FLOPs) to represent computational complexity, the number of parameters to describe model complexity, and mIoU to evaluate model performance.

Taking Area 5 of the S3DIS [31] dataset as an example, Table 12 shows that our method has the lowest FLOPs, while only PointNet++ [18] has fewer parameters than our proposed method. This demonstrates that our approach is worthy of recognition in terms of computational and model complexity. Although our method's performance did not exceed LG-Net [71], our model parameters and FLOPs are at least two times smaller than LG-Net. Overall, our method shows satisfactory performance and achieves a better balance between accuracy and complexity.

V. CONCLUSION

This paper presents a novel approach for point cloud semantic segmentation called Local-Global Feature Fusing (LGFF-Net). Our proposed method includes a Local Feature Aggregation module that enables the interaction of geometric and semantic features. Additionally, we use a simple and effective Global Feature Extraction module, and local and global features are hierarchically fused in the U-Shape segmentation network. We evaluated our proposed model's performance on four public benchmarks and achieved competitive results compared to state-of-the-art methods.

Since LGFF-Net focuses on the mutual enhancement of geometric and semantic features, there is still room for improvement on datasets that only have geometric information. In future work, we will continue to enhance our model to perform well even on datasets with only geometric information. One feasible idea is to reduce the role of semantic information and focus more on the point cloud's geometric structure, changing semantic information from a decisive factor to a secondary influencing factor. This will ensure that our model can achieve better results on datasets with only geometric information. At the same time, if semantic information is added, its performance will be further improved.

REFERENCES

[1] X. Du, Y. Lu, and Q. Chen, "A fast multiplane segmentation algorithm for sparse 3-D LiDAR point clouds by line segment grouping," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.
[2] W. Wang, Y. Fan, Y. Li, X. Li, and S. Tang, "An individual tree segmentation method from mobile mapping point clouds based on improved 3-D morphological analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2777–2790, 2023.
[3] X. Mi, B. Yang, Z. Dong, C. Chen, and J. Gu, "Automated 3D road boundary extraction and vectorization using MLS point clouds," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5287–5297, Jun. 2022.
[4] E. J. C. Nacpil, Z. Wang, and K. Nakano, "Application of physiological sensors for personalization in semi-autonomous driving: A review," *IEEE Sensors J.*, vol. 21, no. 18, pp. 19662–19674, Sep. 2021.
[5] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.
[6] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 722–739, Feb. 2022.
[7] L. Han, T. Zheng, Y. Zhu, L. Xu, and L. Fang, "Live semantic 3D perception for immersive augmented reality," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 5, pp. 2012–2022, May 2020.
[8] J. Xiong, E.-L. Hsiang, Z. He, T. Zhan, and S.-T. Wu, "Augmented reality and virtual reality displays: Emerging technologies and future perspectives," *Light, Sci. Appl.*, vol. 10, no. 1, p. 216, Oct. 2021.
[9] F. Gul, W. Rahiman, and S. S. N. Alhady, "A comprehensive study for robot navigation techniques," *Cogent Eng.*, vol. 6, no. 1, Jan. 2019, Art. no. 1632046.
[10] K. Zhu and T. Zhang, "Deep reinforcement learning based mobile robot navigation: A review," *Tsinghua Sci. Technol.*, vol. 26, no. 5, pp. 674–691, Oct. 2021.
[11] S. Halder and K. Afsari, "Robots in inspection and monitoring of buildings and infrastructure: A systematic review," *Appl. Sci.*, vol. 13, no. 4, p. 2304, Feb. 2023.
[12] 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)*, Sep./Oct. 2015, pp. 922–928.
[13] 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. 6620–6629.
[14] T. Hackel, N. Savinov, L. Ladicky, J. 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. IV-1/W1, p. 91, Jun. 2017.
[15] X. Wei, R. Yu, and J. Sun, "View-GCN: View-based graph convolutional network for 3D shape analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1847–1856.
[16] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks," *Comput. Graph.*, vol. 71, pp. 189–198, Apr. 2018.
[17] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.
[18] 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. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 30, Dec. 2017, pp. 5099–5018.

[19] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Oct. 2019.

[20] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11105–11114.

[21] H. Shuai, X. Xu, and Q. Liu, "Backward attentive fusing network with local aggregation classifier for 3D point cloud semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 4973–4984, 2021.

[22] S. Fan, Q. Dong, F. Zhu, Y. Lv, P. Ye, and F. Wang, "SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14499–14508.

[23] G. Wang, Q. Zhai, and H. Liu, "Cross self-attention network for 3D point cloud," *Knowl.-Based Syst.*, vol. 247, Jul. 2022, Art. no. 108769.

[24] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16239–16248.

[25] S. Qiu, S. Anwar, and N. Barnes, "Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1757–1767.

[26] Z. Zeng, Y. Xu, Z. Xie, W. Tang, J. Wan, and W. Wu, "LEARD-Net: Semantic segmentation for large-scale point cloud scene," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102953.

[27] Q. Hu, B. Yang, G. Fang, Y. Guo, A. Leonardis, N. Trigoni, and A. Markham, "SQN: Weakly-supervised semantic segmentation of large-scale 3D point clouds," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2022, pp. 600–619.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun./Jul. 2016, pp. 770–778.

[29] M. Chen, Q. Hu, Z. Yu, H. Thomas, A. Feng, Y. Hou, K. McCullough, F. Ren, and L. Soibelman, "STPLS3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset," 2022, *arXiv:2203.09065*.

[30] W. Tan, N. Qin, L. Ma, Y. Li, J. Du, G. Cai, K. Yang, and J. Li, "Toronto-3D: A large-scale mobile LiDAR dataset for semantic segmentation of urban roadways," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 797–806.

[31] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2D-3D-semantic data for indoor scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2017, pp. 1–9.

[32] 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. 2432–2443.

[33] T. Rabbani, F. A. van den Heuvel, and G. Vosselman, "Segmentation of point clouds using smoothness constraint," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 36, no. 5, pp. 248–253, 2012.

[34] R. Hulik, M. Spanel, P. Smrz, and Z. Materna, "Continuous plane detection in point-cloud data based on 3D Hough transform," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 86–97, Jan. 2014.

[35] L. Landrieu, H. Raguet, B. Vallet, C. Mallet, and M. Weinmann, "A structured regularization framework for spatially smoothing semantic labelings of 3D point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 132, pp. 102–118, Oct. 2017.

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

[37] Z. Yang and L. Wang, "Learning relationships for multi-view 3D object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7505–7514.

[38] W. Wang, Y. Xu, Y. Ren, and G. Wang, "Parsing of urban facades from 3D point clouds based on a novel multi-view domain," *Photogramm. Eng. Remote Sens.*, vol. 87, no. 4, pp. 283–293, Apr. 2021.

[39] B. Graham, M. Engelcke, and L. van der Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 9224–9232.

[40] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on X-transformed points," in *Proc. 32nd Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 31, Dec. 2018, pp. 820–830.

[41] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6411–6420.

[42] M. Xu, R. Ding, H. Zhao, and X. Qi, "PAConv: Position adaptive convolution with dynamic kernel assembling on point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3172–3181.

[43] Z. Du, H. Ye, and F. Cao, "A novel local–global graph convolutional method for point cloud semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2022.3155282.

[44] C. Wang, X. Ning, L. Sun, L. Zhang, W. Li, and X. Bai, "Learning discriminative features by covering local geometric space for point cloud analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5703215.

[45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[46] D. Nie, R. Lan, L. Wang, and X. Ren, "Pyramid architecture for multi-scale processing in point cloud segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17263–17273.

[47] Y. Lin, G. Vosselman, Y. Cao, and M. Y. Yang, "Local and global encoder network for semantic segmentation of airborne laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 176, pp. 151–168, Jun. 2021.

[48] Q. Yuan and H. Z. M. Shafri, "Multi-modal feature fusion network with adaptive center point detector for building instance extraction," *Remote Sens.*, vol. 14, no. 19, p. 4920, Oct. 2022.

[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[51] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9297–9307.

[52] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal ConvNets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3070–3079.

[53] Y. Zheng, X. Xu, J. Zhou, and J. Lu, "PointRas: Uncertainty-aware multi-resolution learning for point cloud segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 6002–6016, 2022.

[54] W. Han, C. Wen, C. Wang, X. Li, and Q. Li, "Point2node: Correlation learning of dynamic-node for point cloud feature modeling," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 10925–10932.

[55] L. Ma, Y. Li, J. Li, W. Tan, Y. Yu, and M. A. Chapman, "Multi-scale point-wise convolutional neural networks for 3D object segmentation from LiDAR point clouds in large-scale environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 821–836, Feb. 2021.

[56] Y. Li, L. Ma, Z. Zhong, D. Cao, and J. Li, "TGNet: Geometric graph CNN on 3-D point cloud segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3588–3600, May 2020.

[57] B. Rim, A. Lee, and M. Hong, "Semantic segmentation of large-scale outdoor point clouds by encoder–decoder shared MLPs with multiple losses," *Remote Sens.*, vol. 13, no. 16, p. 3121, Aug. 2021.

[58] J. Du, G. Cai, Z. Wang, S. Huang, J. Su, J. Marcato Junior, J. Smit, and J. Li, "ResDLPS-Net: Joint residual-dense optimization for large-scale point cloud semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 182, pp. 37–51, Dec. 2021.

[59] Z. Zeng, Y. Xu, Z. Xie, J. Wan, W. Wu, and W. Dai, "RG-GCN: A random graph based on graph convolution network for point cloud semantic segmentation," *Remote Sens.*, vol. 14, no. 16, p. 4055, Aug. 2022.

[60] J. Chen, Y. Zhao, C. Meng, and Y. Liu, "Multi-feature aggregation for semantic segmentation of an urban scene point cloud," *Remote Sens.*, vol. 14, no. 20, p. 5134, Oct. 2022.

[61] Y. Xu, W. Tang, Z. Zeng, W. Wu, J. Wan, H. Guo, and Z. Xie, "NeiEA-NET: Semantic segmentation of large-scale point cloud scene via neighbor enhancement and aggregation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 119, May 2023, Art. no. 103285.

[62] Y. Wang, W. Wang, J. Liu, T. Chen, S. Wang, B. Yu, and X. Qin, "Framework for geometric information extraction and digital modeling from LiDAR data of road scenarios," *Remote Sens.*, vol. 15, no. 3, p. 576, Jan. 2023.

[63] Y. Ma, Y. Guo, H. Liu, Y. Lei, and G. Wen, "Global context reasoning for semantic segmentation of 3D point clouds," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2920–2929.
[64] H. Zhao, L. Jiang, C. Fu, and J. Jia, "PointWeb: Enhancing local neighborhood features for point cloud processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5560–5568.
[65] S. Deng and Q. Dong, "GA-NET: Global attention network for point cloud semantic segmentation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1300–1304, 2021.
[66] H. Lei, N. Akhtar, and A. Mian, "Spherical kernel for efficient graph convolution on 3D point clouds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3664–3680, Oct. 2021.
[67] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, Jun. 2021.
[68] W. Zhou, X. Zhang, X. Hao, D. Wang, and Y. He, "Multi point-voxel convolution (MPVConv) for deep learning on point clouds," *Comput. Graph.*, vol. 112, pp. 72–80, May 2023.
[69] J. Chen, B. Kakillioglu, and S. Velipasalar, "Background-aware 3-D point cloud segmentation with dynamic point feature aggregation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5703112.
[70] Y. Fan, K. Liao, Y. Xiao, M. Lu, and W. Yan, "3D point cloud semantic segmentation system based on lightweight FPConv," *IEEE Access*, vol. 11, pp. 31767–31777, 2023.
[71] Y. Zhao, X. Ma, B. Hu, Q. Zhang, M. Ye, and G. Zhou, "A large-scale point cloud semantic segmentation network via local dual features and global correlations," *Comput. Graph.*, vol. 111, pp. 133–144, Apr. 2023.

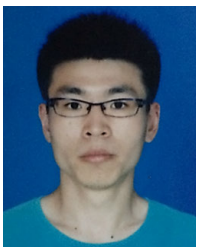
YUANWEI BI received the bachelor's degree in engineering from Yantai University, Yantai, China, in 1993, and the M.E. degree in science from Jilin University, Changchun, China, in 2001.

He has been teaching and conducting research with the Department of Software Engineering, School of Computer and Control Engineering, Yantai University, for many years, as an Associate Professor. His research interests include computer vision, embedded systems, and software engineering. His work in computer vision focuses on image processing, object detection and tracking, and 3D reconstruction. He has also conducted research in the area of embedded systems, particularly in the design and optimization of embedded software and hardware systems.

LUJIAN ZHANG received the bachelor's degree in engineering from Yantai University, Yantai, China, in 2021, where he is currently pursuing the master's degree in computer engineering. His research interests include computer vision, point cloud semantic segmentation, and target detection.

YAOWEN LIU received the bachelor's degree in engineering from Jining University, Jining, China, in 2021. He is currently pursuing the master's degree in computer engineering with Yantai University, Yantai, China. His research interests include computer vision and 3D model registration.

YANSEN HUANG received the bachelor's degree in engineering from Jining University, Jining, China, in 2021. He is currently pursuing the master's degree in computer engineering with Yantai University, Yantai, China. His research interests include computer vision, point cloud analysis, and machine learning.



HAO LIU received the B.E. degree in telecommunication engineering from Shandong Agricultural University, Taian, China, in 2017, and the Ph.D. degree in information science and engineering from Shandong University, Qingdao, China, in 2022. He has been a Lecturer with the School of Computer Science and Control Engineering, Yantai University, since July 2022. His research interests include 3D point clouds compression and processing.