LFT-Net: Local Feature Transformer Network for Point Clouds Analysis

Yongbin Gao, Xuebing Liu, Jun Li, *Member, IEEE*, Zhijun Fang, *Senior Member, IEEE*, Xiaoyan Jiang, and Kazi Mohammed Saidul Huq, *Senior Member, IEEE*

***Abstract*— 6G network enables the rapid connection of autonomous vehicles, the generated internet of vehicles establishes a large-scale point cloud, which requires automatic point cloud analysis to build an intelligent transportation system in terms of the 3D object detection and segmentation. Recently, a great variety of deep convolution networks have been proposed for 3D data analysis, making significant progress in the application of deep learning in 3D computer vision. Inspired by the application of transformer network in 2D computer visual tasks, and in order to increase the expression ability of local fine-grained features, we propose an effective local feature transformer network to learn local feature information and correlations between point clouds. Our network is adaptive to the arrangement of set elements through transformer module, so it is suitable for the feature extraction of local point clouds. In addition, experimental results demonstrate that our LFT-network outperforms the state-of-the-art in 3D model classification tasks on ModelNet40 dataset and segmentation tasks on S3DIS dataset.**

***Index Terms*— 6G, point cloud, 3D computer vision, transfomer, classification, segmentation.**

I. Introduction

IN THE development of the fifth-generation (5G) wireless communication network, Heterogeneous Networks (HetNets) optimized by OAA is a promising solution to realize the 5G network in terms of energy efficiency and capacity [1]. Through the development of the 5G network, the Internet of Things (IOT) as an advanced technology is transforming the Internet into a fully integrated future Internet. It includes Internet of Vehicles (IoV) [2] or IoT applications with other automation terminals, such as UAVs carrying advanced algorithms [3]. In the future, the sixth-generation (6G) wireless communication network can greatly improve the system performance through Reconfigurable intelligent surface (RIS) [4], [5], which is expected to integrate ground, air and sea communications into a more reliable, fast and robust network. Large-scale 3D point cloud analysis is a fundamental task for autonomous driving under 6G network since the understanding of the 3D scenes is essential for intelligent transportation system. The visual research on point cloud and the ability of machine to perceive 3D scene information have always been the research hotspots. In recent years, with the rapid development of low-cost depth sensors and lidars, the acquisition of 3D point cloud data has become easier. Compared with the two-dimensional image data, the three-dimensional point cloud data have unique advantages. For example, it can more abundantly and accurately express the spatial position, attitude and geometric shape information of the object. It is less affected by the change of light intensity, imaging distance and viewpoint change, and there is no projection transformation and occlusion in the two-dimensional image in the large scene. Different from convolutional neural networks in processing 2D structured data such as natural language and image, the disorder and unstructured characteristics of 3D point cloud data make the typical convolution structure unable to be directly applied to unstructured data.

Some of the previous work has made convolution easier to manipulate by structuralizing 3D scene data. For example, the object is represented as a regular voxel in space [6], [7], the 3D point cloud data is decomposed into multi-view images of multi-angle 2D images [8] or based on spherical projection [9], [10], which makes convolution easier to operate. Inspired by PointNet [11], other point-based methods [12], [13] directly throw 3D point cloud into the training network, and use shared MLP and symmetric pooling functions to obtain features on point cloud data. Point-based method learns the relative relationship between points by constructing modules to deal with the disorder of point cloud. For example, the pointSIFT module [13] in encodes the direction of points in point cloud. Recently, more and more networks [14], [15] designs are using the corresponding features of point cloud local spatial structure. The idea of improvement is basically how to better obtain local information. They consider capturing subtle geo-metric structure from tiny localities, gathering localities through continuous down-sampling, in order to obtain more

Manuscript received April 16, 2021; revised October 30, 2021; accepted December 14, 2021. This work was supported in part by the Science and Technology Innovation Action Plan of Shanghai Science and Technology Commission for Social Development Project under Grant 21DZ1204900, in part by the International Collaborative Research Program of Guangdong Science and Technology Department under Grant 2020A0505100061, in part by the Guangzhou Municipal Science and Technology Project under Grant 202102010416, in part by Guangzhou University-The Hong Kong University of Science and Technology (GZU-HKUST) Joint Research Program under Grant YH202110, and in part by the Natural Science Foundation of Guangxi Provence under Grant 2019GXNSFAA245049. The Associate Editor for this article was S. Mumtaz. *(Corresponding authors: Jun Li; Zhijun Fang.)*

Yongbin Gao, Xuebing Liu, Zhijun Fang, and Xiaoyan Jiang are with the School of Electronic and Electronics Engineering, Shanghai University of Engineering Science, Shanghai 201620, China (e-mail: gaoyongbin@sues.edu.cn; zjfang@sues.edu.cn).

Jun Li is with the Research Center of Intelligent Communication Engineering, School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China (e-mail: lijun52018@gzhu.edu.cn).

Kazi Mohammed Saidul Huq is with the Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd CF37 1DL, U.K. (e-mail: kazi.huq@southwales.ac.uk).

Digital Object Identifier 10.1109/TITS.2022.3140355

high-dimensional abstract features, until we obtain the whole global features. The expression of point cloud has two fatal shortcomings. First, the noise cannot be avoided. Second, the amount of data is too large. In the design process of this module, calculation is a serious problem. Therefore, some current work also has some limitations. So at present some work also has some limitations. For example, in [14], if the number of shells too much, the network can not achieve good real-time, too small number will affect the accuracy. Reference [15] uses random sampling in order to be able to process large-scale point clouds, so the accuracy of the network may be unstable. Because random sampling cannot guarantee the quality of points.

Inspired by the amazing effect of the application of the Transformer model [16] proposed by natural language processing in the filed computer vision. The Transformer model in the NLP domain can obtain the corresponding relationship between the source and the target words, and for a complete sentence, it can also effectively obtain the dependence between the source itself and the words. In our task, we can capture the dependencies between points of the source itself through the self-attention mechanism of Transformer model. Therefore, we use the self-attention mechanism of Transformer model to design the local feature transformer unit and trans-pooling unit, and prove that the Transformer model is suitable for feature extraction of point cloud through experiments. The purpose of local feature transformer model is to capture the dependence of local point clouds and learn high-dimensional features through forward neural network. trans-pooling avoids the loss of a large number of feature information for symmetric functions such as max-pooling. Our designed local feature transformer unit and trans-pooling unit are different from the Transformer model proposed in the field of NLP. The main difference is that in the field of NLP, especially the machine translation task is a sequential operation process, so the encoder needs a corresponding decoder. This makes the model complex and requires large computation. In our task, we only need to extract high-dimensional features of local point clouds, so we give up the decoder part and only use the encoder to learn point cloud features. On the other hand, due to the complex 3d structure and the difference base number of object point in space. Such as semantic information of doors and windows edge structure or the juncture of various objects is difficult to learn. The number of ceiling, wall and floor points is much more than that of desks and chairs. Following the work done by Lin *et al.* [17], we apply weighted multi-classification focus loss function to measure loss in classification and segmentation.

The main contributions of this paper are summarized as follows.

1) We propose a local feature transformer module to extract the local correlations of point cloud, based on which, the global abstract features can be learned.
2) We propose a trans-pooling model to highlight the informative points, which preserves more useful features compared with the other pooling operations.
3) We apply the focal loss function to the classification and segmentation task of point cloud to extract the geometric features of complex 3d regions that are difficult to learn and balance the base number of points.
4) We design a novel point cloud analysis framework based on the proposed local feature transformer module and trans-pooling model, we perform a large number of experiments and ablation studies in terms of classification and segmentation tasks. Experiments show that LFT-Net with local region learning ability achieves the state-of-the-art performance for shape classification and scene semantic understanding tasks.

II. Related Work

A. Transformer

Transformer model was first proposed in the field of natural language processing. Later, it was applied in the field of computer vision and reached state-of-the-art on multiple tasks.

1) For NLP: Transformer model [16] has been pro-posed in the field of natural language processing and is widely used in text classification, machine translation and other tasks. This model with self-attention mechanism assigns different weights to the input at the encoding stage. Devlin *et al.* [18] introduced Transformer model to propose a Bidirectional Encoder Representation from Transformers (BERT), which can be jointly mediated with context semantics at all levels. The BERT model discards the traditional RNN (recurrent neural network) and CNN (convolutional neural network) structure, and solves the thorny long-term dependence problem in NLP by Attention mechanism. Recently, a large number of natural language processing frame-works based on the Transformer model, such as Ro-BERT(Robustly Optimized BERT Pre-training) [19] and Bio-BERT [20], have been proposed. The scalability of Transformer model has a profound impact on the field of natural language processing.

2) For CV: Self-attention mechanism makes Trans-former model in the field of natural language processing brilliant success, which has attracted wide attention in computer vision. Now, the Transformer model has been successfully applied to other tasks in the field of 2d computer vision, such as image recognition [21], target detection [22], image super-resolution [23] and other tasks [24]–[27]. Existing frameworks usually apply local or global attention, through convolution or linear layer vectorization features. For example, in the large-scale dataset ImageNet [28], the dependency relation-ship between long distance pixel features cannot be obtained by sliding on the image through a fixed-size convolution kernel. For example, in image classification tasks, the local relation layer proposed by Hu *et al.* [29] can adaptively determine the aggregation weight of local pixels compositional relations. This method has stronger modeling and reasoning ability than ordinary sliding convolution. Similar to image classification, DERT [22] uses the Transformer model to establish the relationship between objects when predicting a set of objects. DERT embeds the Trans-former model into the target detection backbone network to eliminate the dependence of the network on the manual design modules region proposal network (RPN) and non-maximal suppression (NMS).

B. 3D Point Cloud Segmentation

In recent years, the methods of point cloud feature extraction based on learning include projection-based, voxel-based and point-based schemes, which are described in detail as follows.

1) Projection-Based Methods: Projection-based methods mainly convert irregular point clouds into regular representations in Euclidean space, and then use the success of deep learning in 2D vision. Such as multi-view [8], [30], first convert 3D point cloud into multiple 2D images and then use the multi-stream neural network to learn the generated projection image, and finally fuse the features to form the final output. In the related methods, Lawin *et al.* [31] tried to project the 3D point cloud model onto the 2D plane by using multiple virtual cameras, and then used multi-stream FCN to predict the pixel-level scores of multiple 2D images, and finally fused the prediction results of multiple 2D images to obtain the semantic labels of each point. In summary, the method has strong scalability and can decompose large-scale point cloud from multiple perspectives for processing, but the selection of perspective has a seriously influence on the prediction results. In addition, due to the projection process, the loss of depth geometry information will seriously affect the recognition performance. On the other hand, regarding the spherical representation method, Milioto *et al.* [32] proposed RangeNet++ for the real-time semantic segmentation task of LiDAR point cloud. Firstly, point cloud data are converted to range image, and 2D full convolution semantic segmentation is performed on range image. Then, 2D semantic labels are converted to 3D point cloud. Fast GPU KNN search for all data points to remove undesired discretization and inference artifacts based on range image 3D post-processing.

2) Voxel-Based Methods: The method of transforming irregular point clouds into regular representation also transforms point clouds into dense grids [6], [33], [34], and then use 3D convolutional neural networks to learn. 3D convolution neural network has strong applicability. Point clouds of different sizes can be extracted by convolution kernels of different sizes in 3D convolution network. In Fully-Convolutional Point Networks [7] proposed by Rethage *et al.*, the point cloud is transformed into an ordered structure through internal processing, and then accurate local geometry is captured in local areas at different levels. The context information is collected through weighted average pooling, and finally the information is combined according to the hierarchical structure for learning. Although the Voxel-based method has achieved leading results in the semantic segmentation of 3D point clouds by using the strategy of fully convolutional neural network to collect features from coarse to fine, the underestimation of the sparsity of 3D data leads to high computational and memory consumption, especially when dealing with large point clouds.

3) Point-Based Methods: Inspired by pointNet [11]/ PointNet++ [35], huge amounts of neural networks [12]–[15], [36] have been proposed in recent years in the direction of point cloud semantic under-standing. These networks usually regard point cloud as a set embedded in continuous space, and learn the local features of point cloud by replacing invariant operators. For example, ECC [37] uses dynamic edge-conditioned filters to model the relationship between local points, allowing the network to learn local and global features better. KCNet [38] defined the point set kernel as a set of learnable 3D points, which jointly responded to a set of adjacent data points according to the geometric relationship measured by the kernel correlation, and then used the local high-dimensional feature to aggregate the repeated features from the 3D position on the nearest neighbor graph. In KPConv [39], a set of kernel points that calculate convolution weights by Euclidean distance is proposed. The kernel points perform feature updating by different weight matrices. These networks that rely on share MLP are not consider other expensive and complex operations, and are highly applicable to point cloud semantic understanding. Based on the Recurrent Neural Networks method, Engelmann *et al.* [40] proposed that by learning the features of multi-scale blocks and grid blocks, the features of different scales were connected as the input of the consolidation unit (CU). Through a series of CUs, the features of each point and its adjacent points were repeatedly strengthened.

III. LFT-NET

LFT-Net completes the down-sampling stage through local feature transformer model and trans-pooling unit learning features, and then restores the initial resolution through interpolation. In this section, we take the object classification, object part segmentation and large scene semantic segmentation of 3D point cloud as examples to introduce the structure of local feature transformer model and trans-pooling and their representation ability on point cloud in detail. Especially for the local feature transformer model, we use the self-attention mechanism to learn the point cloud features that are more representative. Then we also introduce Multi Focal-Loss with Weight used to deal with data imbalance. Finally, we detail the entire network architecture.

A. Local Feature Transformer Model

The local feature transformer model is shown in Fig. 1. Local feature transformer model is mainly composed of local position encoding and self-attention layer. The local feature transform model performs an embedding operation on the coordinates of the input point cloud through a local position encoding unit. The self-attention layer computes three new matrices, query, key, and value by share MLP. The features of query, key and value are entered into the feedforward neural network after matrix operations.

The point cloud as input contains information such as spatial coordinates and initial features (such as RGB color, normal vector and other additional features). Firstly, the point cloud generates several local neighborhoods by KNN algorithm. We set multiple sets of parameter k for ablation experiment to verify the influence of neighborhood size on local feature transformer model (cf. Table IV). For each local neighborhood, local position encoding unit encodes position based on each neighborhood center point, and generates position coding with $d^{\wedge}$ channel for each point. $d^{\wedge}$ is 4 in the experiment, including relative position encoding and Euclidean distance. Then, the

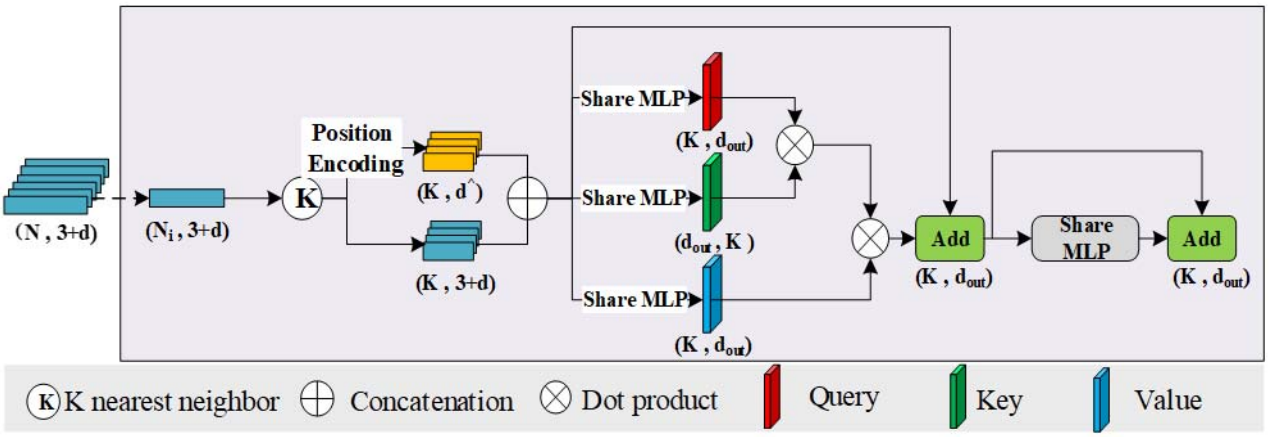

Fig. 1. The proposed local feature Transformer model shows the location space coding block of the feature and the self-attention learning mechanism based on local spatial geometric features and context.

encoded information is connected with the initial information and sent to the self-attention layer. Self-attention layer is mainly composed of three components, including query encoder, key encoder and value encoder. The three groups of encoders can heuristically learn the geometric characteristics of local point clouds. Finally, the high dimensional features learned from the self-attention layer are fed into a feedforward neural network. After several down-samplings with different resolutions and multiple local feature transformer models, the local point cloud features are continuously aggregated into the global features of point cloud. This model can describe the correlation of local area points as the basis for learning high-dimensional abstract features of point cloud. Next, we introduce local position encoding unit and self-attention layer in detail.

1) Local Position Encoding: For each point N_i in the point cloud, generate a local neighborhood by K-nearest neighbors (KNN) algorithm For K points in neighborhood N(i)$\{N_i^1, N_i^2, N_i^3 \ldots\ldots, N_i^k\}$, we use the following formula (1) for position coding:

$$P_i^k = \left\langle \left(N_i - N_i^k\right) \oplus \left\| N_i - N_i^k \right\| \right\rangle \tag{1}$$

where N_i, N_i^k does not contain other features (RGB color, normal, etc.), only use three-dimensional coordinates of points. Where $(N_i - N_i^k)$ calculates the relative coordinates of the relative center point N_i of the K points in the neighborhood, $\left\| N_i - N_i^k \right\|$ calculates the Euclidean distance between the points in the neighborhood and the center point, and finally we output the 4 channels position encoding information of the local point cloud by operation $\oplus$ connecting the two parts.

2) Self-Attention Layer: we connect the position information coding with the original feature through the vector self-attention layer, and the formula is described as follows:

$$D^K = \sum_{k \in \mathbb{N}(i)} P_i^k \oplus N_i^k, \left(D^K \in \mathbb{R}^{K*d}\right) \tag{2}$$

$$D_{\text{self}-att}^K = \text{SoftMax}\left[\alpha\left(D^K\right) \odot \beta\left(D^K\right)^T\right] \odot \gamma\left(D^K\right) \tag{3}$$

For each set of point cloud neighborhood, we use α,β,γ three groups of learning MLP for self-attention feature learning. So each point can generate three different matrices, but the matrices are the same size, they are query matrices, key matrices and value matrices. Use Query and Key multiplication to compute a score matrix for each point. Spontaneous learning point cloud features are obtained by multiplying score matrix and value matrix. The local point cloud feature output is $D_{\text{self -att}}^K$, and then we pass $D_{\text{self -att}}^K$ through a feed-forward neural network (including MLP, Batchnorm layer and Relu layer). In order to better preserve the details of the input point cloud features, we refer to ResNet [41] of He *et al.*, using additional skip connection as a dilated residual block at the point cloud feature self-attention learning and feed-forward neural network, respectively.

B. Trans-Pooling

Point cloud is different from pixel arrangement in image or voxel arrangement in 3D mesh. Point cloud is a set of sequential variability points. A network that consumes a three-dimensional point set needs to keep the order of the point set unchanged to ensure that the features extracted are invariant each time.

Existing works [11], [13], [35] usually use symmetry functions such as max/mean/sum pooling to reduce dimension and retain main features to solve the disorder issue of point cloud. But this will discard most of the learned features. By contrast, we use self-attention pooling function without additional information to heuristically preserve features, the structure of trans-pooling is shown in Fig. 2:

For the features output by local feature transformer model, $F_i = \{f_i^1, f_i^2, f_i^3, \ldots\ldots, f_i^k\}$, we use formula (4) to automatically weight local features.

$$f_i = \sum \text{SoftMax}\left[\eta\left(F_i\right) \odot \theta\left(F_i\right)^T\right] \odot F_i \tag{4}$$

where η and θ generate query and key matrices by sharing MLP and using self-attention mechanism to learn attention score of feature by matrix multiplication. The attention score after SoftMax is regarded as a mask for automatic feature

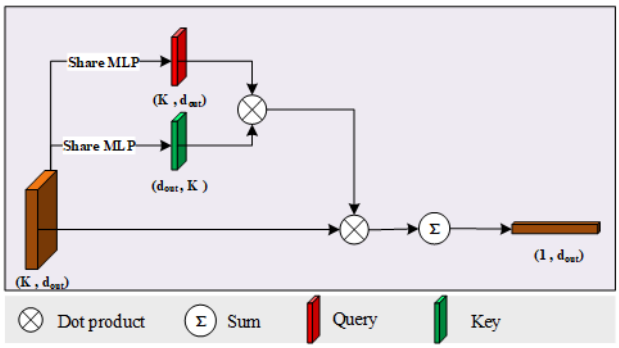

Fig. 2. The proposed Trans-pooling module is a self-attention weighted aggregation mechanism based on local spatial geometric features.

selection. A d_{out} dimension vector is generated by multiplying the mask and point cloud features and adding the results according to the channel (the d_{out} dimension of four trans-pool units are 64, 128, 256, 512, respectively, according to the sampling degree of the network). In summary, trans-pooling unit is a powerful mechanism for gathering features by considering the mutual dependence of local neighborhood points. Our trans-pooling unit will aggregate the spatial geometric features of K points in the neighborhood of N_i into an information feature vector f_i, and then use the end-to-end training mode to effectively train the proposed network architecture.

C. Multi Focal-Loss With Weight

There are problems of class-imbalance and difficult to learn geometric features of complex regions in point cloud semantic understanding tasks. For example, floor, ceiling and table, door training sample number is very different. It is difficult to learn the features of the edge of doors and Windows and the connection between tables and chairs. Inspired by the unbalanced solution of extreme foreground and background category encountered in the training process of intensive detector by Lin *et al.* [17]. We apply the focus loss function to point cloud classification and segmentation by reshaping the standard cross entropy loss function. The formula of Multi Focal-Loss with Weight is as follows:

$$\mathrm{MFL}\left(p_t\right) = -\frac{N - N_t}{N}\left(1 - p_t\right)^{\gamma} \log\left(p_t\right) \tag{5}$$

where N is the total number, N_t is the number of t class, and the current class coefficient is generated by $\frac{N-N_t}{N}$. By reducing the weight of the major class sample data and increasing the weight of the minor class sample data to balance the data distribution, the purpose of this is to control the importance of the current category in the back propagation. This additional cost can make the classifier more concerned about minor class samples. On this basis, $(1 - p_t)^{\gamma}$ in the multi focal loss function dominates the sample points that are difficult to learn by reducing the weight of easy-to-class samples. Where p_t is the accuracy, so the meaning of factor representation is the higher the accuracy, the smaller the entire loss value. We call the parameter γ decay coefficient, and the higher the accuracy, the worse the class decays. This makes the low accuracy class can occupy most of the loss and dominate the training.

D. Network Architecture

We implement LFT-Net by stacking Local Feature Transformer model and Trans-pool model. The detailed architecture is shown in Fig. 3. The down-sampling of feature encoder in LFT-Net for semantic segmentation and classification is divided into four stages. We follow the farthest sampling method in the work [42], [43] to gradually reduce the point set by four times the sampling rate. Therefore, the cardinality of the point set generated in each stage is [N/4, N/16, N/64, N/256], where N is the number of input points. Each stagepoint set generates a neighborhood through the K-nearest neighbor method to learn features through the local feature transformer model and the trans-pooling function that we construct. The number of stages, sampling rate and feature size can vary according to the application program. In the experimental design, we use four down-sampling, the data feature dimensions in the four down-sampling process are 64,128,256,512, respectively.

For the classification task, the input point cloud is subjected to multi-stage down-sampling and then the global average pooling aggregation feature is used to obtain the global feature of the three-dimensional object, and then the prediction probability is output through two linear layers. For the segmentation task, the down-sampled feature passed through a shared multi-layer perceptron. Then for each layer in the decoder, we refer to the interpolation method of RandLA-Net [15]. Firstly, KNN algorithm is used to find a nearest neighbor for each query point in the point set of the previous stage, and then the feature set of the point is updated through the nearest neighbor interpolation. Among we adopt the U-Net design [44], and the encoder is combined with the symmetric decoder, that is the feature mapping of up-sampling is connected with the feature mapping of down-sampling. In the network, we use four encoder and four decoder layers to learn the feature of each point. Finally, the shared multi-layer perceptron is used to adjust the feature size to predict the semantic label of each point.

IV. Experiments

For the proposed LFT-Net, we effectively evaluate multiple tasks on the point cloud datasets. For example, 3D object classification task, part segmentation task, and large scene semantic segmentation task. The Classification task is relatively simple, only need to classify the 3D point cloud model of independent objects correctly. Part tasks require dividing different parts of an object. In contrast, the segmentation of large-scale scene cloud is complex, and different objects need to be segmented from the point cloud scene of millions of levels. For point cloud object shape classification tasks, we use the widely used ModelNet40 dataset [45]. For part-level segmentation, we use shapenet part dataset [46] to evaluate. In addition, in the semantic segmentation task in 3D large scene, we use the challenging Stanford large-scale 3D indoor space (S3DIS) dataset [47]. All the comparative test data in the experiment were taken from the comparative experiment paper.

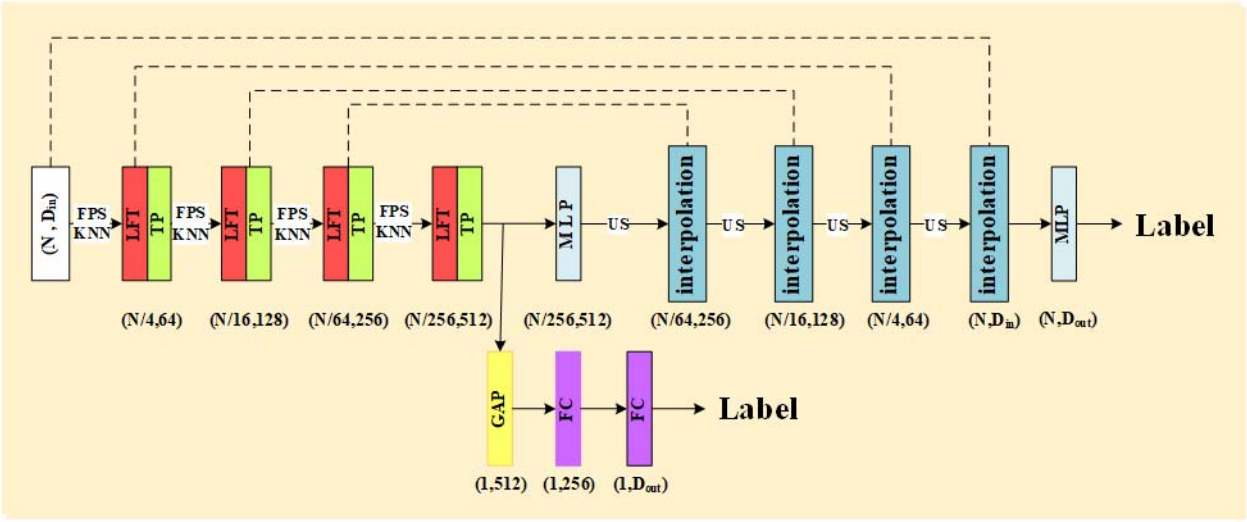

Fig. 3. The detailed architecture of LFT-Net for semantic segmentation(top) and 3D object classification(bottom). (N, D) represent the number and feature size of points, respectively. FPS: farthest point sampling. LFT: Local Feature Transformer module. TP: Trans-pooling module. MLP: shared multi-layer perceptron. US: up-sampling. interpolation: feature interpolation. GAP: Global average pooling. FC: full connection layer.

In our experiments, the initial learning rate of point cloud classification is set to 0.01 on the ModelNet40 dataset, and the initial learning rate of point cloud segmentation task is set to 0.05 on the S3DIS dataset. After each epoch, the initial learning rate decreases by 1%. Using a random gradient descent optimizer with momentum value of 0.9, and the number of K-nearest neighbor points k is set to 16. All experiments were performed on NVIDIA QUADRO RTX 8000 GPU.

A. 3D Model Classification

ModelNet40 [45] dataset contains 40 categories, including 12311 CAD models. We trained 9843 models and tested 2468 models. We conduct a unified sampling of 2048 points and normal vectors for each object. We used the mean accuracy (mAcc) of each category and the overall accuracy (OA) of all categories as evaluation metrics for the experiment.

As Table I shows, our network outperforms other advanced methods in mAcc and OA metrics. The average accuracy and overall accuracy on ModelNet40 dataset are 89.7% and 93.2% respectively, which are better than those on voxel-based networks VoxNet [48], 3DShapeNets [45] and some networks [35], [39] based on learning point features. The focus of both [35] and [39] is also to extract the local features of the point cloud, so we believe that the local feature transformer model has a stronger ability to describe the local point cloud. From the classification experiment, we can perceive that the global feature of point cloud is effectively aggregated through LFT-Net.

B. Part Segmentation

ShapeNet part Dataset [46] is a large annotated 3D shape dataset. It contains a variety of semantic types of 3D models and provides detailed annotation information for each model, such as consistent rigid alignment, parts and bilateral symmetric planes, physical dimensions, and semantic annotation

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE MODELNET40 CLASSIFICATION DATASET

Method	input	mAcc	OA
MVCNN [49]	image	-	90.1
VoxNet [48]	voxel	83.0	85.9
3DShapeNets [45]	voxel	77.3	84.7
PointNet [11]	point	86.2	89.2
PointNet++ [35]	point+normal	-	91.9
DGCNN [50]	point	90.2	92.2
KPConv [39]	point	-	92.9
PointWeb [12]	point	89.4	92.3
PointConv [43]	point+normal	-	92.5
RS-CNN [51]	point	-	92.9
PointASNL [52]	point	-	92.9
LFT-Net(ours)	point+normal	**89.7**	**93.2**

of sub-parts. We evaluated 16 categories in the ShapeNet parts dataset with Intersection-over-Union (IoU) for each object and the average IoU value of part instance for all shapes.

The results of Part Segmentation are shown in Table II. We reported each category and average IoU (%) score. The average IOU reached 86.2%, which was similar to the effect of one work [51] of extracting topological constraint relations from point clouds. Average IoU and most categories of data results are at the most advanced level in part segmentation. Part segmentation is a more difficult task than shape classification. We find that our network can also segment the junction of objects better. This is due to the ability of self-attention mechanism to capture information from local to global to obtain greater receptive field and context information. The results of LFT-Net part segmentation on shapenet dataset can reach the best in most categories. But some categories such as motorbike, skateboard and so on may fail to achieve optimal performance because of too few training samples, but they can also approach state-of-the-art. In Fig. 4, we give the visualization results of some data models. It can be seen that although the prediction of some objects is challenging, the network prediction is still reasonable and reliable.

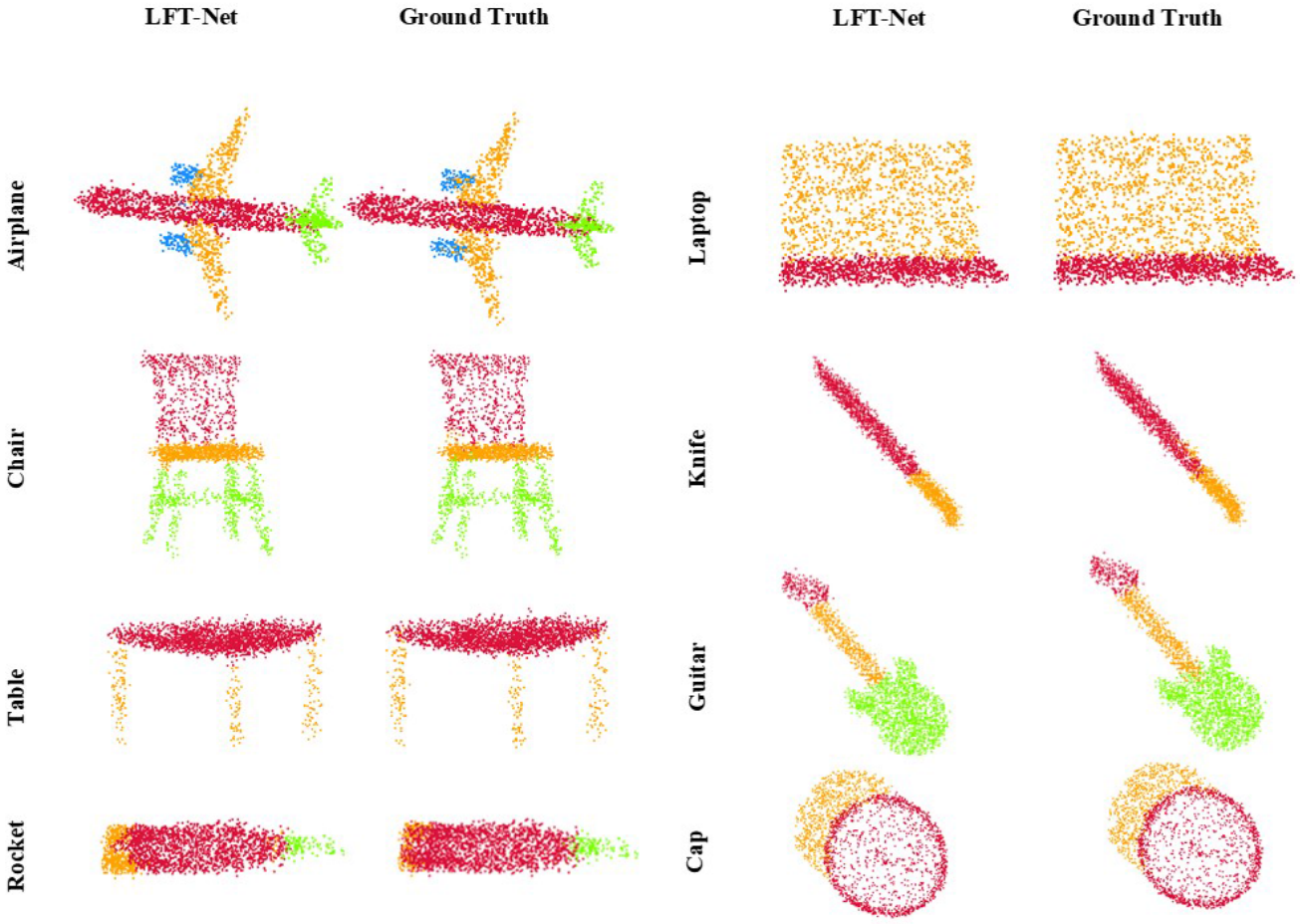

Fig. 4. LFT-Net visualization results of parts segmentation on ShapeNet part point cloud dataset. The left column is the network segmentation result and the right is the ground truth value of the dataset.

TABLE II
PART SEGMENTATION EVALUATION ON THE SHAPENET PART DATASET

Method	mIou	airplane	bag	cap	car	chair	earphone	guitar	knife	lamp	laptop	motorbike	mug	pistol	rocket	skateboard	table
PointNet [11]	83.7	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9	80.8	95.3	65.2	93.0	81.2	57.9	72.8	80.6
PointNet++ [35]	85.1	82.4	79.0	87.7	77.3	90.8	71.8	91.0	85.9	83.7	95.3	71.6	94.1	81.3	58.7	76.4	82.6
SO-Net [53]	84.9	82.8	77.8	88.0	77.3	90.6	73.5	90.7	83.9	82.8	94.8	69.1	94.2	80.9	53.1	72.9	83.0
PointASNL [52]	86.1	**84.1**	84.7	87.9	**79.7**	92.2	73.7	91.0	87.2	84.2	95.8	**74.4**	**95.2**	81.0	63.0	76.3	83.2
RS-CNN [51]	**86.2**	83.5	**84.8**	88.8	79.6	91.2	**81.1**	91.6	88.4	**86.0**	**96.0**	73.7	94.1	83.4	60.5	**77.7**	83.6
DGCNN [50]	85.2	84.0	83.4	86.7	77.8	90.6	74.7	91.2	87.5	82.8	95.7	66.3	94.9	81.1	63.5	74.5	82.6
LFT-Net(ours)	**86.2**	83.0	83.9	**90.9**	79.4	**93.1**	71.4	**92.5**	**88.6**	85.7	95.9	69.3	94.2	**85.0**	**65.6**	74.6	**85.5**

C. 3D Semantic Segmentation

The S3DIS [47] dataset consists of 272 rooms from 6 area of three different buildings. Each point cloud data contains $20 \times 15 \times 5$ m indoor scenes in the real three-dimensional space. Each point in the data scanning is assigned 13 categories (ceiling, floor, table, etc.) semantic labels. We used the area-5 as a test set in the experiment of the dataset, because area-5 is not in the same building as other areas, and there are some differences between the objects in area-5 and other areas. For evaluation metrics, we use average on-class precision (mAcc) and average on-class intersection (mIou).

The experimental results of semantic segmentation are shown in Table III. In mAcc and mIou evaluation metrics, LFT-Net reached 76.2% and 65.2% respectively, and exceeded previous work by multiple percentage points in multiple categories. For sofa, chair, table and other indoor independent object accuracy did not achieve good results may be due to the sensor point cloud data collected in three-dimensional space unevenly distributed, and in the point cloud sampling process object three-dimensional shape is destroyed. Experimental results show that LFT-Net outperforms voxel-based methods (such as SegCloud [54]), graph-based methods (such as SPGraph [55]) and MLPs-based methods (such as PointNet [34] and Point-CNN [42] for weighting and permutation of input features by X-transformation matrix, etc.). In particular, the description ability of local feature transformer unit and trans-pooling unit is also useful for large point cloud datasets. We believe that LFT-Net can learn a larger dataset. Because its network module combines strong description ability and strong autonomous learning.

Fig. 5 shows the visualization results of LFT-Net in S3DIS area-5 dataset test. We can see that the prediction results of complex regions are very close to the ground truth, which

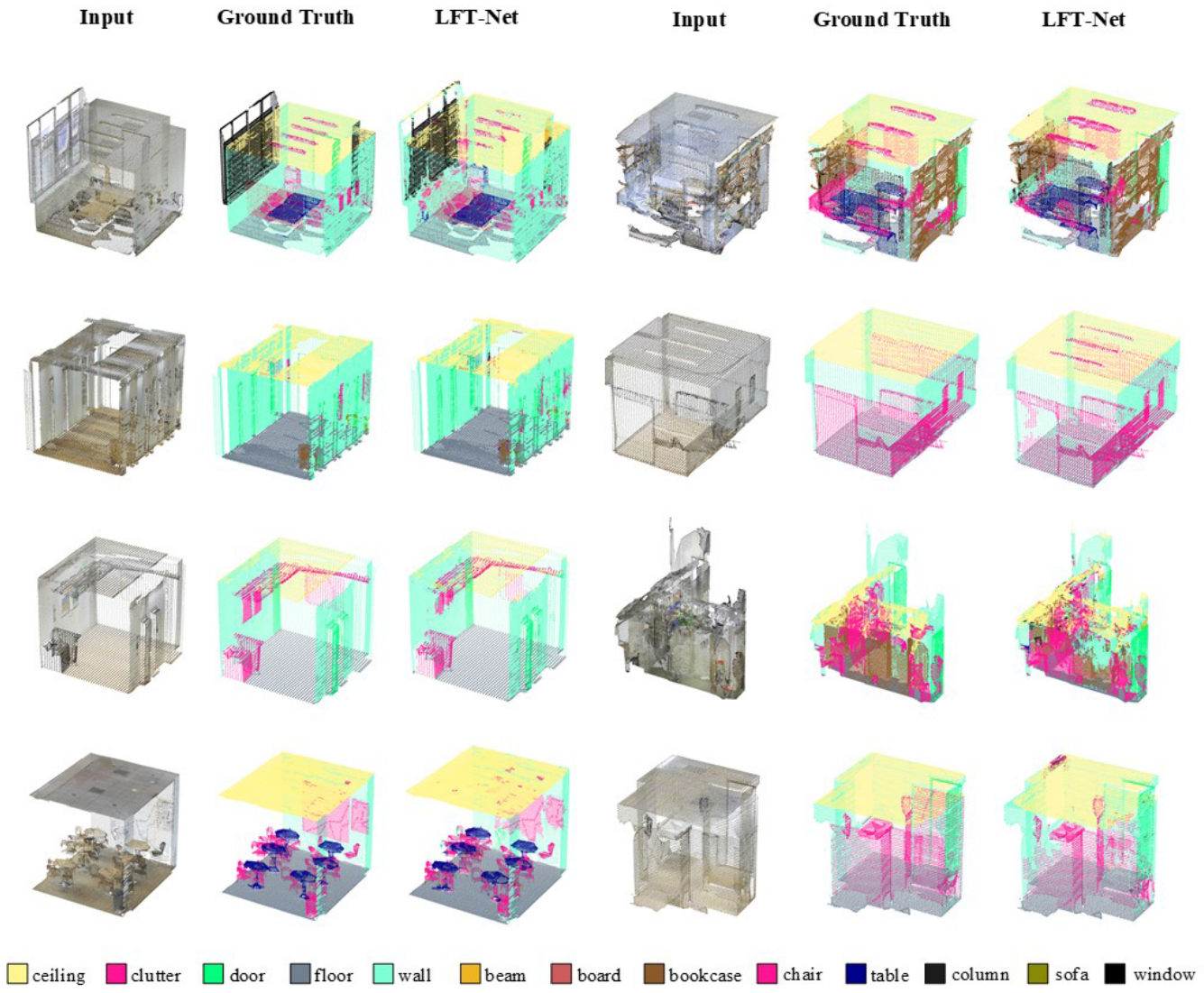

Fig. 5. Visualization of semantic segmentation results for partial room testing in S3DIS Area-5 Dataset.

TABLE III
SEMANTIC SEGMENTATION EVALUATION ON S3DIS AREA-5

Method	mAcc	mIou	ceil	floor	wall	beam	col	wind	door	table	chair	sofa	book	board	clut
SegCloud [54]	57.4	48.9	90.1	96.1	69.9	0.0	18.4	38.4	23.1	70.4	75.9	40.9	58.4	13.0	41.6
PointNet [11]	49.0	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	59.0	52.6	5.9	40.3	26.4	33.2
PointCNN [42]	63.9	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 [12]	66.6	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
ParamConv [55]	67.0	58.3	92.3	96.2	75.9	**0.3**	6.0	69.5	63.5	66.9	65.6	47.3	68.9	59.1	46.2
SPGraph [56]	66.5	58.0	89.4	96.9	78.1	0.0	**42.8**	48.9	61.6	**84.7**	75.4	**69.8**	52.6	2.1	52.2
LFT-Net(ours)	**76.2**	**65.2**	**92.8**	96.1	**81.9**	0.0	37.6	**70.3**	70.4	73.2	76.0	40.9	**78.8**	**71.0**	**58.2**

shows that the module we designed can show good performance in large scene point cloud feature learning.

D. Ablation Study

In order to verify the rationality and effectiveness of our design modules, we conducted multiple ablation studies on the local feature transformer module and the trans-pooling module. All ablation studies were conducted on the semantic segmentation task based on the S3DIS dataset. We trained in five other areas except Area-5 and tested in Area-5.

1) Number of Neighbor: In the process of network down-sampling, k is used to determine the scale of local neighborhood of each point. Different sizes of k have an important impact on network learning. We conducted four groups of ablation studies on the value of k, and the experimental results are shown in Table IV. When k is set to 4 or 8, the neighborhood is small, and the network cannot learn the fine features of the local region to fit the 3d object, indicating that our local feature transformer model can effectively use the features of adjacent points in learning local features. When k is set to 16, we know from the experimental results that

TABLE IV
ABLATION STUDY OF LOCAL NEIGHBORHOOD NUMBER K

k	mAcc	mIou
4	68.0	54.3
8	72.4	61.7
16	**76.2**	**65.2**
32	75.0	64.5

TABLE V
ABLATION STUDY OF THE LOCAL FEATURE TRANSFORMER MODULE POSITION ENCODING

Position encoding	mAcc	mIou
none	60.3	49.0
relative coordinates	71.4	60.8
distance	69.8	57.0
all	**76.2**	**65.2**

the model achieves the best results in point cloud semantic information prediction. By further increasing the value of k, we set k to 32 to generate a larger neighborhood. From the experimental results, the prediction accuracy is not high when k is 16. We speculate that there may be many points with far relative distance and low correlation in local point clouds, which may have a negative impact on local feature learning, thereby reducing the accuracy of the model. It shows that the learning of local features needs to be controlled within a certain range. In addition, increasing the k value will increase the network training time and cause unnecessary time overhead.

2) Position Encoding: We performed different ablation studies on two kinds of position coding information in local feature transformer model. The experimental results are shown in Table V. Obviously, without any location coding information, the local neighborhood of point cloud will lose the local geometric structure of the center point N_i, resulting in local feature transformer model unable to learn features more effectively. And in the transformer model [16], position coding is very important to the attention of data points. When we only retain relative coordinate coding or relative distance coding, the performance of the network model will be significantly improved. In the last set of experiments, we combined two location coding methods to improve the performance more than using only one location coding method.

3) Pooling Type: In previous works [11], [35], [42], symmetric pooling functions such as max/mean/sum are usually used to overcome point cloud disorder. For the local point set feature output $F_i = \{f_i^1, f_i^2, f_i^3, \ldots\ldots, f_i^k\}$ by local feature transformer model, the formulas of max/mean/sum are as follows:

$$f_i\{x_1, x_2, \ldots, x_n\} = \max\left(F_i\{x_1, x_2, \ldots, x_n\}\right) \quad (6)$$

$$f_i\{x_1, x_2, \ldots, x_n\} = \sum\left(F_i\{x_1, x_2, \ldots, x_n\}\right) \quad (7)$$

$$f_i\{x_1, x_2, \ldots, x_n\} = \frac{\sum\left(F_i\{x_1, x_2, \ldots, x_n\}\right)}{k} \quad (8)$$

We can see that the max function (formula (6)) selects the maximum feature of the point in the local neighborhood to ensure the order of the point cloud. In the pooling phase, the feature values on the feature dimension (x_i) except for the maximum point feature are discarded, which leads to additional loss of useful information. The mean function (formula (7)) and sum function (formula (8)) will use all the feature information, but fundamentally, these two types of functions treat all the feature values equally. In other words, they cannot assign more weight to important feature values to ensure the efficiency of network learning. In the final experiment, we use the trans-pooling function, which can learn and aggregate the features of all points in the local point cloud, and can heuristically use more effective feature points. In contrast, the max/mean/sum symmetry function is widely used to solve the point cloud disorder, it is difficult to gather all the point cloud features with weight, and it is inevitable to lose some useful feature information, so their performance may be suboptimal.

TABLE VI
ABLATION STUDY OF VARIOUS POOLING TYPES

Pooling type	mAcc	mIOU
max pooling	73.1	64.1
mean pooling	70.4	60.8
sum pooling	72.5	63.2
trans-pooling	**76.2**	**65.2**

V. CONCLUSION

Transformer discards the traditional RNN structure and fundamentally uses the self-attention layer to model the set. The emergence of Transformer model has changed the field of natural language processing and made remarkable achievements in the field of computer vision. In this paper, we continue to study in this aspect, the Transformer module is applied to point cloud local feature extraction and pooling layer, proposed LFT-Net to effectively segment point cloud. Local Feature Trans-former model is essentially a set operator of a disordered set, which learns the features of local space sets by self-attention, and Trans-pooling compresses features by aggregating local features. We have proved through experiments on multiple benchmarks that through the training of massive data, the Transformer model is advanced and powerful in the learning of point cloud features. We expect our work can further optimize the performance of Transformer model in point cloud feature learning and its application in 3D target detection and tracking direction tasks.

REFERENCES

[1] M. Ali, S. Qaisar, M. Naeem, and S. Mumtaz, "Energy efficient resource allocation in D2D-assisted heterogeneous networks with relays," *IEEE Access*, vol. 4, pp. 4902–4911, 2016.
[2] S. Goudarzi, N. Kama, M. H. Anisi, S. Zeadally, and S. Mumtaz, "Data collection using unmanned aerial vehicles for Internet of Things platforms," *Comput. Elect. Eng.*, vol. 75, pp. 1–15, May 2019.
[3] B. Ji *et al.*, "Survey on the Internet of Vehicles: Network architectures and applications," *IEEE Commun. Standards Mag.*, vol. 4, no. 1, pp. 34–41, Mar. 2020.
[4] J. Zhang, J. Liu, S. Ma, C.-K. Wen, and S. Jin, "Large system achievable rate analysis of RIS-assisted MIMO wireless communication with statistical CSIT," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5572–5585, Sep. 2021.
[5] K. Xu, J. Zhang, X. Yang, S. Ma, and G. Yang, "On the sum-rate of RIS-assisted MIMO multiple-access channels over spatially correlated Rician fading," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8228–8241, Dec. 2021.
[6] B. Graham, M. Engelcke, and L. V. D. Maaten, "3D semantic segmentation with submanifold sparse convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9224–9232.

[7] D. Rethage, J. Wald, J. Sturm, N. Navab, and F. Tombari, "Fully-convolutional point networks for large-scale point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 596–611.

[8] A. Boulch, B. Le Saux, and N. Audebert, "Unstructured point cloud semantic labeling using deep segmentation networks," *3DOR@ Eurographics*, vol. 3, pp. 1–8, Apr. 2017.

[9] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1887–1893.

[10] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4376–4382.

[11] 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. 652–660.

[12] H. Zhao, L. Jiang, C.-W. 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. 5565–5573.

[13] M. Jiang, Y. Wu, T. Zhao, Z. Zhao, and C. Lu, "PointSIFT: A SIFT-like network module for 3D point cloud semantic segmentation," 2018, *arXiv:1807.00652*.

[14] Z. Zhang, B.-S. Hua, and S.-K. Yeung, "ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1607–1616.

[15] Q. Hu *et al.*, "RandLA-Net: Efficient semantic segmentation of large-scale point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11108–11117.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[19] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[20] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 213–229.

[23] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5791–5800.

[24] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.

[25] W. Su *et al.*, "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*.

[26] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8808–8817.

[27] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021, *arXiv:2102.04432*.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[29] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.

[30] B. Yang, W. Luo, and R. Urtasun, "PIXOR: Real-time 3D object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7652–7660.

[31] F. J. Lawin, M. Danelljan, P. Tosteberg, G. Bhat, F. S. Khan, and M. Felsberg, "Deep projective 3D semantic segmentation," in *Proc. 17th Int. Conf. Comput. Anal. Images Patterns (CAIP)*. Ystad, Sweden: Springer, 2017, pp. 95–107.

[32] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet++: Fast and accurate LiDAR semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 4213–4220.

[33] H.-Y. Meng, L. Gao, Y.-K. Lai, and D. Manocha, "VV-Net: Voxel VAE net with group convolutions for point cloud segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8500–8508.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.

[36] L.-Z. Chen, X.-Y. Li, D.-P. Fan, K. Wang, S.-P. Lu, and M.-M. Cheng, "LSANet: Feature learning on point sets by local spatial aware layer," 2019, *arXiv:1905.05442*.

[37] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3693–3702.

[38] Y. Shen, C. Feng, Y. Yang, and D. Tian, "Mining point cloud local structures by kernel correlation and graph pooling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4548–4557.

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

[40] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe, "Exploring spatial context for 3D semantic segmentation of point clouds," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 716–724.

[41] 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. 2016, pp. 770–778.

[42] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on χ-transformed points," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 820–830.

[43] W. Wu, Z. Qi, and L. Fuxin, "PointConv: Deep convolutional networks on 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9621–9630.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[45] Z. Wu *et al.*, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[46] L. Yi *et al.*, "A scalable active framework for region annotation in 3D shape collections," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, Nov. 2016.

[47] I. Armeni *et al.*, "3D semantic parsing of large-scale indoor spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1534–1543.

[48] 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. 2015, pp. 922–928.

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

[50] 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, 2019.

[51] Y. Liu, B. Fan, S. Xiang, and C. Pan, "Relation-shape convolutional neural network for point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8895–8904.

[52] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5589–5598.

[53] J. Li, B. M. Chen, and G. H. Lee, "SO-Net: Self-organizing network for point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9397–9406.

[54] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "SEGCloud: Semantic segmentation of 3D point clouds," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 537–547.

[55] S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky, and R. Urtasun, "Deep parametric continuous convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2589–2597.

[56] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4558–4567.

Yongbin Gao received the Ph.D. degree from Jeonbuk National University, South Korea. He is currently an Associate Professor with the School of Electronic and Electrical Engineering and the Vice Director of the Next Generation Intelligent Research Center, Shanghai University of Engineering Science, Shanghai, China. He has published 30 SCI articles in prestigious journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Information Science*, and *Pattern Recognition Letters*, in the area of image processing, patter recognition, and computer vision.

Xuebing Liu is currently pursuing the master's degree with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. His research interests include point cloud scene understanding, computer vision, and machine learning.

Jun Li (Member, IEEE) received the B.S. degree from the South Central University for Nationalities, Wuhan, China, in 2009, and the Ph.D. degree from Chonbuk National University, Jeonju, South Korea, in 2016. He is currently an Associate Professor with Guangzhou University, Guangzhou, China. He has published more than 50 papers in refereed journals and conference proceedings. His research interests include spatial modulation and OFDM with index modulation. He serves as a Reviewer for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Zhijun Fang (Senior Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. He is currently a Professor and the Dean with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. His current research interests include image processing, video coding, and pattern recognition. He was the General Chair of the Joint Conference on Harmonious Human Machine Environment (HHME) in 2013 and the General Co-Chair of the International Symposium on Information Technology Convergence (ISITC) in 2014, 2015, 2016, and 2017. He received the "Hundred, Thousand and Ten Thousand Talents Project" in China. He received several major program project of the National Natural Science Foundation of China and the National Key Research and Development Project of Ministry of Science and Technology of China.

Xiaoyan Jiang received the Ph.D. degree in computer science from the Friedrich-Schiller University of Jena, Germany, in 2015. She is currently an Associate Professor with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. Her research interests include 3D vision, including multi-object tracking, semantic segmentation, and SLAM. She has published more than 30 papers in the field of computer vision. She holds the fund from the National Natural Science Foundation (NSFC) of China and attended several essential projects funded by Shanghai Science and Technology Committee and NSFC. She is also cooperated with enterprises aiming to design AI solutions for practical scenarios.

Kazi Mohammed Saidul Huq (Senior Member, IEEE) was a Senior Research Fellow with the Faculty of Computing, Engineering, and Science, University of South Wales, Pontypridd, U.K. He has served as a project coordinator for major international research projects. He has authored several publications, including papers in high-impact international conferences and journals, books, and book chapters. He has vast experiences in 5G networks. His research interests include 5G, 6G paradigm, communication haul, D2D, green radio, higher frequency communication (mmWave and THz), the IoT, MAC-PHY, RRM, and URLLC. He is a Founding Member of the IEEE 1932.1 Standard for Licensed/Unlicensed Spectrum Interoperability in Wireless Mobile Networks. He is an Associate Editor of *IET Quantum Communication* journal. He was also the Lead Guest Editor of *IEEE Communications Magazine* Feature Topic on THz Communications in the June 2018 Issue.