ALS Point Cloud Semantic Segmentation Based on Graph Convolution and Transformer With Elevation Attention

Shuowen Huang, Qingwu Hu, Pengcheng Zhao, Jiayuan Li, Mingyao Ai, and Shaohua Wang

***Abstract*—Semantic segmentation of airborne point clouds is crucial for 3D scene reconstruction and remote sensing in surveying applications. Current deep learning methods for point clouds primarily focus on effectively aggregating local neighborhood information. However, they often overlook the fusion of global context information and elevation features, which are vital for airborne point clouds. In this study, we propose Dense-LGEANet, a novel network with dense connected architecture and multiscale feature supervision based on our designed LGEA module. The key component of our LGEA module is the combination of the graph convolution block and the transformer block with elevation attention. It can effectively fuse local neighborhood information and global context information to improve the accuracy of semantic segmentation of airborne point cloud. Moreover, the designed dense connected network architecture can enhance the feature extraction capability for point cloud objects at different scales by facilitating interactions between multiple up-sampling and down-sampling layers. We have conducted multiple experiments on the public point cloud dataset. Experimental results show that our method can achieve an mIoU of 58.5% and an mF1 of 72.0% on the ISPRS Vaihingen 3D dataset, while an mIoU of 67.2% and an mF1 of 78.3% on the LASDU dataset. It demonstrates the superior performance of our network and the effectiveness of the proposed feature enhancement module and network architecture.**

***Index Terms*—Airborne laser scanning (ALS), graph convolution, point cloud, semantic segmentation, transformer.**

I. Introduction

POINT cloud is a compact and reliable digital representation of real world in three dimensions. It comprises numerous 3D points that store valuable information, including location coordinates, color, normal vectors, and reflectance intensity, which can provide rich spatial information for complex environments [1]. Therefore, point cloud is suitable for realistic representation of large scene. In recent years, the development and popularity of light detection and ranging (LiDAR) equipment have made it increasingly convenient to obtain point cloud data from various scenes. According to the scanning platform and scene, point cloud can be categorized into terrestrial laser scanning (TLS) point clouds, vehicle laser scanning (VLS) point clouds, and airborne laser scanning (ALS) point clouds. Among them, ALS point clouds have wider usage in the field of geographic information due to their high scanning platform, wide coverage, and ability to generate digital elevation models. So they are widely employed in urban planning [2], power line detection [3], forest surveying [4], and other domains [5]. All of these applications require the extraction of semantic information from airborne point cloud data. Therefore, semantic segmentation of airborne point clouds has been an important research topic for the photogrammetry and remote sensing community.

Semantic segmentation for ALS point clouds has unique characteristics compared to TLS and VLS point clouds, such as a large number of geometric instances, different scale between categories, the variability in elevation distribution, complex spatial structure, and so on. Consequently, it is challenging to achieve fully automatic and accurate semantic segmentation for ALS point clouds. The semantic segmentation methods of ALS point clouds typically fall into two categories: traditional machine learning methods and deep learning methods. Over the past few decades, traditional machine learning methods have made significant advancements. These methods primarily focus on extracting various manual features to capture the local geometric structure of the point cloud. They employ the classifiers like random forests and support vector machines to predict the label of each point [6], [7]. However, due to the limitation of low dimensional manual features, these algorithms often fail to deliver satisfactory semantic segmentation results in complex point cloud scenarios. In order to improve the accuracy of semantic segmentation, scholars have integrated graphical models such as Markov random field (MRF) and conditional random field (CRF) with machine learning models. For example, the work in [8] proposes a coarse-to-fine MRF method for ground point cloud segmentation, which develops a feature extraction algorithm for rough segmentation, and then uses MRF for global adjustment. Wolf et al. [9] use random forest method to initialize the unary potentials of a densely CRF, so that it can perform fast semantic segmentation of point clouds. Yang et al. [10] design a continuous CRF convolution method for segmentation, which

Manuscript received 12 August 2023; revised 2 November 2023 and 2 December 2023; accepted 14 December 2023. Date of publication 28 December 2023; date of current version 16 January 2024. This work was supported in part by the National Key R&D Program of China under Grant 2021YFB2600401, in part by the National Natural Science Foundation of China under Grant 42101441, and in part by the Special Fund of Hubei Luojia Laboratory under Grant 230700007. *(Corresponding author: Qingwu Hu.)*

The authors are with the School of Remote Sensing and Information Engineering, Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China (e-mail: huangshuowen@whu.edu.cn; huqw@whu.edu.cn; pengcheng.zhao@whu.edu.cn; ljy_whu_2012@whu.edu.cn; aimingyao@whu.edu.cn; shwang@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3347224

can effectively capture the structure of features in large-scale point cloud.

In recent years, deep learning algorithms have demonstrated their effectiveness in extracting high dimensional semantic features for 2D image semantic segmentation tasks [11]. Taking inspiration from this success, researchers have started to explore the use of neural networks to learn discriminative high-dimensional features for 3D point cloud semantic segmentation. However, 3D point clouds present unique challenges due to their irregular spatial distribution, occlusion, and uneven density. In addition, the extra dimension of point clouds significantly increases the complexity of the features. These factors make feature extraction and fusion of 3D point cloud data difficult. To address these challenges, researchers have attempted to convert point clouds into images or voxels [12], and learn point cloud features using well-established 2D or 3D convolutional neural networks. However, these approaches inevitably lead to the loss of feature information, and the accuracy and speed of these methods are limited by the size of the image and voxel resolutions. PointNet [13] is the first network that operates directly on the original point cloud, and has been promoted and applied to point cloud classification, semantic segmentation, and target detection. Since then, neural networks that directly process the original input point clouds have become popular in research. These networks can be categorized into four main types: MLP (multilayer perceptron)-based methods, convolution-based methods, graph-based methods, and transformer-based methods.

MLP-based methods usually employ shared MLPs to extract features independently for each point and aggregate these features through pooling operations. PointNet++ [14] develops multiple sampling and grouping techniques to capture more local information. However, PointNet++ only extracts information from points within the local region and does not fully capture the structural relationships between points in that region. To improve the performance of networks in complex scenes, RandLANet [15] proposes a local feature aggregation module that can effectively preserve geometric details while gradually increasing the perceptual field of points. So-Net [16] uses self-organizing feature mapping to analyze the distribution of point clouds to achieve a permutation-invariant network for point cloud semantic segmentation. PointASNL [17] introduces the adaptive sampling module to adaptively adjust the coordinates of the initial sampling points to make it more suitable for feature learning with intrinsic geometric characteristics. SCF-Net [18] aggregates local key features using a dual-distance attention pool block that considers both geometric and feature distances. In order to enhance the segmentation capability of point cloud scene boundaries, BAAF-Net [19] designs a bilateral filter module to capture local geometric and semantic information. And LGS-Net [20] uses a parallel attention fusion module that focuses on geometric structure and semantic information to reduce the ambiguity of features and improve the mining of local geometric structure information.

Convolution-based methods have been devoted to designing effective convolution operations for disordered and nonuniform point clouds. For example, PointCNN [21] introduces the χ-Conv module to aggregate information on the spatial structure and local features of points. PointConv [22] multiplies the weights obtained from the local relative coordinates with the inverse density coefficients as a new weight function for feature learning. RS-CNN [23] explicitly encodes the geometric relationship of points and proposes a new convolution operator relational shape convolution. KPConv [24] applies the convolution weights defined in Euclidean space to the input points around the kernel points, thus proposing a module for point continuous convolution. Moreover, DenseKPNet [25] extracts initial geometric features from coarse to fine through multiscale kernel point convolution, and uses dense connections to learn expressive local geometric features. PAConv [26] constructs the convolution kernels by dynamically combining weight matrices stored in a weight library, which reduces the complexity of the model while maintaining the ability to capture important features.

Graph-based approaches aim to represent points as nodes in a graph, and establish edges based on the relationships between these points. As a natural representation of point clouds, a graph can effectively encode local geometric structure through graph convolution. SPG [27] first segments large point clouds into meaningful target shapes using an unsupervised method, then constructs a superpoint graph and extracts features on these superpoints using PointNet. DGCNN [28] proposes EdgeConv, which aggregates local neighborhood information by graph edge convolution. DGANet [29] employs an improved K-nearest neighbor search algorithm for constructing a dilated graph, which allows the network to learn local features with maximum receptive field during convolution. AdaptConv [30] develops an adaptive convolution kernel that can dynamically extract feature information of each point, thereby establishing diverse connections between different points in the local neighborhood. DDGCN [31] constructs a dynamic neighborhood graph by obtaining the similarity matrix of the point cloud to further encode local features in the point cloud.

Transformer-based methods are inspired by the self-attention mechanism's success in natural language processing and image processing. Some researchers have developed point cloud processing networks based on transformers to improve segmentation accuracy. For instance, MLMSPT [32] leverages a multiscale transformer to capture relationships between different features while aggregating information from various levels of contextual information at each scale. Point transformer [33] uses vector self-attention and subtraction relations to compute the importance of edge points, which improves segmentation accuracy and network extensibility. PCT [34] proposes a permutation-invariant point cloud transformer based on offset attention, which is suitable for unstructured point cloud learning in irregular domain. Stratified transformer [35] samples neighborhood dense points and remote sparse points as keys, expanding the effective receptive field of the model. Fast point transformer [36] designs a lightweight transformer network that encodes continuous three-dimensional coordinates based on a voxel hashing architecture to improve computing efficiency.

The above-mentioned methods have made significant advancements in the field of point cloud semantic segmentation. However, challenges still exist in the ALS point cloud semantic segmentation. First, most of them have focused on how to aggregate the local neighborhood information of points. However, airborne point clouds usually cover a larger geographic range than TLS and VLS point clouds, making it difficult to encode global context information for large scenes. Although previous research has attempted to enhance semantic segmentation accuracy by increasing the network's receptive field [37], it still lacks the fusion of global features. And then, the objects in airborne point clouds usually exhibit extreme scale variations, and the single encoding and decoding architecture used in most networks cannot fully exploit feature information at all scale levels.

Second, as mentioned earlier, different category in the airborne point cloud usually have different elevation distributions. And previous studies have proved that elevation information is a feature information that cannot be ignored in airborne point cloud semantic segmentation [38], [39]. However, most of the existing methods simply encode the elevation information and add it into the input features, which may ignore the importance of elevation features within the encoding layer. Consequently, it is still worthwhile to explore techniques for incorporating elevation attention into the global feature representation.

To tackle the above-mentioned challenges, we present a novel network architecture for ALS point cloud semantic segmentation. Specifically, in order to encode the global feature and elevation information, we design a feature enhancement module, referred as local and global feature enhancement with elevation attention (LGEA) module, which comprises the graph convolution block and the transformer block. The graph convolution block facilitates effective aggregation of local position and feature information within the point cloud. Simultaneously, the transformer block captures global information through the utilization of the self-attention mechanism based on elevation attention. In addition, to enable comprehensive utilization of information at various levels, we construct a dense hierarchical architecture named Dense-LGEANet driven by multiscale loss based on LGEA module, which can improve perception ability of multiple categories at different scales.

In brief, our contributions can be summarized as follows.

1) We propose the LGEA module, a novel point cloud feature enhancement module. Compared with previous approaches, this module effectively integrates local and global feature information and enhances object perception through elevation attention.
2) We design Dense-LGEANet, a dedicated network designed for ALS point cloud semantic segmentation. It can provide effective supervision of point cloud features at multiple scales, thus improving the classification accuracy of objects with varying resolutions in airborne point clouds.
3) We evaluate our approach on publicly available airborne point cloud datasets, including the ISPRS Vaihingen 3D dataset, LASDU dataset, and WHU point cloud dataset. The results demonstrate the significant performance improvement achieved by our method.

The subsequent sections of this article are organized as follows. Section II presents the detailed introduction of the proposed LGEA module and Dense-LGEANet. In Section III, extensive experiments are conducted on the ISPRS Vaihingen 3D dataset and the LASDU dataset to show the superior performance of our method. Furthermore, a comprehensive ablation experiment and generalization test are performed to validate the effectiveness of our proposed module and network. Finally, Section IV concludes this article and provides some potential future research directions.

II. Methodology

In this section, we first introduce the proposed LGEA module. The LGEA module includes graph convolution block, transformer block, and elevation attention block. We will introduce them separately. Then, we analyze the structure of the proposed Dense-LGEANet network. Finally, the designed loss function of the network is given.

A. LGEA Module

As illustrated in Fig. 1, the LGEA module takes the original point features as input and leverages the graph convolution block and transformer block to extract local and global features of the point cloud, respectively. The elevation attention block is also incorporated to enhance the perception ability of the features. Finally, the LGEA module fuses the features output by graph convolution block and transformer block to obtain the enhanced features as output.

1) Graph Convolution Block: To construct the graph, for each point p_i in point cloud $P = \{p_i | i = 1, \ldots, n\}$, its nearest k points $\{p^1, \ldots, p^j, \ldots p^k\}$ is calculated. Then a directed graph can be denoted as $G = (V, E)$, where the vertices $V = \{V_i | i = 1, \ldots, n\}$ and edges $E \subseteq V^2$. Considering a point $p_i \subseteq R^3$ with its corresponding feature $f_i \subseteq R^d$ as input, the local information of the point can be easily represented by a graph. The edge feature for each connection can be expressed as follows:

$$e_{ij} = g(\varphi(p_i, p_j) || \beta(f_i, f_k)) \tag{1}$$

where function $g()$ comprises a set of learnable parameters, typically implemented as a shared MLP. The $||$ indicates concatenation operation. Function $\varphi()$ and $\beta()$ are employed for position encoding and graph feature encoding, respectively. Specifically, their definitions are as follows.

a) Position encoding: Geometric information is the most important attribute of point cloud, which can be intuitively reflected by the position information of the point. Therefore, it is necessary to encode the position information to get the spatial distribution of points. Inspired by RandLA-Net [15], we encode the positional relationship between points through the nearest neighbor points of spatial coordinates. Specifically, for each point $p_{i,}$ which is regarded as the central point, its relative position relationship can be encoded by its KNN neighboring

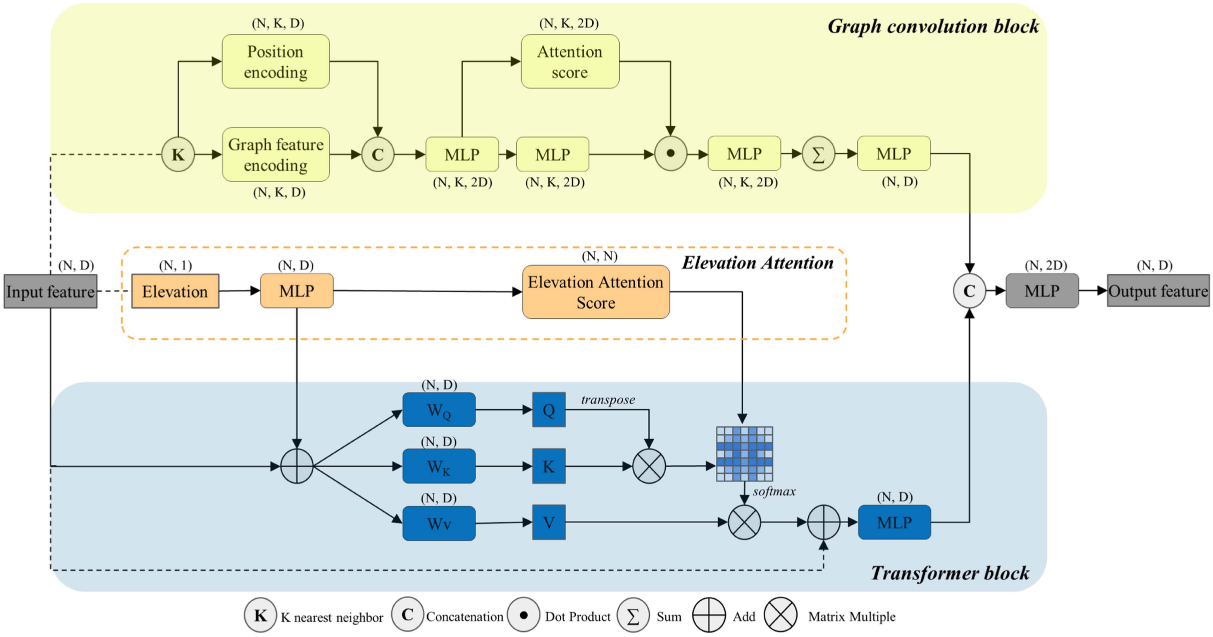


Fig. 1. Proposed LGEA module.

points as follows:

$$\varphi(p_i, p_j) = MLP(p_i||(p_j - p_i)||(p_i, p_j)_2) \tag{2}$$

where p_i and p_j represent the coordinates of the central point and its KNN neighboring points, respectively, and $(p_i, p_j)_2$ denotes the Euclidean distance between p_i and p_j.

b) Graph feature encoding: Different from position encoding, the graph feature encoding layer takes the feature encoded by the preceding network layer as input. Following the approach of DGCNN [28], the layer can be defined as follows:

$$\beta(f_i, f_k) = MLP(f_i||(f_k - f_i) \parallel (f_i, f_k)_2) \tag{3}$$

where the coordinates of the central point and its neighboring points in feature space are denoted as f_i and f_k, respectively, while the feature distance between them is represented as $(f_i, f_k)_2$.

c) Attention score: After obtaining the feature e_{ij} of each edge, a local aggregation operation is needed to update the feature of the center point, so as to achieve a similar effect as the convolution operation. Max pooling is the most commonly used aggregation operation, but it will lose some features of points in large-scale point cloud semantic segmentation, resulting in loss of precision, which has been demonstrated in [15], [40], and [41]. Following it, in order to preserve important features, an attention score is calculated for each edge as follows:

$$a_{ij} = \rho(e_{ij}) \tag{4}$$

where function $\rho()$ is a shared MLP layer followed by a softmax activation. Subsequently, to aggregate the features of neighboring points, the graph features are weighted by the attention scores and summed as follows:

$$\bar{f_i} = \alpha\left(\sum_{j=1}^{k} \delta\left(\gamma\left(e_{ij}\right) \cdot a_{ij}\right)\right) \tag{5}$$

where $\bar{f_i}$ represents the enhanced feature output after the graph convolution block. $\alpha()$, $\delta()$ and $\gamma()$ are MLP layer. $\cdot$ is dot product, and $\sum_{j=1}^{k}$ is sum operation.

2) Transformer Block: Transformer is a network architecture commonly composed of three modules: input feature embedding, positional encoding, and self-attention. The core component is self-attention, which can generate attention features based on the input features of the global context. To begin with, the input feature f_{in} can acquire three learnable weight matrices W_Q, W_K, and W_V through MLP. Consequently, the matrices Query(Q), Key(K), and Value(V) can be expressed in the following manner:

$$\begin{cases} Q = f_{\text{in}} \times W_Q \\ K = f_{\text{in}} \times W_K \\ V = f_{\text{in}} \times W_V \end{cases} \tag{6}$$

where $\times$ is matrix multiple.

Next, the attention weight between any point feature can then be obtained by matching the Q and the K matrix

$$A = \text{softmax}\left(\frac{Q \times K}{\sqrt{C}}\right) \tag{7}$$

where A is the generated attention map, and C represents the dimension of Q and K. The attention feature is then defined as the product of A and V

$$f_{sa} = A \times V. \tag{8}$$

Finally, the output feature by transformer block can be defined as follows:

$$f_{\text{out}} = MLP(f_{sa}) + f_{\text{in}}. \tag{9}$$

It should be noted that the designed transformer block differ from the previous methods in that the encoding of neighborhood

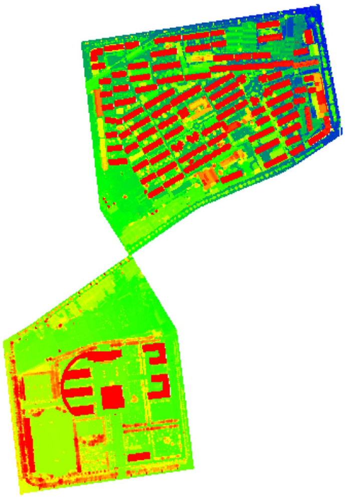


Fig. 2. LASDU dataset colored by elevation.

features is discarded, because the graph convolution block mentioned earlier can well encode the local neighborhood features of the point cloud. And then, elevation feature and elevation self-attention feature are added to the block, which makes the network more suitable for semantic segmentation of airborne point clouds in large scenes. We will introduce it in the next section.

3) Elevation Attention: Point clouds captured from airborne environments exhibit distinct characteristics compared to those captured indoors. And the height distribution of various point cloud classes follows a consistent pattern. For instance, as illustrated in Fig. 2, roof points within the LASDU dataset usually have higher elevations, while ground and low vegetation points tend to have lower elevations. In order to integrate this attention into the global feature scale, we compute the elevation attention score using the following formula:

$$f_{ele} = MLP(Z) \tag{10}$$

$$S_{ele} = softmax\left(f_{ele} \times f_{ele}^{T}\right) \tag{11}$$

$$A_{after} = A_{before} + S_{ele} \tag{12}$$

where Z is the z coordinates of the input point set, f_{ele}^{T} is the transpose of elevation coding feature f_{ele}, S_{ele} is the elevation attention score, and A_{before} is the is the attention weight obtained in (7). Moreover, the elevation coding feature f_{ele} is added to f_{in_before} to increase the sensitivity of input features for elevation

$$f_{in\_after} = f_{in\_before} + f_{ele}. \tag{13}$$

B. Network Architecture of Dense-LGEANet

Dense connection and deep supervision have been successfully applied in 2D image segmentation such as UNet++ [42], and has gradually attracted the attention in the field of 3D point cloud [25], [43], [44]. It can train the network with richer features to improve the segmentation performance of different size objects. Inspired by it, we design the Dense-LGEANet, which is built upon the LGEA module and shown in Fig. 3. The network is a dense hierarchical network by constructing skip pathways between down-sampling and up-sampling layers. The down sampling layers use the farthest point sampling algorithm to reduce the number of point cloud and encode the feature by convolution module, while up sampling layers restore point cloud to its original dimension by nearest neighbor interpolation to output per-point prediction. Following PointNet++, the down sampling layers and up sampling layers can be defined as set abstraction (SA) layers and feature propagation (FP) layers, respectively. Consequently, each feature $F_{i,j}$ extracted by different layer can be expressed as follows:

$$F_{i,j} = \begin{cases} SA(F_{i-1,j}), j = 0 \\ MLP([F_{i,k}]_{k=0}^{j-1} || FP(F_{i+1,j-1})), i = 0, j > 0 \\ MLP([F_{i,k}]_{k=0}^{j-1} || FP(F_{i+1,j-1}) || SA(F_{i-1,j-1})), \\ i > 0, j > 0 \end{cases} \tag{14}$$

where MLP denotes multilayer perceptron, and ‖ means the concatenation operation. It is worth noting that we use multiple SA operations between different layer to facilitate the interaction between features at different scales. Meanwhile, to optimize computational efficiency, we selectively incorporate the LGEA module into the $j = 0$ layer. Because of the dense connectivity of the entire network, the features extracted by the LGEA module can be effectively propagated to every layer.

C. Multiscale Loss Function

The distribution of object scales in airborne point clouds is uneven, such as the roof is much larger than the car. To enhance the accuracy of object segmentation across different scales, the network uses a multiscale feature supervision training approach. Specifically, the features F(0,3), F(1,2), F(2,1), and F(3,0) pass through the fully connected layer to predict the confidence scores for all candidate semantic categories, respectively. In addition, deep supervision can be achieved by utilizing labels of different resolutions $\{L_0 \in R^{N\times C}, L_1 \in R^{N/4\times C}, L_2 \in R^{N/16\times C}, L_3 \in R^{N/64\times C}\}$. The multiscale feature loss function based on the cross-entropy loss can be defined as follows:

$$L_{mf\_loss} = \sum_{i=0}^{3} \lambda_i L_{i\_loss} = \sum_{i=0}^{3} \sum_{j=1}^{N_i} \sum_{c=1}^{C} \lambda_i \left(L_i^{cj} \log \widehat{F}_i^{cj} + \left(1 - L_i^{cj}\right) \log\left(1 - \widehat{F}_i^{cj}\right)\right) \tag{15}$$

where L_{i_loss} represents the loss of the different feature, while λ_i serves as a weight hyper-parameter. C stands for the total number of categories, while N_i signifies the count of points within the respective category.

III. Experiments and Results

A. Datasets and Configuration

ISPRS Vaihingen 3D dataset [45] and LASDU dataset [46] are used for our experiments. The ISPRS Vaihingen 3D dataset

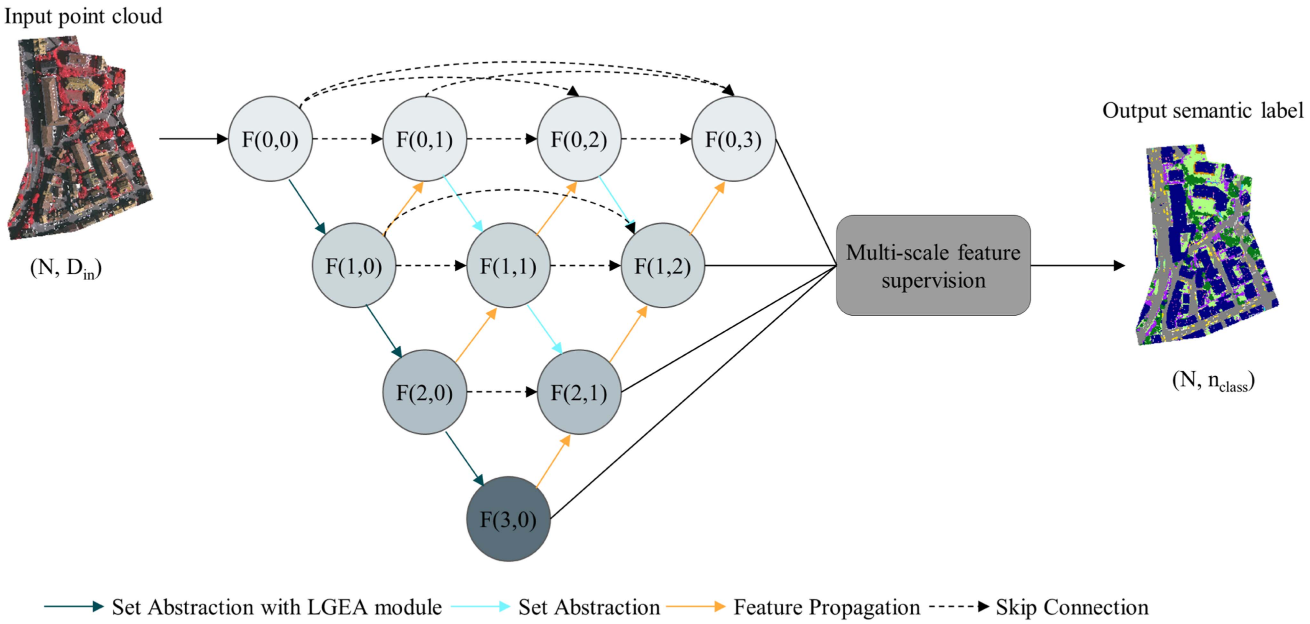


Fig. 3. Architecture of the proposed network.

TABLE I
CLASS DISTRIBUTION OF THE ISPRS VAIHINGEN 3D DATASET

Class	Training Set	Test Set
Powerline	546	600
Low Vegetation	180850	98690
Impervious Surfaces	193723	101986
Car	4614	3708
Fence	12070	7422
Roof	152045	109048
Facade	27250	11224
Shrub	47605	24818
Tree	135173	54226
Total	753876	411722

TABLE II
CLASS DISTRIBUTION OF THE LASDU DATASET

Class	Training Set	Test Set
Ground	704425	637257
Building	508479	395109
Tree	204775	108466
Low Vegetation	210495	192051
Artifact	66738	53061
Total	1694912	1385944

comprises point clouds obtained from ALS in three distinct regions, encompassing a total of nine classifications: powerline, low vegetation, impervious surfaces, car, fence, roof, facade, shrub, and tree. The data is collected above Stuttgart, Germany, at an approximate altitude of 500 m, with a flying angle of around 45°. The point density averages at approximately 6.7 points per square meter. Each point entry includes various fields such as XYZ coordinates, reflectivity, repeat count information, and label. The training set consists of 753876 points, while the testing set contains 410722 points. Table I provides an overview of the category distribution within the dataset. Following RFFS-Net [37], we partition the entire scene into regular cuboid blocks measuring 30 m × 30 m horizontally. During network training, we sample 4096 points from each block as input, whereas in the test phase, all points are utilized to calculate accuracy.

LASDU is a large-scale airborne point cloud dataset acquired by the use of a Leica ALS70 system onboard an aircraft with a flying height of about 1200 m. The point density is approximately 3–4 points per square meter. It encompasses four distinct regions, namely Sections I–IV, with semantic categories including ground, buildings, trees, low vegetation, and artifacts. A comprehensive breakdown of the category distribution can be found in Table II. Sections II and III are used as the training set, and the remaining Sections I and IV are selected as our test set. Due to the relatively lower point density in LASDU, the dataset is partitioned into regular cuboid blocks measuring 50 m × 50 m in the horizontal dimensions. During network training, 4096 points are sampled from each block and fed into the network.

The WHU point cloud dataset is used to test the generalization capability, including info point cloud and Luojia Hill point cloud. The area corresponding to the data is the Department of Information Science and Luojia Hill, Wuhan University. It is collected using the Luojia Yiyun FT1500 LiDAR system [49], a UAV-based device utilizing rotating mirror scanning. The point clouds are highly dense, precise with common types of campus elements, covering with lots of trees and buildings. However, it has no semantic labels, so we provide the orthophoto for qualitative evaluation of test result.

In our experiment, we configure the graph convolution block to utilize 16 nearest points, and the hyper-parameters for the weight in the multiscale feature loss function are set to $\{\lambda_0 = 0.5, \lambda_1 = 0.2, \lambda_2 = 0.2, \lambda_3 = 0.1\}$. The batch size is set to 16, and the model is trained for 500 epochs with a learning rate of

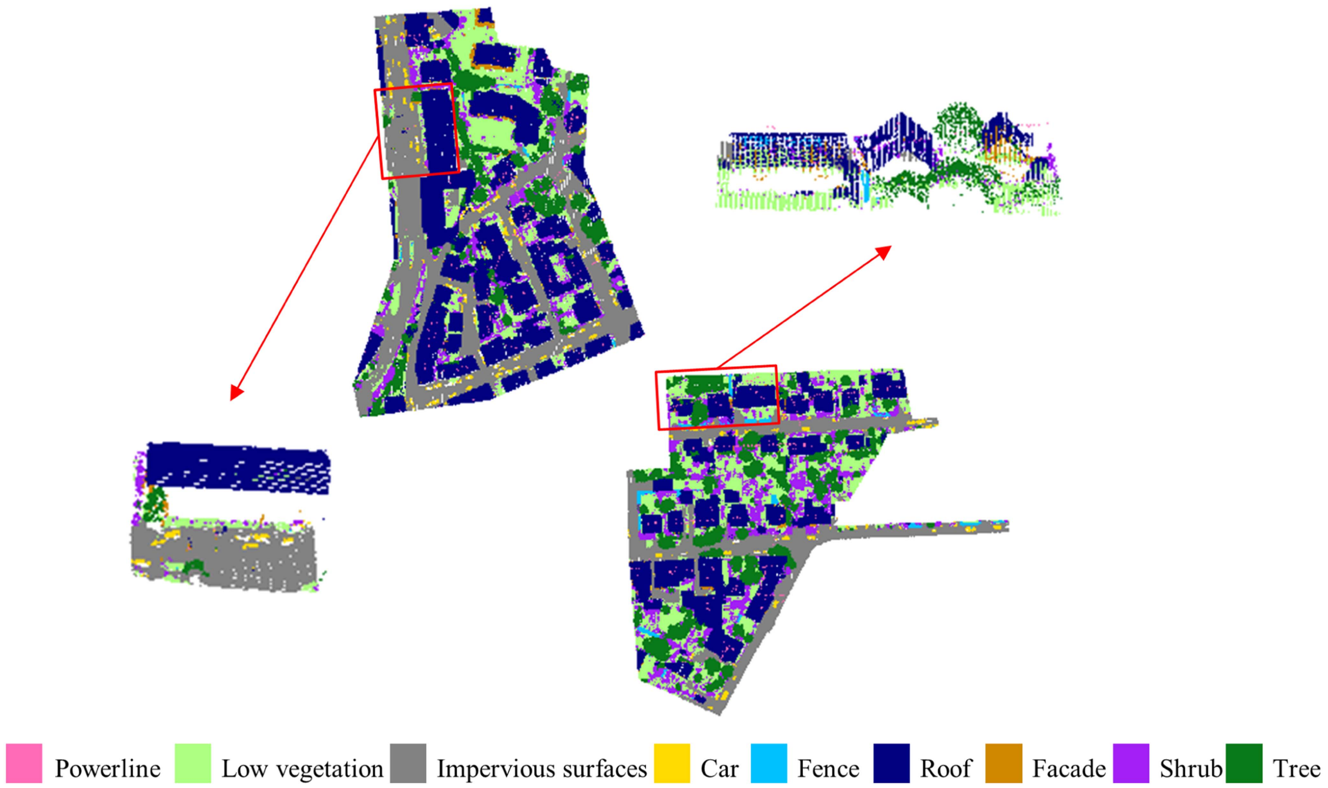


Fig. 4. Visualization of the classification results achieved by proposed Dense-LGEANet on the ISPRS Vaihingen 3D dataset.

0.002 for each training run. To minimize the loss function, we employ the Adam optimizer with a weight decay of 0.002. All experiments are conducted on a single NVIDIA A40 GPU. Following the previous work [37], we employ the overall accuracy (OA), IoU score, and F1 score as evaluation metrics to assess the performance of our proposed Dense-LGEANet. The OA represents the percentage of correctly classified points out of the total points. The IoU score measures the ratio of the intersection to the union of predicted and true values for the same class, while the F1 score is the harmonic average of accuracy and recall for each category. The IoU and F1 scores are particularly suitable for evaluating performance when dealing with imbalanced category distributions, as they assess performance on a per-category basis. The mathematical definitions for each indicator are provided in the following formula:

$$precision = \frac{TP}{TP + FP} \tag{16}$$

$$recall = \frac{TP}{TP + FN} \tag{17}$$

$$IoU score = \frac{TP}{TP + FP + FN} \tag{18}$$

$$F1score = 2 \times \frac{precision \times recall}{precision + recall} \tag{19}$$

where TP represents true positives, FP represents false positives, and FN represents false negatives. The mean F1 score (mF1) and mean IoU (mIoU) can be calculated as the mean of the F1 scores and IoU scores across all categories, respectively.

B. Semantic Segmentation Results

1) Results on the ISPRS Vaihingen 3D Dataset: Fig. 4 illustrates the visualization of the prediction outcomes achieved by our Dense-LGEANet on the ISPRS Vaihingen 3D dataset. It can be seen that most of the points are classified correctly. Notably, the red boxes indicate the classification results for representative regions. Our network demonstrates excellent performance in classifying both large-scale point cloud objects such as roofs, and small-scale point cloud objects including cars, fences, and powerlines. This capability stems from the integration of our well-designed LGEA module, which comprises a graph convolution block for capturing local details and a transformer module for encoding global information. Through the application of multiscale feature supervision training, the network becomes efficient in learning local features across various scales, as well as global features.

Table III presents a comprehensive comparison of the classification performance achieved by our Dense-LGEANet on the ISPRS Vaihingen 3D dataset against several other methods, including PointNet++ [11], PointSIFT [47], PointCNN [21], DGCNN [28], KPConv [24], RandLA-Net [15], SCF-Net [18], and RFFS-Net [37]. The table clearly demonstrates the superiority of our proposed neural network in terms of two metrics, including mF1 and mIoU. Compared to the baseline method of PointNet++, our network achieved improvements of 2.0% in OA, 6.4% in mF1, and 6.5% in mIoU. Furthermore, when compared to the state-of-the-art performance of RFFS-Net, Dense-LGEANet demonstrated improvements of 0.4% and

TABLE III
QUANTITATIVE RESULTS OF THE DIFFERENT APPROACHES ON THE ISPRS VAIHINGEN 3D DATASET

Method	Powerline	Low_veg	Imp_surf	Car	Fence	Roof	Facade	Shrub	Tree	OA	mF1	mIoU
PointNet++	57.9	79.6	90.6	66.1	31.5	91.6	54.3	41.6	77.0	81.2	65.6	52.0
PointSIFT	55.7	80.7	90.9	77.8	30.5	92.5	56.9	44.4	79.6	82.2	67.7	54.6
PointCNN	61.5	**82.7**	91.8	75.8	35.9	92.7	57.8	**49.1**	78.1	83.3	69.5	56.3
DGCNN	44.6	71.2	81.8	42.0	11.8	93.8	**64.3**	46.4	81.7	78.3	59.7	46.8
KPConv	63.1	82.3	91.4	72.5	25.2	94.4	60.3	44.9	81.2	**83.7**	68.4	55.7
RandLA-Net	68.8	82.1	91.3	76.6	43.8	91.1	61.9	45.2	77.4	82.1	70.9	57.4
SCF-Net	64.2	81.5	90.8	73.9	35.2	93.6	61.5	43.4	**82.6**	83.2	69.8	56.8
RFFS-Net	**75.5**	80.0	90.5	78.5	**45.5**	92.7	57.9	48.3	75.7	82.1	71.6	58.2
Ours	72.2	80.8	**92.1**	**78.7**	41.2	**93.9**	61.5	47.2	80.5	83.2	**72.0**	**58.5**

The bold values mean the highest value of the current indicator.

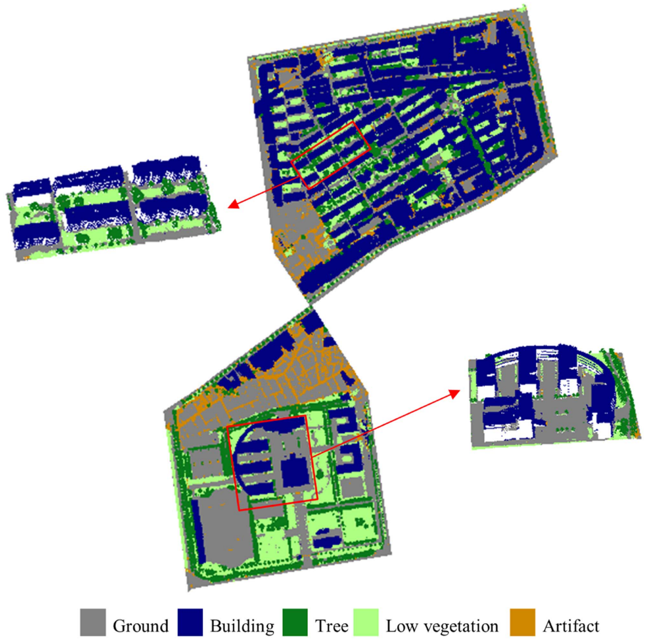


Fig. 5. Visualization of the classification results achieved by proposed Dense-LGEANet on the LASDU dataset.

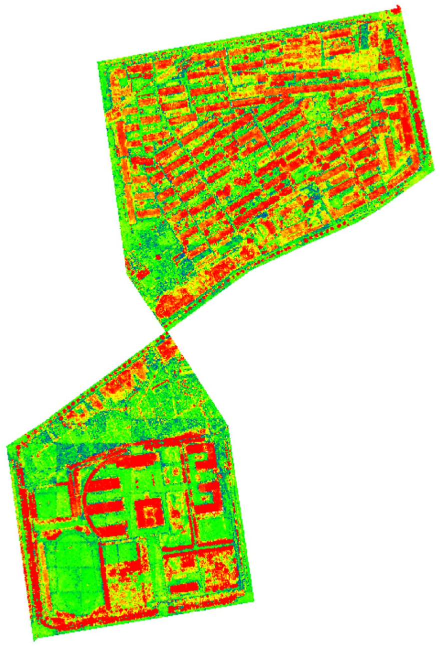


Fig. 6. Visualization of the point cloud feature on the LASDU dataset.

0.3% in mF1 and mIoU, respectively. Notably, our network achieved the highest classification performance in categories such as impervious surfaces and roofs. This improvement can be attributed to the incorporation of elevation attention, which enhances the differentiation between these two feature types. In addition, our method gets favorable results in challenging categories such as powerline, car, and facade, where their points are sparse and samples are limited. Because of the ability to enhance each other between the features of different layers, Dense-LGEANet can capture local features of different scales to improve the performance of the category with little label.

2) Results on the LASDU Dataset: The visualization of prediction results for the LASDU dataset achieved by our Dense-LGEANet is depicted in Fig. 5. It shows that most of points are correctly distinguished. The details of the scene's classification are highlighted by the red box. It can be seen that both the building points with regular structure and the tree points with irregular structure have been accurately classified. This demonstrates the network's ability to perceive the overall structure as well as the local structure. Furthermore, we visualize the last layer of features in the network as shown in Fig. 6. The maximum value in each point feature channel is selected as the feature response value. It can be seen that due to the use of global transformer feature and elevation attention, the Dense-LGEANet has larger feature response value for the objects with greater height variations, and thus improves the perceptual ability of the network.

A performance comparison between our method and other approaches on the LASDU dataset is presented in Table IV. The methods included for comparison are PointNet++ [14], PointCNN [21], DGCNN [28], KPConv [24], PosPool [48], PointConv [22], and RFFS-Net [37]. Our Dense-LGEANet demonstrates the best classification performance, outperforming the other advanced methods in terms of mF1 and mIoU metrics. Moreover, our network demonstrates superior performance in four out of the five categories, including ground, building, tree, and low vegetation, which can prove the high performance of our method. It is worth noting that all of the methods have relatively low classification accuracy on the artifact class of the LASDU dataset. Because this category is composed of wall, fence, light

TABLE IV
QUANTITATIVE RESULTS OF THE DIFFERENT APPROACHES ON THE LASDU DATASET

Method	Ground	Building	Tree	Low_veg	Artifact	OA	mF1	mIoU
PointNet++	87.7	90.6	82.0	63.2	31.3	82.8	71.0	59.0
PointCNN	89.3	92.8	84.0	62.8	31.7	85.0	72.1	60.9
DGCNN	90.5	93.2	81.6	63.3	37.1	85.5	73.1	61.6
KPConv	89.1	93.4	83.2	59.7	31.9	83.7	71.5	60.2
PosPool	88.3	93.7	83.9	61.0	38.3	83.5	73.0	61.4
PointConv	89.6	94.3	84.6	67.5	36.4	85.9	74.5	63.4
RFFS-Net	90.9	95.4	86.8	71.0	**44.4**	87.1	77.7	66.9
Ours	**91.4**	**95.6**	**88.0**	**72.5**	43.9	**87.7**	**78.3**	**67.2**

The bold values mean the highest value of the current indicator.

TABLE V
ABLATION STUDY OF THE CORE MODULES OF DENSE-LGEANET

Model	GCB	TB	EA	DC	mF1	mIoU
A	√				66.9	54.3
B		√			58.1	48.8
C	√	√			68.9	55.2
D	√	√	√		70.9	57.0
E				√	69.5	56.2
F	√	√	√	√	**72.0**	**58.5**

The bold values mean the highest value of the current indicator.

TABLE VI
EXPERIMENT RESULTS OF DIFFERENT ELEVATION ATTENTION COMBINATION METHODS

Model	Q	K	V	SA	mF1	mIoU
A					68.9	55.2
B	√	√			69.1	55.3
C			√		69.5	55.9
D				√	69.8	56.6
E	√	√	√	√	**70.9**	**57.0**

The bold values mean the highest value of the current indicator.

pole, vehicle, and other artificial, the spatial and morphological distribution of these objects are quite different, which brings challenge to the training of the network.

However, our network's classification accuracy in this category is only second to that of RFFS-Net, which can be attributed to the network's ability to extract the global receptive field.

C. Ablation Study

In this section, we conduct ablation studies on the core components, the integration of elevation attention features and dense connected network architecture to demonstrate the effectiveness of our Dense-LGEANet using the ISPRS Vaihingen 3D dataset.

1) Ablation Study of the Core Components: To verify which component plays a key role in our network, we conducted some ablation experiments as shown in Table V. Specifically, GCB means graph convolution block, TB means transformer block, EA means elevation attention, DC means dense connected network architecture with multiscale feature supervision. Without the addition of any modules, our network structure is the same as the baseline of PointNet++. As can be seen from the table, by adding the graph convolution block in model A, the mF1 and mIoU experience increase by 1.3% and 2.3%, respectively, which shows its ability to effectively aggregate local features. However, when solely relying on the transformer block, the classification accuracy significantly decreases, indicating that encoding only global information is not suitable for large-scale point cloud semantic segmentation. When local information encoded in GCB is combined with global information encoded in TB, the performance of the model can be further improved, as shown in model C. The complete LGEA module, comprising the GCB, TB, and EA, achieves an mIoU of 57.0%, further validating the effectiveness of our proposed module. In addition, our approach achieves the best performance when employing the dense connected network architecture, as shown in Model F.

Fig. 7 shows the detailed classification results of different models. Subfigure (b) shows that the baseline method easily confuses roof points with some tall tree points when no modules are added. This situation improves when a dense connected network architecture is used, but there is still a small proportion of facade points and powerline points that are misclassified, as shown in (c). The full model uses the transformer encoding mechanism to fuse global context information and adds elevation attention, so that it performs well in classifying objects at different scales, such as powerline and façade. The results are basically consistent with the ground truth.

2) Ablation Study of How Elevation Attention Features Are Combined: We conducted comparison experiments to explore different methods of combining elevation attention features within our transformer block. These experiments were performed on a network without dense connected architecture. In the transformer's core self-attention mechanism, the query items are represented by Q, the data items by V, and the corresponding keys by K. The final self-attention score (SA) is a weighted sum of all data items, which is a global result. Table VI illustrates the addition of global elevation coding features to different parts of the transformer, with model E achieving the best results. This result is also reasonable, because adding elevation features to each item of self-attention makes them have the same modal features, which is better for query in global context information. At the same time, the combination of elevation information is equivalent to adding an implicit position code, which is more conducive to the learning of attention scores.

3) Ablation Study of the Dense Connected Network Architecture: In order to explore the most reasonable dense network structure design, we conducted multiple sets of ablation experiments, as shown in Table VII. In particular, 3L represents the 3-layer network structure, 4L represents the 4-layer network

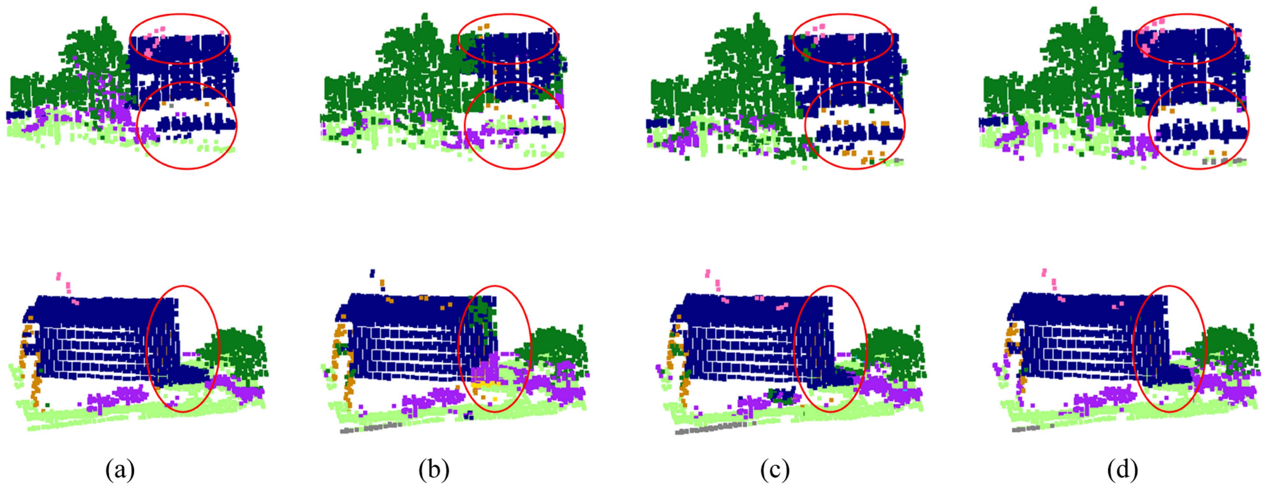


Fig. 7. Detailed results on the on the ISPRS Vaihingen 3D dataset with different models. (a) Ground truth. (b) Baseline. (c) Ours without LGEA module. (d) Ours.

TABLE VII
ABLATION STUDY OF THE DENSE CONNECTED NETWORK ARCHITECTURE

Model	3L	4L	5L	DR	MS	mF1	mIoU
A	√					65.0	50.7
B	√			√		65.1	51.2
C	√			√	√	66.3	53.0
D		√				68.7	55.2
E		√		√		69.9	56.0
F		√		√	√	**72.0**	**58.5**
G			√			68.5	54.8
H			√	√		68.8	55.5
I			√	√	√	70.6	57.5

The bold values mean the highest value of the current indicator.

structure, 5L represents the 5-layer network structure, DR represents dense residual connection, and MS represents the multiscale loss function. As can be seen from the table, fewer network structure layers cause the accuracy of semantic segmentation to drop rapidly, but using DR and MS can improve the performance to some extent. This also verifies their effectiveness, but in comparison, the performance improvement brought by MS is more significant than that of DR, which can also be observed in other network structures with different layers. This shows that in a densely connected network architecture, multiscale loss function can promote the network's learning of features at different scales, thereby improving performance. Furthermore, the network with a 4-layer structure achieved the best results in our experiments. Although there is not much performance gap with the network with a 5-layer structure, more network layers mean more parameters and it is easier to cause the problem of overfitting.

D. Generalization Capability of the Proposed Model

To demonstrate the generalization capabilities of our proposed model, we conduct additional experiment on WHU point cloud dataset. In the experiment, we directly employ a pretrained model derived from the LASDU dataset to classify the points within the WHU point cloud dataset without retraining. It should be noted that unlike the point clouds in the LASDU dataset, the WHU point cloud dataset has a large difference in the elevation distribution of building points and does not have any semantic labels, which requires higher generalization performance of the test model. The corresponding orthophoto and semantic segmentation results are presented in Figs. 8 and 9, respectively. It can be seen that our model can classify most of the point clouds correctly, and the classification results can roughly correspond to the orthophoto. However, in the classification results of RFFS-Net, some building points are misclassified as artifact points, which may be caused by the inconsistent elevation of the buildings in the area. Our model overcomes this challenge by incorporating elevation attention and global context information. Fig. 10 provides a detailed view of the semantic segmentation results obtained by our test model. It shows that our model successfully distinguishes between buildings, trees, and artifacts (vehicles) adjacent to buildings, which verifies the strong

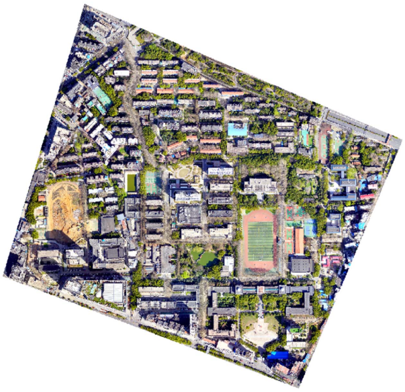


Fig. 8. Orthophoto corresponding to the WHU info point cloud.

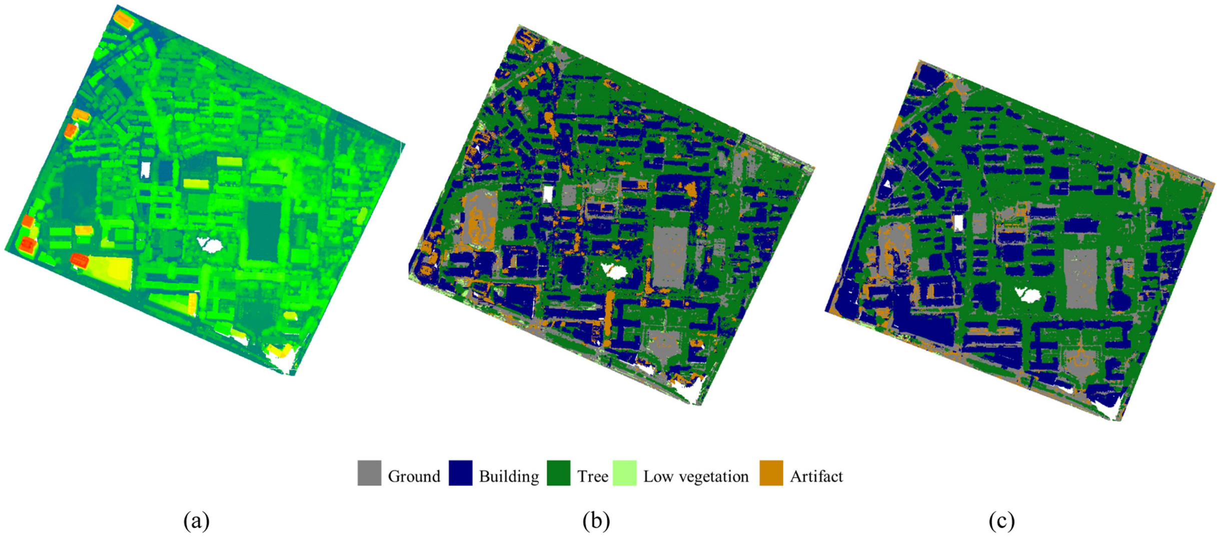


Fig. 9. Generalization validation results tested on the WHU info point cloud. (a) Input point cloud. (b) Result of RFFS-Net. (c) Result of ours.

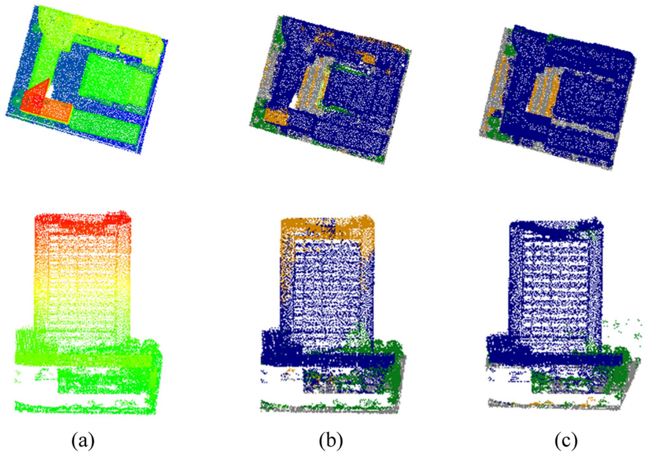


Fig. 10. Detailed results on the WHU info point cloud. (a) Raw point cloud. (b) Result of RFFS-Net. (c) Result of ours.

generalization ability of our proposed method. Figs. 11 and 12 show the classification results of the Luojia Hill point cloud. Compared with the info point cloud, it has greater elevation variation, but our network still classifies most areas correctly, which further validates the effectiveness of our proposed feature enhancement module.

E. Discussion

In the experiment, we compare the performance of different models on the ISPRS Vaihingen 3D dataset, LASDU dataset, and WHU point cloud dataset. The experimental results show that our method gets excellent results on different datasets. It can achieve an mIoU of 58.5% and an mF1 of 72.0% on the ISPRS Vaihingen 3D dataset, while an mIoU of 67.2% and an mF1 of 78.3% on the LASDU dataset. Our method also achieves good results on some challenging categories, such as powerline, car, facade, artifact, etc. Through detailed ablation experiments, we conclude that the main reason for the improvement in model performance is the addition of the LGEA feature enhancement module and the densely connected network structure. The global context information obtained by the transformer and elevation attention in the LEGA module improve the network's perception of challenging objects and enhance the generalization ability of the model. However, it should be noted that global information obtained by transformer block cannot be used alone. Only combined with local information obtained by the graph convolution block, it can achieve better results. Moreover, through ablation experiments on dense connected network architecture, we find that too few network layers will cause network performance to decline rapidly, but more network layers do not mean a significant improvement in performance because it will cause training difficulties. Consequently, designing a network structure with a reasonable number of layers is worth exploring.

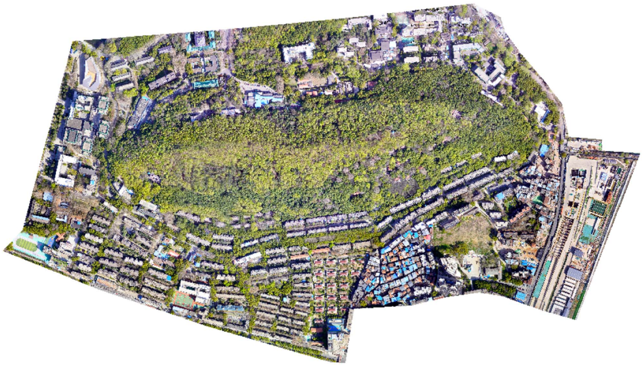


Fig. 11. Orthophoto corresponding to the WHU Luojia Hill point cloud.

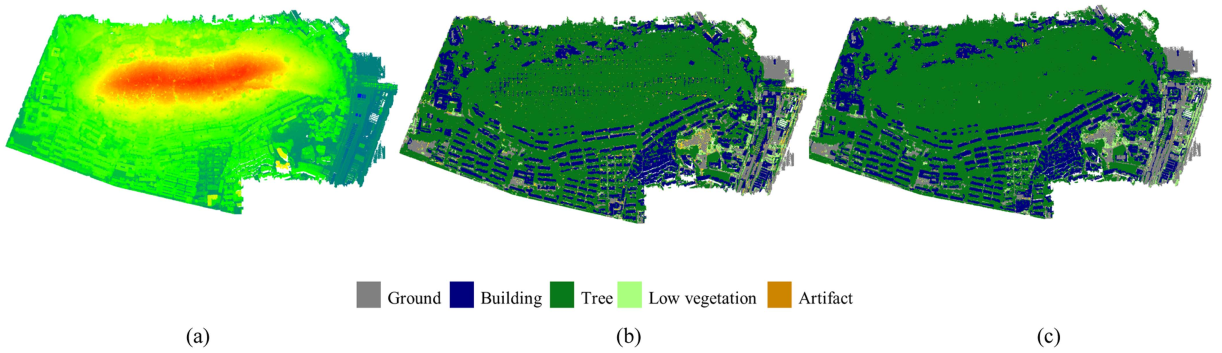


Fig. 12. Generalization validation results tested on the WHU Luojia Hill point cloud. (a) Input point cloud. (b) Result of RFFS-Net. (c) Result of ours.

IV. Conclusion

In this research, we present a novel Dense-LGEANet model designed specifically for semantic segmentation of airborne point clouds. In order to fuse local features and global context information and improve the semantic segmentation accuracy of large-scale airborne point clouds, we propose the LEGA module, which includes graph convolution block and transformer block. The position encoding, graph feature encoding, and attention score in the graph convolution block provide effective encoding of local neighborhood information of point clouds. Meanwhile, transformer combined with elevation attention effectively improves the network's ability to perceive object features with complex structures and differences in elevation distribution. Through dense connected network architecture and multiscale loss function supervision, our method effectively achieves semantic segmentation of airborne point clouds, and has excellent performance on the ISPRS Vaihingen 3D dataset, LASDU dataset, and the WHU point cloud dataset. However, it should be noted that our network also has limitations, such as the global information extraction of the transformer and the use of dense connected network architecture will inevitably increase the computational complexity.

Moving forward, our future work aims to design a more lightweight transformer module to reduce network parameters. In addition, we intend to explore methods for training the network with fewer labeled data or even in the absence of labeled data, considering the high costs associated with data labeling and training.

References

[1] J. Li, P. Shi, Q. Hu, and Y. Zhang, "QGORE: Quadratic-time guaranteed outlier removal for point cloud registration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11136–11151, Sep. 2023.

[2] Y. Gao et al., "SHREC 2023: Point cloud change detection for city scenes," *Comput. Graph.*, vol. 115, pp. 35–42, 2023.

[3] C. Nardinocchi, M. Balsi, and S. Esposito, "Fully automatic point cloud analysis for powerline corridor mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8637–8648, Dec. 2020.

[4] H. Weiser, J. Schäfer, L. Winiwarter, N. Krašovec, F. E. Fassnacht, and B. Höfle, "Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests," *Earth Syst. Sci. Data*, vol. 14, pp. 2989–3012, 2022.

[5] X. Duan, Q. Hu, P. Zhao, F. Yu, and M. Ai, "A low-drift and real-time localisation and mapping method for handheld hemispherical view LiDAR-IMU integration system," *Photogrammetric Rec.*, vol. 38, pp. 176–196, 2023.

[6] M. Mohamed, S. Morsy, and A. El-Shazly, "Improvement of 3D LiDAR point cloud classification of urban road environment based on random forest classifier," *Geocarto Int.*, vol. 37, pp. 15604–15626, 2022.

[7] C. Sothe et al., "Tree species classification in a highly diverse subtropical forest integrating UAV-based photogrammetric point cloud and hyperspectral data," *Remote Sens.*, vol. 11, 2019, Art. no. 1338.

[8] W. Huang et al., "A fast point cloud ground segmentation approach based on coarse-to-fine Markov random field," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7841–7854, Jul. 2022.

[9] D. Wolf, J. Prankl, and M. Vincze, "Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 4867–4873.

[10] F. Yang, F. Davoine, H. Wang, and Z. Jin, "Continuous conditional random field convolution for point cloud segmentation," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108357.

[11] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[12] H. Qiu, B. Yu, and D. Tao, "GFNet: Geometric flow network for 3D point cloud semantic segmentation," 2022, *arXiv:2207.02605*.

[13] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

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

[15] Q. Hu et al., "Learning semantic segmentation of large-scale point clouds with random sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8338–8354, Nov. 2022.

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

[17] 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.*, 2020, pp. 5589–5598.

[18] 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.*, 2021, pp. 14504–14513.

[19] 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.*, 2021, pp. 1757–1767.

[20] Y. Shao, G. Tong, and H. Peng, "Mining local geometric structure for large-scale 3D point clouds semantic segmentation," *Neurocomputing*, vol. 500, pp. 191–202, 2022.

[21] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on X-transformed points," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 828–838.

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

[23] 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.*, 2019, pp. 8895–8904.

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

[25] Y. Li, X. Li, Z. Zhang, F. Shuang, Q. Lin, and J. Jiang, “DenseKPNET: Dense kernel point convolutional neural networks for point cloud semantic segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5702913.

[26] 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.*, 2021, pp. 3173–3182.

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

[28] 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, pp. 1–12, 2019.

[29] J. Wan, Z. Xie, Y. Xu, Z. Zeng, D. Yuan, and Q. Qiu, “DGANet: A dilated graph attention-based network for local feature extraction on 3D point clouds,” *Remote Sens.*, vol. 13, 2021, Art. no. 3484.

[30] H. Zhou, Y. Feng, M. Fang, M. Wei, J. Qin, and T. Lu, “Adaptive graph convolution for point cloud analysis,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4965–4974.

[31] L. Chen and Q. Zhang, “DDGCN: Graph convolution network based on direction and distance for point cloud learning,” *Vis. Comput.*, vol. 39, pp. 863–873, 2023.

[32] Q. Zhong and X. Han, “Point cloud learning with transformer,” 2021, *arXiv:2104.13636*.

[33] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, “Point transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16259–16268.

[34] M. Guo, J. Cai, Z. Liu, T. Mu, R. R. Martin, and S. Hu, “PCT: Point cloud transformer,” *Comput. Vis. Media*, vol. 7, pp. 187–199, 2021.

[35] X. Lai et al., “Stratified transformer for 3D point cloud segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8500–8509.

[36] C. Park, Y. Jeong, M. Cho, and J. Park, “Fast point transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16949–16958.

[37] Y. Mao et al., “Beyond single receptive field: A receptive field fusion-and-stratification network for airborne laser scanning point cloud classification,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 45–61, 2022.

[38] W. Li, F. Wang, and G. Xia, “A geometry-attentional network for ALS point cloud classification,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 164, pp. 26–40, 2020.

[39] R. Huang, Y. Xu, and U. Stilla, “GraNet: Global relation-aware attentional network for semantic segmentation of ALS point clouds,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 1–20, 2021.

[40] Y. Xu et al., “NeiEA-NET: Semantic segmentation of large-scale point cloud scene via neighbor enhancement and aggregation,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 119, 2023, Art. no. 103285.

[41] W. Jing, W. Zhang, L. Li, D. Di, G. Chen, and J. Wang, “AGNet: An attention-based graph network for point cloud classification and segmentation,” *Remote Sens.*, vol. 14, 2022, Art. no. 1036.

[42] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[43] Z. Zeng, Q. Hu, Z. Xie, J. Zhou, and Y. Xu, “Small but mighty: Enhancing 3D point clouds semantic segmentation with U-next framework,” 2023, *arXiv:2304.00749*.

[44] 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.*, 2022, pp. 17284–17294.

[45] J. Niemeyer, F. Rottensteiner, and U. Soergel, “Contextual classification of lidar data and building object detection in urban areas,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 152–165, 2014.

[46] Z. Ye et al., “LASDU: A large-scale aerial lidar dataset for semantic labeling in dense urban areas,” *ISPRS Int. J. Geo-Information*, vol. 9, no. 7, p. 450, 2020.

[47] 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*.

[48] Z. Liu, H. Hu, Y. Cao, Z. Zhang, and X. Tong, “A closer look at local aggregation operators in point cloud analysis,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 326–342.

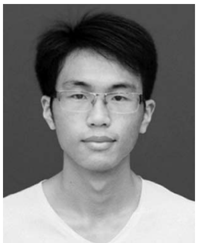
[49] H. Zhou, Q. Mao, Y. Song, A. Wu, and X. Hu, “Analysis of internal angle error of UAV LiDAR based on rotating mirror scanning,” *Remote Sens.*, vol. 14, 2022, Art. no. 5260.

Shuowen Huang was born in China, in 1997. He is currently working toward the Ph.D. degree in Surveying and mapping engineering with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include point cloud data process, computer vision, and 3D deep learning.

Qingwu Hu received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include 3S integrated mobile measurement, point cloud intelligent processing, laser vision fusion navigation, and digital cultural heritage.

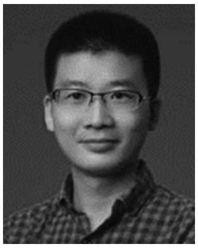


Pengcheng Zhao received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2020.

He is currently working as an Experimentalist with the School of Remote Sensing and Information Engineering, Wuhan University. He focuses on the research of intelligent processing of point cloud data.

Jiayuan Li received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include photogrammetry, machine vision, and remote sensing image processing.



Mingyao Ai received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2020.

He is currently working as an Experimentalist with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include three-dimensional modeling and remote-sensing image processing.

Shaohua Wang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2010.

He is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include remote sensing, GIS, and software engineering.