LRAD-Net: An Improved Lightweight Network for Building Extraction From Remote Sensing Images

Jiabin Liu, Huaigang Huang, Hanxiao Sun, Zhifeng Wu, *Member, IEEE*, and Renbo Luo, *Member, IEEE*

***Abstract*—The building extraction method of remote sensing images that uses deep learning algorithms can solve the problems of low efficiency and poor effect of traditional methods during feature extraction. Although some semantic segmentation networks proposed recently can achieve good segmentation performance in extracting buildings, their huge parameters and large amount of calculation lead to great obstacles in practical application. Therefore, we propose a lightweight network (named LRAD-Net) for building extraction from remote sensing images. LRAD-Net can be divided into two stages: encoding and decoding. In the encoding stage, the lightweight RegNet network with 600 million flop (600 MF) is finally selected as our feature extraction backbone net though lots of experimental comparisons. Then, a multiscale depthwise separable atrous spatial pyramid pooling structure is proposed to extract more comprehensive and important details of buildings. In the decoding stage, the squeeze-and-excitation attention mechanism is applied innovatively to redistribute the channel weights before fusing feature maps with low-level details and high-level semantics, thus can enrich the local and global information of the buildings. What's more, a lightweight residual block with polarized self-attention is proposed, it can incorporate features extracted from the space of maps and different channels with a small number of parameters, and improve the accuracy of recovering building boundary. In order to verify the effectiveness and robustness of proposed LRAD-Net, we conduct experiments on a self-annotated UAV dataset with higher resolution and three public datasets (the WHU aerial image dataset, the WHU satellite image dataset and the Inria aerial image dataset). Compared with several representative networks, LRAD-Net can extract more details of building, and has smaller number of parameters, faster computing speed, stronger generalization ability, which can improve the training speed of the network without affecting the building extraction effect and accuracy.**

Manuscript received 17 August 2022; revised 23 November 2022; accepted 13 December 2022. Date of publication 15 December 2022; date of current version 23 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41801250 and in part by the Team Project of Guangdong Provincial Natural Science Foundation under Grant 2018B030312004. *(Corresponding author: Renbo Luo.)*

Jiabin Liu and Hanxiao Sun are with the School of Geographic Science and Remote Sensing, Guangzhou University, Guangzhou 510006, China (e-mail: zz05231638@163.com; sunhanxiao199825@163.com).

Huaigang Huang is with the Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: hhgteddy@163.com).

Zhifeng Wu is with the School of Geography and Remote Sensing, and MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area, Guangzhou University, Guangzhou 510006, China, and also with the Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou 511458, China (e-mail: zfwu@gzhu.edu.cn).

Renbo Luo is with the School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China, and also with the Guangdong Province Engineering Technology Research Center for Geographical Conditions Monitoring and Comprehensive Analysis, Guangzhou 510006, China (e-mail: luorb@gzhu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3229460

***Index Terms*—Building extraction, channel attentional mechanism, high-resolution remote sensing image, semantic segmentation.**

I. Introduction

BUILDINGS are primary spaces for human life and play an important role in the development of humans and society. Recently, building extraction from remote sensing images has been widely used in smart city construction, land use surveys, military target reconnaissance, and other fields of study. The distribution of buildings also has a high reference value when evaluating the population and development of a region and understanding the historical origin of a region [1]. Traditional building extraction methods first extract the statistical features of remote sensing images using specific feature extraction algorithms and then use hand designed classifiers to extract buildings. Traditional methods can be roughly divided into two types: one method is based on the cell values of remote sensing images, and the other is object-oriented classification [2], [3], [4]. For buildings in high-resolution remote sensing images, the traditional approach is to extract buildings from optical images using spectral, textural, geometric and shading features [5], [6], [7]. From the analysis of principle, the current traditional methods of building extraction can be classified as extraction based on edge and corner point extraction [8], [9], extraction based on area segmentation [10], extraction based on building features and integration of various methods, such as digital elevation model based on auxiliary information [11], [12], [13], [14], [15]. Traditional extraction methods have many limitations with complex images. For buildings with different shapes, sizes and environments, it is difficult to obtain high accuracies and good generalizability.

With the development of deep learning technology, convolutional neural networks (CNN) have been applied to the field of building extraction due to their powerful automatic feature extraction capability [16]. Compared with traditional methods, CNN methods are much more efficient and accurate. However, CNN-based segmentation methods still have some drawbacks [17], which hinder the wide application of CNN. First, the storage overhead is large, such that the storage space required by sliding window-based CNN methods will increase markedly according to the number and size of sliding windows. The computation of convolution for each pixel block one by one is computationally repetitive because adjacent pixel blocks typically have repetitive parts, which markedly reduces efficiency. Second, as the depth of the network increases, the number of required parameters increases exponentially, which is not conducive to practical applications. These problems were somewhat solved with the proposal of a fully convolutional neural network (FCN) by Long et al. [18]. The FCN reduces the number of parameters and improves the perceptual field of the neural network. Due to these advantages, FCN-based models have been widely used for other tasks [19], [20], [21]. However, FCN does not consider global contextual information, and excessive down sampling operations also lose information of some low-level features and are insensitive to details in images.

To solve these problems, segmentation models that use low-level features with rich spatial information have been developed. Typical model structures include the improved FCN-based model U-Net proposed by Olaf et al. [22] and SegNet proposed by Badrinarayanan et al. [23]. U-Net can achieve good results in medical image segmentation, and U-Net ++ [24] and U-Net3+ [25] networks that fuse multilevel features have been derived. Since 2014, the Deeplab series of networks were proposed in [26], [27], [28], [29], and [30], where the most effective Deeplabv3+ network was inspired by the depthwise separable convolution [31], [32], [33] with the encoder-decoder structure. Deeplabv3+ designed a simple encoder-decoder structure and improves the Xception [34] and ASPP modules to improve extraction accuracy.

Recently, the model based on FCN and its variants have been widely used for the task of building extraction in remote sensing images and achieved good results. For example, Zhang et al. [35] proposed Gaussian expansion convolution and embedded it into a hierarchical dense fusion structure to form a dense hierarchical spatial Gaussian pool (dense HSGP). Dense HSGP has the advantages of the original expanded convolution and retains more contextual information while providing richer perceptual fields and higher feature extraction capability in the model. In [36], the authors proposed a dense residual neural network, it combines the densely connected CNN and residual network structures, which can enable the full integration of the underlying features with the high-level features. In order to alleviate the influence of the background of the irrelevant feature region, a net with attention block and multiple losses (AMUNet) was presented in [37]. Zhu et al. [38] developed an E-D-Net to solve the problem that the building boundary extraction is not obvious by introducing the cascading network. Meanwhile, a multiple attending path neural network was proposed by Zhu et al [39]. This net can learn multi-scale features through multiparallel paths and refine discontinuous building footprint by using attention mechanism and pyramid pool module.

With the continuous superposition of the number of model layers, the number of network parameters and computational complexity become huge, which has a serious impact on the practical applications. Therefore, more researchers try to reduce the number of parameters and simplify computational complexity of the model, and proposed some lightweight models, such as RFA-UNet [40], ARC-Net [41], and DAN-Net [42]. RFA-UNet introduced the attention mechanism to reweight the features at different stages before the feature fusion, so as to make up for the semantic differences before the features. ARC-Net used residual blocks with asymmetric convolution and atrous convolution to reduce the number of parameters in the model and speed up the calculation. DAN-Net combined lightweight network DenseNet and spatial attention fusion module to effectively extract high-level feature information and suppress noise. In 2022, Huang et al. [43] proposed RSR-Net based on the U-Net architecture, which has improved RegNet basic units by incorporating attention mechanism.

All of these networks achieved marked successes in building extraction, but how to find the balance between the accuracy and speed of building extraction is still necessary to be studied. Therefore, in order to reduce number of network parameters and make the model more efficient in practical application, we design a new lightweight network which named LRAD-Net based on encoder-decoder structure. LRAD-Net can achieve a good performance with a small number of parameters and computation, and can achieve a good balance between speed and accuracy. The main innovation points and work are summarized as follows.

1) A lightweight residual block with polarized self-attention (LPA) is proposed, it can extract fused features both from space and channels, with smaller parameters and higher accuracy of recovering building boundary.
2) We present a new depthwise separable atrous spatial pyramid pooling (DSASPP) module, it can make full use of the context information of the original remote sensing images, enlarge receptive field without changing the map shape, and improve the capability of network multi-resolution feature extraction.
3) In order to extract more details of the building, we fuse feature maps with the low-level detail and high-level semantics in the decoder. Before the fusion, the squeeze-and-excitation (SE) attention mechanism is used innovatively to improve the feature weight of the building, so as to improve the performance of extracting the building.
4) As the spatial resolution of used public datasets are 0.3 m, in order to validate the robustness of LRAD-Net, a new building data set with higher spatial resolution 0.1 m is labeled for building extraction evaluation and analysis, which we refer to as the "self-annotated UAV dataset."

II. Proposed LRAD-Net

This section details our proposed LRAD-Net, and the specific structure of LRAD-Net is shown in Fig. 1. First, considering the computational efficiency of the network, we take the network structure searched by RegNet with the computational complexity of 600 million flops (600 MF) as our feature extraction backbone, and record it as RegNet-600. Second, SE [44] block is added before the feature fusion, which can enhance the sensitivity of the network to the channel and improve the accuracy effectively under the condition that only a few network parameters are added. Third, due to the inconsistency of building scales in remote sensing images, the process of single path extraction of semantic features fails to make full use of image context information. Therefore, we proposed DSASPP module to enrich semantic information. Finally, an LPA is proposed and used in the decoder. Compared to the traditional decoder that

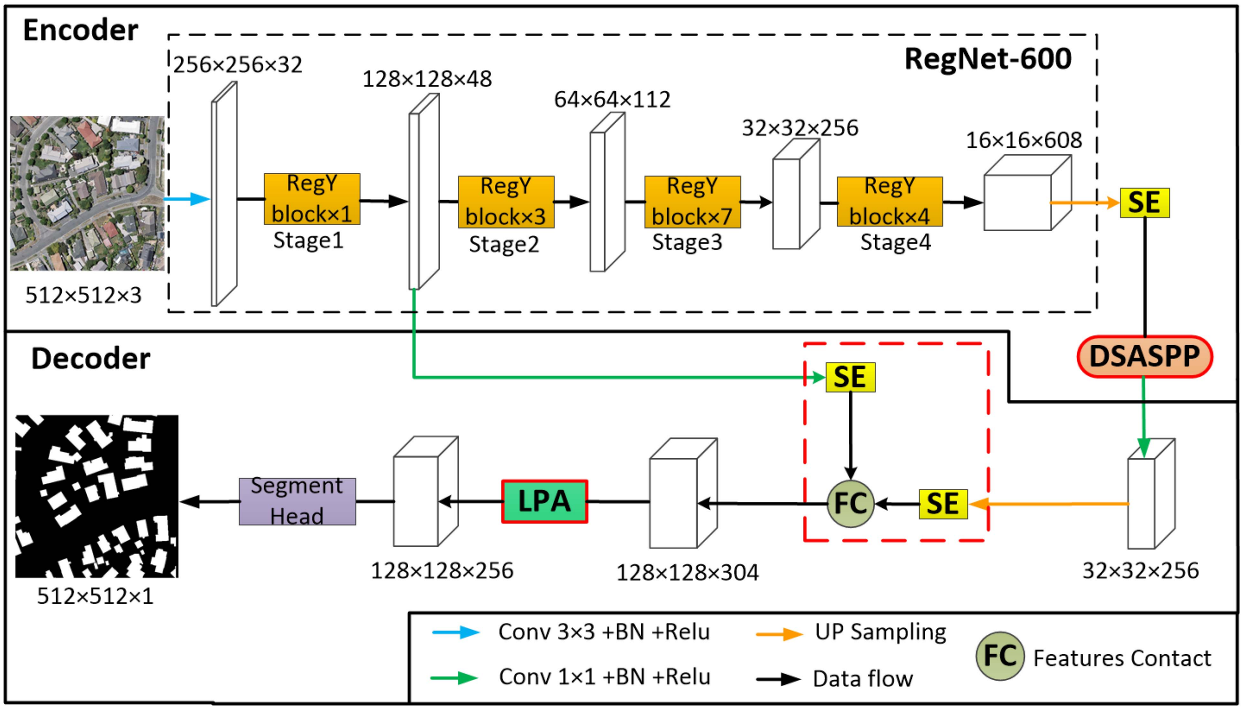

Fig. 1. Structure of LRAD-Net. The SE means squeeze and excitation attention mechanism, DSASPP represents depthwise separable atrous spatial pyramid pooling structure, LPA is a lightweight residual block with polarized self-attention.

uses two sets of 3×3 standard convolution layers [34], the LPA block can reduce the number of parameters and floating point of operations (FLOPs) while in extracting important features from the relevant spaces and channels to recover building boundaries with greater accuracy.

A. Encoder

1) RegNet-600: In 2020, RegNet [45] was proposed as a network that combines manual design with a neural structure search [46], [47], [48]. We select the network structure searched by RegNet under the computational complexity of 600 MF as our feature extraction backbone, and name it RegNet-600. The structure of RegNet-600 is shown in Fig. 2.

As shown in Fig. 2(a), RegNet-600 consists of a set of stem layers and four stages. As shown in Fig. 2(b), each stage consists of a series of stacked RegY blocks. By passing each stage, the height and width of the input feature matrix is reduced by half. The RegY block primarily consists of a residual structure with group convolution, and a SE module is added between the convolution layers, its detailed structure in shown in Fig. 2(c). In the RegY block, the first 1×1 convolution layer can reduce the features dimension, thus reducing the network parameters and the number of calculations. The 3×3 group convolution layer is used to extract textural features, where S is the stride of the convolution and G is the number of groups. g stands for the group width of each group in the group convolution.

2) Depthwise Separable Atrous Spatial Pyramid Pooling (DSASPP): DSASPP takes the advantage of depthwise separable atrous convolution and spatial pyramid pooling [49], it can enlarge receptive field without changing the maps shape and enhance the network multiscales feature extraction ability. As depthwise separable convolution can greatly compress the number of parameters and computation of the model while maintaining similar, we use depthwise separable convolution to build our DSASPP module.

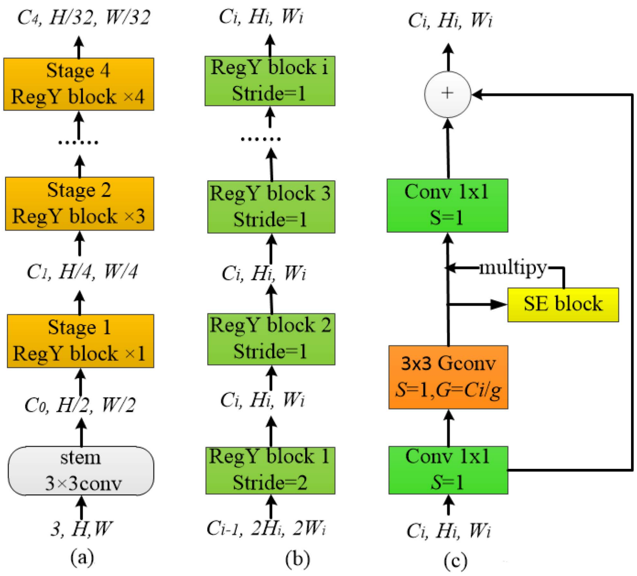

Fig. 2. A concrete illustration of RegNet-600. Gconv stands for group convolution and G is the number of groups, S is the stride of the convolution, g stands for the group width of each group in the group convolution.

AS is shown in Fig. 3, the DSASPP consists of a set of 1×1 convolution; four groups of 3×3 depthwise separable atrous convolutions with atrous rates of {6, 10,14,18}; and a global

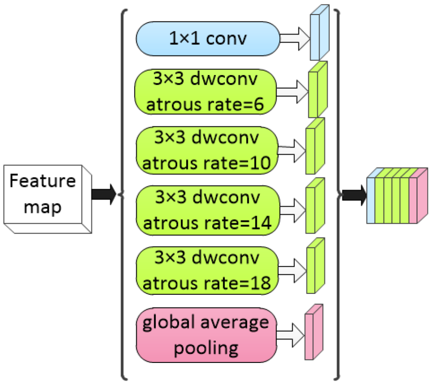
Feature map
1×1 conv
3×3 dwconv atrous rate=6
3×3 dwconv atrous rate=10
3×3 dwconv atrous rate=14
3×3 dwconv atrous rate=18
global average pooling

Fig. 3. Structure of DSASPP. The dwconv stands for deepthwise separable convolution.

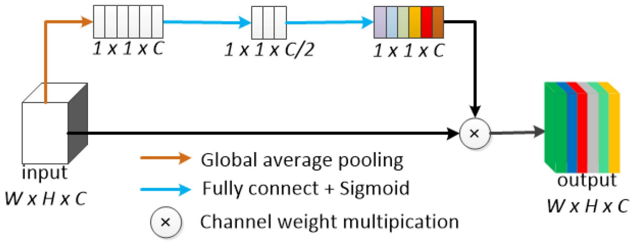

Fig. 4. Structure of SE block.

average pooling layer. Thus, the network receptive field is magnified without losing detailed information and increasing computational complexity, the output of each convolution includes a wide range of information, and the multiscale features can be captured. As a result, the targets with deferent sizes can be segmented well by the proposed four groups of atrous rate convolution kernels. In the building extraction application, DSASPP can extract more multiscale features and better segment buildings with different sizes.

3) SE Attention: In order to improve the extraction accuracy of buildings from remote sensing images, SE blocks are added after the feature extraction backbone network and before the fusion of deep and shallow layer features. Through SE block, the convolution operation of the network can focus on the extraction of building features and ignore the existence of irrelevant features. SE block operation can adjust the channel information of the input feature map, increase the weight of the building information in the feature map, and the network can pay more attention to the building information of the image, so as to complete the building extraction and reconstruction more efficiently. The SE block structure is shown in Fig. 4.

The procedure of the SE module can be divided into two steps. In the first step, we obtain a vector with a global receptive field through a global average pooling layer. The equation is as follows:

$$z = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} f_{in}\left(i, j\right) \tag{1}$$

where f_{in} is the input feature, and H and W represent the height and width of the input feature, respectively. Equation (1)

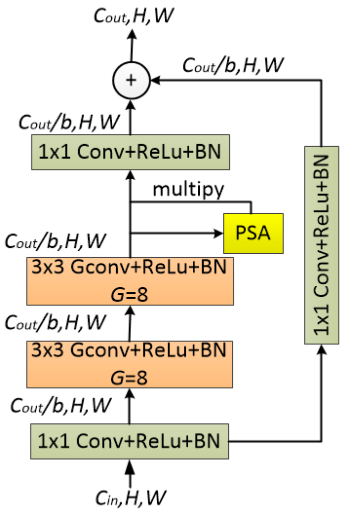

Fig. 5. Structure of proposed LPA block. Gconv stands for group convolution and G for grouping number, C_{in}, C_{out} represent the input channels, output channels, and b is bottleneck ratio.

changes the input with size $(H \times W \times C)$ into output with size $(1 \times 1 \times C)$, the real number z contains global feature information.

In the second step, z is used to generate weights for each feature channel through two fully connected layers, and its equation is as follows:

$$s = \partial\left(L_2 \sigma\left(L_1 z\right)\right) \tag{2}$$

where L_1 represents the first fully connected layer, L_2 represents the second fully connected layer, and σ represents the activation function. According to (2), s can be obtained to express the correlation between feature channels.

B. Decoder

The structure of decoder can be seen in Fig. 1. Its input is composed of two parts: one is the low-level features from the output in the first-stage layer of RegNet-600, which has 48 channels; the other part is the high-level features obtained by DSASPP module, which has 256 channels. The two parts are fused together after passing through the SE module to form a new feature map with 304 channels. This new feature map contains meaningful semantic information and building boundary information. Then, we input the fused feature maps into the LPA module for feature extraction, and obtain a feature map with 256 channels. Finally, the segmentation result can be obtained through a 1×1 convolution layer and up sampling.

1) Lightweight Residual Bottleneck With Polarized Self-Attention (LPA) Block: The proposed LPA block consists of three parts: the residual bottleneck structure, PSA block and group convolution. The structure of LPA block is shown in Fig. 5. C_{in}, C_{out}, H, and W represent the input channels, output channels, length, and width of the feature graph, respectively. b represents bottleneck ratio, which means that the channel of the output characteristic matrix is reduced to $1/b$ of the input

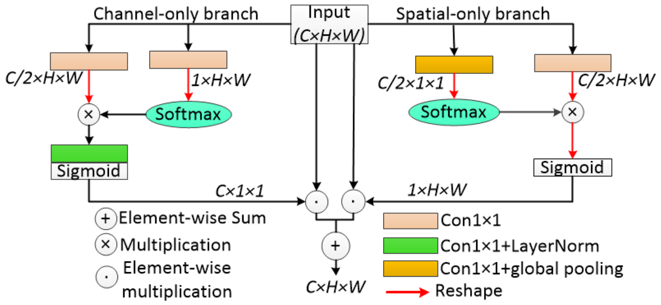

Fig. 6. Structure of PSA block.

characteristic matrix channel. G is the number of groups in a grouping convolution. In this article, b is 1 and G is 8.

As shown in the Fig. 5, LPA module includes a main branch and a shortcut branch. The main branch first reduces the dimension through a 1×1 convolution layer to reduce the number of network parameters. Then in order to improve the feature extraction capability, we use two groups of 3×3 group convolution layers for feature extraction, and finally connect a 1×1 convolution layer. In order to reduce the information loss caused by dimension reduction, PSA module is applied following group convolution. At the end of LPA, the different features extracted from the main branch and the shortcut branch are fused. Compared with the method using two sets of 3×3 standard convolution, the LPA module can improve the ability of network feature extraction with fewer parameters.

2) Polarized Self-Attention (PSA): PSA [50] block maintains a relatively high resolution in channel and spatial dimensions, which can reduce the information loss caused by dimension reduction. The PSA module is a lightweight plug and play module that can improve the performance of semantic segmentation tasks. The details of PSA can be seen from Fig. 6.

As shown in the Fig. 6, PSA block includes channel-only and spatial-only branch. Each branch is divided into two parts. The PSA module first uses 1×1 convolution to fully collapse the features in one dimension (like channel dimension) while maintaining high resolution in the orthogonal dimension (like spatial dimension). For compressed dimensions, PSA uses the softmax normalization function to enhance its information to improve the dynamic range of attention. Finally, the sigmoid function is used for dynamic mapping.

Compared with other attention mechanisms in CNNs (such as convolutional block attention module [51] and efficient channel attention [52]), PSA can maintain a higher resolution in attention calculation and capture long-distance dependencies at a lower computational overhead. In addition, in the channel and space branches, PSA can improve the performance of building extraction task by using softmax-sigmoid joint function to adjust and optimize the focus weights.

3) Group Convolution: The process of grouping convolution can be divided into three steps. We start by defining some common notations. X is the input feature map with size (C_1, H, W), Y is the output feature map with size (C_2, H, W). C_1 represents the number of input channels, H, W represent the width and height of the input, respectively. G represents the number of groups. C_2 represents the number of output channels. k stands for convolution kernel size.

In the first step, the X with size (C_1, H, W) is divided into G parts, we use X_i to represent the feature map of i*th* part. the size of X_i is as follows:

$$\text{size}\,(X_i) = \left(\frac{C_1}{G}, H, W\right). \tag{3}$$

The second step is to convolve X_i and W_i to get Y_i. W_i is the group convolution kernel size with $(\frac{c_2}{G}, k, k)$, Y_i is output feature map of i*th* part by the operation of group convolution, The size of Y_i as follows:

$$\text{size}\ \ (Y_i) = \left(\frac{C_2}{G}, H, W\right). \tag{4}$$

The third step is to concat the Y_i with size $(\frac{C2}{G}, H, W)$, then get Y with size (C_2, H, W)

$$Y\ = \sum_{i=1}^{G} Y_i. \tag{5}$$

III. Dataset and Laboratory Environment

A. Dataset Selection

To demonstrate the feasibility and generalizability of the proposed LRAD-Net in practical applications, we perform experiments on four datasets: the WHU aerial image dataset [51], WHU satellite image dataset [51], Inria aerial image dataset [52] and a custom GZHU UAV image dataset. Several sample cases from these four datasets are shown in Fig. 7.

As shown in Fig. 6, the various colors and changeable environments in the WHU satellite image dataset and Inria aerial image dataset make the task of extracting buildings more difficult compared to the WHU aerial image dataset and GZHU UAV image dataset.

The WHU aerial imagery dataset contains 8189 pictures with a resolution of 0.3 m, and each picture has a spatial coverage of 512 × 512 pixels. This dataset is divided into three sets: the training set, which contains 130 500 buildings (4736 pictures); the validation set, which contains 14 500 buildings (1036 pictures); and the test set, which contains 42 000 buildings (2416 pictures).

The WHU satellite dataset after pretreatment and random grouping, training sets (3135) and test sets (903) are obtained.

The Inria remote sensing dataset includes the areas with different landforms and building types, such as the highly dense urban center area dominated by high-rise buildings, the suburban area dominated by low-rise buildings and the mountainous area sparsely distributed buildings, which can effectively test the accuracy and robustness of the extraction of model buildings.

The image quality and label quality of the above three public datasets are relatively high, but the quality of the partcial dataset cannot be the same as that of the WHU dataset in applications. Therefore, to test the performance of LRAD-Net in practical applications, we use UAV images to make a building dataset in actual applications. In this dataset, UAV images of parts of Haizhu District, Guangzhou, in 2012 were selected to make

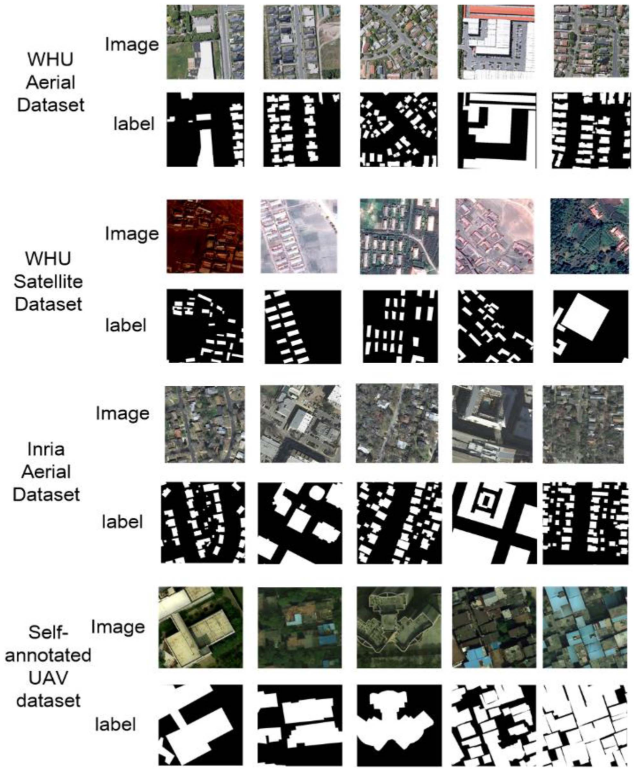

Fig. 7. Demonstration of four datasets.

building labels. The UAV image products selected for this self-annotated dataset are red, green and blue band images with a spatial resolution of 0.1 m. We select 13 high-quality drone images, each of which included different buildings (e.g., single-story houses, multistory houses, schools, irregular buildings), as well as many nonbuilding areas. The size of each image was 10 401×10 401, and a dataset is made. After completing the dataset, we trim images and labels twice with a size of 512×512, which can be input into the network in bulk. The completed dataset consisted of 6854 samples, which is divided into training sets (4798 samples) and test sets (2056 samples).

B. Experimental Environment Configuration and Evaluation Metrics

The experimental environment is a Windows 10 system, the hardware parameter CPU is i5 9500, the running memory is 64 G, the GPU is an NVIDIA Geoforce RTX 3080, the video memory is 10 G, and the deep learning framework is PyTorch 1.8.1. To make the models fit faster, the backbone networks of all models in this article are initialized with parameters that have been pretrained by ImageNet, and the remaining network parts are initialized with random parameters. To ensure the fairness and authenticity of the network, the hyperparameter settings of all networks in this article are consistent. The optimization algorithm uses the Adam optimizer with default parameter settings, the loss function uses Dice loss, the initial learning rate is set to 0.0001, and the batch size is set to 8. A total of 80 epochs of training are conducted. Each batch of images is randomly rotated horizontally and vertically, flipped 90° and scaled for data enhancement. The image size as input into the network is 512×512.

We use the evaluation metrics to measure the effectiveness of LRAD-Net: intersection over union (IoU); precision; recall; and F-score. The formulae of these indicators are as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{TP}+\mathrm{FP}} \tag{6}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

$$\text{F-score} = \frac{2\cdot\mathrm{Precision}\ \cdot\mathrm{recall}}{\mathrm{Precision}+\mathrm{recall}} \tag{9}$$

where TP refers to the number of positive samples (buildings) predicted to be positive samples; FP is the number of negative samples (nonbuildings) predicted to be positive; TN means the number of negative samples expected to be negative samples; and FN denotes the number of positive samples predicted to be negative.

IV. Experimental Results and Analysis

In this section, we first design ablation experiments to test the impact of different modules on network performance with WHU aerial image dataset. Second, to evaluate the feasibility and robustness of LRAD-Net in building extraction tasks, we use the four datasets mentioned in Section III to do several experiments and compare it with several the state-of-art networks.

A. Ablation Experiment

In order to explore the contribution of different modules (e.g., DSASPP module, SE blcok, LPA block) in improving LRAD-Net performance, we conduct ablation experiments on WHU aerial image dataset. We first build the baseline model based on LRAD-Net. In the baseline, we remove the SE block and replaced DSASPP with the ASPP (the atrous convolution with atrous rates of $\{6, 12, 18\}$) module used in Deeplabv3+. Then we replace the LPA block with two sets of 3×3 standard convolution layers.

On the basic of the baseline, the DSASPP block is named (*D*), the SE attention module is represented by (*S*), and the LPA block is called (*L*). Performance was evaluated on IoU, Precision, Recall, F-score, FLOPs and parameter. The results of the ablation study are given in Table I .

The results in Table I show that DSASPP increases the IoU and precision of the network by 0.39% and 0.25%, respectively. At the same time, compared with ASPP module, DSASPP module can slightly reduce the number of network parameters and computational complexity. This shows that depthwise separable convolution can achieve better performance while reducing the number of model parameters and computation. Adding the SE attention module after feature extraction of the backbone network and before the feature fusion improves the performance of building extraction and increased the IoU by 0.58%. Using LPA module instead of two sets of 3×3 standard

TABLE I
INFLUENCE OF DIFFERENT MODULES ON BUILDING EXTRACTION PERFORMANCE ON THE WHU AERIAL IMAGE DATASET

Method	IoU (%)	Precision (%)	Recall (%)	F-score (%)	FLOPs(G)	Parameters(M)
Baseline	87.63	92.96	93.61	92.91	31.62	11.65
Baseline+D	88.02	93.21	**93.85**	93.23	28.04	8.13
Baseline+S	88.21	93.57	93.66	93.36	31.63	11.66
Baseline+L	88.52	94.01	93.47	93.55	16.94	10.76
LRAD-Net	**88.89**	**94.21**	93.74	**93.86**	**13.36**	**7.30**

TABLE II
EFFECT OF SAMPLING RATES ON SEGMENTATION PERFORMANCE OF THE NETWORK

sampling rate	IoU (%)	Precision (%)	Recall (%)	F-score (%)	FLOPs(G)	Parameters(M)
[6,12,18]	88.58	93.71	93.45	93.59	**13.13**	**7.11**
[6,12,18,24]	88.39	94.05	93.26	93.51	13.37	7.30
[6,10,14,18]	**88.89**	**94.21**	**93.74**	**93.86**	13.36	7.30

TABLE III
QUANTITATIVE COMPARISON OF THE PROPOSED LRAD-NET WITH VARIOUS MODELS ON THE WHU AERIAL IMAGE DATASET

Method	IoU (%)	Precision (%)	Recall (%)	F-score (%)	FLOPs(G)	Parameters(M)
Deeplabv3+(Res50)	88.10	94.02	93.66	93.61	69.44	40.42
Deeplabv3+(MobileNetv3)	86.35	93.14	92.08	92.53	30.48	11.79
PANet(Res50)	86.86	93.58	92.98	91.17	34.84	24.26
U-Net(Res50)	87.52	93.25	93.37	93.25	42.75	32.52
PSPNet(Res50)	85.60	92.07	92.22	91.68	22.12	29.81
BiSeNet	86.52	93.07	92.11	92.35	15.22	13.42
BiSeNetV2	86.62	93.45	91.92	92.47	17.74	**5.19**
LRAD-Net	**88.89**	**94.21**	**93.74**	**93.86**	**13.36**	7.30

convolution increases IoU by nearly 1% and reduces the amount of computation by nearly 1 time. Compared with the baseline, LRAD-Net in IoU and precision index increased by 1.26% and 1.25%, respectively, which indicates that the model can achieve better performance.

In order to explore the influence of sampling rate in DSASPP on model performance, three sets of comparison experiments were designed based on LRAD-Net, and the sampling rate of atrous convolution in DSASPP structure was set as {612,18}, {612,18,24}, {610,14,18}, and test results are given in Table II. As can be seen from Table II, LRAD-Net with the sampling rate of {610,14,18} has a higher score overall than the sampling rate of {612,18,24}, which indicates that compared with DSASPP module with a large sampling rate, appropriately reducing the sampling rate can make the features extracted by atrous convolution better fit the character of the buildings in remote sensing images. Compared with the LRAD-Net with a sampling rate of {612,18}, the combination of {610,14,18} can improve the precision by 0.5% and the IoU by 0.32% on the basis of only increasing the number of parameters by 0.19M. Therefore, LRAD-Net selected the atrous rate combination of {610,14,18} to ensure the optimal performance of the model.

B. LRAD-Net Network Performance Experiment

To verify the segmentation performance of LRAD-Net, we conduct experiments on the four datasets mentioned in Section III. Here, four common classical networks (U-Net, Deeplabv3+ and PANet [53], PSPNet [54]) and two recently proposed lightweight networks (BiseNet [55] and BisenetV2 [56]) are used for comparison. In the classic network, we use ResNet50 as the backbone network of U-Net, Deeplabv3+, PANet, and PSPNet. At the same time, the lightweight network MoblieNetv3 [57] is used as the encoder for the experiment. Experiments show that LRAD-Net can achieve high accuracy with less parameters. The experimental results are shown in Tables III–VI.

As given in Table III, on the WHU aerial image dataset, although the IoU of Deeplabv3+(Res50) is higher than U-Net (Res50), PANet(Res50) and PSPNet(Res50) with 0.58%, 1.24%, and 2.50%, respectively, it is 0.79% lower than LRAD-Net. The precision and F-score of LRAD-Net are also significantly higher than these compared networks. Compared with the lightweight networks BiseNet and BiSeNetV2, LRAD-Net is 2.37% and 2.27% higher in IoU and 1.14% and 0.76% higher in precision. It can also be seen from Table III that Parameters of

TABLE VI
QUANTITATIVE COMPARISON OF THE PROPOSED LRAD-NET WITH VARIOUS MODELS ON SELF-ANNOTATED UAV DATASET

Method	IoU (%)	Precision (%)	Recall (%)	F-score (%)	FLOPs(G)	Parameters(M)
Deeplabv3+(Res50)	80.44	89.24	89.33	88.12	69.44	40.42
Deeplabv3+(MobileNetv3)	80.06	88.58	89.50	87.97	30.48	11.79
PANet(Res50)	79.38	**91.37**	89.78	87.48	34.84	24.26
U-Net(Res50)	80.08	88.02	**92.64**	88.40	42.75	32.52
PSPNet (Res50)	72.69	83.22	85.24	83.01	22.12	29.81
BiSeNet	77.77	86.55	88.70	86.55	15.22	13.42
BiSeNetV2	77.44	86.15	88.53	85.71	17.74	**5.19**
LRAD-Net	**81.76**	89.29	90.81	**89.22**	**13.36**	7.30

TABLE IV
QUANTITATIVE COMPARISON OF THE PROPOSED LRAD-NET WITH VARIOUS MODELS ON THE WHU SATELLITE IMAGE DATASET

Method	IoU (%)	Precision (%)	Recall (%)	F-score (%)	FLOPs(G)	Parameters(M)
DeepLabv3+(Res50)	68.79	83.43	79.77	81.31	69.44	40.42
DeepLabv3+(MobileNetv3)	69.95	83.08	81.51	82.20	30.48	11.79
PANet(Res50)	67.42	81.98	80.64	80.21	34.84	24.26
U-Net(Res50)	69.26	83.04	81.81	81.46	42.75	32.52
PSPNet(Res50)	68.55	79.31	82.62	81.22	22.12	29.81
BiSeNet	71.11	83.59	83.06	83.24	15.22	13.42
BiSeNetV2	70.70	**83.62**	82.62	83.06	17.74	**5.19**
LRAD-Net	**71.22**	82.67	**83.67**	**83.39**	**13.36**	7.30

TABLE V
QUANTITATIVE COMPARISON OF THE PROPOSED LRAD-NET WITH VARIOUS MODELS ON THE INRIA AERIAL IMAGE DATASET

Method	IoU (%)	Precision (%)	Recall (%)	F-score (%)	FLOPs(G)	Parameters(M)
Deeplabv3+(Res50)	79.19	89.33	88.67	87.97	69.44	40.42
Deeplabv3+(MobileNetv3)	79.02	88.95	87.39	87.86	30.48	11.79
PANet(Res50)	78.67	**89.41**	88.59	87.65	34.84	24.26
U-Net(Res50)	78.76	88.21	**89.43**	87.71	42.75	32.52
PSPNet(ResNet50)	75.69	86.82	85.52	85.73	22.12	29.81
BiSeNet	77.40	86.03	88.14	86.89	15.22	13.42
BiSeNetV2	76.68	85.58	87.74	86.33	17.74	**5.19**
LRAD-Net	**79.82**	89.39	88.79	**88.37**	**13.36**	7.30

LRAD-Net is only 7.30M, compared with U-Net, Deeplabv3+, PANet, and PSPNet, the parameters and Flops of LRAD-Net are 3×4 times lower. This shows that LRAD-Net not only has a significant improvement in model performance, but also has a significant advantage in the number of parameters and FlOPs. Compared with BiSeNet and BiSeNetV2 network, the number of parameters of LRAD-Net has no significant difference, but its precision index such as IoU is significantly higher. According to the comprehensive accuracy index and application index, LRAD-Net outperforms the above seven networks in building extraction.

It can be seen from Tables IV and V, in the WHU satellite Image dataset and the Inria aerial image dataset, IoU of all networks are generally not high. In the WHU satellite dataset, lightweight models, such as Deeplabv3+(MobileNetv3) and BiSeNet have better extraction effect than larger models, such as Deeplabv3+(Res50) and U-Net(Res50). This is because WHU satellite dataset has a small amount of data and sparse distribution of buildings and there are obvious mismarks and omissions, which affect the performance of the deep network model. Compared with Deeplabv3+(MobileNetv3) and BiSenetV2, the IoU of LRAD-Net increased by 1.27% and 0.52%, respectively.

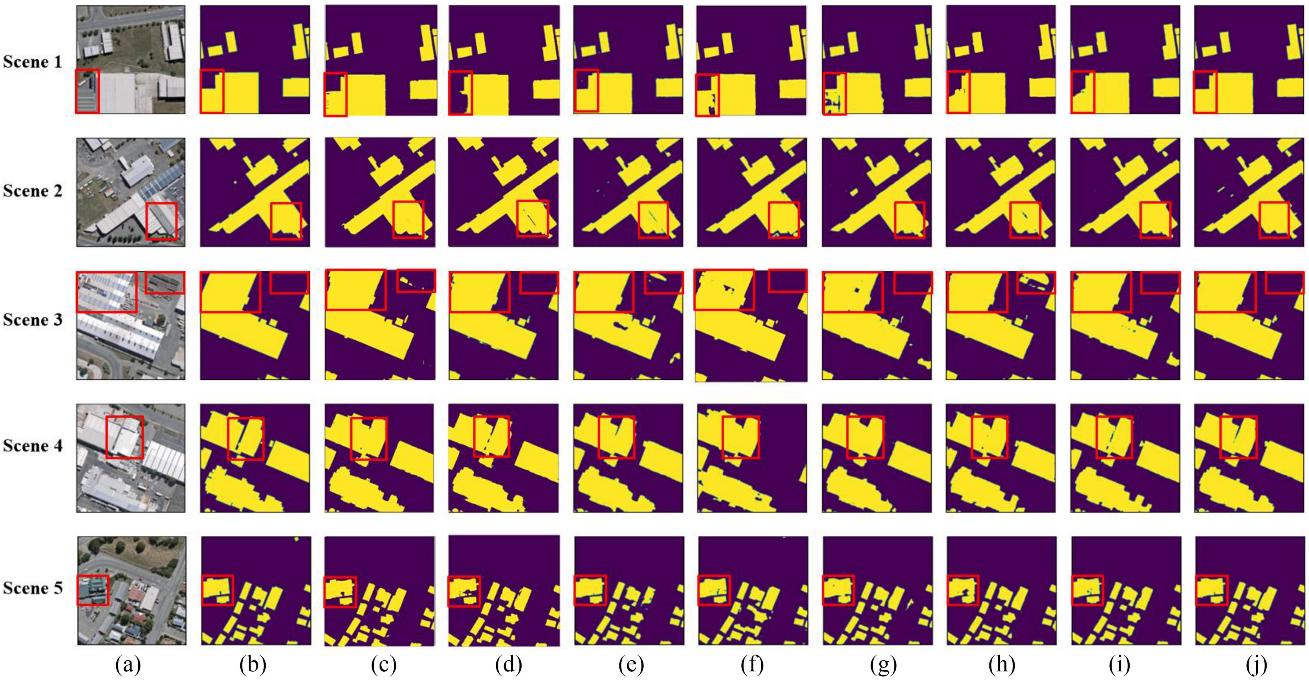

Fig. 8. Examples of the building extraction results by our proposed LRAD-Net, and various models on the WHU aerial image dataset. (a) Image. (b) Label. (c) Deeplabv3+(Res50). (d) Deeplabv3+(Mbv3). (e) PANet(Res50). (f) U-Net(Res50). (g) PSPNet(R-es50). (h) BiSeNet. (i) BiSeNetV2. (j) LRAD-Net.

This is because LRAD-Net integrates the attention mechanism in the feature fusion stage, so that the network can pay more attention to the architectural information in the image, and DSASPP module and LRA module can extract more effective and accurate feature information.

As the spatial resolution of the datasets used above are 0.3 m, some details of the buildings with small size may be ignored, the advantages of our proposed network are difficult to highlight. In order to validate the robustness of LRAD-Net, we do more experiments with a self-annotated building dataset with spatial resolution 0.1 m. Table VI gives the quantitative comparison of the proposed LRAD-Net with other models on the self-annotated UAV dataset. It is clear that LRAD-Net has higher IoU than BiSeNet and BiSeNetV2 with 3.99% and 4.32%, respectively. Compared with U-Net(Res50), PANet(Res50), Deeplabv3+(Res50), and PSPNet(Res50), it has a very small number of parameters.

It can be seen from Tables III–VI that the performance of different networks on different datasets is different. In general, complex networks, such as Deeplabv3+(Res50) and U-Net(Res50) perform better than lightweight networks, such as Bisenet on datasets with large amount of data and good quality, while WHU satellite image dataset with small amount of data and inaccurate image labels perform poorly. Compared with these networks, LRAD-Net can achieve the best performance on different data sets, which fully demonstrates the superiority of LRAD-Net network.

In order to better observe the specific performance of the n-etwork in building segmentation, Figs. 8–11 show the specific effect of LRAD-Net and other networks in extracting buildings on four datasets. The letter numbers in the figure r-efer to (a) input picture, (b) image label, (c) Deeplabv3+(ResNet50), (d) Deeplabv3+(Mobilenetv3), (e) PANet(ResNet50), (f) U-Net(ResNet50), (g) PSPNet(ResNet50), (H) BiSeNet, (I) BiSeN-etV2, and (j) LRAD-Net.

It can be seen from scene 1 of Fig. 8 that Deeplabv3+(Res50), U-Net(Res50) and LRAD-Net can completely extract large buildings in WHU aerial image dataset, while Deeplabv3+(MobileNetv3) and PSPNet(Res50) have obvious leakage phenomenon. From the marks in scene 2, we can see that for irregular buildings, LRAD-Net can accurately extract the outline of the building, and there is no obvious void phenomenon. As can be seen from scenes 3, 4, and 5, for areas with dense buildings, LRAD-Net can accurately extract the boundaries of adjacent buildings, with relatively few misclassification phenomena.

According to scenes 1 and 4 in Fig. 9, when extracting a single regular building from the WHU satellite image dataset, U-Net(MobileNetv3), BiSeNet and LRAD-Net have good performance, while the other five networks have poor effects. PANet(Res50), PSPNet(Res50), and BiSeNetV2 have serious missing scores and misclassification phenomenon, this is due to the simple network structure of PANet(Res50), PSPNet(Res50), and BiSeNetV2 cannot fully extract the deep features of buildings. As can be seen from the red areas in scenes 2 and 3, there are close adjacent buildings in the image, and there is shadow occlusion, which leads to the network misjudging the boundary of the building, resulting in poor segmentation effect. However, this phenomenon is not obvious in LRAD-Net.

In the Fig. 10, LRAD-Net can also perform better segmentation effect in the face of scenes with no obvious color difference. It can be seen from scene 1 in Fig. 10 that LRAD-Net can effectively extract irregular buildings and reduce the occurrence of misclassification and voids.

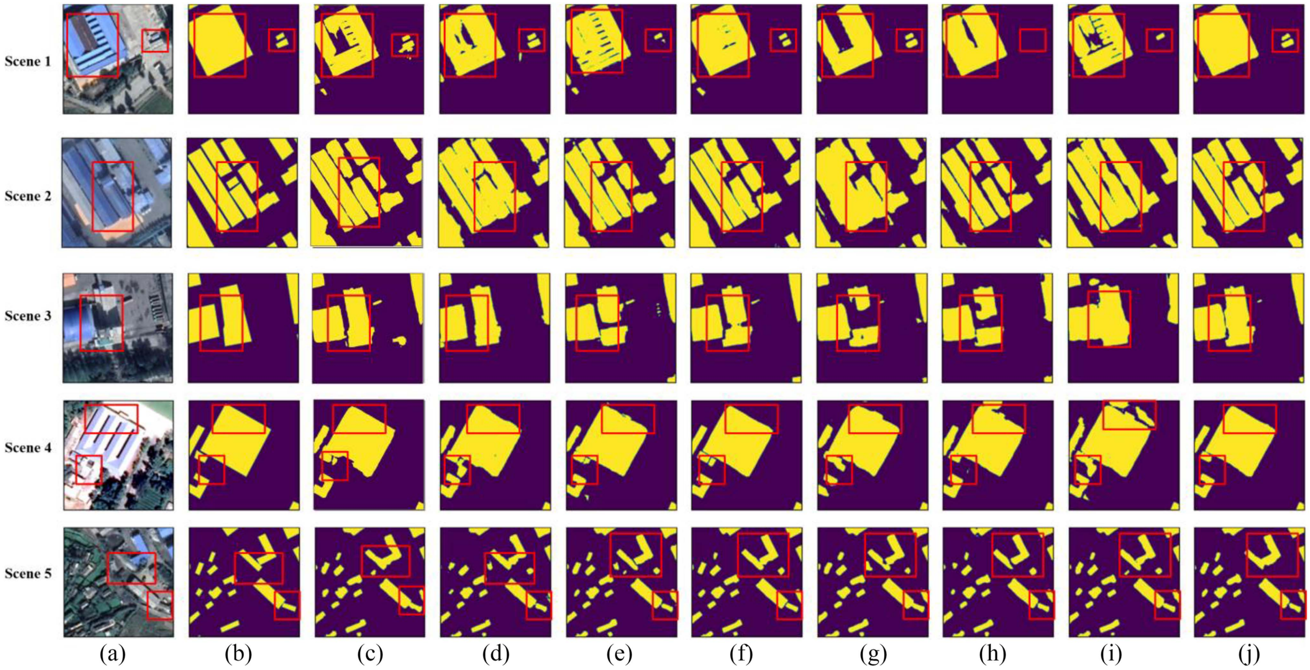

Fig. 9. Examples of the building extraction results by our proposed LRAD-Net, and various models on the WHU satellite image dataset. (a) Image. (b) Label. (c) Deeplabv3+(Res50). (d) Deeplabv3+(Mbv3). (e) PANet(Res50). (f) U-Net(Res50). (g) PSPNe(R-es50). (h) BiSeNet. (i)BiSeNetV2. (j) LRAD-Net.

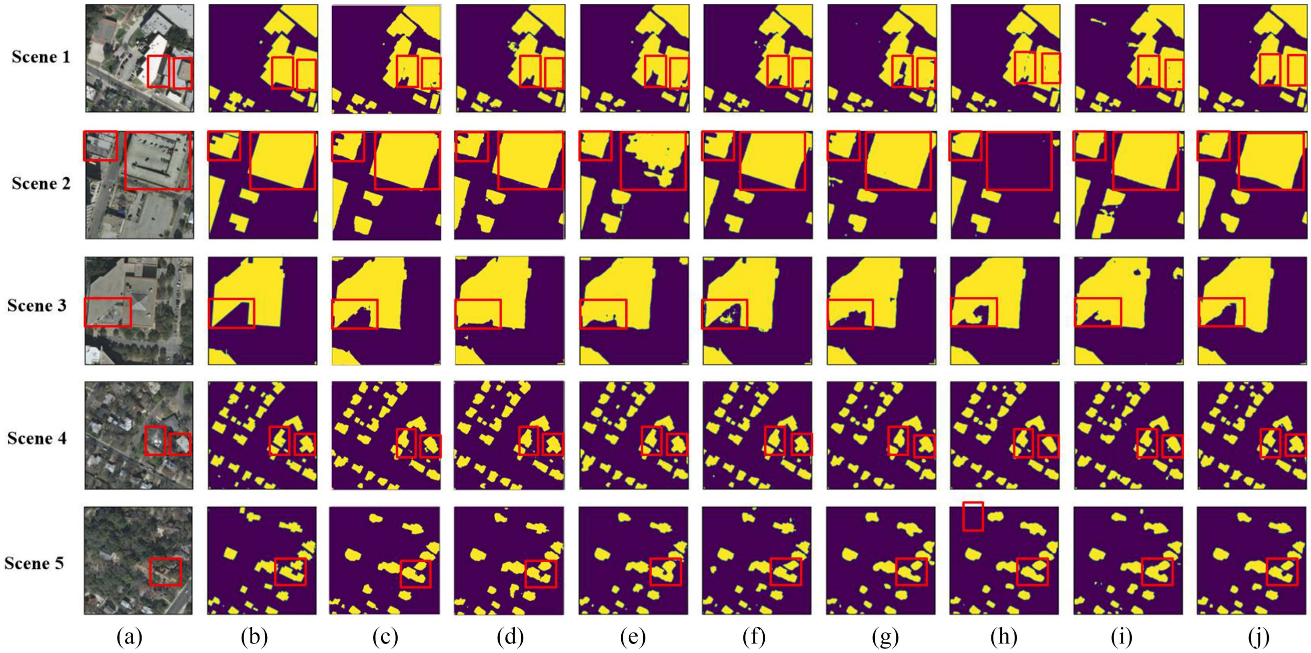

Fig. 10. Examples of the building extraction results by our proposed LRAD-Net, and various models on the Inria aerial image dataset. (a) Image. (b) Label. (c) Deeplabv3+(Res50). (d) Deeplabv3+(Mbv3). (e) PANet(Res50). (f) U-Net(Res50). (g) PS-PNet(Res50). (h) BiSeNet. (i) BiSeNetV2. (j) LRAD-Net.

We can see in scenario 1 in Fig. 11, on the self-annotated UAV dataset with higher a spatial resolution, PANet(Res50), PSPNet(Res50), and BiSeNetV2 networks misclassified a large area of ground into buildings, which is caused by the high spatial resolution of images and the similarity of spectral information between buildings and the ground. LRAD-Net can extract buildings accurately without obvious misclassification. It can also be seen from scenes 23 and 4 in Fig. 11 that LRAD-Net can also extract buildings well on images with different luminance.

In *general*, compared with other networks, LRAD-Net has obvious advantages in extracting medium and large buildings, and also has certain advantages in identifying small building groups.

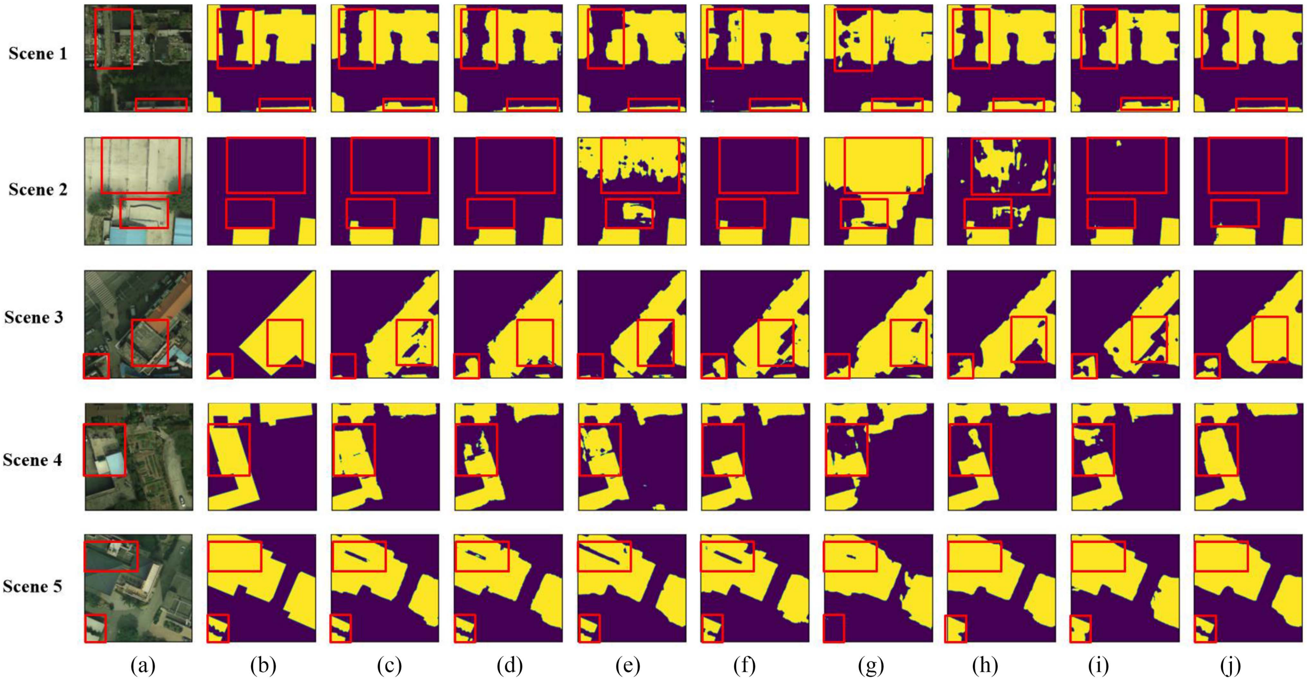

Fig. 11. Examples of the building extraction results by our proposed LRAD-Net, and various models on self-annotated UAV dataset. (a) Image. (b) Label. (c) Deeplabv3+(Res50). (d) Deeplabv3+(Mbv3). (e) PANet(Res50). (f) U-Net(Res50). (g) PSPNet(Re-s50). (h) BiSeNet. (i) BiSeNetV2. (j) LRAD-Net.

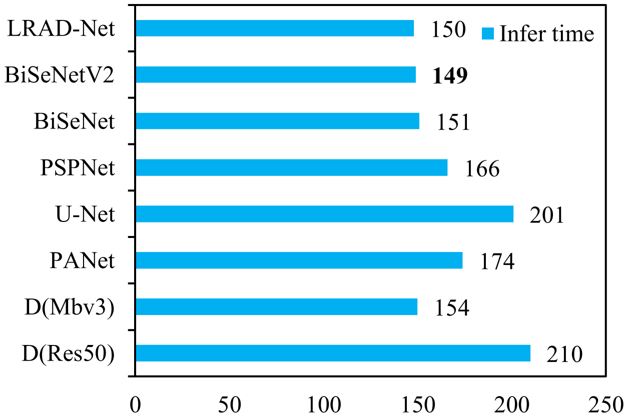

Fig. 12. Comparison of networks prediction speed.

C. LRAD-Net Network Time Efficiency Experiment

Due to the infer speed of a model on a specific hardware, it is affected by several factors, such as hardware features, software implementation, and system environment, in addition to the parameters and FLOPs. Therefore, to verify the inference speed of LRAD-Net, we have tested the speed of LRAD-Net against seven other networks to predict a single three-channel image (512×512) on an NVIDIA GeoForce RTX 3080. The result is shown in Fig. 12.

As can be seen from Fig. 12, the speed of BiSeNet, BiSeNetV2, and LRAD-Net is almost the same. Compared with U-Net(Res50), Deeplabv3+(Res50), and other networks, they have obvious advantages in speed, which is of great help to practical applications. Combined with the accuracy of the four datasets, LRAD-Net can achieve a good balance between accuracy and speed.

V. Conclusion

Based on encoding/decoding structure, this article proposes a lightweight building segmentation network LRAD-Net. LRAD-Net makes best use of the advantages of Reg600 network, SE module, and proposed DSASPP and LPA module, DSASPP expands the sensitivity field through parallel sampling with depthwise separable atrous convolution of multistage sampling rate, enrich semantic information, and avoids the problem of segmentation errors caused by falling into local features. SE attention module can improve the weight of building features and reduce the interference caused by noise information during feature fusion of LRAD-Net. The LPA module in decoder can extract more building information with fewer parameters and improve the accuracy of building extraction. Compared with the commonly used semantic segmentation network, LRAD-Net can achieve a balance between precision and speed, and has better help for practical applications. Although LRAD-Net performs better than common semantic segmentation networks in reasoning speed, it is still far from ideal, how to further improve its reasoning speed will be our future work.

References

[1] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[2] G. Peng, W. Jie, L. Yu, Y. Zhao, and J. Chen, "Finer resolution observation and monitoring of global land cover: First mapping results with landsat tm and etm+ data," *Int. J. Remote Sens.*, vol. 34, pp. 2607–2654, 2013.

[3] J. Harken and R. Sugumaran, "Classification of Iowa wetlands using an airborne hyperspectral image: A comparison of the spectral angle mapper classifier and an object-oriented approach," *Can. J. Remote Sens.*, vol. 31, pp. 167–174, 2005.

[4] S. R. Kim, W. K. Lee, D. A. Kwak, G. S. Biging, and H. K. Cho, "Forest cover classification by optimal segmentation of high resolution satellite imagery," *Sensors*, vol. 11, pp. 1943–1958, 2011.

[5] J. Du et al., "A novel framework for 2.5-D building contouring from large-scale residential scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4121–4145, Jun. 2019.

[6] D. Chaudhuri, N. K. Kushwaha, A. Samal, and R. C. Agarwal, "Automatic building detection from high-resolution satellite images based on morphology and internal gray variance," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, pp. 1767–1779, 2016.

[7] N. L. Gavankar and S. K. Ghosh, "Automatic building footprint extraction from high-resolution satellite image using mathematical morphology," *Eur. J. Remote Sens.*, vol. 51, pp. 182–193, 2018.

[8] N. Shrivastava and P. K. Rai, "Remote-sensing the urban area: Automatic building extraction based on multiresolution segmentation and classification," *Fac. Social Sci. Humanities*, pp. 1–16, 2015.

[9] C. R. Jung and R. Schramm, "Rectangle detection based on a windowed Hough transform," in *Proc. 17th Braz. Symp. Comput. Graph. Image Process.*, 2004, pp. 13–120.

[10] W. Cui and Y. Zhang, "An effective graph-based hierarchy image segmentation," *Intell. Automat. Soft Comput.*, vol. 17, pp. 969–981, 2011.

[11] H. Sportouche, F. Tupin, and L. Denise, "Building extraction and 3D reconstruction in urban areas from high-resolution optical and SAR imagery," in *Proc. Joint Urban Remote Sens. Event*, 2009, pp. 1–11.

[12] S. Fasahat, S. Teng, A. Mohammad, and G. Lu, "A robust gradient based method for building extraction from lidar and photogrammetric imagery," *Sensors*, vol. 16, 2016, Art. no. 1110.

[13] M. Teimouri, M. Mokhtarzade, and M. J. Valadan Zoej, "Optimal fusion of optical and sar high-resolution images for semiautomatic building detection," *Mapping Sci. Remote Sens.*, vol. 53, pp. 45–62, 2016.

[14] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image," *J. Multimedia*, vol. 9, pp. 181–188, 2014.

[15] L. A. Yan, Z. B. Lin, B. Kt, B. Hs, and A. Ml, "Morphological house extraction from digital surface model and imagery of high-density residential areas," *Eng. Remote Sens. J. Amer. Soc. Photogramm.*, vol. 82, pp. 21–29, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Assoc. Comput. Mach.*, vol. 60, pp. 84–90, 2017.

[17] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 640–651, 2015.

[19] X. Bian, S. N. Lim, and N. Zhou, "Multiscale fully convolutional network with application to industrial inspection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–8.

[20] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 621–626.

[21] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4724–4732.

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

[23] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning For Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11.

[25] H. Huang, L. Lin, R. Tong, H. Hu, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1055–1059.

[26] J. Hui, M. Du, X. Ye, Q. Qin, and J. Sui, "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 786–790, May 2019, doi: 10.1109/LGRS.2018.2880986.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *Proc. Int. Conf. Learn. Representations*, May 7–9, 2015.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1706.05587

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[31] J. Jin., A. Dundar, and E. Culurciello, "Flattened convolutional neural networks for feedforward acceleration," 2014, *arXiv:1412.5474*.

[32] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[33] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.

[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[35] Y. Zhang, W. Gong, J. Sun, and W. Li, "Dense-HSGP: Dense Gaussian-based context pooling for very high-resolution building extraction," in *Image Signal Processing Remote Sensing XXV*. Bellingham, WA, USA: SPIE, 2019.

[36] M. Chen, J. Wu, L. Liu, W. Zhao, and R. Du, "DR-net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, 2021, Art. no. 294.

[37] M. Guo, H. Liu, Y. Xu, and Y. Huang, "Building extraction based on U-Net with an attention block and multiple losses," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1400.

[38] Y. Zhu, Z. Liang, J. Yan, G. Chen, and X. Wang, "E-D-Net: Automatic building extraction from high-resolution aerial images with boundary information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4595–4606, 2021, doi: 10.1109/JSTARS.2021.3073994.

[39] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[40] Z. Ye, Y. Fu, M. Gan, J. Deng, A. Comber, and K. Wang, "Building Extractio-n from very high resolution aerial imagery using joint attention deep neu-ral network," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2970.

[41] Y. Liu et al., "ARC-Net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020, doi: 10.1109/ACCESS.2020.3015701.

[42] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1768.

[43] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614812, doi: 10.1109/TGRS.2021.3131331.

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

[45] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10428–10436.

[46] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[47] M. Tan et al., "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2820–2828.

[48] T. Elsken, J. H. M. Thomas, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2007, 2019.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, pp. 1904–1916, Sep. 2015.

[50] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," 2021, *arXiv:2107.00782*.

[51] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[52] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.

[53] W. Guangming et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, 2018, Art. no. 407.

[54] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017; pp. 3226–3229.

[55] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:180510180*.

[56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 6230–6239.

[57] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, *Bisenet: Bilateral Segmentation Network for Real-Time Semantic Segmentation*. Cham, Switzerland: Springer, 2018.

[58] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.

[59] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.

Jiabin Liu received the B.S. degree in remote sensing science and technology from Chang'an University, Xian, China, in 2020. He is currently working toward the M.S. degree in surveying and mapping engineering with Guangzhou University, Guangzhou, China.

His major research interests include remote sensing image processing and computer vision.

Huaigang Huang received the B.S. degree in civil engineering from Wuyi University, Jiangmen, China, in 2016, the M.S. degree in building and civil engineering from Guangzhou University, Guangzhou, China, in 2021. He is currently working toward the Ph.D. degree in geomatics engineering from the University of Calgary, Calgary, Canada.

His major research interests include remote sensing image processing and computer vision.

Hanxiao Sun received the B.S. degree in software engineering from Qingdao University of Science and Technology, Qingdao, China, in 2020. She is currently working toward the M.S. degree in surveying and mapping engineering with Guangzhou University, Guangzhou, China.

Her major research interests are remote sensing image processing and multi-source data fusion.

Zhifeng Wu (Member, IEEE) received the B.S. degree in geography education from Hunan Normal University, Changsha, China, in 1992, the M.S. degree in physical geography from South China Normal University, Guangzhou, China, in 1995, and the Ph.D. degree in GIS from the Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences, Beijing, China, in 2002.

He is currently a Professor with the School of Geographical Sciences and Remote Sensing, Guangzhou University, Guangzhou. His research interests include urban remote sensing, land ecological remote sensing, spatiotemporal data analysis, and monitoring and assessment of natural resources.

Renbo Luo (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control theory and engineering from the South China University of Technology, Guangzhou, China, in 2010 and 2017, respectively, and the Ph.D. degree in computer science engineering from Ghent University, Ghent, Belgium, in 2017.

He is currently a Lecturer with the School of Geographical Sciences and Remote Sensing, Guangzhou University, Guangzhou. His research interests include image processing, object recognition, data fusion, and urban remote sensing.