Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction

Wenjing Deng, Qian Shi, *Senior Member, IEEE*, and Jun Li, *Fellow, IEEE*

***Abstract*—Rapidly developing remote sensing technology provides massive data for urban planning, mapping, and disaster management. As a carrier of human productive activities, buildings are essential to both urban dynamic monitoring and suburban construction inspection. Fully-convolutional-network-based methods have provided a paradigm for automatically extracting buildings from high-resolution imagery. However, high intraclass variance and complexity are two problems in building extraction. It is hard to identify different scales of buildings by using a single receptive field. For this purpose, in this article, we use the stable encoder– decoder architecture, combined with a grid-based attention gate and *atrous* spatial pyramid pooling module, to capture and restore features progressively and effectively. A modified ResNet50 encoder is also applied to extract features. The proposed method could learn gated features and distinguish buildings from complex surroundings such as trees. We evaluate our model on two building datasets, WHU aerial building dataset and our DB UAV rural building dataset. Experiments show that our model outperforms other five most recent models. The results also exhibit great potential for extracting buildings with different scales and validate the effectiveness of deep learning in practical scenarios.**

***Index Terms*—Attention gate (AG), building extraction, deep learning, fully convolutional networks (FCNs), semantic segmentation.**

I. Introduction

Since 1980s, building extraction has been an essential research goal in remote sensing [1]. Especially with the acceleration of urbanization and the need of urban planning, a precise and immediate extraction of buildings becomes critical.

Conventional methods often focus on designing distinguishable features and using simple classifiers such as the Bayesian classifier to extract buildings [2]. The selected features contain spectral and spatial contextual characteristics such as shape, edge, and height. These features were often used to better recognize buildings. For example, an integrated strategy including structural, contextual, and spectral information was used for identifying buildings in [3]. Aytekin *et al.* proposed an automatic and unsupervised method based on morphological and length parameters to detect buildings in complex urban environments [4]. Considering the properties of the building itself, some shadow-based methods were proposed to help extract buildings [5]–[7]. A morphological building/shadow index was also developed to detect buildings in high-resolution imagery that was widely used in later researches [8]. Ok *et al.* proposed a new fuzzy landscape generation approach to model the directional spatial relationship between buildings and their shadows [9]. In general, these traditional methods based on handcraft features often need prior knowledge and could only solve specific tasks that cannot be widely applied in automatic identification of buildings.

Nowadays, convolutional neural networks (CNNs) are heavily used in image recognition tasks such as image classification, object detection, and semantic segmentation [10]–[12]. Meanwhile, huge progress has been done in remote sensing image processing, including building extraction and road detection. The development in deep learning has created a new shift in learning features automatically. The classic CNN models, such as AlexNet, VGG, GoogLeNet, ResNet, and DenseNet [13]–[17], have achieved enormous success in classification tasks. In 2014, the emergence of fully convolutional network (FCN) has boosted CNN architectures for dense predictions without any fully connected layers [18], thus providing semantic image segmentation a paradigm for all the subsequent state-of-the-art approaches, such as U-Net [19], SegNet [20], a series of Deeplab networks [21]–[24], GCN [25], and DFN [26]. Accordingly, these FCN-based models and their variants are widely used in remote sensing tasks [27]–[30].

With regard to semantic segmentation in building footprints extraction, a lot of FCN-based methods have been proposed [31]–[34]. Li *et al.* evaluated the FCN on building extraction and compared it with some conventional methods, which showed the validity of the FCN [35]. In the study of building extraction, many methods also have been designed [36]–[38]. Xu *et al.* proposed a method termed as “ResUNet” that combined U-Net and ResNet, together with guided filters to extract buildings [39]. Huang *et al.* developed an end-to-end trainable gated residual refinement network that fused high-resolution aerial images and LiDAR point clouds [40]. To solve the problem of multiscale information capture, Liu *et al.* designed a spatial residual inception module to preserve details and used large kernels to capture the context information [41]. Compared to the original U-Net, Ji *et al.*

Manuscript received October 16, 2020; revised January 5, 2021; accepted January 30, 2021. Date of publication February 9, 2021; date of current version March 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 42030111, in part by the National Natural Science Foundation of China under Grant 61771496, in part by Guangdong Natural Science Foundation under Grant 2019A1515011057, and in part by the National Natural Science Foundation of China under Grant 61976234. *(Corresponding author: Qian Shi.)*

The authors are with the Guangdong Provincial Key Laboratory of Urbanization and Geosimulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: dengwj33@mail2.sysu.edu.cn; shixi5@mail.sysu.edu.cn; lijun48@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3058097

proposed a Siamese U-Net sharing weight in two branches and improved the segmentation accuracy [42]. Feng *et al.* presented an enhanced deep convolutional network with a postprocessing method through the superpixel-based conditional random fields in building extraction [43]. Besides, the residual refinement module was used in [44] to serve as a postprocessing part in extracting buildings. In [45], a multiscale aggregation strategy in the prediction level and the two postprocessing methods are introduced to refine the segmentation maps. Despite that these recent FCN-based methods demonstrate obvious advantages over the existing aerial building datasets, two aspects in building extraction still exist. The first one is the high intraclass variance of buildings and the low interclass difference between buildings and other nonbuilding objects. The other one is the scale invariance of buildings under many complex scenarios.

In this work, we propose an encoder–decoder architecture for automatical building extraction through a modified ResNet-50 architecture with an *atrous* spatial pyramid pooling (ASPP) module and attention gate (AG) mechanism. On the one hand, the grid-based AG is a gating signal conditioned to buildings' spatial information without adding a large number of additional parameters, while the ASPP module contribute to the integrity of large-scale buildings. The proposed method that make use of these two modules not only can catch the large-scale features but also pay more attention to small buildings to a certain extent as well. On the other hand, the integration of the two modules shows effectiveness on different and complex environments in the task of building extraction.

The remainder of this article is organized as follows. Section II demonstrates the proposed network architecture. Section III presents four experiments and the corresponding analysis on two datasets with other FCN models. Finally, Section IV concludes this article with some remarks and hints at plausible future research lines.

II. Network

A. Encoder–Decoder Architecture

The encoder–decoder architecture is a simple yet effective structure in semantic segmentation. The encoder part gradually reduces the spatial dimension through the pooling layer, obtaining shallow to deep features. The decoder gradually recovers the object details and spatial dimension through deconvolution or interpolation. It also ensures that the output has the same dimension as the input image to achieve end-to-end training. There are usually shortcut connections from the encoder to the decoder to help decoder recover the object details better. FCN, U-Net, and SegNet are popular methods within this class.

However, the simple form of the encoder–decoder structure often loses the global information because of directly fusing the deepest feature without any other possible operation. This leads to misclassified pixels within the buildings of large scale when it comes to the dense classifications of high-resolution aerial images. Thus, it is necessary to design a network, which can both utilize the stability of the encoder–decoder architecture and avoid missing global features.

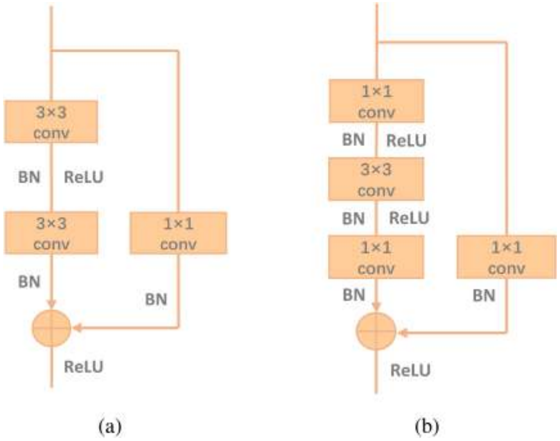

Fig. 1. (a) Basic building block. (b) Bottleneck building block.

B. Residual Block

First proposed for solving gradient disappearance and explosion problems, residual networks have been increasingly applied to semantic segmentation. In [16], it is shown that the accuracy increases with the depth, overcoming the optimization difficulties. Besides, it was demonstrated that in [15] and [16], compared with the plain net, the residual net converge faster under the same layers. In this work, we chose ResNet-50 as our backbone to make the network converge faster as well as learn deep features.

The residual blocks consist of two paths, including the residual part and the identity shortcut, with the ability to smooth the backward and forward flow of information. The residual part is often stacked by several convolution blocks and the identity shortcut is a way of matching dimensions that usually include a 1×1 convolution operation. Typically, as shown in Fig. 1, there are two types of residual block, i.e., the basic building block and the bottleneck building block. The main difference of these two designs lies in the residual part, two layers for the "basic" and three for the "bottleneck." In the bottleneck building block, two 1×1 convolutions are responsible for first reducing and finally increasing the dimensions, leaving a 3×3 convolution as a bottleneck with a relatively small dimension. This helps the whole architecture to be more efficient. For this reason, we use a modified ResNet-50 architecture as our encoder to learn more feature representations. In our encoder part, the whole layers in ResNet-50 are employed except for the first max pooling layer to improve the resolution of feature maps. In the meantime, the convolutional operation follows a batch normalization (BN) [46] layer and a rectified linear unit (ReLU) [13] activation layer. The whole residual block simplifies the learning process and enhances gradient propagation.

C. Atrous Spatial Pyramid Pooling (ASPP)

Originally proposed in [47], spatial pyramid pooling is a way of combining multilevel features. As the dilated convolution has the ability to increase the receptive field without pooling that may lose location information [23], by adding dilated convolution on the original spatial pyramid pooling block, ASPP can capture the

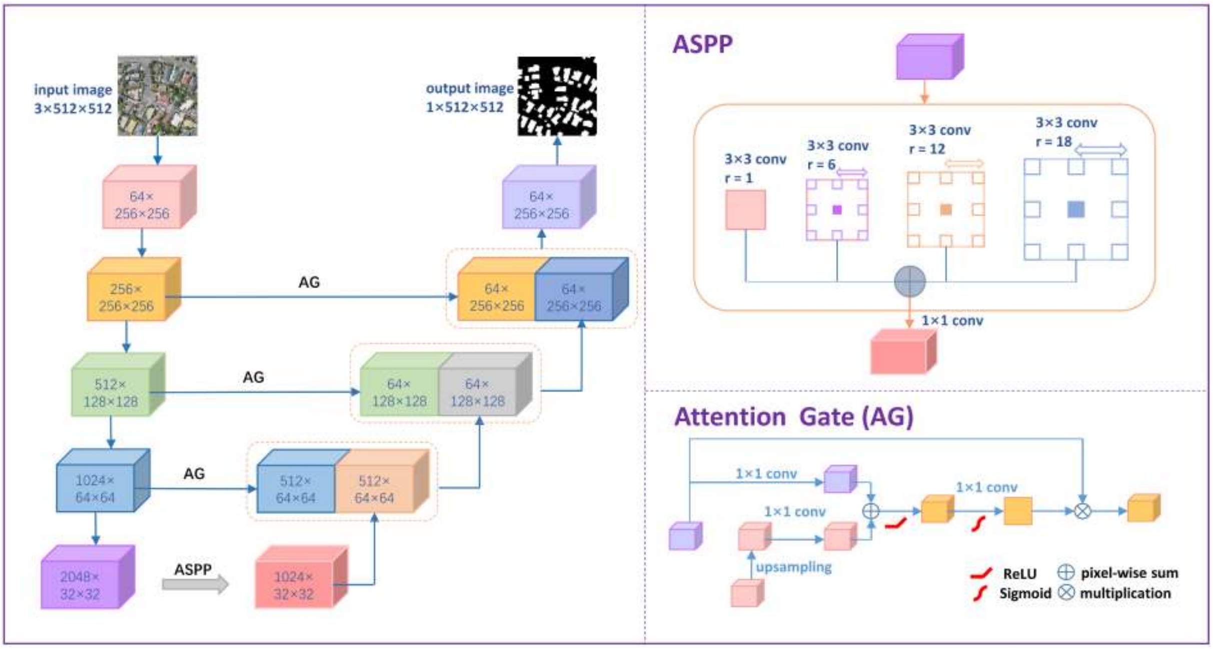

Fig. 2. Overview of the proposed network architecture. The encoder of the network is a modified ResNet-50 design without a first pooling layer. The input image is first downsampled by a 7 × 7 kernel with a stride of two and the following four layers are composed of three, four, six, and seven residual blocks, respectively. Then, a convolution operation is used to reduce the dimensions. As a center bridge, an ASPP module with four different *atrous* convolutions is designed to enlarge the view of field to get a broader context information. During the other skip connection layers, a grid-based AG mechanism is used between the encoder and decoder, which filters the irrelevant background information. Schematics of the AG and the ASPP are shown on the right side of the chart.

context of a center pixel at multiple scales [22]. After multiple parallel dilated convolutions with different dilation rates are computed, a sum operation follows to fuse all the feature maps. As illustrated in Fig. 2, there are four parallel paths with different dilation rates of 1, 6, 12, and 18, respectively. Although dilated convolution can capture broader information, it also brings the problem of much memory consumption in shallow layers. In this case, the ASPP block is used in our network between the encoder part and the decoder part when the resolution of feature maps gets relatively small.

D. Attention Gate

The basic idea of the attention mechanism in computer vision is to allow the system to be able to ignore irrelevant information and focus on the key parts. This could be divided into hard attention and soft attention. For hard attention, the value of each region is either 0 or 1. In this case, the model is a nondifferentiable process and the training process is often done through reinforcement learning [48]. For soft attention, the attention weights of each region could be expressed by some continuous values between 0 and 1. In addition, the attention mechanism can also be divided into spatial and channel aspects from the concern of domain. In this article, the AG is a kind of soft attention that focuses on the spatial domain. It was first applied in medical segmentation. The grid-based AG block provides better attention to salient regions and suppression of irrelevant regions [49]. Besides, it can be easily embedded into the FCN framework and improve the model performance without adding a large number of parameters of network computation.

As shown in Fig. 2, spatial regions are selected by using the contextual information collected from a coarser scale. By multiplying attention coefficients with feature maps that are combined with coarse- and fine-level information, this block could focus on the features that are useful for the final prediction in a specific task. For example, in building extraction tasks, the attention block could overlook those nonbuilding background information such as clutters by giving more weight to deeper feature maps that have higher semantic information. In this experiment, we embed this attention module on the skip connection. Specifically, the AG could filter the features learned by cascading convolutions, and then, concatenate with up-sampling output layer accordingly to get a finer feature map.

E. Architecture Design

As introduced in Section II-B, we use a modified ResNet-50 design as our contracting path. First, a 7 × 7 large kernel convolution (with a stride of 2) is applied to obtain low-level features and reduce the resolution of the input image. We discard the pooling layer in the original ResNet-50 architecture. That is, after the first convolution operation, four stages of the bottleneck residual block are stacked in the spare encoder part. The feature map in the next stage has twice the channels and half the feature map size of the previous stage, except for the first block. After four stages, the output of the encoder has 1/16 the size of the input data, while capturing deeper semantic features. To reduce the model parameters, we add a convolution operation to change the dimensions of feature maps. We consider an ASPP block as the center bridge to connect the encoder and decoder. In such a small resolution feature map, the ASPP module could get more multilevel high semantic information. Besides, the AG is embedded in the skip connection in our network. To be specific, the outputs of up-sampling layers in the decoder part are concatenated with the corresponding encoder output. It is worth noting that the encoder output is a set of weighted feature maps that have learned the importance of features. By using skip connection with AG and concatenation operation, redundant information could be filtered and the extracted features can be fused to get a clear segmentation map. After concatenation, two 3 × 3 convolutions are applied to change dimensions together with the corresponding BN and ReLU layer. At the end of the decoder part, the image size is the same as the input layer, and a convolution with 3 × 3 kernel size is added to get a score map of the pixels. Finally, a sigmoid function is applied with a threshold of 0.5 to segment the buildings and background.

F. Comparison With Other Models

To evaluate the proposed method in building extraction, we make comparison with the typical encoder–decoder network such as SegNet and U-Net, as well as Deeplab v3+. Meanwhile, another network using attention mechanism, SE-U-Net, and the most recent model, MA-FCN in building extraction are also considered. Here follows a brief introduction of these models.

1) SegNet: Brought up by Badrinarayanan *et al.* in 2015 [20], SegNet adds more shortcut connections compared with the FCN. At the same time, indices from maxpooling layers are introduced to the expanding path. This makes SegNet more efficient and significant for the following FCN models.

2) U-Net: U-Net is the most typical encoder–decoder symmetrical architecture that was first proposed by Ronneberger *et al.* in the biomedical image segmentation field [19]. A cascading contraction path, along with the corresponding expansion path, makes up the entire network. This network is usually used as a baseline for comparing all kinds of models. With the skip connection between two paths, this model exhibits a relative stable performance both in medical imaging and remote sensing tasks.

3) Deeplab v3+: Based on a series of Deeplab networks [21]–[23], Chen *et al.* presented an improved version, Deeplab v3+ [24]. By retaining the advantage of *atrous* convolution and pyramid pooling, this network uses Xception [50] as its backbone. To this extent, Deeplab v3+ achieved state-of-the-art performance.

4) Se-U-Net: Inspired by the squeeze and excitation (SE) module for channel recalibration of feature maps for image classification in [51], Roy *et al.* introduced three variants of SE modules for image segmentation, including squeezing spatially and exciting channel-wise (cSE), squeezing channel-wise and exciting spatially (sSE), and concurrent spatial and channel squeeze and excitation (scSE) [52]. These SE modules were incorporated within three FCN models and achieved consistent improvement of performance across all architectures. These SE modules are essentially an attention mechanism. In this article, we use the scSE module on U-Net for comparison.

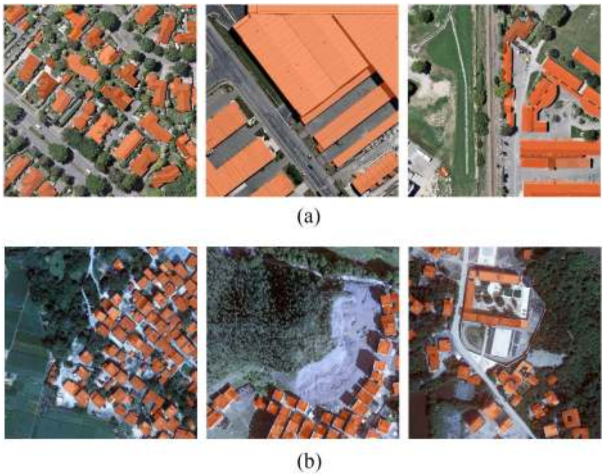

Fig. 3. Samples of (a) WHU building dataset and (b) DB UAV rural building dataset, with orange marks.

5) Ma-Fcn: A multiscale aggregation FCN termed as MA-FCN is one of the most recent networks to extract building pixels [45]. The backbone of the MA-FCN is a four-layer VGG-16 encoder. To make full use of the multiscale feature information, the MA-FCN implements multiscale feature aggregation by concatenating the last convolution layers of each scale in decoding and predict the final result by concatenating these feature maps. At the same time, polygon regularization methods for boundary refinement are also introduced for boundary refinement. This method achieved state-of-the-art results on the WHU dataset.

III. Experiments and Analysis

A. Datasets

To evaluate the performance of our proposed method in different environments, two datasets are conducted in our experiments. The first one is an open dataset, WHU building dataset, and the second is our DB UAV rural building dataset. Fig. 3 shows experimental images of these two datasets.

1) WHU Aerial Building Dataset: This dataset was created by [42], which consists of 18 7000 buildings with a resolution of 0.3 m. The original data from the New Zealand Land Information Services website have been manually edited so the dataset has a rather high quality. It covers about 450 km^2 of New Zealand area with different building shapes and appearances. This dataset contains 8188 tiles 512 × 512 RGB aerial images and is divided into three parts, 4 736 for training, 1 036 for validation, and 2 416 for testing.

2) DB UAV Rural Building Dataset: In this experiment, we collected UAVs aerial orthoimages with 0.2-m spatial resolution, which mainly covers the countryside area of Dianbai county of Guangdong province, China. The whole image is collected using UAV platform and the ground truth of the buildings is manually annotated by experts. Some operations, such as projection transformations and alignment adjustment, are applied.

We downsampled these original images into 0.4-m ground resolution using bilinear interpolation method and cropped them into 23 932 tiles with an image size of 512 × 512. After downsampling, we randomly divided these images into training set, validation set, and testing set at a ratio of 3:1:1. In the bottom row of Fig. 3, we could see that rural buildings have different distributions and characteristics. Compared to the WHU building dataset, the buildings in this dataset are often tiny and hard to recognize. In fact, due to the limitations of drone platform imaging such as atmospheric conditions, light intensity, etc., the quality of the images is worse than those in the WHU aerial dataset. Furthermore, in rural environments, the buildings are often built around the trees and overshadowed by them, which makes the extracting more difficult.

B. Evaluation Metrics

In evaluation part, we use four metrics for pixel-based evaluation, including precision, recall, F1-Score, and intersection over union (IoU). These metrics are widely used as the criteria in semantic segmentation and building extraction. F1 is the weighted average of precision and recall. These four indicators are explained as follows. IoU is a criterion that calculates the ratio of the intersection and union between the predicted category and the real category.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{4}$$

where TP, FP, and FN denote the true positive, false positive, and false negative, respectively.

C. Implementation Details

To enhance network generalization ability and avoid overfitting, data augmentation strategies are implemented. We first use spatial data augmentation methods on both two datasets, including random rotation at different angles (90°, 180°, 270°, 360°) and random mirror flip vertically or horizontally. Considering that DB aerial images from the UAV platform are acquired under different time and illumination conditions, we also use radiometric augmentation on the DB rural building dataset. Spectral augmentation methods include brightness and contrast enhancement, blurs, and gauss noise.

The whole network was implemented using PyTorch with CUDA10.0 and CuDNN7.6. For the parameter setting, we adopt Adam [53] as our optimizer to make the training process converge quickly and the initial learning rate is 0.0001. To fit our GPU memory, batch size is set to 8. For the whole training period, there are 60 epochs on 2× GeForce RTX 2080 Ti and the entire training process on WHU dataset took about 6 h. In our experiments, we use binary cross-entropy loss as the loss function.

TABLE I
QUANTITATIVE RESULTS FOR PRECISION, RECALL, F1-SCORE, AND IOU ON WHU AERIAL BUILDING DATASET

Methods	Precision	Recall	F1-score	IoU
Segnet	0.9209	0.9369	0.9289	0.8672
Deeplab v3+	0.9344	0.9415	0.9379	0.8831
U-Net	0.9461	0.9371	0.9416	0.8896
SE-U-Net	0.9414	**0.9487**	0.9450	0.8957
MA-FCN	0.9444	0.9485	0.9465	0.8984
Proposed	**0.9497**	0.9481	**0.9490**	**0.9029**

TABLE II
QUANTITATIVE RESULTS FOR PRECISION, RECALL, F1-SCORE, AND IOU ON DB UAV RURAL BUILDING DATASET

Methods	Precision	Recall	F1-score	IoU
Segnet	0.8090	0.8244	0.8166	0.6901
Deeplab v3+	0.7601	0.8513	0.8032	0.6711
U-Net	0.7897	**0.8885**	0.8362	0.7185
SE-U-Net	0.8215	0.8665	0.8434	0.7293
MA-FCN	0.8375	0.8727	0.8547	0.7463
Proposed	**0.8387**	0.8803	**0.8590**	**0.7531**

D. Experiments and Analysis

In this part, we first compared our method with the aforementioned competitors on building extraction on two datasets. Then, the ablation study on the WHU dataset was conducted to validate the effectiveness of each module. Finally, we explored the impact of different backbone on the whole network. Here follows the detailed experiments.

1) Experiments on WHU Aerial Building Dataset: Table I shows the obtained numerical comparisons, including precision, recall, F1-score, and IoU, and best records are marked with bold. Our method outperforms other four methods on IoU and F1-score metric, obtaining 0.9029 and 0.9490, respectively.

Fig. 4 further displays the prediction visualization results of these comparative models. It can be observed that in large buildings, the network with multiscale feature module, such as our proposed method, Deeplab v3+, and MA-FCN, hs better performance with regard to the integrity of buildings. This is because of the aggregation of multiscale features, which lead to broader context information. Visualization result of our proposed method has a rather good improvement of detecting both large and small buildings, as well as accurate edge information. By introducing the ASPP module between the encoder and decoder part, our method could catch the multiscale information of these buildings and retain a complete edge and detailed information.

2) Experiments on DB UAV Rural Building Dataset: In this section, we compare our network with other state-of-the-art semantic segmentation methods under the same training configuration of 4 108 testing images totally on our DB UAV rural building dataset. Compared to the WHU building dataset, the images in this dataset are foggy and lower quality due to the operating platform. All tests in Table II have been conducted with radiometric augmentation before training as discussed in the previous section. As shown in Table II, our methods exhibits

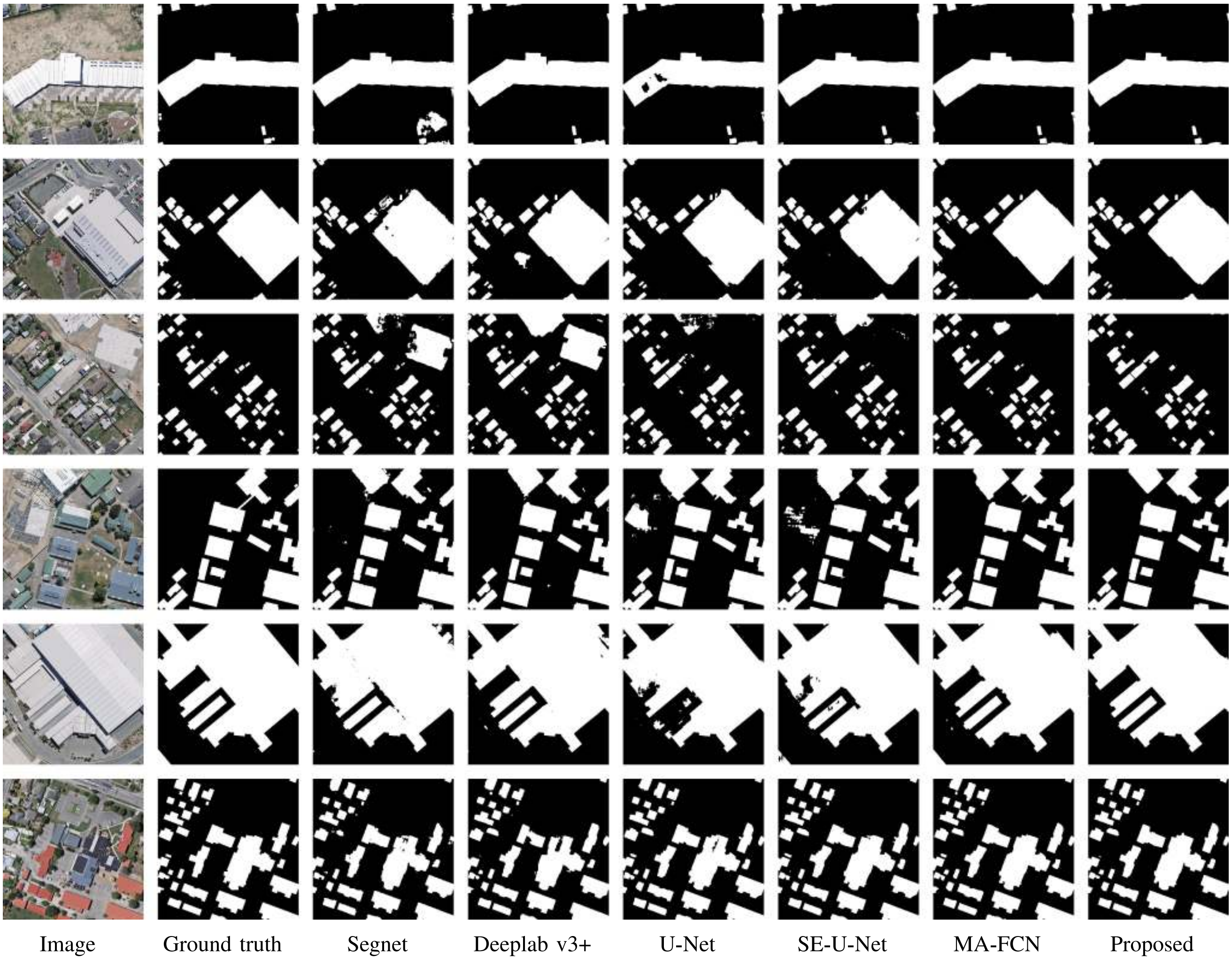

Fig. 4. Visualization of the obtained results by different models on WHU aerial building dataset.

better or comparable numerical results compared to others. Specifically, our method has higher F1-Score and IoU value that outperforms the original U-Net model by 2.28% and 3.46%, respectively. The highest values are highlighted in bold.

Apart from the numerical comparison, we also present visualization results. Fig. 5 represents several sets of binary prediction examples of our network and compared methods in rural environments. Different from WHU dataset, these examples of buildings have irregular distributions and different roof materials. Some buildings are overshadowed by trees and some roofs are even covered with moss, which increases the difficulty of extraction. For illustrative purposes, we have selected two representative samples with the corresponding results and zoomed them in Fig. 6. More obvious area is marked with red rectangles.

According to the visualization results, our proposed method, SE-U-Net and MA-FCN, performed better than others as the whole, especially in the recognition of edges, tiny, and shady buildings. By introducing the grid-based AG in building extraction, spatial regions are selected by highlighting salient and context information collected from a coarser scale. Thus, information extracted from a coarse scale could be used in gating to filter irrelevant and noisy responses in skip connections. At the same time, the ASPP module can help to catch broader context information that contributes to recognize large-scale buildings. In general, with the grid-attention technique and the ASPP bridge, our model could achieve a better and more precise extraction result.

3) Ablation Experiments: To better show the influence of the ASPP module and the AG, we conducted the ablation studies on WHU dataset and quantified the results. First, we conducted the baseline experiment without any module, which is a modified ResNet-50 architecture. Then, we added the grid-based AG mechanism on this baseline between the encoder and the decoder. Finally, the baseline experiment with an ASPP module used as a bridge to get broader context information was conducted.

Numerical results are shown in Table III, and the best values are marked with bold. Compared to the baseline experiment and the other two experiments with only one module added, the proposed network achieved relatively higher values of the four metrics. With the two modules, the proposed method outperformed the baseline by 1.62 % on IoU and 0.9 % on F1-score. Despite this, it should be noted that the AG module and the ASPP module can both improve the recall metric. It can be

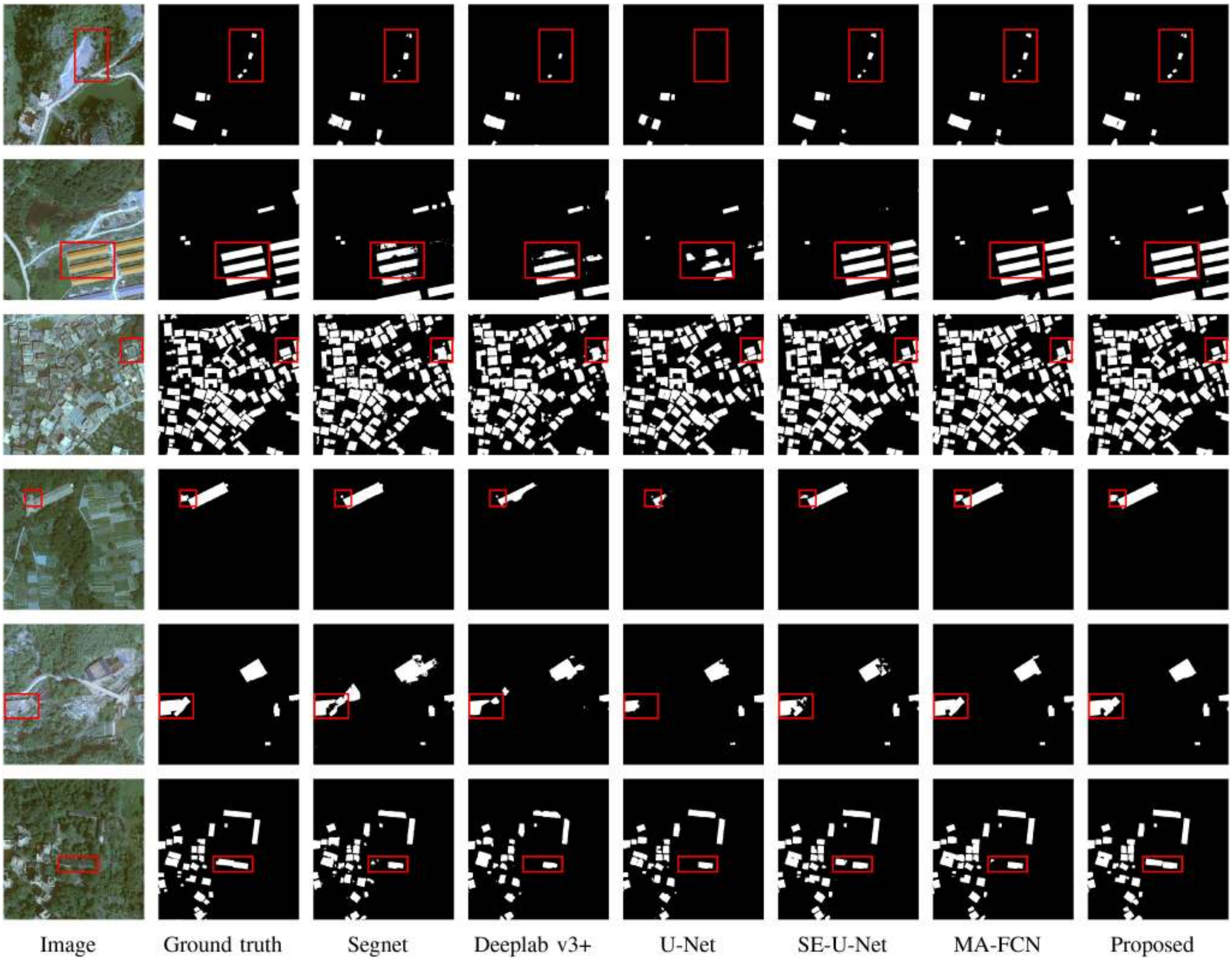

Fig. 5. Visualization of the results obtained by different state-of-the-art models and our model on DB UAV rural building dataset.

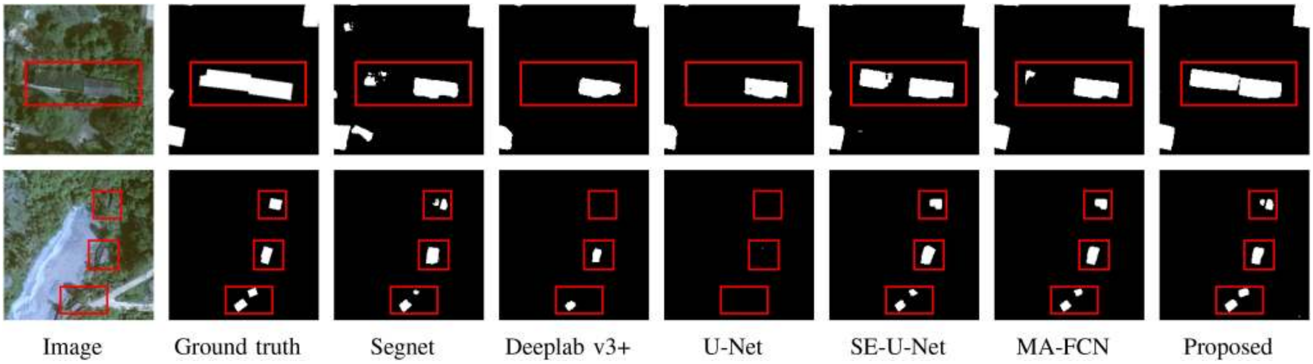

Fig. 6. Presentation of the characteristics of buildings in rural environments and the corresponding prediction results of comparing models.

inferred that, in the building extraction task, the modules can both reduce the probability of false-negative predictions. This may be because of the broader context information from the ASPP module and the salient features of buildings from the grid-based spatial AG mechanism. Besides, the baseline with the ASPP module performs better than the AG with regard to the gain in accuracy. Accordingly, in relation to the network complexity, it also brings more parameters than the AG module.

Visualization results of the ablation experiments are in Fig. 7. The top row of Fig. 6 is an example filled by a large building and the bottom one is an example mainly filled with small buildings. We could see that the baseline model with only an AG module are more likely to misclassify the category (such as the hole in it), especially in the large building recognition. Meanwhile, while predicting small buildings, this method can capture them accurately and precisely. In contrast, the baseline network with

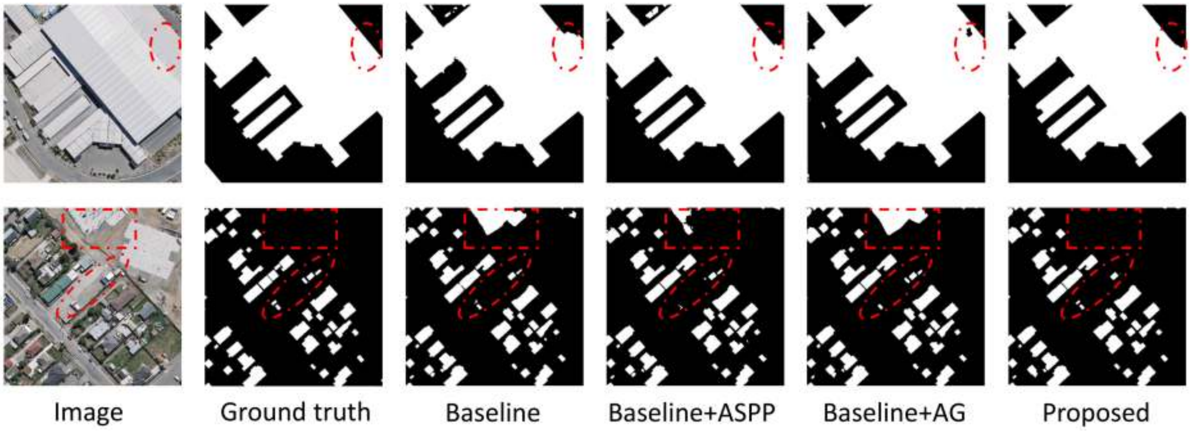

Fig. 7. Visualization results of the ablation studies.

TABLE III
QUANTITATIVE RESULTS OF THE ABLATION STUDIES ON WHU AERIAL BUILDING DATASET, INCLUDING FOUR EXPERIMENTS, I.E., BASELINE, BASELINE+AG, BASELINE+ASPP, AND PROPOSED METHOD

Method	Precision	Recall	F1-score	IoU
baseline	0.9490	0.9311	0.9400	0.8867
baseline+AG	0.9472	0.9419	0.9446	0.8950
baseline+ASPP	0.9473	0.9481	0.9477	0.9006
proposed	**0.9497**	**0.9482**	**0.9490**	**0.9029**

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT BACKBONES FOR PRECISION, RECALL, F1-SCORE, AND IOU ON WHU AERIAL BUILDING DATASET

Encoder Design	Precision	Recall	F1-score	IoU
Modified ResNet-18	0.9413	0.9513	0.9462	0.8980
Modified ResNet-34	0.9496	0.9461	0.9479	0.9009
Modified ResNet-50	**0.9497**	**0.9481**	**0.9490**	**0.9029**

an ASPP module is more likely to catch the context information and predict the right category. Accordingly, the details of the building are not maintained well. In this case, the proposed method that makes use of these two modules can both catch the multiscale features and pay attention to small buildings to a certain extent as well. More specifically, the ASPP module used in the coarse layer can get more global context information and the AG modules in the relatively finer layers can highlight salient information in local regions.

4) Different Backbone Comparison: In this article, we use a modified ResNet-50 as our encoder path. We also explore the performance of ResNet-18 and ResNet-34 as encoder part. Similarly, we removed the first max pooling layer to ensure the resolution of feature maps. Same as ResNet-50, there are four stages in ResNet-18 and ResNet-34 and every stage consists of several blocks. In ResNet-18, every stage contains two basic building blocks, and in ResNet-34, the four stages contain three, four, four, and six basic building blocks, respectively. ResNet-50 has the same building blocks as ResNet-34 but it uses bottleneck building block instead of basic building block as its component unit. Table IV recorded the prediction results obtained for the testing data of WHU aerial building dataset. The highest values are marked with bold. Compared with ResNet-18, ResNet-34 and ResNet-50 achieve better results, getting 0.9009 and 0.9029 on IoU metric, respectively. It can be concluded that the stack of convolutional layers contributes to the final results, which exhibits about 0.49% IoU improvement. At the same time, the stack of bottleneck building blocks provides a slight improvement as compared to the basic building block. Despite this, with the complexity of the network structure, the amount of network parameters and running time will increase accordingly.

IV. CONCLUSION

FCNs have shown great potential in semantic segmentation of buildings. In this article, a new encoder–decoder architecture (combined with AG and an ASPP module) is proposed to detect various scales of buildings under complex circumstances. Simultaneously, different augmentation methods are applied to enhance the generalization ability of the model and tackle the radiation difference problem observed in rural building datasets. This spatial-wise AG mechanism is applied to highlight salient regions and restraint irrelevant information. This avoids misclassifying those background objects that have similar spectral features as buildings, such as concrete road and courtyard wall. We use the ASPP module as a bridge between the encoder and the decoder, to capture multiscale features of the objects. This innovative contribution allows us to take into account both small- and large-scale buildings. Because of the capacity of our method to distinguish buildings under complex environments and extract multiscale features of buildings, our network exhibits high value in practical scenarios.

As the semantic labeling extraction of buildings is only part of buildings extraction, in future work, we will further extract the vector boundaries of buildings based on our method to provide structured individual building polygons for practical applications. Meanwhile, in relation to the parameters of the two modules, we will consider the balance between the network complexity and accuracy gains in the future.

REFERENCES

[1] A. Mishra, A. Pandey, and A. S. Baghel, “Building detection and extraction techniques: A review,” in *Proc. 3rd IEEE Int. Conf. Comput. Sustain. Glob. Develop.*, 2016, pp. 3816–3821.

[2] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2013.

[3] X. Jin and C. H. Davis, "Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information," *Eurasip J. Appl. Signal Process.*, no. 14, vol. 2005, pp. 2196–2206, 2005.

[4] Ö. Aytekin, A. Erener, I. Ulusoy, and H. S. Düzgün, "Automatic and unsupervised building extraction in complex urban environments from multi spectral satellite imagery," in *Proc. 4th Int. Conf. Recent Adv. Space Technol.*, 2009, pp. 287–291.

[5] L. Hu, J. Zheng, and F. Gao, "A building extraction method using shadow in high resolution multispectral images," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2011, pp. 1862–1865.

[6] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image," *J. Multimedia*, vol. 9, no. 1, pp. 181–188, 2014.

[7] X. Gao, M. Wang, Y. Yang, and G. Li, "Building extraction from RGB VHR images using shifted shadow algorithm," *IEEE Access*, vol. 6, pp. 22034–22045, 2018.

[8] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[9] A. O. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Using ImageNet pre-trained networks," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[12] P. Du, E. Li, J. Xia, A. Samat, and X. Bai, "Feature and model level fusion of pretrained CNN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2600–2611, Aug. 2019.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016. [Online] Available: https://arxiv.org/abs/1602.07261

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

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

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

[20] 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.

[21] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014. [Online] Available: https://arxiv.org/abs/1412.7062

[22] 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.

[23] 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

[24] 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. Vision*, 2018, pp. 801–818.

[25] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4353–4361.

[26] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, vol. 1, pp. 1857–1866.

[27] S. Liu, Q. Shi, and L. Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3018879.

[28] Q. Shi *et al.*, "Domain adaption for fine-grained urban village extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1430–1434, Aug. 2020.

[29] C. Kyrkou and T. Theocharides, "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1687–1699, Mar. 2020.

[30] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, May 2020.

[31] Y. Tan, S. Xiong, and Y. Li, "Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3988–4004, Nov. 2018.

[32] R. Davari Majd, M. Momeni, and P. Moallem, "Transferable object-based framework based on deep convolutional neural networks for building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2627–2635, Aug. 2019.

[33] S. Chen, W. Shi, M. Zhou, M. Zhang, and P. Chen, "Automatic building extraction via adaptive iterative segmentation with LiDAR data and high spatial resolution imagery fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2081–2095, May 2020.

[34] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-Driven multitask parallel attention network for building extraction in high-resolution remote sensing images," in *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3014312.

[35] Y. Li, B. He, T. Long, and X. Bai, "Evaluation the performance of fully convolutional networks for building extraction compared with shallow models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 850–853.

[36] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[37] X. Li, X. Yao, and Y. Fang, "Building-a-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3680–3687, Oct. 2018.

[38] Y. Xie *et al.*, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, Apr. 2020.

[39] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, pp. 144–161, 2018.

[40] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, no. 9, pp. 91–105, 2019.

[41] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, 2019.

[42] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[43] W. Feng, H. Sui, L. Hua, and C. Xu, "Improved deep fully convolutional network with superpixel-based conditional random fields for building extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 52–55.

[44] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 6, pp. 1–18, 2020.

[45] S. Wei, S. Ji, and M. Lu, "From aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[46] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: https://arxiv.org/abs/1502.03167

[47] 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, no. 9, pp. 1904–1916, Sep. 2015.

[48] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *Adv. Neural Inf. Process. Syst.*, vol. 3, no. 1, pp. 2204–2212, 2014.

[49] O. Oktay *et al.*, "Attention u-Net: Learning where to look for the pancreas," 2018. [online] Available: https://arxiv.org/abs/1804.03999

[50] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

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

[52] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2018* (Lecture Notes in Computer Science), vol. 11070. Cham, Switzerland: Springer, 2018, pp. 421–429.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

Wenjing Deng received the B.S. degree in geographical information science from the Southwest University, Chongqing, China, in 2018. She is currently working toward the M.S. degree in cartography and geographical information system with Sun Yat-sen University, Guangzhou, China.

Her research interests include semantic segmentation and scene classification for high-spatial resolution remote sensing imagery.

Qian Shi (Senior Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2015.

She is currently an Associate Professor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. Her research interests include remote sensing image classification, including deep learning, active learning, and transfer learning.

Jun Li (Fellow, IEEE) received the B.S. degree in geographic information systems from Hunan Normal University, Changsha, China, in 2004, the M.E. degree in remote sensing from Peking University, Beijing, China, in 2007, and the Ph.D. degree in electrical engineering from the Instituto de Telecomunicações, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal, in 2011.

She is currently a Full Professor with Sun Yat-sen University, Guangzhou, China, where she founded her own research group: Hyperspectral Calibration and Learning (HCL), in 2013. Since then, she has obtained several prestigious funding grants at the national and international level. She has authored and co-authored more than 160 journal citation report papers, 60 international conference papers, and a book chapter.

Prof. Li is serving as the Editor-in-Chief for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.