A Multiscale and Multipath Network With Boundary Enhancement for Building Footprint Extraction From Remotely Sensed Imagery

Hua Zhang ©, Xiangcheng Zheng ©, Nanshan Zheng ©, and Wenzhong Shi ©

Abstract—Due to its high efficiency and low cost, automatic extraction of building footprints from remotely sensed imagery has long been an important means to obtain building footprint information, which can be easily implemented using existing fully convolutional network (FCN)-based methods. However, such methods suffer from imperfections and thus accurately extracting building footprints from remotely sensed imagery remains a challenging task. For example, cascaded convolutions generally cannot preserve the spatial details well, leading to blurred boundaries and omission of small buildings. Insufficient multiscale features fusion without considering semantic gaps between different level features could yield misclassification. In addition, the fixed receptive field always produces discontinuous holey extracted large buildings. To this end, we propose a novel multiscale and multipath network with boundary enhancement (MMB-Net) that accurately extracts building footprints from remotely sensed imagery. More specially, a parallel multipath feature extraction module is first designed to capture high spatial information-preserved multiscale features with less semantic distances. In addition, the receptive field is enlarged and broadened by a multiscale features enhancement module. Then, an attention-based multiscale features fusion module is built to appropriately aggregate multiscale features. Lastly, a spatial enhancement module is presented to refine the extracted building boundaries by capturing boundary information from low-level features. The proposed MMB-Net has been tested on two benchmark data sets together with other SOTA approaches. The results show that MMB-Net can achieve promising building footprints extraction performances and it outperforms the SOTA methods. The implementation of MMB-Net is available at.

Index Terms—Attention mechanism, boundary enhancement, building footprint extraction, fully convolutional network (FCN), semantic gap.

I. INTRODUCTION

BUILDING footprint information is important for urban planning, land management, illegal building monitoring, etc. [1], [2], [3]. With the great advancement of remote sensing technologies, now high spatial resolution remote sensing images can provide rich semantic and spatial details information for ground objects, which makes buildings in the images recognizable. Accurate delineation of building footprints from remotely sensed imagery has become feasible and aroused growing attention [1], [2]. Although manual interpretation can achieve accurate building footprints from remotely sensed imagery, it is time-consuming and inefficient. Therefore, it is crucial to develop automatic, accurate, and efficient methods for extracting building footprints from images. So far, numerous algorithms have been proposed to automatically extract building footprints from images, they can be roughly divided into traditional handcrafted feature-based methods and deep learning-based methods.

For the traditional methods, based on such extracted features as spectrum, shape, and texture from images by using the manual designed feature operators, recent advances have made great achievements but capturing the above-mentioned handcrafted features usually requires a strong prior knowledge. Moreover, the empirically designed features have limited generalization ability due to that these handcrafted features vary with different sensor types, building structures and light conditions, etc. As a result, the traditional methods have limitations in generalization, and are only suitable for some specific tasks.

More recently, with the rapid development of deep learning in computer vision [1], more and more deep learning-based building extraction methods have been proposed owing to its excellent feature representation. Among the deep learning-based methods, the fully convolutional network (FCN)-based building footprints extraction algorithms have shown promising results and have been proven effective in building footprints extraction [10]. Deriving from the FCN architectures, the encoder-decoder-based architectures have also been widely applied in building footprints extraction [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. In which the multiscale features are extracted by the decoder through cascaded convolutions, and then the spatial size of feature map is recovered gradually through the upsampling operations of the decoder. The methods can achieve
better performance than FCN-based methods, particularly on the boundary localization through the spatial details recovered by the decoder. Nevertheless, due to the fact that there exists large within-class and small between-class variance in pixel values, similarity between buildings and nonbuildings, and great various size of buildings in the images, the existing encoder-decoder-based methods still suffer from limitations in the accurate extraction of building footprints, including blurred boundaries, omission of small buildings and holes in large buildings, caused by the insufficient multiscale capturing, aggregation of multiscale features and boundary information losses [21]. To solve these problems, many strategies have been explored.

Improvement in the ability of extracting multiscale features can help to accurately extract building footprints. Ji et al. [11] provided the SR-FCN model for accurate building footprint extraction from aerial and satellite imagery, in which two atrous convolutions were used to expand the receptive fields for multiscale features extraction. SIU-Net [12] applied weights-shared branches for inputting multiscale images to enhance the multiscale features extraction. PSPNet [13] was proposed to capture multiscale features and global context information through a pyramid pooling module. Kang et al. [14] provided the EU-Net for extracting buildings from optical remote sensing images, where a deep spatial pyramids pooling module was designed for multiscale features extraction. The atrous spatial pyramid pooling module was developed to capture multiscale contextual information with filters at multiple sampling rates and receptive fields [15], [16], [17]. UNEt-AP [18] was developed for building footprint extraction from remotely sensed imagery, in which the multiscale features were captured by the atrous convolution with up-sampled filters. Different from the above-mentioned encoder-decoder-based methods, HRNet [19] was provided to extract multiscale features through a high-resolution CNN with parallel feature extraction paths, and achieved promising semantic segmentation. Inspired by the HRNet, Zhu et al. [20] proposed the MAP-Net for building footprint extraction, where a multiparallel paths structure was applied to extract spatial localization-preserved multiscale features for accurate extraction of building footprints. A densely upsampling convolution was applied to enable DE-Net to capture spatial information in feature maps [21]. These methods can improve the accuracy of extraction of building footprints by capturing multiscale features based on the designed modules, while high-resolution features may not be preserved enough, resulting in the blurred boundaries and omission of small buildings. In addition, the semantic gaps between features from different levels are not considered, which may degrade feature aggregation.

It is important to aggregate features learned from different levels to achieve multiscale features for accurate building footprints extraction. In the encoder-decoder framework, low-level features contain more detailed information, while high-level features include more semantic information, they are complementary for accurate segmentation [22]. Therefore, the skip connection was usually used to aggregate these features learned from different levels, in which the features extracted by the encoder can be fused more flexibly in the decoder, and has been proved effective in recovering fine-grained details information [22], [23], [24]. Based on UNet [22] and residual network, a new deep residual learning serial segmentation network was proposed for building footprints extraction from images by combing skip connection and residual representation [25]. ResUNet-a was proposed by Diakogiannis et al. [26], which took UNet as encoder-decoder backbone, and combined residual connections module, pyramid scene parsing pooling module, and multitasking module. Moreover, attention mechanism [27], [28], [29], [30], [31], [32], [33] has been proved helpful for multiscale futures fusion through capturing global relations with long-range dependences in spatial or channel aspects, which effectively improves the performance of segmentation. To enhance spatial details and prevent irrelevant information in the low-level feature map, a new dense attention network was proposed by Yang et al. [34]. Pan et al. [35] provided a new generative adversarial network to achieve fine multiscale features combining spatial and channel attention mechanisms. The above-mentioned methods can perform well in most cases, but the fusion of deep and shallow features is insufficient. For example, in the UNet architecture, only plain skip connections from encoding part to the decoding part can help to recover the location information, lacking different hierarchical feature maps from the encoder which can be merged to the decode module. Furthermore, due to the semantic and resolution gaps between features from different levels, introducing low-level features with coarse information may bring redundant information or noise, resulting in inaccurate boundary localization and unrecognized small buildings.

To solve the problem of extracted building footprints with blurred boundaries, some studies have been conducted to predict more accurate boundaries by introducing boundary information or postprocessing techniques. Bischke et al. [36] proposed a multitask learning network which used the signed distance function as the output represent for producing fine-grained segmentations with precise boundaries. To obtain accurate boundaries, Pan et al. [37] provided the PEGNet which combined an atrous module and edge-region detection module containing dilated edge information. EA-Net [38] was designed to generate accurate buildings from aerial images by introducing the edge perception networks. To improve the detection accuracy of edges, based on multitask learning and dense D-LinkNet, Xia et al. [39] proposed a CNN framework which adopts full-scale skip connections and edge guidance module used to improve the location accuracy of buildings. Shrestha et al. [40] designed an improved FCN for building footprints extraction from images, in which the CRF-based postprocessing was adopted. Wei et al. [41] proposed a deep neural network framework for automatically extracting building footprint, in which the polygon regularization was performed on the initial segmentation result by a multiscale aggregation FCN to obtain a polygonized map. To reduce the noise and sharpen the boundary of the buildings, Zhou et al. [42] presented BOMSC-Net to yield continuous building footprints by considering boundary optimization based on a multiscale context-aware module. To obtain precise boundaries of segmentation masks, Jung et al. [43] designed a boundary enhancement (BE) module for enhancing the boundaries of segmentation mask in the network, where edge features were extracted by using holistically nested edge detection method. To overcome
the limitations in boundary accuracy of extracted buildings from images, the active contour model and CNN were combined to extract precise buildings using remote sensing images and LiDAR data [44]. However, most of the above-mentioned models require more accuracy boundaries information, auxiliary data, or sophisticated structures. While edges only occupy a tiny part of one image, it is difficult in obtaining accurate edges due to poor spatial resolution, spectral variation, and noisy pixels. Furthermore, the postprocessing will also reduce the performance of the model. In fact, losing spatial details is the main reason causing blurred boundaries. Although many strategies had been provided to aggregate the low-level features to recover spatial details, the noisy information is also introduced at the same time, resulting in inaccurate features representation, because the coarse low-level features contain not only spatial details information but also noise information. However, research to solve these problems is lacking.

To address the aforementioned problems, this article presents a novel multiscale and multipath network with boundary enhancement (MMB-Net), which can extract accurate building footprints with accurate boundaries from remotely sensed imagery. First, a parallel multipath module is constructed to extract multiscale features with characteristics of spatial localization preserved, high-level semantic information, and less semantic gap through serial convolutional blocks with fixed number of convolutional layers in each path. Then, a multiscale features enhancement module is introduced to enlarge and broaden the receptive field for more multiscale features extraction. Next, an attention-based multiscale features fusion (AMFF) module is built to appropriately aggregate features learned from multilevel for optimal multiscale features fusion by bridging semantic gaps and considering different importance of different level features. Lastly, a spatial enhancement module is embedded to capture boundary information for refining the extracted building boundaries. Experimental results on the SpaceNet building detection data set and WHU aerial building data set demonstrate that the proposed method can address the problems of blurred boundaries, omission of small buildings, and holes in large buildings to some extent, and outperform other SOTA algorithms.

The main contributions of this work can be summarized as follows.

1) A novel semantic segmentation network MMB-Net for automatic building footprints extraction is proposed in this study.
2) A novel parallel multipath module is proved effective for multiscale features extraction, which can preserve the spatial details information and alleviate the semantic distances between features learned from different levels.
3) Toward bridging semantic gaps and considering different importance of features learned from different levels, an attention-based fusion module is devised to achieve optimal multiscale features aggregation.
4) To enhance the extracted building boundaries, a spatial enhancement block (SEB) is presented to capture boundaries from low-level features.

The rest of this article is organized as follows. Section II provides the detailed structure of the MMB-Net. In Section III, the performance of the proposed approach is illustrated through experiments. Finally, Section IV draws the conclusion.

II. PROPOSED METHOD

In this section, the MMB-Net is presented for extracting building footprints from remotely sensed imagery. First, an overview of the proposed method is provided to introduce the general motivation and architecture, and then the details of main modules are described one by one.

A. Overview of the Proposed MMB-Net

It is essential to extract robust multiscale features and abundant contextual information for accurately extracting building footprints, cascaded convolutions and pooling operations are commonly utilized in most FCN-based building footprints extraction methods. In which the receptive field is enlarged and more intrinsic features are learned, while it is inevitable to lose the spatial details information, especially the serious loss of boundary information, resulting in blurred boundaries. To improve the accuracy of prediction results, some effective network architectures have been constructed to combine complementary cues of low-level features containing more detailed location information and the high-level features with more semantic information through skip connections or other methods. However, features from low-level layers are too noisy to provide sufficient high-resolution semantic guidance, some background “noisy” features may be introduced when fusing features learned from different levels, which will degrade the robustness of features and cause semantic inconsistency. In addition, the semantic and resolutions gaps exist between the low-level and high-level features for their significant differences in both spatial distributions and semantic information, which will disturb the multiscale features fusion. Thus, directly fusing these low-level and high-level features is less effective. Moreover, it is not enough for large buildings to obtain global dependence information by increasing the receptive field through merely repeated stacked convolutions. To this end, we construct the MMB-Net, as shown in Fig. 1, toward effectively extract homogeneous building footprint with accurate boundaries. It mainly includes the following components.

1) A parallel multipath module for extracting spatial detail-preserved multiscale features, alleviating the semantic gap through the same number of convolutional layers in each branch.
2) A multiscale feature enhancement (MFE) block for enlarging the receptive field of the network and capturing more multiscale features simultaneously.
3) To effectively exploit the multiscale features learned from different branches, an AMFF block for gradually fusing the multilevel features learned from its parallel branches in a coarse-to-fine manner.
4) A SEB for extracting the spatial complementary information of boundaries from low-level features to improve the boundary accuracy.

It can be seen from Fig. 1, the main stages are as follows: first, a stem block is used to down-sample the resolution, subsequently
Fig. 1. Architecture of the proposed MMB-Net, which is composed of four modules. (a) Parallel multipath feature extraction module. (b) Multiscale features enhancement module. (c) AMFF module. (d) SFE module. The stem block consists of two stride convolution layers to extract feature maps and decrease the spatial resolution of the input image. The Conv block is a residual block consists of two convolution layers, a shortcut fuses input to output through element-wise addition, and BN and ReLU.

four parallel paths are designed to extract multiscale features with the same number of convolutional layers in each path, while preserving local details and alleviating the semantic gaps between different level features. Then, the MFE block is provided to enlarge the receptive field and enhance the multiscale features for extracting building footprints of different sizes robustly. Next, to bridge the semantic and gaps between multilevel feature maps and aggregate multiscale features efficiently, the AMFF block is constructed to selectively aggregate multilevel features gradually. Lastly, to further improve the boundary accuracy of building footprints, the SFE block is presented to remove the noise and keep the boundary information in low-level features.

At the end of MMB-Net, the multiscale features extracted from the four paths and SFE block are aggregated for obtaining pixel-wise predictions by using element-wise sum. In Fig. 1, C, H, and W denote the numbers of channels, height, and width of the input image, respectively.

B. Parallel Multipath Feature Extraction Module

Generally, in the traditional encoder-decoder network, the encoder part is usually explored to achieve multiscale features by aggregating multilevel dense features, and enlarge the receptive field by cascaded convolutions operations, but it may result in semantic gap and loss of spatial information. Solving the above-mentioned problems is crucial for precise semantic segmentation. Inspired by the HRNet [19] which presents a parallel multipath architecture with high-resolution representation to extract multiscale features, we construct the parallel multipath module to extract multiscale features from images (as shown in Fig. 1), in which each branch has the same number of convolutional layers, in this way, features extracted from each path are independent with fixed scales, the high-resolution features can also be maintained by Path 1 instead of restoring the resolution by an upsampling operation, and alleviating the semantic gap between multilevel features simultaneously. Compared with the original encoder-decoder network, the parallel multipath feature extraction module can extract multiscale features containing richer high-level semantic information and more accurate spatial information.

As shown in Fig. 1, first, in order to decrease the computational complexity and alleviate the noise contained in the low-level feature maps simultaneously, a stem block is designed to down-sampling the features before applying the parallel multipath feature extraction module, as illustrated in Fig. 2(a), it consists of two stride $3 \times 3$ convolution layers to down-sample resolutions and a $1 \times 1$ convolution layer to extract feature maps with 128 channels and decreases the spatial resolution of the input image to $1/4$, and subsequently the four parallel feature extraction branches. Then, the features are encoded parallel (resolution and channels remain the same) and downstream (double-downsampled resolution and double channels), the downstream encoding consists of a max-pooling layer to down-sample resolutions and a $1 \times 1$ convolution layer to decrease channels of feature maps. Each path is composed of the same amount of conv blocks, in this way, adjacent encoding branch extracts multiscale features with less semantic distance. Fig. 2(b) shows the Conv block, which includes a residual block consists of two $3 \times 3$ convolution layers, a shortcut fuses input to output through element-wise sum, and batch normalization (BN) and rectified linear unit (ReLU). The parallel multipath feature extraction module is composed of four parallel paths for
extracting multiscale feature maps, the resolutions of extracted feature maps are 1/4, 1/8, 1/16, and 1/32 of the original input image, and the corresponding channels are 128, 256, 512, and 1024 in each path.

Compared with the typical encoder-decoder network, the MMB-Net employs four parallel branches to extract and aggregate features at multiple scales. Differently from the HRNet, there are no repeated multiresolution fusions during the features extraction in each path, by this way, the high-resolution spatial location information can be preserved without being weaken by fusing the high-level features with low spatial resolution including noise, which is important for accurately extracting smaller buildings and the boundary. Another difference is that a multiscale features fusion module will be applied to aggregate the multiscale features learned from the four parallel branches.

C. Multiscale Features Enhancement Module

As described in Section II.B, multiscale features for building footprint extraction are captured through the parallel multipath feature extraction module, and the highest scaling rate of the feature map reaches 1/32, which can ensure a large receptive field for the network to some extent. However, buildings’ sizes vary in images, and buildings’ representations show significant differences in images. For small buildings, a fine scale features with a small neighbor are enough, while large buildings require a coarse-scale feature containing a large neighbor. In other words, the network’s receptive field needs to be adjusted according to the sizes of buildings.

Dilated convolution can solve the above-mentioned problems to a certain extent [13], it can effectively enlarge the receptive field of network without increasing network parameters and reducing image resolution. In which, different sizes of receptive fields can be achieved by setting different dilated rates, and then multiscale features can be obtained to capture multiscale context information. Inspired by the dilated convolution and inception structure [15], this study proposes a MFE module, which can capture broader and deeper semantic features by constructing multilevel associative branches with multiscale dilated convolutions, so as to enhance the expression of high-level semantic features for building footprint extraction. As shown in Fig. 3, the dilated convolutions are stacked in a cascaded manner. The proposed MFE module contains five cascaded branches, the first four branches are designed to extract multiscale features information, in which, different dilated convolution numbers are used to obtain four different branches of receptive fields which are $3 \times 3$, $9 \times 7$, $7 \times 9$ and $13 \times 13$, respectively. At the same time, in order to prevent the gradient disappearance of multiple convolution stacks in the process of network training and the grid effect caused by cascaded dilated convolutions, inspired by the residual networks [45], we designed the fifth branch similar to residual mapping. Finally, by fusing the multiscale features learned from the five branches, the MFE module can enhance multiscale features discriminability and robustness, also enlarge the receptive field of MMB-Net to extract global semantic information for alleviate the number of holes exist in the extracted building footprints in a certain.

D. Attention-Based Multiscale Features Fusion Module

Features learned from different branches of the MMB-Net contain multiscale features information, it is essential to aggregate them to achieve multiscale context information for precise building footprints extraction. In the typical encoder-decoder architecture, multiscale features are obtained by fusing via a simple sum ignoring the different importance of different scales. These methods have achieved great success, while the semantic and resolution gaps existing between low-level and high-level features limit the performance of multiscale features fusion. Although in Section II.B, the four parallel paths can reduce the semantic gaps between features from different levels to a certain extent for that each path uses the same number of convolutions, the semantic and resolution gaps still exist for their different resolutions. Furthermore, due to the semantic and resolution gaps
between features from different levels, redundant information or noise may be introduced to reduce the extraction accuracy.

In order to further alleviate the inconsistencies between feature learned from different levels, a new dual attention architecture is designed to optimize the fused features from two aspects of channel and spatial. Concretely, to ensure robust feature fusion, a new attention-based fusion method is proposed to progressively fuse the features from neighbor branches in a bottom-up manner to shrink the semantic and resolution gaps. As shown in Fig. 4, the high-level feature XH is upsampled to the same resolution as the low-level feature XL, and concatenated with XL. Then a dual attention block is designed to enhance the fused features by reducing the semantic and resolution gaps between features learned different level. Finally, the fused feature XF is obtained by using the conv block based on the enhanced features.

The designed dual attention block in the AMFF module is illustrated in Fig. 5. In Fig. 5, the concatenated high-level and low-level from neighbor branches are taken as input features F, and then S is produced through the channel attention module (CAM) and spatial attention module (SAM) in turns. At the same time, to enhance the fitting ability of the model, the residual connection is added to the original features, the last output feature Fref is generated by fusing S and F using element-wise sum. The dual attention feature fusion module consists of CAM and SAM.

1) Channel Attention Module: The CAM focuses on “which channel” is useful in an input feature map by modeling the dependency of features along the channel using squeeze-and-excitation method. First, to avoid the interference of spatial location information and reduce the amount of calculation, the average pooling and maximum pooling are used to compress the spatial dimension of input features simultaneously, and then the generated features are forwarded to a shared network for obtaining two feature vectors. Next, the channel attention weight vector MCA(F) can be achieved by concatenating these two feature vectors and normalizing them with sigmoid function. Finally, the output feature map FCA is captured through fusing MCA(F) and F by using element-wise product. Here, the shared network is a multilayer perceptron (MLP) network with one hidden layer, the size of hidden activation layer is \( r^c/r \times 1 \times 1 \), and \( r \) is set to 16. The CAM is described as follows:

\[
F_{CA} = \sigma(MLP(P_{a}(F)) + MLP(P_{m}(F))) \otimes F \quad (1)
\]

where \( F \in \mathbb{R}^{C \times H \times W} \) denotes the input features, \( \sigma \) is the sigmoid function, MLP is a MLP network, \( P_{a}(F) \) and \( P_{m}(F) \) are the generated features by the average pooling and maximum pooling, respectively, \( \otimes \) represents the element-wise product operation, and \( F_{CA} \) is the intermediate output features through the CAM.

2) Spatial Attention Module: The SAM explores “where” is a useful part, which focuses on the effective pixels and suppresses the ineffective pixels. Similar to the CAM, based on the input features \( F_{CA} \), MCS(FCA) and ASC(FCA) are first generated by using average-pooling and max-pooling operations, respectively. Then, they are concatenated and convolved by a standard convolution with a kernel of \( 7 \times 7 \), and the spatial attention weight vector \( MSA(F_{CA}) \) is achieved by normalizing them with the sigmoid function. Finally, the output feature map \( S \) is captured through fusing \( MSA(F_{CA}) \) and \( F_{CA} \) by using element-wise product. The weighting process of SAM is as follows:

\[
S = \sigma(Conv_{7 \times 7}(Concat(MCS(F_{CA}),ASC(F_{CA}))) \otimes F_{CA} \quad (2)
\]

where \( S \) is the output features through the SAM, \( Conv_{7 \times 7} \) denotes a convolution operation with kernel of \( 7 \times 7 \), Concat represents the operation using element-wise sum, MCS(FCA) and ASC(FCA) are the generated features by the average pooling and maximum pooling, respectively.

Lastly, the final output feature \( F_{ref} \) is generated by fusing \( MSA(F_{CA}) \) and \( F \) using element-wise sum.

\[
F_{ref} = S + F. \quad (3)
\]

The AMFF module selectively aggregated appropriate multilevel features to enrich the multiscale features for precise building footprint extraction, and bridge the semantic and resolution gaps between low-level and high-level features.

E. Spatial Features Enhancement Module

As described earlier, in the proposed MMB-Net, multilevel features at different scales for building footprint extraction are captured through the parallel multipath feature extraction module, and then the attention-based fusion method is employed to obtain the multiscale features by progressively fusing the features from neighbor branches in a bottom-up manner. Although the achieved multiscale features contain rich high-level semantic information and accurate spatial information, the cascaded convolutional operations used in the MMB-Net also result in a loss of boundary information. To recover the boundary information, some methods directly concatenate features from very low-level layers with high-level layers or through skip connections, However, the features in the very low-level layers not only contain high accurate boundary information but also contain “noisy” information, which may negatively affect the robustness of high-level features. Thus, how to relieve the noise information and keep boundary information in the very low-level layer are vital when recovering spatial details. To achieve this, a spatial features enhancement (SFE) module for boundary refinement, as shown in Fig. 6, is designed and employed in the network.

First, the input image \( L \) is obtained by downsampling the original input image (as shown in Fig. 1) to 1/4, and then...
features X, as the very low-level features, are generated by a conv block. It should be noted that the output features through Path 1 are selected as the very low-level features due to the fact that it owns features with the highest resolution without any upsampling operation among the four paths. The features H, produced by the MFE module as the high-level features, are fed into a $1 \times 1$ convolution and then upsampled to the same resolution of L to obtain features $H_1$.

Second, to achieve the significance $S$ of building area, as we know, the low-level features contain more local region and edge information for buildings, and redundant or “noisy” features mostly locate inside the buildings, while the high-level features contain more discriminative information for the building segmentation. Thus, these high-level features can be used as guidance to remove these redundant or “noisy” features and keep the boundary information in the low-level features. Therefore, the significance $S$ of building area is described as follows:

$$S = \sigma(H_1) = \frac{1}{1 + \exp(-H_1)} \quad (4)$$

where $S \in [0, 1]$ represents the significance of building, its value is proportional to the degree of significance of building. And based on $S$, the image residuals $R$ are calculated as follows:

$$R = 1 - S = 1 - \frac{1}{1 + \exp(-H_1)} \quad (5)$$

where $R \in [0, 1]$, when feature locates inside the building, the value of $R$ is equal to or close to 0, which is useful to suppress the “noisy” features and keep the boundary information in the low-level features.

Finally, image residuals $R$ and the low-level features $X$ is fused to obtain $X_1$ by using element-wise product, and $X_2$ is generated based on $X_1$ using a conv block for better aggregation. Furthermore, a residual connection is added to the low-level features $X$ to improve the fitting ability of the network, the last boundary features $Y$ is generated by fusing $X$ and $X_2$ using element-wise sum.

$$X_1 = R \odot X \quad (6)$$

$$Y = X_2 + X \quad (7)$$

By applying the SFE module, the proposed MMB-Net well-captured spatial details near building boundaries in the very low-level features, which is beneficial for building boundary refinement.
III. EXPERIMENTAL RESULTS

A. Data Sets

To evaluate the performance of the proposed MMB-Net, two different data sets are employed in this study, i.e., the SpaceNet building detection data set [46] and the WHU aerial building data set [12].

1) SpaceNet Building Detection Data Set: The SpaceNet building detection data set consists of WorldView-3 satellite images with a resolution of 0.3 m, covering six regions including Rio de Janeiro, Vegas, Paris, Shanghai, Khartoum, and Atlanta. In this study, the subsets of Las Vegas and Shanghai were selected to evaluate the performance of all compared networks. The selected regions cover about 1216 km$^2$, containing 243382 independent building footprints. We further cut the image into 650 × 650 pixels, and divided them into the training set, test set, and verification set according to the ratio of 6:2:2. Some samples are shown in Fig. 7.

2) WHU Aerial Building Data Set: The WHU aerial building data set contains more than 187000 buildings with various architectures and purposes, the aerial data set is about 450 km$^2$ located in Christchurch, New Zealand. The original ground resolution of the aerial images is 0.3 m, each image is composed of red (R), green (G), and blue (B) wavelengths. In this data set, images are cropped into 512 × 512 tiles, all the cropped images include training, test, and validation with the number of 4736, 2416, and 1036, respectively. Some samples of them are shown in Fig. 8.

B. Experimental Setting

In our experiments, all networks used the same settings as follows: the adaptive moment estimation (Adam) optimizer was adopted with the initial learning rate of 0.0001, batch size was set to 4, and decreased the learning rate by 0.5 times every 30 epochs. We conducted 100 epochs for all methods on the WHU data set and SpaceNet data set. Besides, the standard CrossEntropyLoss was employed as the segmentation loss function. All networks were carried out with Pytorch 1.2.0 and python 3.7.9, and were checked using a single GPU Tesla P100 16 GB with 256 GB memory.

To validate the effectiveness of the proposed approach, according to the main architecture and modules of the proposed MMB-Net, some SOTA methods, including UNet, SegNet, UNet++, PSPNet, DeeplabV3+, and HRNetV2, were adopted to compare with the proposed network on the two data sets. Among them, UNet, SegNet, and UNet++ belong to typical encoder-decoder architecture, in which, low-level features and high-level features are fused to achieve high resolution and more semantic features in the low-to-high upsample process progressively. In the PSPNet, a feature pyramid pooling module is embedded to incorporate multiscale context by aggregating features at different scales. While the Deeplabv3+ combines dilated convolution with spatial pyramid pooling to fuse multiscale features from atrous convolutions with different dilated rates in a parallel manner. Different from the above-mentioned methods, the HRNetV2 learned semantically strong and spatially precise representations by starting start from a high-resolution convolution stream, gradually adding high-to-low resolution convolution streams one by one, and fusing the multiresolution streams in parallel.

C. Experimental Results

As the metrics to evaluate the performance of network, recall, precision, F1 score, and IoU are applied to quantitatively assess the performance of MMB-Net and other compared networks. Equations are described as follows:

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{8}
\]

\[
\text{precision} = \frac{TP}{TP + FP} \tag{9}
\]

\[
F1 = \frac{2 \times \text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \tag{10}
\]
where TP denotes the true prediction (correctly labeled building pixels) on a positive sample, FP denotes the false prediction (pixels which are mislabeled as buildings) on a positive sample, TN is the true prediction (pixels that are correctly labeled as nonbuilding) on a negative sample, FN is the false prediction (pixels which are mislabeled as nonbuildings or labeled as missed building) on a negative sample. Recall represents the percentage of TP over the total positive samples, Precision indicates the percentage of TP in total positive prediction, the F1-score is calculated by the weighted average of precision and recall, and IoU is calculated by the average percentage of overlap of the prediction and ground truth over the union whole image set.

1) Results on the SpaceNet Building Detection Data Set: Fig. 9 visualizes segmentation results of the SpaceNet building detection data set derived from the UNet, DeeplabV3+, PSPNet, HRNetV2, and MMB-Net (the results of SegNet and UNet++ are not provided for the page capacity limit, but the quantitative results are listed in Table I), respectively. As we can see from the results in Fig. 9, owing to existing mixed pixels’ spectral variation, and background complexity in the image, UNet produces segmentation result containing a lot of noise and shows weaker performance than those by the other four approaches. The reasons may be that cascaded convolutions and pooling operations will result in semantic gap and some spatial information losing. And the highest scaling rate of feature map only reaches 1/16 in the UNet, a small receptive field for the network also leads to insufficient features extraction. Furthermore, insufficient utilization of feature maps from different layers may be occurred for only using the plain skip connections in the UNet without carefully considering the semantic and resolution gaps. While DeeplabV3+ and PSPNet product little

| Method   | IoU (%) | Precision (%) | Recall (%) | F1 (%)  |
|----------|---------|---------------|------------|---------|
| UNet     | 76.97   | 88.64         | 85.39      | 86.96   |
| SegNet   | 76.19   | 89.94         | 83.28      | 86.49   |
| UNet++   | 77.16   | 87.78         | 86.45      | 87.11   |
| PSPNet   | 78.76   | 87.36         | **88.89**  | 88.12   |
| HRNetV2  | 80.37   | 90.24         | 88.02      | 89.12   |
| DeeplabV3+ | 80.72  | 90.12         | 88.55      | 89.33   |
| MMB-Net (ours) | **81.02**  | **91.05**  | 88.03      | **89.51**  |
noise due to that the embedded pyramid pooling module and dilated convolution can enlarge the receptive field and extract more high semantic features. The HRNetV2 also can enlarge its receptive field through deep convolution for capturing high semantic features. The proposed MMB-Net produces the least noise, the reason may be that the highest scaling rate of the feature can reach 1/32 through cascade conv blocks, and the embedded MFE module further enlarge the receptive field and extract more high semantic features. In addition, the parallel multipath feature extraction module and AMFF module also can alleviate the semantic and resolution gaps to preserve detailed spatial information to some extent.

It can be seen from Fig. 9, many mislabeled pixels exist in the segmentation result maps. For example, in marked areas A and D, many road pixels are misclassified as building pixels, and some building pixels are mislabeled as background in marked areas B and E. This phenomenon is the most serious in the results of UNet, because the semantic and resolution gaps between different-level features may harm the multiscale features fusion, and the plain skip connections also result in insufficient utilization of feature maps from different layers. DeeplabV3+ and PSPNet can enhance high semantic features and multiscale features through the dilated convolution and pyramid pooling module. And the HRNetV2 achieves rich semantic and high spatial representations by connecting the high-to-low resolution convolution streams in parallel with repeatedly exchange the information across resolutions. Thus, the deeplabV3, PSPNet, and HRNetV2 can produce few misclassified pixels than the UNet, but many mislabeled pixels still exist due to that the semantic and resolution gaps are not well-handled. As for the proposed MMB-Net, most of the pixels are correctly labeled, the reason is that many strategies are employed. First, the parallel multipath feature extraction module is used to extract multiscale features, which can alleviate semantic gap and preserve detailed spatial information to some extent. Second, the MFE module is applied to enlarge receptive field for extracting abundant multiscale features. Last, the AMFF block is employed to reduce the semantic and resolution gaps between features from different levels for alleviating the inconsistencies between features learned from different levels.

As shown in Fig. 9, we also find that the MMB-Net obtains the most accurate boundaries among the five approaches. Taking the marked areas A and E as an example, adjacent buildings can hardly be divided in the results of UNet, DeeplabV3+, PSPNet, and HRNetV2, while they can be well-extracted with accurate boundaries by the MMB-Net. And the extraction results in all marked areas A, B, C, D, and E indicate that the MMB-Net products more accurate building boundaries than the other four methods. The reason is that the cascaded convolutions and pooling operations lead to loss of spatial information, although fusion of low-level features with detailed location information, the skip connections, dilated convolution, and high-resolution convolution stream are introduced in their network architectures, the semantic and resolution gaps exist between different-level features are not well-considered and alleviated, redundant information or noise may be introduced to the fusion stage. While in our proposed MMB-Net, not only the parallel multipath feature extraction module and AMFF block are employed to alleviate the semantic and resolution gaps, but also a special SEB block is designed to capture the spatial complementary information of boundaries from low-level features for the refinement of building boundaries.

Table I lists the quantitative evaluation results. As we can see from Table I, UNet yields lower segmentation accuracies than these of other six approaches. Amongst all methods, MMB-Net achieves most of the greatest accuracy. Compared with UNet, after applying a parallel multipath feature extraction module, MFE, AMFF, and SFE modules, the prediction accuracy is significantly improved, MMB-Net obtains 81.02%, 91.05%, 88.03%, and 89.51% with respect to IoU, Precision, Recall, and F1, respectively, the accuracy gains of MMB-Net over UNet are 4.05%, 2.41%, 2.64%, and 2.55%, respectively.

2) Results on the WHU Aerial Building Data Set: Fig. 10 shows the building footprint extraction results of the WHU building data set by the UNet, DeeplabV3+, PSPNet, HRNetV2, and MMB-Net, respectively. Visually, as shown in Fig. 10, in the second row of segment results maps, mislabeled pixels are particularly serious in the results of UNet, DeeplabV3+, PSPNet, and HRNetV2, due to g similar colors and textures between round and building pixels, many ground pixels are incorrectly classified as buildings, while MMB-Net almost accurately distinguish ground and buildings with similar colors and textures, which indicates the MMB-Net can capture more discriminative information for the building segmentation. In addition, from Fig. 10, we can find a lot of small buildings cannot be extracted by the UNet, DeeplabV3+, PSPNet, and HRNetV2, many buildings are labeled as backgrounds by mistake, while the proposed approach can accurately extract them. It shows that the MMB-Net can obtain richer context and higher resolution information better than other methods due to that some strategies are utilized in it.

As illustrated in Fig. 10, many adjacent buildings close together are labeled as a whole by the UNet, DeeplabV3+, PSPNet, and HRNetV2, this can be seen from the illustration of marked areas A and D, UNet shows the weakest performance, the two adjacent buildings are misclassified as one, DeeplabV3+, PSPNet, and HRNetV2 achieved better results compared with UNet, but the adjacent building is not completely separated. While our proposed MMB-Net can divide the adjacent buildings well. The reason is that multiple strategies are applied in the MMB-Net to extract rich and accurate high-level semantic with high-resolution features. And the AMFF module also ensures the effective multiscale features fusion.

Obviously, as shown in Fig. 10, compared with the ground truth, the MMB-Net has achieved the most performance of producing homogeneous segmentation maps with accurate boundaries among the all compared methods. Taking marked area D as an example, the building boundary is seriously blurred in the results of UNet, DeeplabV3+, PSPNet, and HRNetV2, while the boundary is clear in the result of MMB-Net. The main reason may be that details losing, insufficient features extraction, or insufficient utilization of feature maps from different level features in the UNet, DeeplabV3+, PSPNet, and HRNetV2. The MMB-Net embeds multiple strategies, including
the parallel multipath feature extraction module and AMFF block to bridge semantic and resolution gaps for capturing the most representative features. Furthermore, in order to solve the problem of segmentation result with poor boundaries aroused by the convolution and pooling operations in MMB-Net, the SEB block is provided for capturing geometric information of building boundary to produce fine-grained segmentations with precise boundaries.

It also can be illustrated by Table II, MMB-Net has the greatest accuracy with 91.14%, 95.95%, 94.98%, and 95.37% with respect to IoU, Precision, Recall, and F1, respectively. Especially, compared with UNet, the IoU, Precision, Recall, and F1 increased by 2.99%, 3.04%, 0.27%, and 1.67%, respectively.

In summary, conclusions can be drawn from Figs. 9 and 10, the proposed MMB-Net outperforms the other compared methods obviously, it exhibits good performance for building extraction, especially for accurately extracting small building, completely extracting large buildings, and exactly refining the boundaries of buildings. The quantitative evaluation results in Tables I and II showed the proposed method achieved the highest IoU. In a word, the superiority of this method is verified both visually and quantitatively.

3) Comparison With Recent Methods: We compare the proposed MMB-Net with the recent building footprint extraction approaches on the WHU aerial building data set, including DE-Net [21], EU-Net [14], MA-FCN [41], MAP-Net [20], and BOMSC-Net [42]. Table III shows the quantitative comparison results of MMB-Net and the recent methods on the WHU aerial building data set, and the highest scores are marked in bold.

DE-Net is a FCN for building extraction, in which the lately segmentation techniques, including downsampling, encoding, compressing, and upsampling, were introduced and the approach achieved about 90.12% in IoU. By applying the dense spatial pyramid pooling and focal loss modules, EU-Net obtained an IoU of 90.56%. MA-FCN achieved a higher IoU of 90.70% than that of EU-Net through embedding multiple scale aggregation of feature pyramids and two postprocessing strategies to refine
the segmentation results. Different from the above-mentioned methods, MAP-Net utilized a multiparallel path to learn spatial localization preserved multiscale features, it slightly outperformed the MA-FCN, and achieved 90.86% in IoU. More recently, BOMSC-Net was designed for building extraction based on the boundary optimization and multiscale context awareness, which achieved 90.15% with respect to IoU.

From Table III, we can see that the IoU, Precision, and F1-score of MMB-Net outperformed all tested methods from the recent studies. Compared with the DE-Net, EU-Net, MA-FCN, MAP-Net, and BOMSC-Net, the proposed approach achieved 1.02%, 0.58%, 0.44%, 0.28%, and 0.99% improvement in IoU. The combination of baseline + P + M achieved an improvement of 2.8%, 2.73%, and 2.99%, respectively, on IOU. The combination of baseline + P + A + M + S, namely MMB-Net, achieved the best accuracies, and improvements on IoU and F1-score over baseline are 4.051% and 2.55%, respectively.

4) Ablation Study: To further evaluate the contributions of different modules combined in the proposed MMB-NET, the ablation experiments were carried out on the WHU aerial building data set and the SpaceNet data set, respectively. As shown in Table VI, we took the UNet with the Conv block [as shown in Fig. 2(b)] encoder as the baseline. The baseline was improved with the parallel multipath feature extraction module, which is denoted as baseline + P. Here, A represents the AMFF module, M is the MFE module, and S denotes the SFE module. Table IV listed the ablation experimental results on the SpaceNet data set, from which we can find that the baseline achieved a 76.97% in IoU and 86.96% in F1-score. When the parallel multipath feature extraction module was introduced into it, the segmentation achieved a 1.17% improvement in IoU and a 0.77% improvement in F1-score, this indicates that more rich high-level semantic information and accurate spatial information can be captured by the parallel multipath feature extraction module rather than the original encoder-decoder network in the baseline. Embedding the AMFF module to the network again achieved 0.34% and 0.22% improvements on IoU and F1-score, respectively. This proved that the AMFF can ensure robust feature fusion by reducing the semantic and resolution gaps between features learned different level and considering different importance of different level features. Furthermore, the applied SFE module improved IoU by 1.57% and F1-score by 0.97% in the extraction results. The reason is that the blurred building boundaries can be refined using the SFE module which can capture the boundary information from low-level features. Adding the MFE module to the baseline + P, it obtained an 0.96% of IoU improvement, which indicated MFE module can enlarge the receptive field of the network and capture more multiscale features simultaneously. At last, baseline + P + A + M + S, namely MMB-Net, achieved the best accuracies, and improvements on IoU and F1-score over baseline are 4.051% and 2.55%, respectively.

As shown in Table V, after adding each module one by one, compared with the baseline, it has increased by 2.04%, 2.54%, 2.8%, 2.73%, and 2.99%, respectively, on IoU. The combination of baseline + P + A + M + S achieved the highest scores on IoU and F1-score on the WHU aerial building data set. The experimental results demonstrate the effectiveness of our proposed methods.

5) Parameter Analysis: To validate the efficiency of the proposed MMB-Net, we also compared training parameters and IoU score of different modules on the WHU data set. As shown in Table VI, the baseline has the lowest complexity but with poor performance. Due to the increase in the number and depth of convolution layers of the network, the number of model parameters of “Baseline + P” has increased by 7.97 M compared with the basic network, but achieved a 2.04% improvement in IoU. The MFE module is another main factor that makes the parameter quantity larger due to the multilevel associative branches with multiscale dilated convolutions. S module can achieve 0.26% improvement in IoU with little computation increase of 0.28 M. Although the increase in the parameters of MMB-Net reaches 14.61 M, it yields a 2.99% improvement on IoU.
IV. CONCLUSION

A novel MMB-Net is proposed in this study to solve the problems existing in the automatic extraction of building footprints from remotely sensed imagery, including blurred boundaries, omission of small buildings, holes in large buildings, and crowded building combination. With the proposed MMB-Net, rich high-level semantic information and accurate spatial details information can be captured by the parallel multipath feature extraction, which allows it to extract building footprints with accurate edges and detect small buildings; the receptive field of network is enlarged and multiscale features are further enhanced by the MFE module; thus, the global dependence information is obtained for suppressing the holes in the large buildings; the AMFF module ensures appropriate aggregation of features by bridging the semantic and resolution gaps between features learned from different levels and considering different importance of different level features; the SEB module captures the boundaries information from low-level features for the refinement of building boundaries. Two experiments are conducted to test the performances of MMB-Net. Compared with existing classic semantic segmentation approaches, MMB-Net is more accurate by visual and quantitative evaluation. Moreover, it outperforms the most recent building extraction methods with higher accuracy on the WHU aerial building data set. Therefore, MMB-Net is an effective approach to extract building footprints from remotely sensed imagery.

In future work, we will investigate how to obtain vector polygons from the extracted building footprints. On the other hand, auxiliary data will be explored to provide supplementary information for the extractions of building footprints.

ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous reviewers whose insightful suggestions have significantly improved this article.

REFERENCES

[1] X. X. Zhu et al., “Deep learning in remote sensing: A comprehensive review and list of resources,” IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, Dec. 2017.
[2] X. Huang, W. Yuan, J. Li, and L. Zhang, “A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 10, no. 2, pp. 654–668, Feb. 2017.
[3] B., Sirmacek, C. Unsalan, “Urban-area and building detection using SIFT keypoints and graph theory,” IEEE Trans. Geosci. Remote Sens., vol. 47, no. 4, pp. 1156–1167, Apr. 2009.
[4] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, “An efficient approach for automatic rectangular building extraction from very high-resolution optical satellite imagery,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 3, pp. 487–491, Mar. 2015.
[5] 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.
[6] J. Ingliada, “Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features,” ISPRS J. Photogrammetry Remote Sens., vol. 62, pp. 236–248, Aug. 2007.
[7] S. Du, F. Zhang, and X. Zhang, “Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach,” ISPRS J. Photogrammetry Remote Sens., vol. 105, pp. 107–119, Jul. 2015.
[8] E. Li, S. Xu, W. Meng, and X. Zhang, “Building extraction from remotely sensed images by integrating saliency cue,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 3, pp. 906–919, Mar. 2017.
[9] Z. Li, W. Shi, Q. Wang, and Z. Miao, “Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 2, pp. 883–899, Feb. 2015.
[10] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3431–3440.
[11] L. C. Chen, S. Wei, and M. Lu, “A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery,” Int. J. Remote Sens., vol. 40, no. 9, pp. 3308–3322, May 2019.
[12] 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.
[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2881–2890.
[14] W. Kang, Y. Xiang, F. Wang, and H. You, “EU-net: An efficient fully convolutional network for building extraction from optical remote sensing images,” Remote Sens., vol. 11, no. 23, Nov. 2019, Art. no. 2813.
[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2818–2826.
[16] 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., Sep. 2018, pp. 833–851.
[17] 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.
[18] K. Rastogi, P. Bodani, and S. Sharma, “Automatic building footprint extraction from very high-resolution imagery using deep learning techniques,” Geocarto Int., vol. 37, pp. 1501–1513, 2022.
[19] J. Wang et al., “Deep high-resolution representation learning for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[20] Q. Zhu, C. Liao, H. Xu, 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.
[21] H. Liu et al., “DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery,” Remote Sens., vol. 11, no. 20, Oct. 2019, Art. no. 2580.
[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. Intervent., 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. Tjahjukh, and J. Liang, “UNet++: Redesigning skip connections to exploit multiscale features in image segmentation,” IEEE Trans. Med. Imag., vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[25] J. Liu, S. Wang, X. Hou, and W. Song, “A deep residual learning serial segmentation network for extracting buildings from remote sensing imagery,” Int. J. Remote Sens., vol. 41, no. 14, pp. 5573–5587, May 2020.
[26] F. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, “RestUNet-a: A deep learning framework for semantic segmentation of remotely sensed data,” ISPRS J. Photogramm. Remote Sens., vol. 162, pp. 94–114, Apr. 2020.
[27] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794–7803.
[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.
[29] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, “Attention branch network: Learning of attention mechanism for visual explanation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 10705–10714.
[30] S. Wu, J. Park, J.-Y. Lee, and S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.
[31] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, “Adaptive pyramid context network for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 7519–7528.
Hua Zhang received the Doctoral degree in cartography and geographical information engineering from the China University of Mining and Technology, Xuzhou, China, in 2012. He is currently an Associate Professor of GIS and Remote Sensing with the School of Environment and Spatial Informatics, China University of Mining and Technology. He has authored or coauthored more than 40 peer-reviewed articles in international journals, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and ISPRS Journal of Photogrammetry and Remote Sensing. His current research interests include multi/hyperspectral and high-resolution remotely sensed images processing, uncertainty in classification, pattern recognition, and remote sensing applications.

Nanshan Zheng received the Ph.D. degree in urban environment from Kyoto University, Kyoto, Japan, in 2009. He is currently a Professor with the School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, China. He has authored or coauthored more than 60 peer-reviewed journal or conference articles. His current research interests include GNSS-R, signal processing, and remote sensing of environment and disaster.

Wenzhong Shi received the Doctoral degree in philosophy from the University of Osnabrück, Vechta, Germany, in 1994. He is currently a Professor of GIS and Remote Sensing with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong. He has authored or coauthored more than 100 SCI papers and 10 books. His current research interests include GIS and remote sensing, uncertainty and spatial data quality control, and image processing for high-resolution satellite images.

Prof. Shi was the recipient of the State Natural Science Award from the State Council of China in 2007 and The Wang Zhizhuo Award from the International Society for Photogrammetry and Remote Sensing in 2012.