HCRB-MSAN: Horizontally Connected Residual Blocks-Based Multiscale Attention Network for Semantic Segmentation of Buildings in HSR Remote Sensing Images

Zhen Li, Zhenxin Zhang, *Member, IEEE*, Dong Chen, *Member, IEEE*, Liqiang Zhang, Lin Zhu, Qiang Wang, Siyun Chen, and Xueli Peng

***Abstract*—Accurate and efficient semantic segmentation of buildings in high spatial resolution (HSR) remote sensing images is the basis for applications such as fine urban management, high-precision mapping, land resource utilization investigation, and human settlement suitability evaluation. The current building extraction methods based on deep learning can obtain high-level abstract features of images. However, due to the limitation of convolution kernel size and the vanishing gradient, the extraction of some buildings is inaccurate, and some small-volume buildings are missing as the network deepens. In this regard, we design a horizontally connected residual blocks-based multiscale attention network to achieve high-quality extraction of buildings in HSR remote sensing image. In this network, we subdivide each residual block by channel grouping and feature horizontal connection to consider the difference and saliency of feature information between channels, and then combine the output features with multiscale attention module to consider the contextual semantic relationship of different regions and integrate multilevel local and global information of buildings. A stepwise up-sampling mechanism is designed in the decoding process to finally achieve precise semantic segmentation of buildings. We conduct experiments on two public datasets and compare the proposed method with state-of-the-art semantic segmentation methods. The experiments show that our method could achieve better building extraction results in HSR remote sensing image, which proves the effectiveness of our proposed method.**

***Index Terms*—Building semantic segmentation, deep learning, horizontally connected residual block, high spatial resolution (HSR) remote sensing image, multiscale attention.**

Manuscript received 30 November 2021; revised 24 April 2022; accepted 28 June 2022. Date of publication 5 July 2022; date of current version 20 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071445 and Grant 41971415, in part by the Beijing Natural Science Foundation under Grant 8212023, in part by the General scientific research projects of Beijing Municipal Commission of Education under Grant KM202010028012, in part by the National Science Fund for Distinguished Young Scholars under Grant 41925006, and in part by the Beijing Outstanding Young Scientist Program under Grant BJJWZYJH01201910028032. *(Zhen Li, Zhenxin Zhang, and Dong Chen contributed equally to this work.) (Corresponding authors: Zhenxin Zhang; Dong Chen.)*

Zhen Li, Zhenxin Zhang, Lin Zhu, Siyun Chen, and Xueli Peng are with the Key Laboratory of 3D Information Acquisition and Application, MOE, Capital Normal University, Beijing 100048, China, and also with the College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China (e-mail: sdlz123@126.com; zhenxin066@163.com; hi-zhulin@163.com; csiyun_hb@163.com; xuelipeng@cnu.edu.cn).

Dong Chen is with the College of Civil Engineering, Nanjing Forestry University, Nanjing 210037, China (e-mail: chendong@njfu.edu.cn).

Liqiang Zhang is with the Geographical Science, and the State Key Laboratory of Remote Sensing Science, Beijing Normal University, Beijing 100875, China (e-mail: zhanglq@bnu.edu.cn).

Qiang Wang is with the School of Geographic and Environmental Sciences, Tianjin Normal University, Tianjin 300387, China (e-mail: wangqiang_study@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3188515

I. Introduction

With the development of aviation and aerospace remote sensing technology, earth observation capabilities have been gradually improved, and people can conveniently obtain large-scale high spatial resolution (HSR) earth observation image data, which contains a large amount of building detail information. Using HSR remote sensing images for rapid and efficient extraction of buildings is the basis for land resource management [1], fine mapping [2], land use change monitoring [3], human settlement suitability assessment [4], and so on. However, HSR remote sensing images also bring some problems such as large amounts of calculation, complex calculation process, and partial information redundancy. Additionally, the structural complexity, large differences in distribution, and surrounding complexity of buildings also cause certain difficulties and challenges to the efficient extraction of buildings in HSR remote sensing images.

The building semantic segmentation from HSR remote sensing images is to label each pixel according to whether the pixel belongs to the type of building. How to efficiently obtain building semantic information from HSR remote sensing image is the foundation and key of its application. Currently, the building extraction algorithms can be classified into the traditional feature-based methods and the deep learning feature-based methods. In the traditional feature-based methods, some researchers have proposed many building extraction algorithms [5]–[10], but most of these algorithms depend on manually designed features, such as geometry [5], texture [6], shading [7], and edge [10]. Besides, the support vector machine [5], AdaBoost [6], conditional random field (CRF) [8], and random forest [9] are also usually employed to label each pixel. However, the complex appearance and spectral information of buildings are easily confused with other categories in HSR remote sensing image. Moreover, different building materials, volumes, and

lighting conditions may also have obvious differences in remote sensing images, which create more difficulties to efficiently label the semantics of buildings. The traditional feature-based methods have some limitations on representing the features of buildings with complex spatial distributions and patterns, and the generalization ability of the traditional model needs to be further improved.

Deep learning can obtain high-level abstract features from spatial data by a multilevel structure to improve classification or detection accuracy [11]. The performance of deep learning feature surpasses and gradually replaces the traditional artificially designed features (e.g., SIFT [12], FAST [13], SURF [14], and ORB [15]) under the background of big data. Based on the deep learning, some researchers apply deep learning technology in many fields, such as the land use change [16], resource management [17], etc. In terms of building extraction in remote sensing images, researchers also try to improve the extraction effect of building feature. Fully convolutional network (FCN) [18] is one of the deep learning semantic segmentation models commonly used [19], which can achieve end-to-end object extraction results. Some researchers also combine multisource data [20], [21], multilayer training sample [22], and CRF in postprocess procedure [23] to boost the effects of building extraction. Besides, the attention mechanism can consider the difference of different regions in feature map [24], [25]. The deep learning methods of extracting building in HSR remote sensing image have not considered the differences, salience, and multilevel fusion enhancement between different channels within the model, and the attention mechanism in network also does not consider the multiscale information of buildings and cannot better solve the problem of extraction buildings with different sizes.

As buildings in HSR remote sensing images have the characteristics of discrete distribution, complexity, different sizes, and multiple details, the traditional semantic segmentation methods are not completely applicable to buildings semantic segmentation of HSR remote sensing images. In this research, we first integrate the structure of channel grouping and horizontal connection and combine them with a multiscale attention mechanism to design a novel deep learning network, which performs the pixel-based semantic segmentation model of remote sensing images for building extraction. The main contributions of the research are summarized as follows.

1) We design a horizontally connected residual blocks-based module of building feature representation in HSR remote sensing images to make the network focus on the feature information enhancement between different channel groups, thereby consolidating building category information and small targets recognition.
2) A novel multiscale attention structure is constructed according to the characteristics of remote sensing images, so that the extracted building features can be integrated with different scales of the features to improve the accuracy of building extraction in HSR remote sensing images.
3) Aiming at the problem of losing information and blurred segmentation boundaries in the decoding process, we propose a horizontally connected residual blocks-based multiscale attention network by combining the shallow layer of features with high spatial resolution stepwise during the decoding process. This network fuses the information between different channel groups and the attention mechanism of different scales to extract building features and increases the weight of small target of buildings, thereby enhancing semantic segmentation of building results.

II. Related Work

In this part, we discuss the related work of building segmentation in remote sensing images, including traditional feature-based building segmentation and deep feature-based building segmentation.

A. Traditional Feature-Based Building Segmentation

The mathematical and geometric relationships between the line features of buildings are considered to extract buildings [26]. And then, Wang *et al.* [27] adopt filter to enhance building edge contrast and extract line segments and present a graph search-based perceptual grouping approach to extract buildings. Additionally, other kind of data source (such as airborne laser scanning data) is combined to assist building extraction in remote sensing image [28]. Some researchers also propose specific algorithms based on building features to improve the building extraction effect. For instance, an algorithm based on the differential morphological profile is designed [29], [30] to construct image profile and extract the buildings by morphological opening and closing operations while varying the size of structuring element. Lee *et al.* [31] use the classification results of IKONOS multispectral images to provide approximate location and shape, and the fine building extraction is carried out through segmentation based on the iterative self-organizing data analysis technique. A novel CRF formula, which incorporates pixel-level information and segment-level information, uses regional consistency and shape features to extract buildings in [8]. However, the features of these methods are based on some rules and cannot achieve end-to-end building extraction results, the representative ability of the feature should be further improved.

B. Deep Feature-Based Building Segmentation

At present, target extraction algorithms have evolved from manually designed feature-based methods to deep feature-based methods. Developed from convolutional neural network (CNN), FCN [18] realizes end-to-end and pixel-level image segmentation for the first time. On this basis, many research works have proposed some modified structures, such as the atrous spatial pyramid pooling [32], pyramid pooling module [33], and encoder–decoder with atrous separable convolution [34]. However, limited by the fixed size of the convolution kernel and the lack of sufficient image context information, these methods should be further improved on the representation of object feature. Recently, some researchers introduce the attention mechanism into many research fields, e.g., machine translation [35], pose estimation [36], image processing [37], [38], video understanding [39], and target tracking [40], [41]. The attention mechanism can capture the context information in the spatial dimension [42] or the channel dimension [43].

Benefited from the progress of deep learning in computer vision, many scholars have proposed some methods for semantic segmentation of buildings in remote sensing image. Some researchers use traditional building features to assist deep learning feature extraction. For example, in order to improve the performance of building segmentation, the symbolic distance function of building boundary is introduced as the output in [44], to enhance representation power of deep learning. To obtain a refined building boundary, Xie *et al.* [45] introduce morphological filtering to enhance the regularity of the boundary under pixel-level segmentation of buildings using CNN. Xu *et al.* [46] preprocess the image by the way of edge enhancement to highlight the pixels at the edge of buildings. Combining manually designed features such as the normalized difference vegetation index and normalized digital surface model, they adapt guided filter to optimize the classification map to remove salt-and-pepper noise. Besides, some modified structures of deep learning are also designed in building segmentation, e.g., Yue *et al.* [47] propose an adaptive network and tree-CNN blocks according to the confusion matrix and the tree-cutting algorithm, to fuse multiscale features and learn the optimal weights of the model; Alshehhi *et al.* [48] design a single patch-based CNN to extract features and combine low-level features with convolutional features in the postprocessing stage; Zhou *et al.* [49] design a feature decoupling module to encode the class co-occurrence relations in the scene, thus improving the segmentation performance. To improve the spatial resolution, Ji *et al.* [50] introduce an atrous/dilated convolution in FCN and combine the hierarchical building features extracted by the network in the decoding stage. To refine discontinuous building footprints, Zhu *et al.* [25] design multiple parallel paths to learn multiscale features and construct the pyramid spatial pooling module in network.

III. Proposed Method

This part mainly introduces our proposed horizontally connected residual blocks-based multiscale attention model for semantic segmentation of buildings in HSR remote sensing image, including the overall network structure and each part of the network (e.g., the horizontally connected residual block structure, the multiscale attention module, and the stepwise up-sampling module).

A. Overall Framework

The deep features contain more abstract information, and the shallow features contain rich spatial details. The traditional ResNet [51] can gradually extract features from the shallow level to the deep level, while these features are transmitted linearly, without considering the complementary and fusion relationship between different channels. With the deepening of the network layer, the phenomenon of vanishing gradient often occurs, which results in the loss of some useful information. Presently, most semantic segmentation networks directly decode the final feature map into a prediction map, which can easily lose spatial details. Aiming at this problem, we propose the horizontally connected residual blocks-based multiscale attention network (HCRB-MSAN) (see Fig. 1) that integrates high-level and low-level feature information, in which the hierarchy and complementarity between different channels are considered to construct the contextual semantic feature. The HCRB-MSAN includes the following three parts:

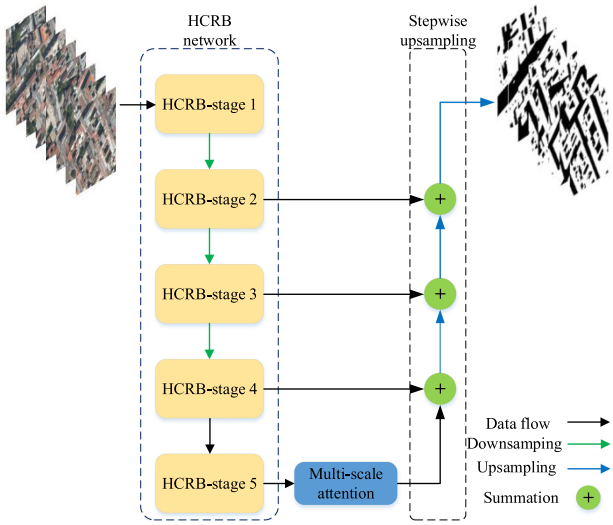


Fig. 1. Overview of the HCRB-MSAN.

1) the horizontally connected residual blocks based on channel grouping, which enables the network to integrate the feature information between different channel groups when extracting features;
2) the multiscale spatial attention module, which gives context information to the features obtained by the HCRB network;
3) the stepwise up-sampling part, in which the low-level features containing rich spatial details are merged during decoding process to obtain prediction results.

As shown in Fig. 1, the channel grouping-based horizontally connected residual block network first extract the high-level and low-level features of the input remote sensing image. The same as the traditional ResNet structure, our designed HCRB network also has five stages for feature extraction, and different stages contain multiple horizontally connected residual blocks (as shown in Fig. 2). However, in each stage of network, the horizontally connected residual block further subdivides the feature map channels into N groups of channels ($N = 4$ in this research) without changing the spatial size of each channel and merge the features between different channel groups through the horizontal connection structure to obtain the features in different scales of receptive fields and realize the joint extraction of global and local features. Then, the feature map is input into the multiscale attention module (see Fig. 3) to construct the feature pixel-level contextual semantic information. Finally, in order to maintain more discriminative spatial detail information, we stepwise fuse different stages of low-level features during decoding process by using bilinear interpolation and employ the rich building semantic information of the high-level feature to generate the rich spatial details of the low-level feature to predict the final building map.

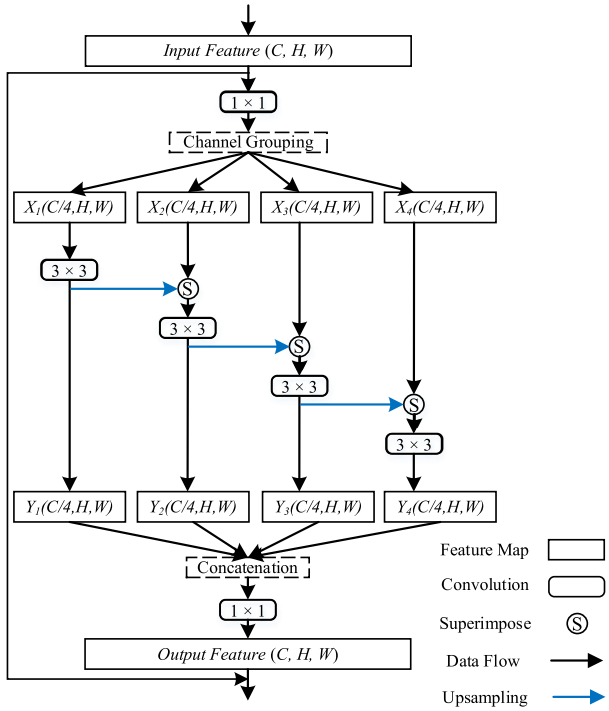


Fig. 2. Channel grouping horizontally connected residual block.

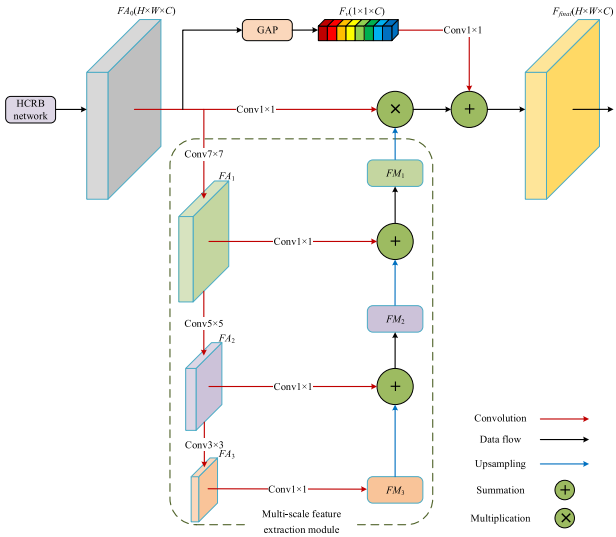


Fig. 3. Multiscale attention module. Each convolution operation is followed by the operation of batch normalization and ReLU activation function.

B. Network Framework

1) Horizontally Connected Residual Block: Previously, many scholars have conducted research on the channel grouping of features [52]–[54], grouped convolutions have been proven to be a successful approach for increasing the performance of network [38], in which, a more flexible way of convolution (multibranch convolution) is proposed. But it did not consider the correlation of features between different channel groups. On this basis, inspired by the Res2Net [55], we first integrate the structure of channel grouping and horizontal connection into the building segmentation of HSR remote sensing images by the construction of HCRB and the extraction of discriminative features of buildings in HSR remote sensing images.

In the HCRB, we evenly group all channels of the feature maps after the operation of 1×1 convolution and perform a convolution process on each group of feature maps. At different stages of the feature extraction network, feature maps have a different number of channels, unlike that the Res2Net manually sets the number of channel groups in residual block and the number of channels in each group, our HCRB structure just manually set the number of channel groups, and the number of channels in each group are divided equally by the feature map channels. In other words, the number of channels in each group will change according to the input feature map channels. Thus, the HCRB structure can maintain more feature information of buildings, and only has one hyperparameter, which leads to a more efficient computation. We also use Res2Net to replace HCRB in the proposed network (named as Res2-MSAN) to obtain the effects of these two blocks (the statistical results are shown in Tables IV–VI) to verify the advantage of our method.

For more details of the HCRB, except for the first group, the input features in each subsequent group are superimposed with the horizontally transmitted convolution features of the previous group. We repeat this operation to the last group of feature maps and finally superimpose the feature maps of all groups sequentially to execute the 1×1 convolution operation to achieve feature fusion of different scales of buildings.

Specifically, the size of feature map in the residual block is set as $H \times W$, and there is a total of C channels. All the feature map channels are equally divided into N subsets of channels ($N = 4$ in Fig. 2). Compared with the original feature map, the feature map in each subset has the same size ($H \times W$) with $\frac{C}{N}$ channels.

We denote the obtained feature map of the ith channel subset as $\mathbf{X_i}$ ($\mathbf{X_i} \in R^{H \times W \times \frac{C}{N}}$, $i = 1, 2, \ldots, N$). For the feature maps of N subsets of channels, the 3×3 convolution kernel are used to perform the convolutional operation, which is represented as $\phi_{3\times3}(\cdot)$ in (1). The $\mathbf{Y_i}$($\mathbf{Y_i} \in R^{H \times W \times \frac{C}{N}}$) of (1) is the output feature map corresponding to the ith channel subset. Except for $\mathbf{X_1}$, each subsequent feature map $\mathbf{X_i}$ ($i = 2, \ldots, N$) is superimposed with the output of the previous feature map [e.g., the $\mathbf{Y_{i-1}}$ in (1)] before performing convolutional operation. Thus, $\mathbf{Y_i}$ can be expressed by the following equation:

$$Y_i = \begin{cases} \phi_{3\times3}(X_i) & i = 1; \\ \phi_{3\times3}(X_i + Y_{i-1}) & 2 \le i \le N. \end{cases} \tag{1}$$

The channels of feature map were further subdivided by channel grouping, the convolution operation of each group of channels can extract the corresponding feature map of each previous channel subset. In this way, by performing such channel grouping and horizontal connection operation in each residual block, different sizes of receptive fields can be combined to obtain multiscale features, which can not only extract the information of different channels, but also realize the joint extraction of global and local features, to achieve robust detection of the scattered building targets.

2) Multiscale Attention Module: Recently, most semantic segmentation networks directly use multilayer linear convolutional networks to extract image features, but multiple

TABLE I
ABLATION EXPERIMENTS OF THE NETWORK STRUCTURE WITH THE BASELINE OF RESNET50

Method	Dataset	Stepwise up-sampling	Multi-scale attention	F_1-score (%)	Building IoU (%)	OA (%)
Baseline (ResNet50)	WHU/ INRIA			90.25/82.81	82.24/70.66	97.89/95.45
MSAN	WHU/ INRIA	√		94.75(↑4.50)/86.45(↑3.64)	90.02(↑7.78)/76.14(↑5.48)	98.84(↑0.95)/96.39(↑0.94)
MSAN	WHU/ INRIA		√	92.54(↑2.29)/84.15(↑1.34)	86.11(↑3.87)/72.64(↑1.98)	98.36(↑0.47)/95.77(↑0.32)
MSAN	WHU/ INRIA	√	√	95.00(↑4.75)/86.50(↑3.69)	90.48(↑8.24)/76.20(↑5.54)	98.90(↑1.01)/96.42(↑0.97)

Note: The symbol "↑" represents the increased value relative to the baseline, similarly hereinafter.

TABLE II
ABLATION EXPERIMENTS OF HORIZONTAL CONNECTION RESIDUAL BLOCK

Method	Dataset	F_1-score (%)	Building IoU(%)	OA(%)
Baseline (Res-MSAN)	WHU /INRIA	95.00 /86.50	90.48 /76.20	98.90 /96.42
Baseline + CG	WHU /INRIA	95.25 (↑0.25) /86.62(↑0.12)	90.92(↑0.44) /76.40(↑0.20)	98.96(↑0.06) /96.42(-0.00)
Baseline + CG + HC (ours)	WHU /INRIA	95.41(↑0.41) /86.90(↑0.40)	91.22(↑0.74) /76.90(↑0.70)	99.00(↑0.10) /96.61(↑0.19)

TABLE III
ABLATION EXPERIMENTS OF DIFFERENT CHANNEL GROUP NUMBERS

Our method	Dataset	F1-score (%)	Building IoU(%)	OA(%)
1 subset	WHU /INRIA	95.00 /86.50	90.48 /76.20	98.90 /96.42
2 subsets	WHU /INRIA	95.13 (↑0.13) /86.64 (↑0.14)	90.70 (↑0.22) /76.43 (↑0.23)	98.94(↑0.04) /96.41(↓0.01)
4 subsets	WHU /INRIA	**95.41 (↑0.41) /86.90 (↑0.4)**	**91.22 (↑0.74) /76.90(↑0.70)**	**99.00(↑0.10) /96.61(↑0.19)**
8 subsets	WHU /INRIA	94.85 (↓0.15) /86.77(↑0.27)	90.20(↓0.28) /76.63(↑0.43)	98.87(↓0.03) /96.42(-0.00)
16 subsets	WHU /INRIA	94.68(↓0.32) /86.19(↓0.31)	89.90(↓0.58) /75.73(↓0.47)	98.85(↓0.05) /96.29(↓0.13)

The bold values represent the best results in ablation experiments.

convolution operations may reduce the spatial detail information of feature maps, resulting in the blurred segmentation boundaries, aliasing, and lack of the extraction in the significant context information.

To solve this problem, inspired by the attention mechanism in images processing [43], [56], we design a multiscale attention module (see Fig. 3) following the last stage of HCRB network to fully use the significant context information, and the extracted building features can also be integrated with different scales of the features. Compared with the feature map of other stage of HCRB network, the output feature map (e.g., the $\mathbf{FA}_0$in Fig. 3) of network contain more abstract high-level information. Later, we extract features $\{\mathbf{FA_i}|i = 12, 3\}$ of different scales step by step through the operation of three convolutions. For the feature map $\mathbf{FA_i}$, we set different sizes of kernel, padding, and stride in the convolution operation to obtain multiscale information. Finally, the height and width of feature map $\mathbf{FA_i}$ are $1/2^i \times H$ and $1/2^i \times W$, respect to the $\mathbf{FA}_0$ ($H \times W$), and the implementation details are shown in the following equation:

$$FA_i = \phi(FA_{i-1}|K_{9-2i}, P_{4-i}, S_2) \tag{2}$$

where $\phi(\cdot)$ denotes the convolution operation for the feature $\mathbf{FA_i}$ ($i = 12$ and 3), *K*, *P*, and *S* denote the kernel, padding, and stride operation, respectively. The subscripts of *K*, *P*, and *S* (like 9-2*i*, 4-*i*, and 2), respectively, denotes the kernel size, padding size, and stride size.

Inspired by the Unet [57] structure, we design the down-top pathway and horizontal connections to generate multiscale features $\{\mathbf{FM_i}|i = 12, 3\}$, and the procedure is as follows:

$$FM_i = \begin{cases} \sigma(\zeta(FA_i) + \varphi(FM_{i+1})) & i = 12, \\ \sigma(\zeta(FA_i)) & i = 3 \end{cases} \tag{3}$$

where $\zeta(\cdot)$ denotes the horizontal connection, implemented by using a 1×1 convolution . The $\sigma(\cdot)$ denotes the operation of batch normalization and ReLU activation function, and $\varphi(\cdot)$ denotes the transposed convolution. By this down-top pathway and horizontal connections, the multiscale feature extraction module can reduce the loss of information and aggregate the contextual semantic information of different scales and make the features more prominent.

We sample $\mathbf{FA}_0$ as a one-dimensional (1-D) vector $\mathbf{F_v}$ through a global average pooling to obtain coarser channel global information, and then introduce a 1×1 convolution to achieve information fusion results of channels. Meanwhile, after performing the 1×1 convolution on $\mathbf{FA}_0$, we multiply $\mathbf{FM}_1$ with $\mathbf{FA}_0$ to endow it with the contextual information weight. Finally, the two results are added together to obtain the salient features containing global context information. The 1-D vector $\mathbf{F_v}$ and output feature $\mathbf{F_{final}}$ can be calculated by the following equations:

$$F_v = \frac{1}{H \times W} \sum_{i=1, j=1}^{H, W} FA_0(i, j), \tag{4}$$

$$F_{\text{final}} = \sigma(\phi_{1\times1}(FA_0)) \times \varphi(FM_1) + \sigma(\phi_{1\times1}(F_v)) \tag{5}$$

where H and W are, respectively, the height and width of feature map $\mathbf{FA}_0$, *i* represents the *i*th row of pixels in $\mathbf{FA}_0$ (*i* = 1, 2, …, *H*), and *j* is the *j*th column of pixels in $\mathbf{FA}_0$(*j* = 1, 2, …, *W*). The $\phi_{1\times1}(\cdot)$ denotes the 1×1 convolution operation, and the meanings of $\sigma(\cdot)$ and $\varphi(\cdot)$ have been denoted in the following equation:

3) Stepwise Up-Sampling Decoder Structure: Some semantic segmentation networks, such as FCN [18] and PSPNet [33], directly perform up-sampling operation when decoding feature maps into prediction maps. Such decoding methods are prone to lose spatial details and affect the final prediction results. Inspired by the U-Net [57] structure, we design an improved way to maintain more feature details. As shown in Fig. 1, in the decoding process, the features of the multiscale spatial attention module are gradually added to the output features of the

TABLE IV
PERFORMANCE OF EACH METHOD ON THE WHU BUILDING DATASET

Method	Precision (%)	Recall (%)	F1-scores (%)	Building IoU (%)	Background IoU(%)	mIoU (%)	OA (%)
U-Net	92.62	93.50	93.06	87.02	98.29	92.65	98.47
ResNet50	90.09	93.35	91.69	84.66	97.93	91.29	98.14
PSPNet	93.41	91.95	92.68	86.35	98.22	92.29	98.40
DeepLabV3	91.80	93.13	92.46	85.98	98.14	92.06	98.33
DANet	94.01	89.71	91.81	84.86	98.05	91.45	98.24
PAN	92.80	91.49	92.14	85.43	98.09	91.76	98.28
SiU-Net	93.80	93.90	93.85	88.40	-	-	-
MA-FCN	94.50	94.20	94.30	89.50	-	-	-
EaNet	94.63	**96.09**	**95.35**	**91.11**	-	-	-
SRI-Net	95.21	93.28	94.23	89.09	-	-	-
Res2-MSAN	96.12	93.63	94.86	90.22	**98.76**	**94.49**	**98.88**
HCRB-SENet	**96.31**	93.15	94.70	89.94	98.72	94.33	98.85
HCRB-MSAN	**96.78**	**94.68**	**95.72**	**91.79**	**98.96**	**95.37**	**99.07**

The bold values with shading represent the best results in comparative experiments and the bold values represent the second best results in comparative experiments.

TABLE V
PERFORMANCE OF EACH METHOD ON THE INRIA DATASET

Method	Precision (%)	Recall (%)	F1-scores (%)	Building IoU (%)	Background IoU(%)	mIoU (%)	OA (%)
U-Net	82.73	80.09	81.39	68.61	94.47	81.54	95.06
ResNet50	85.09	82.13	83.58	71.79	95.11	83.45	95.65
PSPNet	81.63	86.55	84.02	72.45	94.97	83.70	95.56
DeepLabV3	82.08	82.31	82.20	69.77	94.59	82.18	95.19
DANet	82.66	84.37	83.51	71.68	94.93	83.31	95.51
PAN	90.11	77.02	83.05	71.02	95.27	83.15	95.76
SU-Net	84.30	84.90	84.60	73.30	-	-	-
FPCRF	-	-	**87.65**	74.79	-	-	95.81
SRI-Net	85.77	81.46	83.56	71.76	-	-	-
Res2-MSAN	86.12	**86.23**	86.17	75.71	95.78	85.74	96.27
HCRB-SENet	**89.49**	83.41	86.34	**75.96**	**95.99**	**85.98**	**96.44**
HCRB-MSAN	**89.56**	**88.13**	**88.84**	**79.92**	**96.61**	**88.26**	**97.01**

The bold values with shading represent the best results in comparative experiments and the bold values represent the second best results in comparative experiments.

TABLE VI
DIFFERENCES BETWEEN EACH COMPARED METHOD AND OUR METHOD

Method	Channel grouping	Attention	Multi scale attention	Stepwise decoding
U-Net	×	×	×	√
ResNet50	×	×	×	×
PSPNet	×	×	×	×
DeepLabV3	×	×	×	×
DANet	×	√	×	×
PAN	×	√	√	√
SiU-Net	×	×	×	√
MA-FCN	×	×	×	√
EaNet	×	×	×	√
SRI-Net	×	×	×	√
SU-Net	×	×	×	√
FPCRF	×	×	×	×
HCRB-MSAN (Ours)	√	√	√	√

upper HCRB stage. The rich semantic information of high-level features is combined with the spatial information of the shallow features to represent the characteristics of small building detail and boundary.

IV. EXPERIMENTS

To verify the proposed method, we conduct ablation experiments of the network structure to test the performance of each part in the proposed method, and then analyze the applicability of the method on multisource data. The method is also compared with the other state-of-the-art methods to further verify the ability of the proposed network.

We set the input batch size, weight decay, and initial learning rate as 6, 0.0001, and 0.5, respectively. Considering the time efficiency factor, we set 120 epochs in ablation experiments to reduce the time costs, 300 epochs in comparative experiments to get full training model and best results. The stochastic gradient descent optimization method and cross entropy loss function are employed to train the model. During the training process, we perform random horizontal flip and rotation between positive and negative 15° to enhance the training data. The parameters size of the network is 65.39 MB.

A. Data and Hardware Environment

In the experiments, we use two datasets to verify the proposed method. 1) WHU building dataset [58]. The dataset contains two subsets of remote sensing image (aerial and aerospace images). We select aerial image subset to verify the proposed method. The original aerial image data comes from the New Zealand land information service website, located in Christchurch, New Zealand, which includes 187 000 buildings and a total of 8188 pictures. We choose 4736 images as the training set, 1036 images as the validation set, and 2416 images as the test set. Each image has the size of 512 × 512 pixels and a spatial resolution of 0.3 m, including three bands (red, green, and blue). 2) INRIA aerial image labeling dataset [59]. The dataset is provided by INRIA, which covers different urban settlements, including

Austin, Chicago, Kisap County, West Tyrol, and Vienna. The five areas have different building densities with a spatial resolution of 0.3 m. The original dataset only contains 180 labeled images with 5000 × 5000 size. Considering the impact of computer hardware performance, we divide the images into 18 000 blocks in advance and each has 500 × 500 pixels. We set the training data and validation data as the ratio of 8:2.

The hardware environment configuration is as follows: GeForce Titan Xp GPU containing 12GB memory with the speed of 11.4 Gbps and 3840 CUDA cores. The operating frequency is 1.6 GHz, and Ubuntu 16.04 is used in the experiments. The required software packages for the experiments include Python 3.6, CUDA 9.0, cuDNN 7, Pytorch 1.1.0, OpenCV 3, Pandas, and NumPy.

B. Evaluation Metrics

The proposed building extraction network determines each pixel of the input remote sensing image whether it belongs to the class of building or not. Therefore, we adopt the current popular pixel-level evaluation index to quantitatively evaluate our network performance, including intersection over union (IoU), recall (R), precision (P), F_1-score, and overall accuracy (OA). The specific formulas are as follows:

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

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

$$F_1-\mathrm{score} = 2\times\frac{R\times P}{R+P} \tag{8}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{9}$$

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \tag{10}$$

where TP (true positive) represents the number of building pixels that are correctly extracted; TN (true negative) represents the number of nonbuilding pixels that are correctly extracted; FP (false positive) represents the number of pixels predicted as buildings but actually nonbuildings, which is also called false detection; FN (false negative) represents the number of pixels predicted as nonbuildings but actually buildings, which is also called the missed detection. Precision represents the percentage of building pixels that are detected correctly in all pixels detected as buildings. Recall represents the percentage of building pixels that are detected correctly in all real building pixels. We need to weigh the two indicators (precision and recall) comprehensively, which leads to another indicator F_1-score. This is a comprehensive consideration of the harmony value of precision and recall; the IoU is the ratio of intersection and union of pixels detected as buildings and actual buildings, and the OA represents the overall pixel accuracy of all categories.

C. Experiment Analysis

1) Ablation Experiments

a) Network module: In this section, to verify the performance of the stepwise up-sampling decoder structure and multiscale attention structure in the proposed network, we set the ResNet50 as the baseline to conduct the ablation experiments. For the WHU building dataset, the experimental results are shown in Table I. We can see that, just using the stepwise up-sampling decoding structure (as shown in Fig. 1) can obtain 94.75% of F_1-score, 90.02% of IoU, and 98.84% of OA, which are 4.50%, 7.78%, and 0.95% higher than the baseline, respectively. Besides, the method of just using the multiscale attention structure are 2.29%, 3.87%, 0.47% higher than the baseline separately. Meanwhile, when we integrate the two structures together, our network performance is further improved, and the F_1-score, IoU, and OA increase by 4.75%, 8.24%, and 1.01% compared with the baseline, respectively. All of indexes are better than the network just using an independent structure (the stepwise up-sampling or the multiscale attention). We also obtain improvement effects on the INRIA dataset. The results illustrate the effectiveness of the stepwise up-sampling and the multiscale attention, as the proposed structures can combine high-level semantic features with low-level spatial features and give contextual semantics to high-level features through the attention module, which improves the extraction effects of building feature.

Additionally, we use squeeze-and-excitation block of SENet [38] to replace the multiscale attention module in our proposed model (named as HCRB-SENet) to obtain the effects of these two attention mechanisms. The statistical results of comparison between the two attention mechanisms are shown in Tables IV–VI. Experimental results prove the advantage of our method. We reckon the reason is that the task only has two segmentation targets (building or nonbuilding) in the building extraction of HSR remote sensing images, more channel number may produce the disturbance to the extraction of buildings. Our multiscale attention module can extract different levels of information combined with context semantics and effectively acquire the spatial relationship of objects in HSR remote sensing images, so that the effects of spatial attention mechanism (MSAN in our model) are better than channel attention mechanism of SENet.

b) Horizontal connection residual block: In this section, we test the effectiveness of the horizontal connection residual block based on the stepwise up-sampling decoding structure and multiscale attention module. As shown in Table II, we modify the ResNet50 with the stepwise up-sampling decoding structure and the multiscale attention structure as the baseline (Res-MSAN) and explore the influence of HCRB structure on the network on the aspects of channel grouping (CG) and horizontal connection (HC).

1) *Channel grouping:* In the experiments, we averagely divide the channels into four groups. For the WHU building dataset, we can see from Table II that the building extraction effect is improved when we group the channels, with F_1-score increased by 0.25%, IoU increased by 0.44%, and OA increased by 0.06% in comparison with the baseline. For the INRIA dataset, grouping channels can obtain 86.62% of F_1-score, 76.40% of IoU, and 96.42% of OA, which are 0.12%, 0.20% higher than the baseline in F_1-score and IoU. Experiments on these two datasets prove the channel grouping is effective.

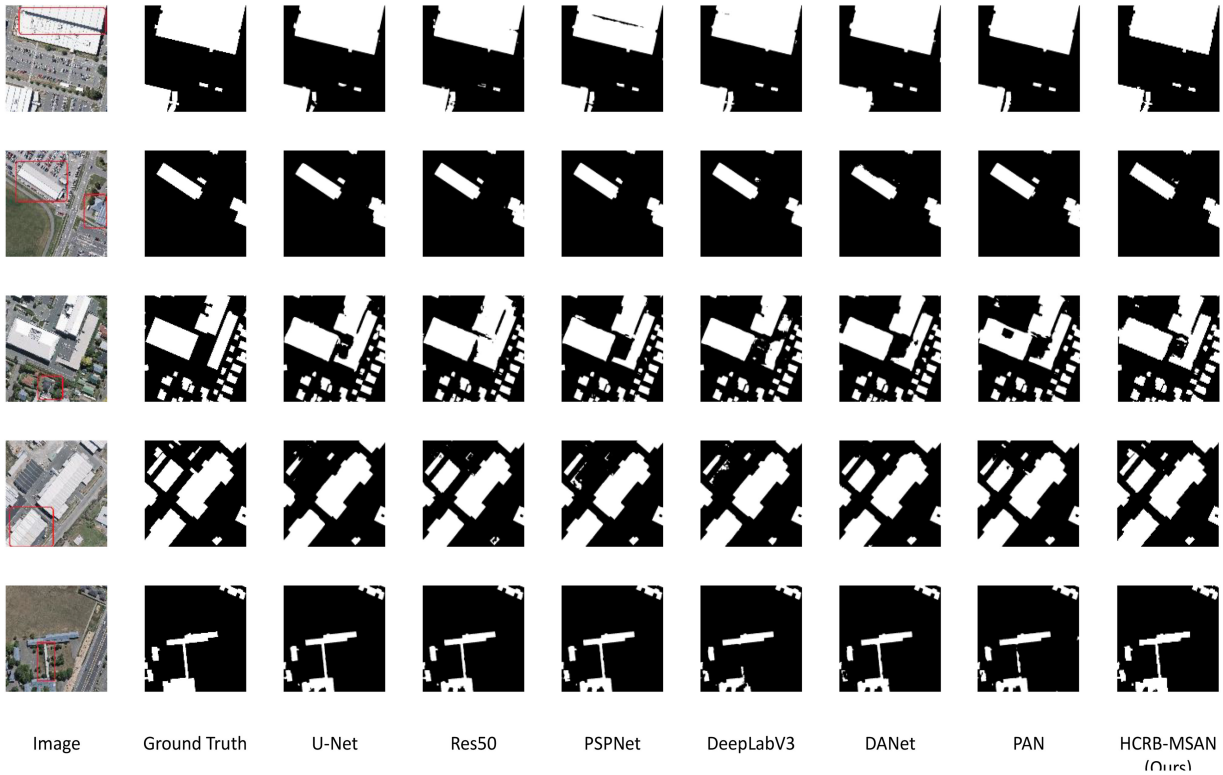


Fig. 4. Visual comparison results of each method on WHU building dataset.

2) *Horizontal connection:* Channel grouping focuses on the ability enhancement of different channel groups to extract feature maps, but does not integrate the information between channel groups, and the features of different channel groups are isolated from each other. To extract prominent features of buildings, we introduce a horizontal connection structure to the grouped feature maps, to superimpose the features between the channel groups and realize the fusion of global and local features. It can be seen from Table II that the effect of building segmentation is improved, reaching to 95.41% of F_1-score, 91.22% of IoU, and 99.00% of OA on WHU building dataset, and 86.90% of F_1-score, 76.90% of building IoU, and 96.61% of OA on INRIA dataset, when we add the horizontal connection structure, which exceeds the effect of only employing the channel grouping. This proves that the horizontal connection structure can effectively improve the network performance of building extraction.
3) *Different channel groups:* We divide the feature map channels into 1, 2, 4, 8, 16 subsets to test the influence of different group numbers on building segmentation of HSR remote sensing images. As shown in Table III, we can see that the building segmentation accuracy basically increases with the augment of channel group number, and the better results can be obtained when the channel group number is equal to 4. When the channel group number further increases, the building segmentation effects decrease, because the task of building segmentation only has two targets (building or nonbuilding), and the excessive number of channel group limits the extraction of category difference information [38]. The effect of different channel grouping numbers on the feature extraction ability may can provide some references for other studies.

2) Comparison With the Other Methods: Currently, many research works have proposed advanced semantic segmentation methods with various network structures, such as U-Net with stepwise up-sampling structure [57], ResNet50 with residual block [51], PSPNet with pyramid pooling module [33], DeepLabV3 with atrous separable convolution [60], DANet with dual attention module [43], PAN with pyramid attention structure [56], SiU-Net with siamese network structure [58], MA-FCN with empirical polygon regularization [61], EaNet with Dice-based edge-aware loss function [62], SRI-Net with spatial residual inception module [63], and SU-Net with scale robust network structure [64]. We conduct the experiments on our proposed method using the above two datasets to verify the performance of our method on building semantic segmentation qualitatively and quantitatively by comparing with these advanced semantic segmentation methods. The test results of comparison after a full training with 300 epochs are listed in Tables IV–V, Table VI shows the differences between our method and each compared method on channel grouping, attention mechanism, multiscale attention, and stepwise decoding structure. Our method can achieve better results compared with the other methods.

Fig. 4 also lists the partial visualization comparison results of each method on the WHU building dataset. Our method is better than the other methods overall and overcomes the shadow

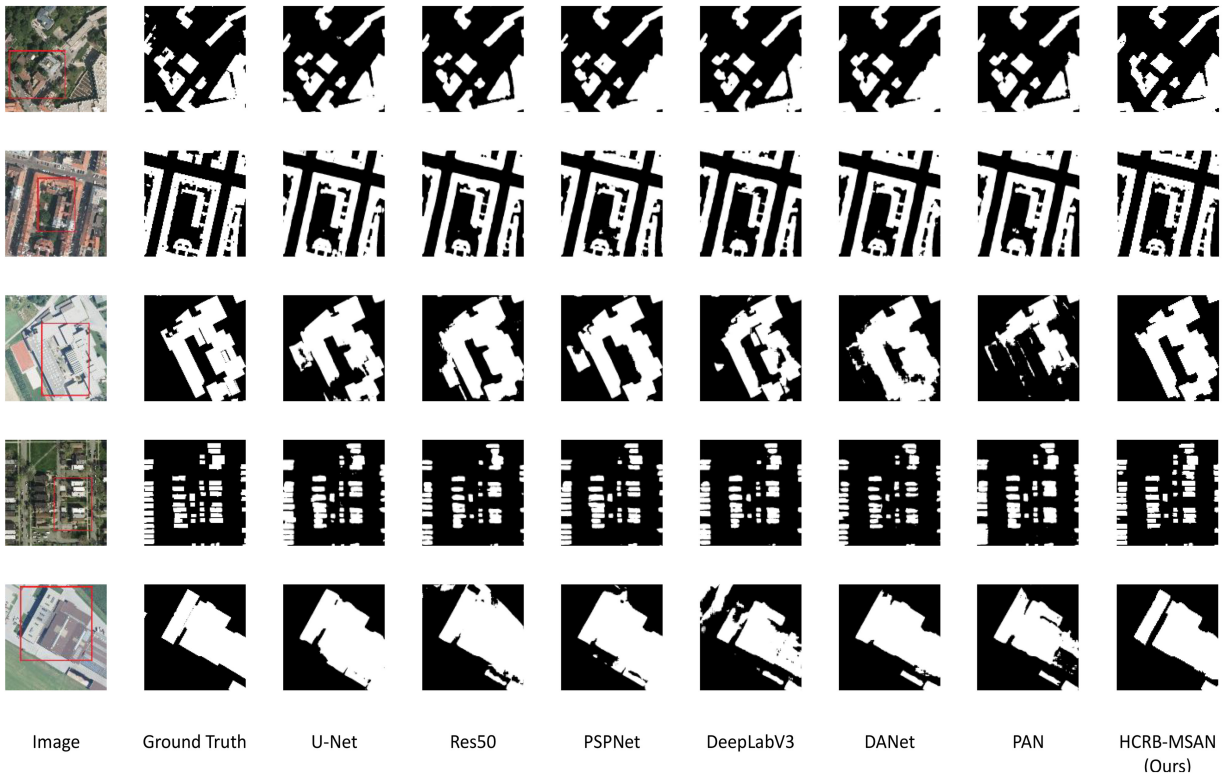


Fig. 5. Visual comparison results of each method on INRIA dataset.

effect to a certain extent (e.g., the building of the box in the first row of subfigures), to obtain more accurate segmentation results in the areas of boundary (e.g., the buildings of the box in the fourth row of subfigures). Besides, for small building targets (e.g., the buildings in the box of the second row of subfigures) and discrete building targets (e.g., the area in the box of the third row of subfigures), our method can also acquire more accurate segmentation results, which proves the excellent performance of our method.

We also perform the experiments on the INRIA dataset to further evaluate the effectiveness of the proposed method. Table V shows the indexes in comparison with the other semantic segmentation methods. Our method can achieve 88.84%, 79.92%, and 97.01% performance on F_1-score, IoU, and OA, respectively, all of them are surpass other methods, which verified the effectiveness of our designed method.

We also compared the visualization results of different methods on the INRIA dataset (see Fig. 5). In the area of dense buildings and small targets (such as the area in the box of the fourth row of subfigures), our method can better identify the buildings than the other compared methods. The network can also achieve efficient segmentation effects for irregularly shaped buildings (such as the area in the box in the third row of subfigures).

To sum up, our method obtains a high-precision extraction effects in the building segmentation of HSR remote sensing images, which proves the robust performance of the proposed method.

V. Conclusion

This article proposes a new method (HCRB-MSAN) on the extraction of buildings from HSR remote sensing images, which combines the channel grouping and horizontal connection inside the residual blocks into the network construction. In the method, the features extracted by the network are fused with a multiscale attention module to fully consider the contextual semantic information of different scales of regions by the integration of multilevel local and global information. Finally, through the stepwise up-sampling decoding, accurate building segmentation results can be obtained. We evaluate this method on two public datasets and compared it with the other state-of-the-art methods. Experiments show that our method has superior building extraction effects.

In future, we will consider automatic enhancement of training data, so that we can achieve more efficient semantic segmentation results of buildings in HSR remote sensing images.

References

[1] S. Hu and L. Wang, "Automated urban land-use classification with remote sensing," *Int. J. Remote Sens.*, vol. 34, no. 3, pp. 790–803, 2013.
[2] S. Ural, E. Hussain, and J. Shan, "Building population mapping with aerial imagery and GIS data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 6, pp. 841–852, 2011.
[3] S. J. Goetz *et al.*, "Monitoring and predicting urban land use change," in *Proc. IEEE Int., Geosci. Remote Sens. Symp.*, 2003, vol. 3, pp. 1567–1569.
[4] H. U. Zui, M. Deng, P. Liu, J. Wang, and Y. Tian, "Niche suitability assessment for human settlement in Hengyang based on GIS," *Trop. Geography*, vol. 31, pp. 211–215, 2011.

[5] J. Inglada, "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, no. 3, pp. 236–248, 2007.

[6] M. Cetin, U. Halici, and Ö. Aytekin, "Building detection in satellite images by textural features and adaboost," in *Proc. IAPR Workshop Pattern Recognit. Remote Sens.*, 2010, pp. 1–4.

[7] J. Peng and Y. Liu, "Model and context-driven building extraction in dense urban aerial images," *Int. J. Remote Sens.*, vol. 26, no. 7, pp. 1289–1307, 2005.

[8] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order CRF," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4483–4495, Aug. 2015.

[9] 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, 2015.

[10] Y. Wei, Z. Zhao, and J. Song, "Urban building extraction from high-resolution satellite panchromatic image using clustering and edge detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2004, vol. 3, pp. 2008–2010.

[11] L. Zhang and L. Zhang, "Deep learning-based classification and reconstruction of residential scenes from large-scale point clouds," *IEEE Trans. Geoence Remote Sens.*, vol. 56, no. 4, pp. 1887–1897, Apr. 2018.

[12] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 20, pp. 91–110, 2004.

[13] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 430–443.

[14] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis. Vol. Part I*, 2006, pp. 404–417.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, 2012, pp. 2564–2571.

[16] P. Zhang, Y. Ke, Z. Zhang, M. Wang, P. Li, and S. Zhang, "Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery," *Sensors*, vol. 18, no. 11, 2018, Art. no. 3717.

[17] D. Li *et al.*, "An image-based hierarchical deep learning framework for coal and gangue detection," *IEEE Access*, vol. 7, pp. 184686–184699, 2019.

[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] Z. Zhong, J. Li, W. Cui, and H. Jiang, "Fully convolutional networks for building and road extraction: Preliminary results," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1591–1594.

[20] K. Bittner, F. Adam, S. Cui, M. Korner, 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.

[21] Z. Cao *et al.*, "End-to-End DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Nov. 2019.

[22] Y. Liu *et al.*, "Multilevel building detection framework in remote sensing images based on convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3688–3700, Oct. 2018.

[23] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.

[24] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[25] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multi 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.

[26] T. Kim and J.-P. Muller, "Development of a graph-based approach for building detection," *Image Vis. Comput.*, vol. 17, no. 1, pp. 3–14, 1999.

[27] 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.

[28] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 62, no. 1, pp. 43–63, 2007.

[29] 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. Adv. Signal Process.*, vol. 2005, no. 14, 2005, Art. no. 745309.

[30] A. K. Shackelford, C. H. Davis, and X. Wang, "Automated 2-D building footprint extraction from high-resolution satellite multispectral imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2004, vol. 3, pp. 1996–1999.

[31] D. S. Lee, J. Shan, and J. S. Bethel, "Class-guided building extraction from ikonos imagery," *Photogrammetric Eng. Remote Sens.*, vol. 69, no. 2, pp. 143–150, 2003.

[32] 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.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

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

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

[36] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5669–5678.

[37] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[38] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[40] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–119.

[41] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-end flow correlation tracking with spatial-temporal attention," in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 548–557.

[42] Y. Li *et al.*, "Attention-guided unified network for panoptic segmentation," in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7026–7035.

[43] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Cong. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[44] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[45] 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.

[46] 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, 2018, Art. no. 144.

[47] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 156, pp. 1–13, 2019.

[48] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, 2017.

[49] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2021.

[50] S. Ji, 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, 2019.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[52] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.
[53] M. Tan and Q. Le, "Mixconv: Mixed depthwise convolutional kernels," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 116.1–116.13.
[54] R. Hang, Q. Liu, and Z. Li, "Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 30, pp. 7256–7265, Aug. 2021.
[55] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[56] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *Comput. Vis. Pattern Recognit.*, 2018, *arXiv:1805.10180*.
[57] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Med. Image Comput. Comput. Assist. Intervention*, pp. 234–241, 2015.
[58] 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.
[59] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.
[60] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[61] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.
[62] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep convnets with edge-aware loss," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 15–28, 2020.
[63] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 830.
[64] S. Ji and S. Wang, "Building extraction via convolutional neural networks from an open remote sensing building dataset," *Acta Geodaetica Cartographica Sin.*, vol. 48, no. 4, 2019, Art. no. 448.

Zhen Li received the bachelor's degree in surveying and mapping engineering from Shandong University of Science and Technology, Qingdao, China, in 2017. He is currently working toward the master's degree in cartography and geographical information engineering in the Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing, China.

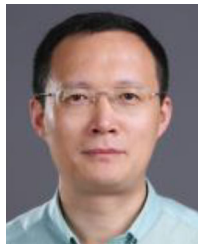
His research interests include deep learning and remote sensing images semantic segmentation.

Zhenxin Zhang (Member, IEEE) received the Ph.D. degree in geoinformatics in the School of Geography, Beijing Normal University, Beijing, China, in 2016.

He is currently an Associate Professor with Beijing Advanced Innovation Center for Imaging Theory and Technology and Key Lab of 3D Information Acquisition and Application, and College of Resource Environment and Tourism, Capital Normal University, Beijing. His research interests include light detection and ranging data processing, quality analysis of geographic information systems, remote sensing image processing, and algorithm development.

Dong Chen (Member, IEEE) received the bachelor's degree in computer science from Qingdao University of Science and Technology, Qingdao, China, the master's degree in cartography and geographical information engineering from Xi'an University of Science and Technology, Xi'an, China, and the Ph.D. degree in geographical information sciences from Beijing Normal University, Beijing, China.

He is currently an Associate Professor with Nanjing Forestry University, Nanjing, China. He is also a Postdoctoral Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. His research interests include image-and LiDAR-based segmentation and reconstruction, full-waveform LiDAR data processing, and related remote sensing applications in the field of forest ecosystems.



Liqiang Zhang received the Ph.D. degree in geoinformatics from the Chinese Academy of Science's Institute of Remote Sensing Applications, Beijing, China, in 2004.

He is currently a Professor with the School of Geography, Beijing Normal University, Beijing. His research interests include remote sensing image processing, 3-D urban reconstruction, and spatial object recognition.

Lin Zhu received the Ph.D. degree in hydrology and water resource from the College of Environment and Resource, Jilin University, Changchun, China, in 2007.

She is currently a Full Professor in the Beijing Laboratory of Water Resources Security, and College of Resource Environment and Tourism, Capital Normal University, Beijing, China. Her research interests include stochastic modeling and spatial analysis.

Qiang Wang received the B.S. degree in GIS from Liaoning Technical University, Fuxin, China, in 2012, and the Ph.D. degree in photogrammetry and remote sensing from China University of Mining & Technology, Beijing, China, in 2018.

He is currently a Lecturer with the School of Geographic and Environment Science, Tianjin Normal University, Tianjin, China. His research interests include image processing, close-range and aerial photogrammetry.

Siyun Chen received the bachelor's degree in geographic information science from Kunming University of Science and Technology, Kunming, China, in 2017. She is currently working toward the master's degree in cartography and geographical information engineering in the Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing, China.

Her research interests include deep learning and vehicle laser scanning point clouds semantic segmentation.

Xueli Peng received the bachelor's degree in geographical information science from Anhui Jianzhu University, Hefei, China, in 2018. She is currently working toward the master's degree in cartography and geographic information systems in the Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing, China.

Her research interests include deep learning and remote sensing images change detection.