# Building Extraction From High Spatial Resolution Remote Sensing Images of Complex Scenes by Combining Region-Line Feature Fusion and OCNN

Dehui Dong, Dongping Ming, *Member, IEEE*, Qihao Weng, *Fellow, IEEE*, Yi Yang, Kun Fang, Lu Xu, Tongyao Du, Yu Zhang, and Ran Liu

***Abstract*—Building extraction from remote sensing imagery has been a research hotspot for some time with the advancement of AI in remote sensing. However, the edges of buildings extracted using existing techniques are commonly broken and inaccurate for the complex scenes in suburban and rural areas. This study proposes a framework for extracting structures by combining region-line feature fusion with object-based convolutional neural networks to solve this problem. First, a building edge detection network known as the Multichannel Attention-based Dense Extreme Inception Network for Edge Detection (MA-DexiNed) is constructed, which is considered more accurate for building edge extraction in complicated image scenes. Second, the probability map of the building edges obtained by MA-DexiNed is refined. According to rule judgment, breakpoints are linked by an edge thinning connection algorithm to obtain single-pixel, contiguous building line features. Third, the geometric boundaries of buildings are obtained by combining region attributes derived by unsupervised image segmentation and line features obtained from deep learning supervised segmentation. Finally, the pretrained AlexNet is employed to identify the class characteristics of buildings. The suggested framework was used for two GF-2 images and one Google Earth image from various regions and with numerous types of complicated scenes. The experimental findings demonstrated that this approach could extract more precise and complete building edges for complex image scenes compared with several existing methods. This advancement results from constrained regional image segmentation using deep semantic edge features. This methodology can offer a benchmark for subsequent building extraction tasks from high resolution imagery.**

***Index Terms*—Building extraction, complex image scenes, convolutional neural network, edge detection network, high spatial resolution imagery, image segmentation, nonurban buildings.**

Manuscript received 28 December 2022; revised 7 March 2023 and 3 April 2023; accepted 29 April 2023. Date of publication 8 May 2023; date of current version 17 May 2023. This work was supported in part by the State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM under Grant 2022-02-14, in part by the Fundamental Research Funds for the Central Universities, and in part by the 2023 Graduate Innovation Fund Project of China University of Geosciences, Beijing under Grant ZD2023YC034. *(Corresponding author: Dongping Ming.)*

Dehui Dong, Dongping Ming, Kun Fang, Lu Xu, Tongyao Du, Yu Zhang, and Ran Liu are with the School of Information Engineering, China University of Geosciences (Beijing), Beijing 100083, China (e-mail: dongdhcugb@163.com; mingdp@cugb.edu.cn; fangkun@cugb.edu.cn; xlirs@cugb.edu.cn; dwoty1998@163.com; zy_email0924@163.com; liuran155@163.com).

Qihao Weng is with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong (e-mail: qihao.weng@polyu.edu.hk).

Yi Yang is with the Chinese Academy of Surveying and Mapping, Beijing 100083, China (e-mail: yangyi@casm.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3273726

## Nomenclature

| | |
|---|---|
| HRS | High spatial resolution remote sensing. |
| DexiNed | Dense extreme inception network for edge detection. |
| CNN | Convolutional neural networks. |
| OCNN | Object-based convolutional neural networks. |
| MRS | MultiResolution segmentation. |
| CBAM | Convolutional block attention module. |
| NMS | Nonmaximum suppression. |

## I. Introduction

THE efficient extraction of building is of great importance in the infrastructure construction of the smart city, urban planning, and measuring the economic development of cities [1]. With the increasing maturity of satellite sensor technology, acquiring HRS images has become more convenient. HRS images contain richer spectral, spatial, and texture information, which creates the potential of obtaining more accurate building using remote sensing technology [2]. Although the identification of buildings from HRS images is a task of binary classification, it is still a challenging topic in the remote sensing community due to the problems of different objects which have the same spectrum and the same objects have different spectrums in HRS images.

The existing research on building extraction usually concentrates on remote sensing images with simple scenes of buildings and relatively high image quality. However, there is still a lack of adequate research for tasks of building extraction on complex image scenes with complicated background information, including suburbs or rural areas [3]. There are various understandings of complex image scenes. Some scholars argue that the complexity of the scenes is induced by the fact that the structure and shape of buildings differ significantly from country to country [4]. Additionally, the impact of surrounding attributes (e.g., trees and billboards) contrast with the buildings, and the surrounding area is extremely low. It has also been argued that the currently existing datasets do not include the complex scenes of modern cities (e.g., overpasses and roundabouts) [5]. Therefore, it is insufficient to be considered a reflection of the real world.

In this study, it is regarded that complex image scenes in building extraction applications should include two meanings: the complexity of the building itself and the complexity of the environment. The intricacy of the building itself refers to the structure's considerable individual variances. Regional and functional differences are frequently the driving forces behind building forms and structure variances. A classic example of this sort of complicated landscape is the suburb [6]. The complexity of the environment refers to the low contrast of buildings in the local area induced by the similarity in spectral features between buildings and their neighboring environment. A typical real world complex scene of this type is a rural area [7]. Furthermore, the spatial resolution of remote sensing images is not as high as the existing datasets (mostly aerial images), and there are many mixed pixels in remote sensing images, which increases the challenges of building extraction. Thus, building extraction in complex scenes from satellite images is more practically challenging and essential.

With the development of deep learning, it has been a trend to introduce deep learning to solve difficult problems in remote sensing. From OCNN [8], semantic segmentation [9] to edge detection networks [10], all of them have shown good performance in the task of building extraction. However, all these single methods have obvious shortcomings. For example, building edges extracted by semantic segmentation are commonly inaccurate and broken [11], these two problems limit the extraction accuracy. It is found that a single edge detection network can obtain accurate edges of objects [12], but it is challenging to acquire contiguous and complete edges [13]. Meanwhile, it is possible to obtain entire edges of objects using OCNN alone; however, since this technique generates segmented objects following clustering, the boundaries of segmented objects are typically not the natural boundaries of objects; therefore, the edges obtained are commonly less precise [14]. To solve the above problems, several scholars have tried to cascade the semantic segmentation network with the edge detection network to constrain the semantic segmentation [15], [16]. This combined method can obtain more accurate building edges to a certain extent. However, this method relies on perfect polygon building samples (in other words, the boundary of the building is accurate and closed). If there are only partially imperfect samples, whether it can still achieve efficient extraction of buildings is a challenge.

In the field of remote sensing, methods for combining edge detection with image segmentation have been proposed for a long time [17], [18]. However, all these methods are implemented using traditional edge detection operators, including Canny [19] and Sobel [20], which are challenging to be helpful in complex image scenes. Currently, some researchers have tried to obtain edges using deep learning methods instead of traditional edge detection operators to achieve edge-constrained image segmentation. For example, Kucharski et al. [21] proposed a new semantic segmentation network to perform separate extraction of cell edges and cell centers, which was then combined with a labeled watershed algorithm to achieve high-precision image segmentation. In this study, we attempt to introduce a combination of deep learning edge detection and OCNN to use in the building extraction task to achieve accurate extraction.

To achieve more accurate segmentation, there is an urgent need to study a methodology that combines the area information obtained by unsupervised segmentation with the edge line features gained through supervised segmentation. Based on such an idea, this article proposes to fuse deep learning edge detection with image region segmentation to enhance the correctness of building geometric boundaries. This study proposes a new method linking deep learning edge detection with OCNN to obtain more accurate and complete building edges from HRS images in complex scenes, which combines building edge line features obtained by the edge detection network and building region features obtained by MRS [22] at the feature level. This method implements a novel idea for image segmentation by combining supervised and unsupervised segmentation.

## II. Related Work

### A. Building Extraction From HRS Images

Currently, approaches for building extraction can be divided into two categories: methods based on shallow features and strategies based on deep learning [23], [24]. The methods based on shallow features primarily adopt the spectral, texture, or spatial characteristics of buildings in remote sensing images to construct morphological indicators [25], [26] or adopt region-based image segmentation methods for building extraction [27], [28]. The morphological methods are susceptible to noisy information; therefore, they are unsuitable for complex image scenes. Image segmentation methods prevent the phenomenon of "salt and pepper noise" well; however, the classification accuracy of these methods depends solely on the segmentation effect. Notably, incorrect segmentation frequently occurs when the edges of buildings are too similar to the neighboring object units (e.g., roads) regarding spectral and textural attributes. All the above methods can efficiently extract simple scenes; however, achieving high accuracy in complex image scenes is challenging. Therefore, it is difficult to use only the shallow features of the image to meet the requirements of practical applications. Thus, how to exploit the deep features of HRS images has become a current research hotspot.

With the advancement of deep learning technology [29], [30], [31], [32], [33], several researchers have tried to apply deep learning models to building recognition and extraction [34], [35], [36]. Previously, the OCNN classification method is generated according to deep learning classification combined with the object-based image segmentation method. This method widely applies to land cover and functional region classification [37], [38], [39], [40]. Subsequently, from the early FCNN [41] to U-Net [42], SegNet [43], DeepLab-V3 [44], the pixel-based semantic segmentation achieves end-to-end recognition of features at various scales [45]. Liu et al. [46] proposed a lightweight network called LRAD-Net for building recognition, which has fewer parameters and faster computational speed. Hui et al. [47] used an enhanced U-Net to extract buildings. A large amount of pixel-level annotations is a prerequisite for semantic segmentation that can achieve high accuracy recognition. To reduce the time of manual annotation, many semi-supervised methods have emerged. Li et al. [48] proposed an instruction

to assign the perturbation to the intermediate feature representations within the encoder of the network, which reduces the misclassification of buildings. Kang et al. [49] proposed a segmentation named PiCoCo for building extraction, which is based on the enforcement of pixelwise contrast and consistency in the learning phase. Recently, edge detection networks have provided novel ideas for building extraction. Lu et al. [10] used RCF for building extraction and proposed a postprocessing algorithm for edge thinning. Xia et al. [50] proposed a semi-supervised deep learning method based on an edge detection network to improve the accuracy of building roof detection with a small number of samples.

In addition, the combination of multiple single methods has been a recent research hotspot [16]. For instance, Marmanis et al [15] fused SegNet with HED for building recognition. Li et al. [17] combined an edge detection operator and marker-based watershed segmentation algorithm to achieve the effect of optimal segmentation. Furthermore, Chen et al. [18] proposed a multiscale segmentation method subject to edge constraints.

### B. Edge Detection

Edge detection is a fundamental problem in image processing and computer vision. The purpose of edge detection is to identify a series of points in a digital image where the brightness varies significantly. The development history of edge detection algorithms can be roughly divided into three stages. In the first stage, edges are usually computed using low-level features (e.g., gradients) of the image. Algorithms like Sobel [20] and Canny [19] are used in many fields due to their simplicity of implementation. In the second stage, edge detection is achieved by learning from manually designed features. Many excellent algorithms have emerged in this stage, such as Pb [51], gPb [52], and Statistical Edges [53]. They all outperformed the algorithms based on low-level features and performed well in a variety of datasets. In the third stage, with the development of deep learning, CNN-based methods gradually become the mainstream of edge detection algorithms. CNN's superb capability of feature mining provides the possibility of extracting deep edge features. For example, networks such as HED [54], RCF [55], and CEDN [56] are widely used in the task of feature extraction. Particularly, the DexiNed [57], one of the most developed edge detection networks currently, can obtain more accurate and fine object edges. Its training method does not necessitate the use of pretraining weights or finetuning. It is worth noting that although remote sensing images often have lower spatial resolution than natural images, they contain rich spectral information. Therefore, it is necessary to develop a new edge detection network that is more suitable for remote sensing tasks, taking into account the data characteristics of remote sensing images.

## III. Methodology

To avoid broken and inexact edges in building extraction tasks from complex image scenes, this study suggests a framework for linking region-line feature fusion and OCNN for building extraction from HRS images. Fig. 1 depicts the architecture of the framework. This method comprises three modules, the first is the *line feature extraction module.* First, a new multichannel attention-based dense extreme inception network for edge detection (MA-DexiNed) is constructed in this study, through which the edge probability map of the building is obtained. Subsequently, the proposed edge thinning connection algorithm based on rule judgment in this study obtains the building line features. Next is the *region feature extraction and region-line feature fusion module*, which applies MRS to obtain the regional characteristics of the building and combine them with the line features to accomplish image segmentation. Finally, the *OCNN building extraction module* trains the AlexNet [58] with manually labeled samples. Then the attributes of the forecasted points are identified, and the most voting algorithm obtains the building extraction results. The key steps in the above process are detailed in the next few subsections.

### A. Edge Detection Network MA-DexiNed

Given the limitations of existing methods for building edge detection tasks in complicated remote sensing scenes, this study offers MA-DexiNed, a novel network based on DexiNed and the multichannel attention model. Fig. 2 depicts the architecture of the network.

As depicted in Fig. 2, the network backbone of MA-DexiNed can be categorized into six major blocks, which are comprised of standard convolutional layers concatenated together. In this study, we eliminate the pooling layer of the last three key blocks of DexiNed and extend the size of the convolutional kernel of this layer, which aims to extend the receptive field for the model to fully guarantee spatial detail information in deeper feature maps. As if the network reaches a significantly deep level, it is straightforward to cause loss of spatial information, which in turn makes the inaccuracy of edge localization in the output image. Additionally, CBAM is inserted in the middle of each set of convolutional layers in the last two significant blocks, through which the network can learn the importance of various feature maps and locations of pixels; therefore, it improves the feature extraction ability of the network. CBAM is a plug-and-play lightweight module built by Woo et al. [59]. This module has been widely used in most basic networks. Thus, all of them have been substantially enhanced. The CBAM comprises a channel attention submodule and a spatial attention submodule. The channel attention submodule comprises maximum pooling and average pooling in parallel. In contrast, the spatial attention submodule includes maximum pooling and average pooling in series. The introduction of CBAM has two main reasons. First, it leverages the excellent spatial feature enhancement ability of CBAM to improve the spatial position information of objects in the deep feature maps. Second, it utilizes CBAM's channel feature enhancement ability to give more attention to the multiband characteristics of remote sensing images.

The MA-DexiNed can input multichannel image data. Remote sensing images generally have more bands (channels) than natural images. Therefore, multiband data has a high prospect for feature mining, particularly the infrared band of remote sensing images, which plays a decisive role in the feature construction of buildings. Therefore, the MA-DexiNed proposed

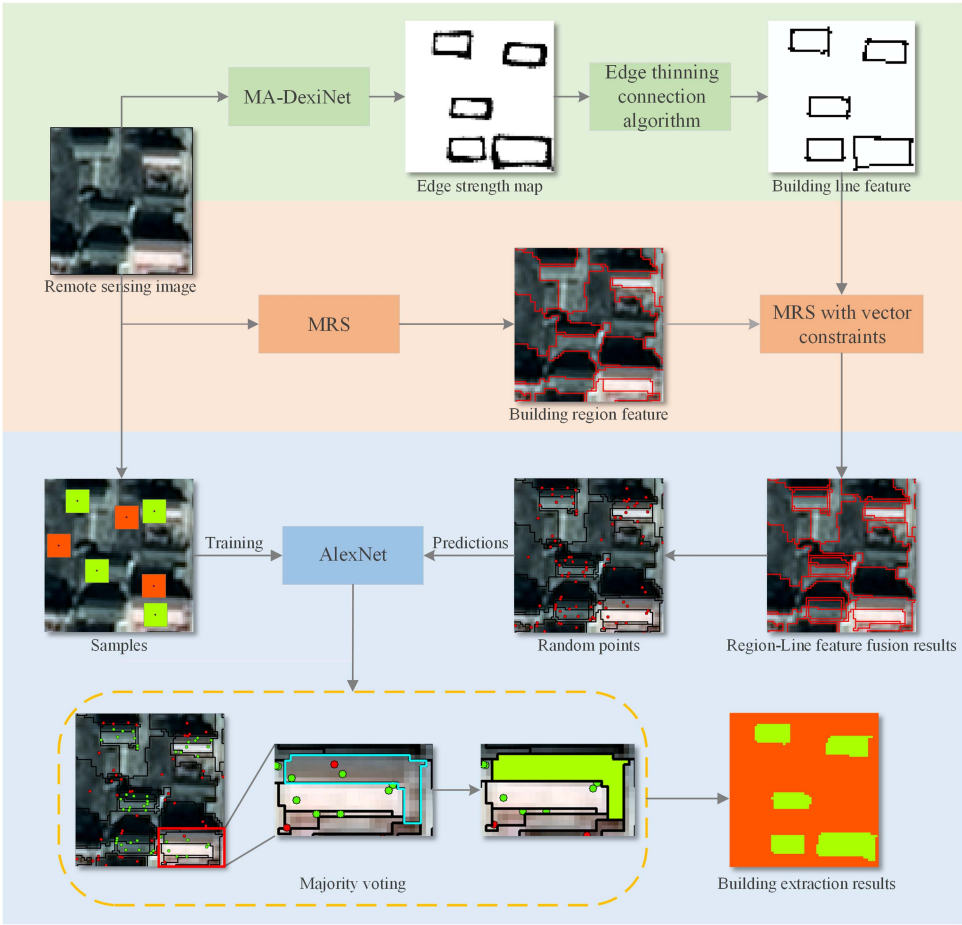


Fig. 1. Process for combining region-line feature fusion with OCNN for building extraction.

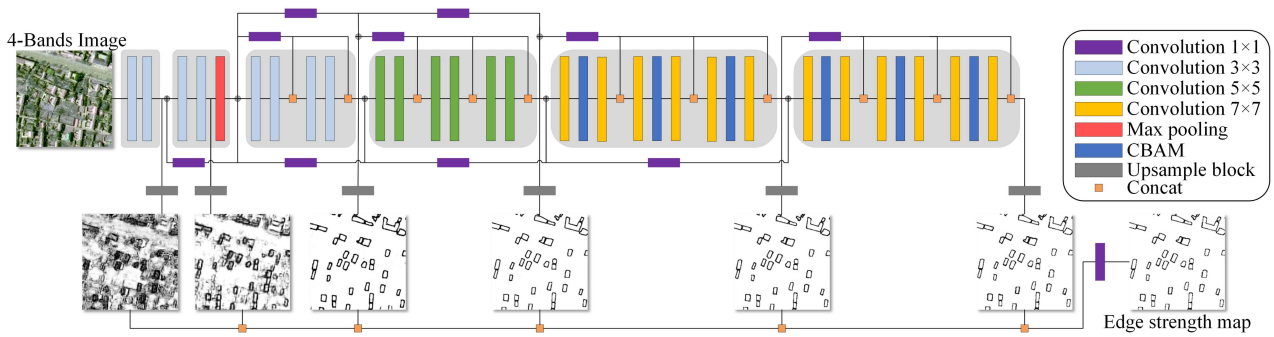


Fig. 2. Architecture of MA-DexiNed.

in this study is uniquely designed to extract building edge features from HRS images in complex scenes.

### *B. Edge Thinning Connection Algorithm Based on Rule Judgment*

MA-DexiNed can obtain a more comprehensive and precise building edge probability map. It is vital to thin the edge probability map to obtain pixel-level building line features. This study uses a rule-based postprocessing algorithm to thin the edge probability map output from the edge detection network. The workflow of the algorithm is depicted in Fig. 3. NMS is commonly employed for post-processing of edge detection [60]. However, the thinning using only NMS results in broken edges, severely impacting future attribute fusion efficacy. Therefore, in this study, after obtaining the preliminary thinning results using NMS, breakpoints are identified and linked based on rules to warrant that the building edges on rule judgment are as complete as possible.

In this study, the eight-neighborhood approach is used for breakpoint identification, i.e., based on the preset breakpoint template, we adopt the sliding window method to correspond

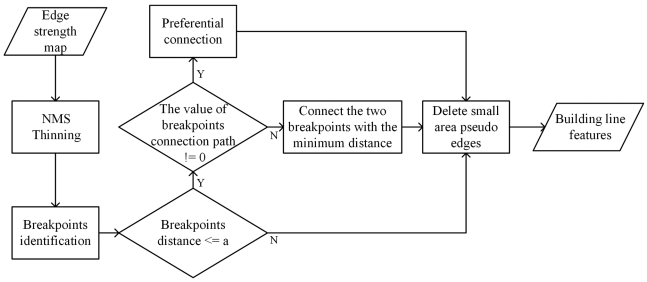


Fig. 3. Workflow of edge thinning connection algorithm based on rule judgment.

one by one to determine whether the center point of the window is a breakpoint. After obtaining the breakpoints of the entire map, the breakpoints that meet the connection conditions are screened based on the distance threshold $a$ between the breakpoints. Breakpoints that are closer to this threshold are preserved for connection. This connection rule is based on the assumption that building units exist in isolation. Hence, the building breakpoints that should be connected are typically near one another. The threshold $a$ is an empirical value. Following many experiments, it has been discovered that the threshold ($a$) is most suitable between 6 and 8. Subsequently, it proceeds to judge whether the saved breakpoints to breakpoints have nonzero edge probability values on their connection paths, i.e., whether there are nonzero values on this path in the original edge probability map. If they exist, the two breakpoints are linked in priority. Otherwise, the two breakpoints with the shortest distance are connected in preference based on the distance criterion. This rule is set given the "invalid rejection" phenomenon of NMS, which means that NMS tends to eliminate some discontinuous building edge points with low probability value in the edge probability map thinning process. However, these points are commonly the weak edges of buildings that are challenging to be found by the edge detection network. Therefore, following the designed connection rules, this study attempts to make the best use of these points to obtain complete and adjacent building edge features. After the above breakpoints connection, the smaller area of pseudoedge connected regions in the image is eliminated to obtain the building line features. The following pseudocode shows the particular process of thinning connection algorithm.

**Algorithm 1:** Edge Thinning Connection Algorithm Based on Rule Judgment.

```
Input: Edge_strength_map; a
Process:
Initial_thinning_map = NMS(Edge_strength_map)
Breakpoints = Breakpoint(Initial_thinning_map)
For i in Breakpoints do
    Candidate_point =
    Range_search(Initial_thinning_map, i, circular, a)
    For j in Candidate_point do
        Route_value_ij =
        Route_search(Edge_strength_map, i, j)
        Distances_ij = Distance(i, j)
    Route_value =
    Route_value.And_array(Route_value_ij)
    Distances = Distances.And_array(Distances_ij)
    If (Route_value = = 0)
        [i, j] = minimum(Distances)
    Else
        [i, j] = minimum(Route_value)
    Thinning_connection_map =
    Connection(Initial_thinning_map, [i, j])
Breakpoints.Delete_array(Breakpoints, [i, j])
Output: Thinning_connection_map
```

### C. Region Feature Extraction and Region-Line Feature Fusion

Image segmentation is an unsupervised method that uses the heterogeneity of the present image pixel and the neighboring pixels for regional segmentation to obtain region attributes of the building. The heterogeneity is a statistical index that does not consider the semantic characteristic among building image pixels. Thus, the buildings obtained by OCNN are moderately complete; however, their edges are imprecise. In contrast, the deep learning edge detection network is a supervised segmentation method, and the obtained line features are semantically informative. These line attributes are also commonly actual building edges; however, this method typically suffers from missed and false detections. The suggested region-line feature fusion method is to complement the merits of the above two methods. The image segmentation and stronger feature mining ability of OCNN are employed to augment the missed and false detection of the edge detection network. The more precise building edges obtained by the edge detection network are used to constrain OCNN. Because the image segmentation algorithm used in this paper is unsupervised, polygon building samples are not required, making it extremely suitable for tasks with a shortage of perfect and sufficient samples.

According to the fractal net evolution algorithm, the image segmentation method used in this study is MRS, which is highly integrated, easy to use, and has strong applicability [14]. The particular steps are depicted in Fig. 1. First, the original remote sensing image is input into MRS to get the preliminary segmentation result, which is the building region feature. The building line feature map and the building region feature are then simultaneously input for the second segmentation by MRS with Vector constraints, i.e., the second MRS is a segmentation of the line feature map, and the region feature is used as a vector constraint to guide the segmentation process to stop when the vector boundary is encountered. Therefore, the second segmentation is an oversegmentation, which is performed only inside the segmented object obtained for the first time. Fig. 4 shows a visual understanding of this process. The irregular solid boxes signify image objects, while the gray boxes denote the initial image objects. In contrast, the red, yellow, and blue boxes represent the new image objects at the building edge pixels, the interior building pixels, and the nonbuilding pixels, respectively, after feature fusion. Finally, the image objects with smaller areas are merged according to the regional color features and

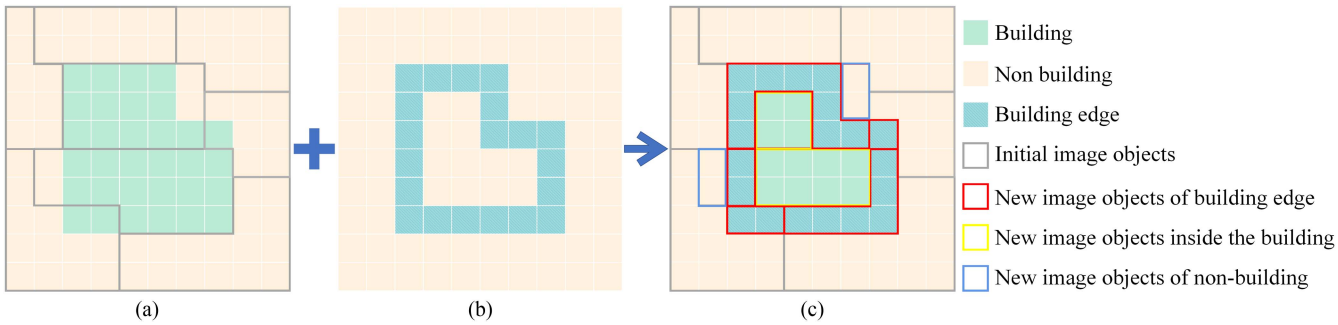


Fig. 4. Illustration of the region-line feature fusion process. Each small box represents a pixel, and different colors represent different objects. (a) Building region features. (b) Building line features. (c) Region-line feature fusion results.

area threshold in order to reduce the number of small regions generated by oversegmentation, and the feature fusion results are obtained. After repeated experiments, the area threshold value between 5 and 8 is commonly suitable.

### D. OCNN Building Extraction

The idea of OCNN is to segment first and subsequently categorize. In this study, AlexNet [58] deep learning convolutional neural network is chosen to denote class features to segmented objects. It is matured in the application of AlexNet in the field of remote sensing for land cover classification as well as functional area recognition [61]. Meanwhile, its efficiency and accuracy could satisfy the requirements of the tasks in this study. AlexNet consists of five convolutional layers, three maximum pooling layers, and two fully connected layers. These structures of AlexNet constitute the feature extractor, which extracts the in-depth features of the input image. The Softmax function after these structures is used as the classifier to classify the deep components extracted by the network. To obtain deeper features of remote sensing images, this study also expands the AlexNet with channels so that it can input multiband remote sensing images, which improves the network's feature mining capability and augments the network's classification presentation.

OCNN considers the class of a point in the object as the class of that object. The basis for using this approach is to verify that the item has a high degree of uniformity. In complex image sceneries, however, building image objects have spectral properties similar to their surrounding nonbuilding image objects. Therefore, the class of a single point within an object is often challenging to precisely identify the class of the entire object. Consequently, we introduce most voting algorithms [14] to augment the classification correctness. The fundamental idea of this algorithm is to allocate the class features with the most number of points to the object according to the forecasted classes of some random points within the object. Finally, achieve image classification and reduce misclassification.

### E. Evaluation Metric

In this study, four classical quantitative evaluation metrics are used to assess the accuracy of the extracted buildings: precision (P), recall (R), overall accuracy (OA), and F1 score. Precision is the proportion of building pixels that are correctly predicted as buildings to all pixels indicated as buildings. Recall, in this study's context, the building binary classification task refers to the proportion of pixels that are accurately predicted as buildings to all pixels that are buildings. The following equations are used to estimate the precision and recall

$$\mathrm{P} = \mathrm{TP}/\left(\mathrm{TP}+\mathrm{FP}\right) \tag{1}$$

$$\mathrm{R} = \mathrm{TP}/\left(\mathrm{TP}+\mathrm{FN}\right) \tag{2}$$

where TP is true positive, demonstrating the number of pixels that are buildings and correctly predicted as buildings. TN is a true negative, indicating the number of pixels that are nonbuildings and correctly predicted as nonbuildings. FP refers to false positives, representing the number of pixels that are really nonbuildings but were wrongly forecasted to be buildings. FN indicates a false negative, representing the number of pixels that are actually buildings but were wrongly forecasted to be nonbuildings. Overall accuracy refers to the proportion of pixels that are correctly forecasted for all pixels. Equation (3) is the calculation of OA. Equation (4) is the calculation of the F1 score, from which it is found that the F1 score is a weighted average of precision and recall. A higher F1 score shows better extraction of the algorithm

$$\mathrm{OA} = \left(\mathrm{TP}+\mathrm{TN}\right)/\left(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}\right) \tag{3}$$

$$\mathrm{F1} = \left(2\times\mathrm{P}\times\mathrm{R}\right)/\left(\mathrm{P}+\mathrm{R}\right). \tag{4}$$

## IV. Experiments and Results

### A. Study Area and Data

To confirm the proposed method's strength, two images from GF-2 and one image from Google Earth are used as the study areas, which are selected from various regions and with multiple types of complex image scenes. Study area A is located in Guangping County, Hebei Province, China. The image was taken on February 25, 2017. Fig. 5(a) indicates the actual color composite image of study area A. Study area A is rural, where buildings are generally small and easily confused with the complex surroundings. Study area B is selected from the remote sensing image of GF-2 in the suburb, which is located in Taipei City, Taiwan. The image was taken on February 3, 2019. Fig. 5(b)

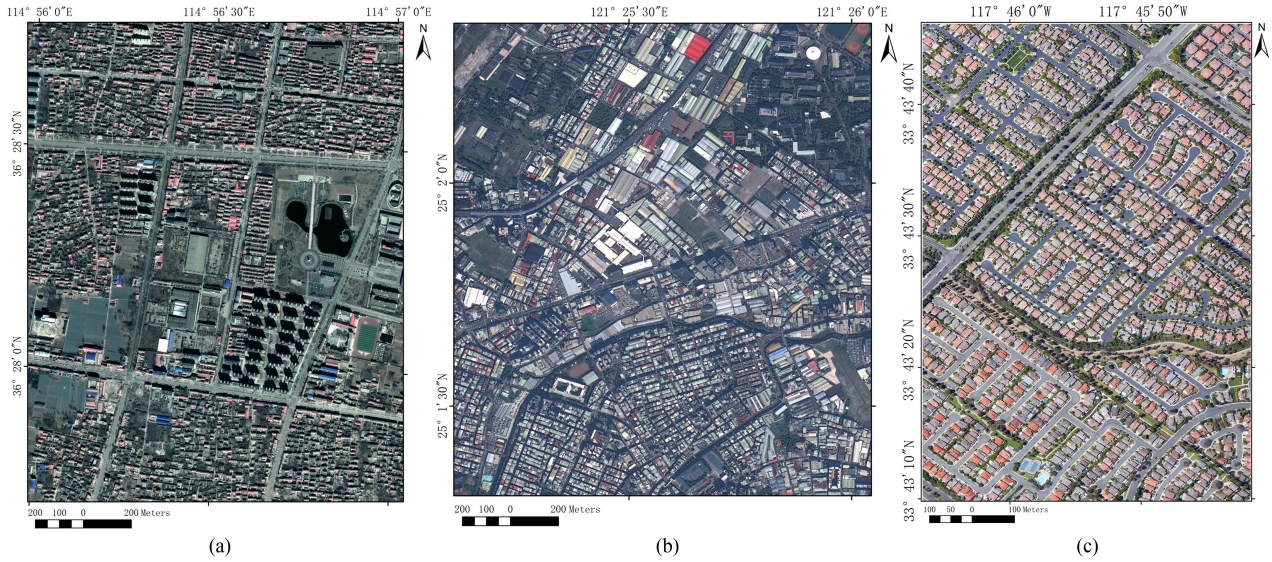


Fig. 5. Remote sensing images of the study area with a resolution of 1 meter. (a) Study area A. (b) Study area B. (c) Study area C.

shows study area B's actual color composite image. There are many factory buildings in this area, with a wide variety of roofing materials and colors. There are also a large number of residential buildings in study area B. These residential buildings are small and dense. It is often difficult to obtain satisfactory results from conventional building extraction methods. To meet the demand for high resolution in this experiment, the image's multispectral and panchromatic bands were fused to obtain HRS fused remote sensing images with a resolution of 1 m. Study area C is situated in Orange County, California, USA. The image was selected on April 3, 2018. Study area C is a suburb with diverse and dense buildings. Furthermore, Fig. 5(c) depicts study area C's actual color composite image. The size of study areas A, B, and C are 1568 × 2000 pixels, 1568 × 1776 pixels, and 1294 × 1878 pixels, respectively.

In this study, various training samples should be constructed for the two deep learning networks of the suggested method. The task of the MA-DexiNed is to extract building edges. For three different complex image scenes of the study area, a quarter area of study area A (located in the top left corner of the image), a quarter area of study area B (located in the center of the image), and a quarter area of study area C (located on the left side of the image) were selected for making samples. The process started with outlining the building edges using ArcGIS 10.2, then converting them into binary raster data and cropping them randomly into square training samples of 304 × 304 pixels (settings for network input). Additionally, data enhancement methods, including flip (horizontal/vertical), rotation (90°/180°/270°), average blur, Gaussian blur, bilateral blur, and adding random noise, were used to expand the number of samples in the cropping process. Finally, the edge detection dataset for each study area contains 6000 sets of images with their corresponding binary labels.

The sample for the AlexNet is a dataset consisting of square image blocks with the class attributes of their central pixels. We employed the concept of uniform sampling to choose a suitable number of building and nonbuilding sample sites for research regions A, B, and C, making the sample more homogenous and speeding up network convergence. Table I shows the specific number of building and nonbuilding samples. These samples were randomly categorized into training and validation datasets, where 80% were selected as training datasets for training the weights and biases of the network. In comparison, the other 20% were used as validation datasets to adjust the hyperparameters of the AlexNet and thus obtained the model with the best prediction results.

TABLE I
NUMBER OF TRAINING SAMPLE POINTS IN THE STUDY AREA

| Class | Buildings | Non-Buildings |
|---|---|---|
| Study Area A | 3546 | 3539 |
| Study Area B | 3578 | 3530 |
| Study Area C | 3850 | 3724 |

To verify the efficiency of the suggested method, 1000 points were randomly selected and allocated features for each of the three study areas as a basis for accuracy verification.

### B. Experiment Parameters

This experiment was conducted on Windows 10 OS, where the CPU is 2.90GHz Core i7-10700, and the GPU is NVIDIA GeForce RTX 3090. The MA-DexiNed and AlexNet were implemented under Pytorch-GPU 1.9.1 and Tensorflow-GPU 1.3.0 deep learning platforms, respectively.

In the training process of MA-DexiNed, Adam was chosen as the optimization algorithm for network weight update. The batch size was set to 4, the initial learning rate was 0.0001, and the total number of epochs was 50. SGD was chosen for AlexNet to update the network weights. Considering the study area's variability in building sizes, we established the image patch size of the input network to 45 × 45 pixels. The batch size

and total epoch number were set to 10 and 100, respectively. Experiments were established to have a learning rate of 0.001 in the first 50 epochs of training and 0.0001 in the last 50 epochs.

### C. Results

*1) Building Line Feature Extraction Results:* The MA-DexiNed was trained using the building edge dataset. Following the completion of the training, the remote sensing images from the research region were fed into the model with the lowest training loss and best confirmation effect to obtain the edge detection results. Fig. 6 depicts the edge probability maps of study areas A, B, and C, respectively. The detection of high-rise buildings is poor in the marked blue areas of Fig. 6(a), because there are no edge samples of high-rise buildings in the edge detection dataset of study area A. Comparing Fig. 6(a1)–(a4) with Fig. 6(A1)–(A4), the detection results are consistent with the manual annotation; however, there is also the phenomenon that the edges of neighboring buildings stick together. The edge detection impact of study areas B and C is similar to that of study area A.

In this study, the edge thinning connection algorithm, according to rule judgment, was used to thin the edge probability map and connect the edge breakpoints. Fig. 7 shows the comparison before and after the algorithm processing. Although the improved MA-DexiNed in this study has obtained finer building edges, the broken building edges and the edge adhesion of neighboring buildings still exist. The marked areas in the figure depict that the algorithm has a more noticeable impact in processing building edge thinning, edge breakpoints connection, and edge adhesion separation.

*2) Region-Line Feature Fusion Results:* In this study, the building line features obtained by edge detection were combined with the building region features obtained by MRS to achieve more accurate building extraction results at the edges. Numerous attempts found that all study areas could obtain better building segmentation results with a scale parameter of 35. Fig. 8 compares the segmentation results before and after the fusion of building features with the same scale parameter. From the marked red areas in the figure, it can be discovered that the method separates the building from its surrounding background with similar spectral characteristics while keeping the original MRS results.

*3) OCNN Extraction Results:* The building dataset was manually labeled and cropped to a uniform size. Subsequently, it was fed into the AlexNet for deep feature extraction. Simultaneously, five random points were generated for each segmented object using the random point generation algorithm. Five random points are sufficient to represent the segmented object. Because the segmentation unit is already sufficiently homogeneous internally, the error tolerance of the algorithm is substantially enhanced. Following the same preprocessing operation as the training sample points, the random points were fed into the model with low training loss and validation loss simultaneously for the prediction. Furthermore, the model generated random points with class properties. Finally, the extraction results

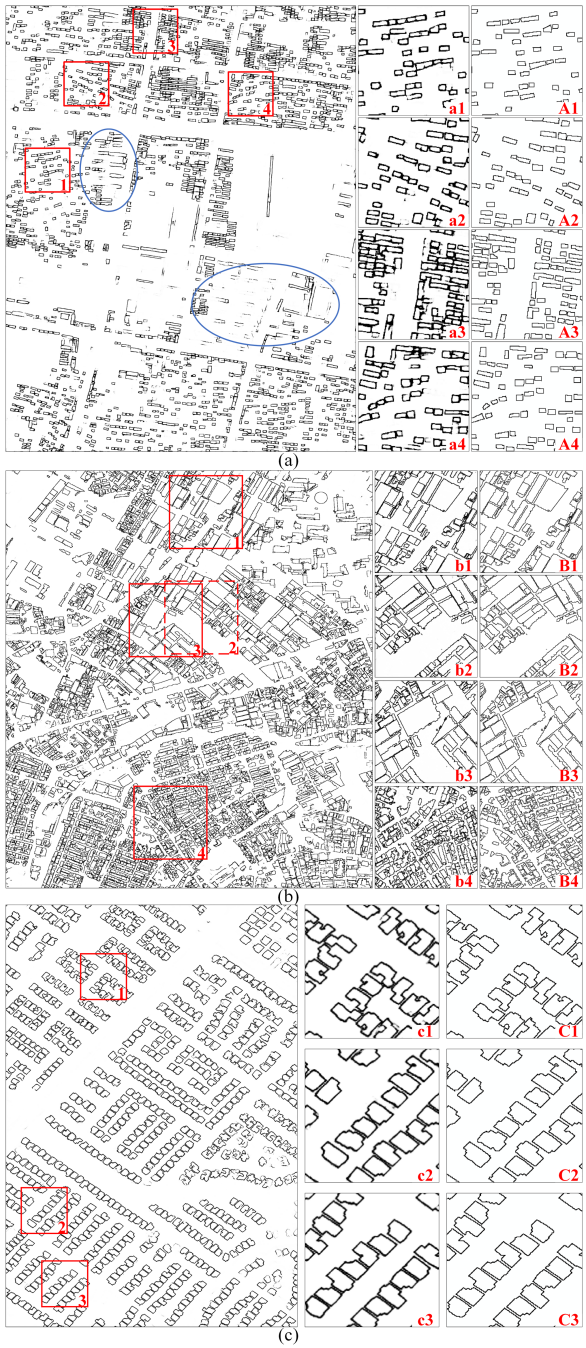


Fig. 6. Building edge detection results. (a)–(c) Edge detection results in study areas A, B, and C. 1–4 subregions and ground truths of different study areas are highlighted and enlarged on the right side.

were obtained by feeding these random locations through the majority voting method. Building extraction findings for the three research regions are shown in Fig. 9. It is discovered in this study that the proposed technique performs better in terms of accuracy and edge integrity of the building edges.

### D. Accuracy Evaluation Result

Table II presents the accuracy evaluation results of building extraction for the three study areas. It is discovered from the

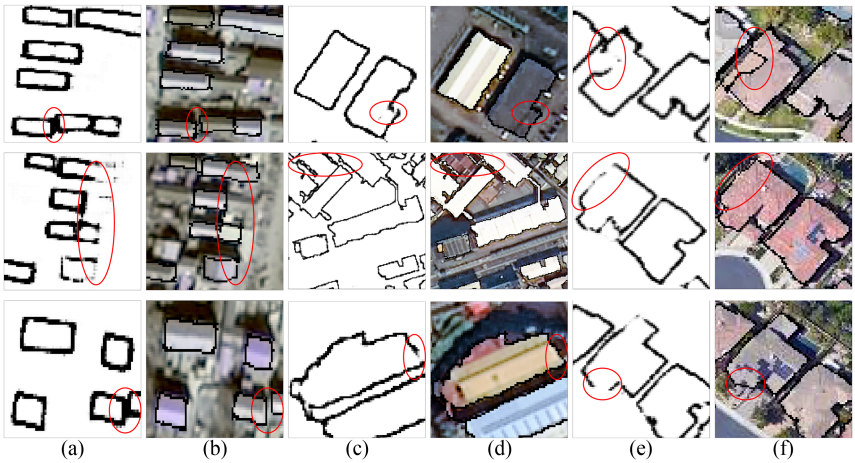


Fig. 7. Comparison of results before and after algorithm processing. (a), (c), and (e) Edge detection results in study areas A, B, and C. (b), (d), and (f) Algorithm processing results in study areas A, B, and C.


Fig. 8. Comparison of results before and after fusion of study area. (a), (c), and (e) Before fusion of study areas A, B, and C. (b), (d), and (f) After fusion of study areas A, B, and C.

table that the precision of all study areas is above 90%; however, the recall is around 80%. The reasons for the low recall are as follows: First, the low recall is a frequent phenomenon for small building recognition in complex image scenes, even when buildings are recognized in complex image scenes using visual interpretation; second, the method proposed in this study obtains more precise segmentation by combining the building region-line features. Therefore, the building line features influence the accuracy. However, in this experiment, the selected edge detection sample production area does not cover a wide enough area, and there are very few high-rise buildings in study area A. This phenomenon leads to insufficient high-rise building samples in the edge detection dataset and poor sample representativeness. So there is a problem of missed detection, which is the main reason for the low recall of the algorithm. The total accuracy of the three study areas is also approximately 90%, especially the

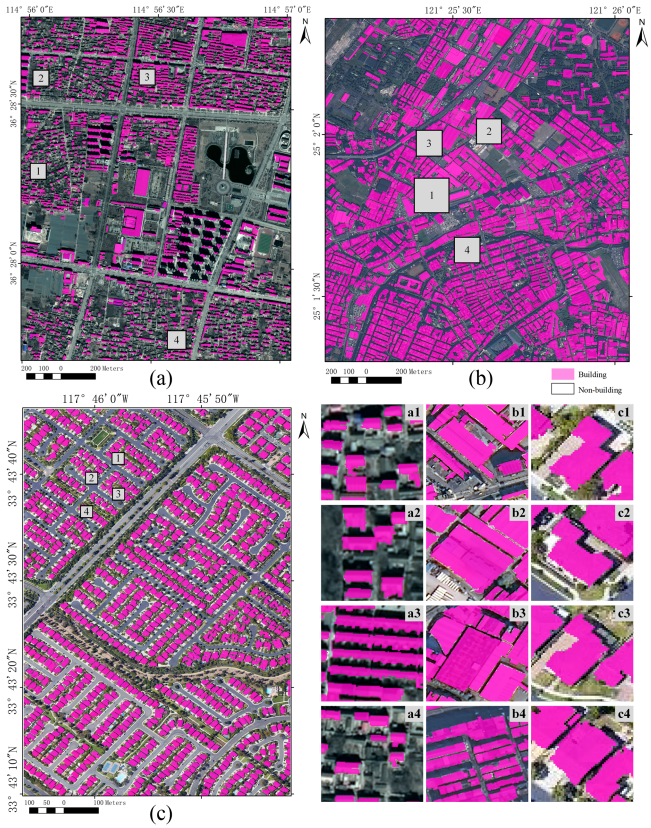


Fig. 9. Building extraction results for the study area. (a) Study area A; (b) study area B; (c) study area C. 1–4 subregions of different study areas are highlighted and enlarged in the lower right corner.

TABLE II
ACCURACY EVALUATION RESULT OF BUILDING EXTRACTION

| Indicators | P(%) | R(%) | OA(%) | F1 |
|---|---|---|---|---|
| Study Area A | 92.00 | 80.50 | 94.80 | 0.86 |
| Study Area B | 90.49 | 82.45 | 88.10 | 0.86 |
| Study Area C | 97.03 | 93.55 | 97.50 | 0.95 |

OA of study area C reaches 97.50%. Additionally, the F1 score reaches 0.95, demonstrating that this study's proposed method can accurately extract the buildings in complex image scenes.

## V. DISCUSSION

### A. *Efficiency of MA-DexiNed for Building Edge Detection*

The MA-DexiNed proposed in this study is enhanced for the problem of easy missed and false detection of small objects. In this section, to verify the effectiveness of MA-DexiNed, it is compared with RCF and DexiNed. The edge detection results of RCF, DexiNed, and MA-DexiNed are shown in Fig. 10, respectively. For the fairness of the experiment, the training and testing datasets used for all the above networks are the same.

The experimental results show that the proposed method in this study can obtain clearer, more refined, and precise building edge features compared with RCF and DexiNed. Among them,


Fig. 10. Comparison of building edge detection results. (a) Remote sensing images. (b) RCF. (c) DexiNed. (d) MA-DexiNed.

comparing the edge detection findings of RCF with other networks, it is discovered that RCF tends to induce the phenomenon of blurred edges and incomplete edges in the face of small and dense rural buildings. Because the structure of RCF is relatively simple and weak in detecting small or complex objects, DexiNed can obtain clearer building edges than RCF. However, the building edges it extracts are substantially broken and omitted. However, because of the dense multilayer structure, the combination of long and short hopping connections and its particular upsampling module make DexiNed perform well in small target edge detection tasks. Furthermore, MA-DexiNed inserts the CBAM module into the network, adjusts the downsampling module, and increases the network's channels. This makes the network more capable of feature extraction and selection. Therefore, comparing the detection results, the proposed network in this study can obtain more accurate and complete building edge features, significantly reducing the missed detection rate of small or complex buildings.

### B. *Effectiveness of Region-Line Feature Fusion for Building Extraction*

With the emergence of deep learning, there are an increasing number of situations of merging classic segmentation methods

TABLE III
ACCURACY EVALUATION RESULTS OF DIFFERENT METHODS

| Study Area | Method | P(%) | R(%) | OA(%) | F1 |
|---|---|---|---|---|---|
| Study Area A | SEEDS-CNN [33] | 62.15 | 55.00 | 84.00 | 0.59 |
| | SLIC-CNN [33] | 65.06 | 58.46 | 87.10 | 0.62 |
| | SLICO-CNN [34] | 63.04 | 58.00 | 84.80 | 0.60 |
| | MRS-CNN [34] | 65.82 | 52.00 | 85.00 | 0.58 |
| | RCF-MRS-CNN | 73.30 | 60.50 | 87.80 | 0.66 |
| | DexiNed-MRS-CNN | 67.03 | 61.39 | 86.40 | 0.64 |
| | **The proposed method** | **92.00** | **80.50** | **94.80** | **0.86** |
| Study Area B | SEEDS-CNN [33] | 84.76 | 75.91 | 83.20 | 0.80 |
| | SLIC-CNN [33] | 82.01 | 74.36 | 80.80 | 0.78 |
| | SLICO-CNN [34] | 82.19 | 76.76 | 81.30 | 0.79 |
| | MRS-CNN [34] | 83.48 | 81.88 | 83.90 | 0.82 |
| | RCF-MRS-CNN | 84.81 | 77.40 | 82.90 | 0.81 |
| | DexiNed-MRS-CNN | 84.67 | 81.24 | 84.20 | 0.83 |
| | **The proposed method** | **90.49** | **82.45** | **88.10** | **0.86** |
| Study Area C | SEEDS-CNN [33] | 77.44 | 82.44 | 88.50 | 0.80 |
| | SLIC-CNN [33] | 75.42 | 80.29 | 87.20 | 0.78 |
| | SLICO-CNN [34] | 77.85 | 80.65 | 88.20 | 0.79 |
| | MRS-CNN [34] | 77.68 | 91.04 | 90.30 | 0.84 |
| | RCF-MRS-CNN | 82.08 | 93.55 | 92.60 | 0.87 |
| | DexiNed-MRS-CNN | 86.96 | 93.19 | 94.30 | 0.90 |
| | **The proposed method** | **97.03** | **93.55** | **97.50** | **0.95** |

with deep learning approaches, all of which have produced reasonably decent results. In this section, the proposed method in this study is compared with the many mature methods, which are SEEDS-CNN [37], SLIC-CNN [37], SLICO-CNN [38], and MRS-CNN [38]. Among them, SEEDS, SLIC, SLICO, and MRS are the more applied image segmentation methods. Additionally, RCF-MRS-CNN and DexiNed-MRS-CNN are added, adopting RCF and DexiNed rather than MA-DexiNed in the line feature extraction module.

Table III shows the accuracy evaluation results obtained from the above seven methods in the building extraction experiments in study areas A, B, and C, respectively. In the accuracy evaluation results of study area A, the precision of the proposed method is 92.00%, the recall is 80.50%, the OA is 94.80%, and the F1 score is 0.86, which is significantly higher than the other methods. Similarly, the accuracy of the proposed method is higher than the other methods in both study areas, B and C. Study Area A is the most rural of the three study regions, with short, decrepit dwellings. Manual interpretation is also difficult to differentiate. Therefore, the buildings in study area A are the most challenging to extract. The proposed method in this study is dedicated to solving the problem of harrowing building extraction in complex image scenes, so the method is more applicable to complex image scenes than simple image scenes. This is why the most noticeable index improvement is observed in the study area A. Additionally, the enhancement of P is higher than that of R in the three study areas. This is caused by the object-based image segmentation method. Because the misclassification caused by a single image segmentation is usually that the segmented object is larger than the actual building, this makes CNN misclassify nonbuildings into buildings more frequently than buildings into nonbuildings when performing the classification. Therefore, the approach enhancement in this study leads to a considerable reduction in the misclassification of non-buildings as buildings. Thus, it leads to a significantly higher enhancement of P than R.

Fig. 11 compares the local findings of building extraction by the above seven methods in the three study areas. From the marked red areas in the figure, it is discovered that the technique with region-line feature fusion can extract the building units with more accurate edges. Simultaneously, the other methods are more obvious misclassification and omission in the corresponding marked areas.

The advantages and disadvantages of different methods are observed by combining the accuracy evaluation results and the visual effect of building extraction. First, by comparing the SEEDS-CNN, SLIC-CNN, SLICO-CNN, and MRS-CNN methods, it is found that the precision of the above methods is generally low in study area A. Because the buildings in study area A are much more difficult to extract than in study areas B and C, obtaining more desirable extraction results without region-line feature fusion is difficult, which the details in Fig. 11 can also confirm. However, the accuracies of the MRS-CNN method in study areas B and C are significantly higher than other methods. Furthermore, because SEEDS, SLIC, and SLICO are three superpixel segmentation algorithms, the size of segmented objects obtained by these algorithms is closer to the size requirement of the receptive field of the CNN. However, this makes the superpixel segmentation highly susceptible to confusing superpixels, and the edge accuracy of their segmented objects is significantly lower than that of the MRS. Therefore, MRS can obtain more accurate image segmentation results than other methods without region-line feature fusion. Second, when MRS-CNN, RCF-MRS-CNN, DexiNed-MRS-CNN, and

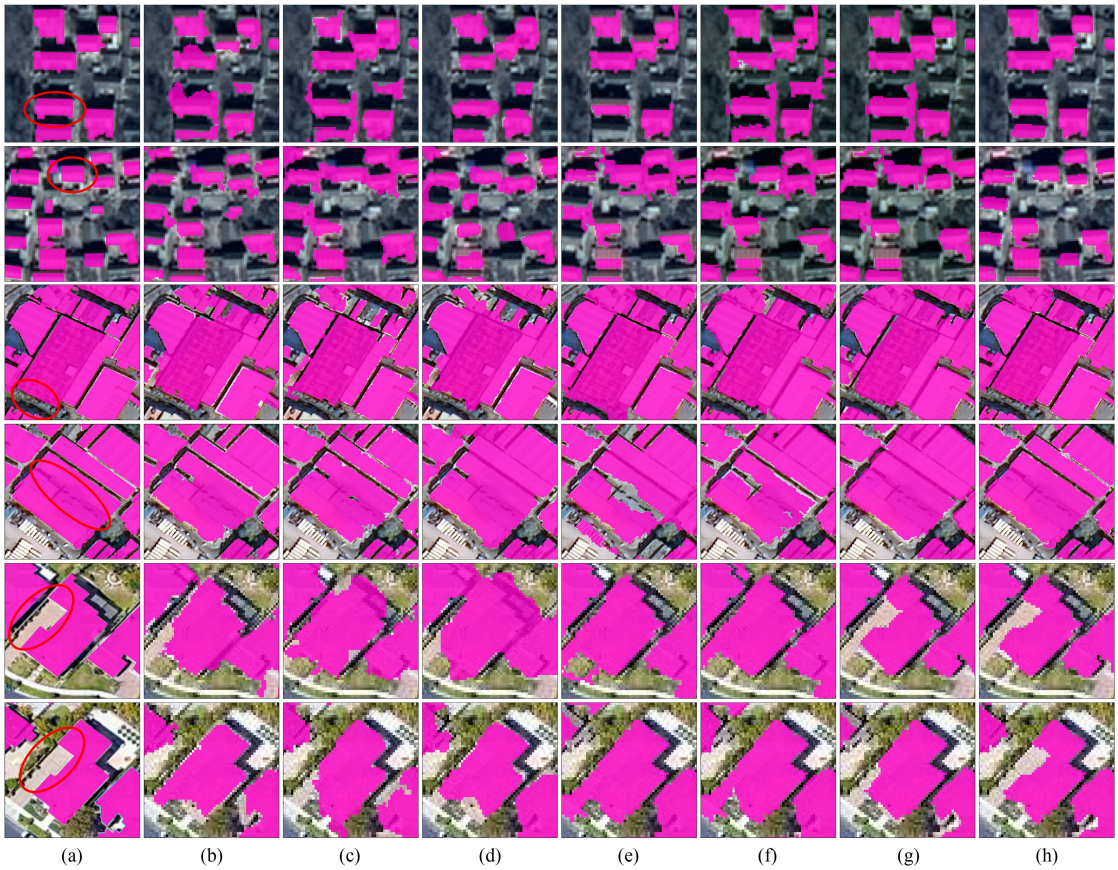


Fig. 11. Comparison of extraction results by different methods. (a) Manual annotation. (b) SEEDS-CNN. (c) SLIC-CNN. (d) SLICO-CNN. (e) MRS-CNN. (f) RCF-MRS-CNN. (g) DexiNed-MRS-CNN. (h) Proposed method. The pink area in the figure shows the buildings extracted by the above methods.

the suggested approach are compared, it is discovered that the region-line feature fusion method can significantly increase the accuracy of building extraction, demonstrating the efficacy of the innovative feature fusion method described in this work. Because the essence of the region-line feature fusion is to perform the oversegmentation based on the initial segmentation result obtained by MRS. This oversegmentation is specifically for the building edges. In other words, this method can separate the edges from the confused segmentation objects at the edges of the buildings. Therefore, this method achieves the effect of optimizing edge segmentation. In addition, in our three study areas, the former two have imperfect samples (only discontinuous edges are available). Thanks to the advantages of the unsupervised MRS, our method can still achieve excellent building extraction results in such cases. Finally, comparing RCF-MRS-CNN, DexiNed-MRS-CNN, and the proposed method in this study, it is found that the accuracies of the proposed method in the three study areas are significantly higher than the other two methods, which again verifies the efficacy of the MA-DexiNed in the task of edge feature extraction.

### *C. Suitability of the Proposed Method for Building Extraction for Complex Image Scenes*

Compared to complex image scenes, buildings in simple image scenes have more precise building contours, usually of a single type. Therefore, extracting structures from complex image scenes is much more challenging than simple image scenes. To verify the method's suitability in this study for different scenes, in this subsection, the ISPRS Potsdam dataset is used to compare study areas A, B, and C. The ISPRS Potsdam dataset is one of the datasets of the ISPRS [62] 2D Semantic Labeling Challenge, which was taken in Potsdam, Germany, with a resolution of 0.05 m and four bands, including NIR. This dataset is of a very high resolution and sparse building distribution, presenting a typical simple image scene. The comparative experiment uses the same data pre-processing as in study areas A, B, and C to generate the same amount of samples.

In Fig. 12, the figures of the first two lines are respectively from study areas A and B, and the figures of the last two lines are from the ISPRS Potsdam dataset. However, comparing the building extraction results of different methods in various scenes, it can be discovered that the proposed method in this study has more obvious advancement than the ISPRS Potsdam dataset regarding edge accuracy and completeness in study areas A and B. It can also be found in Table IV that the precision of the suggested method in study areas A, B, and C is generally improved by more than 10% compared to other methods, and different accuracy indicators also have more substantial improvement. In contrast, this proposed method does not outperform the former when it meets ISPRS Potsdam dataset covered by simple scenes. Its accuracy improvement is limited, implying the superiority of

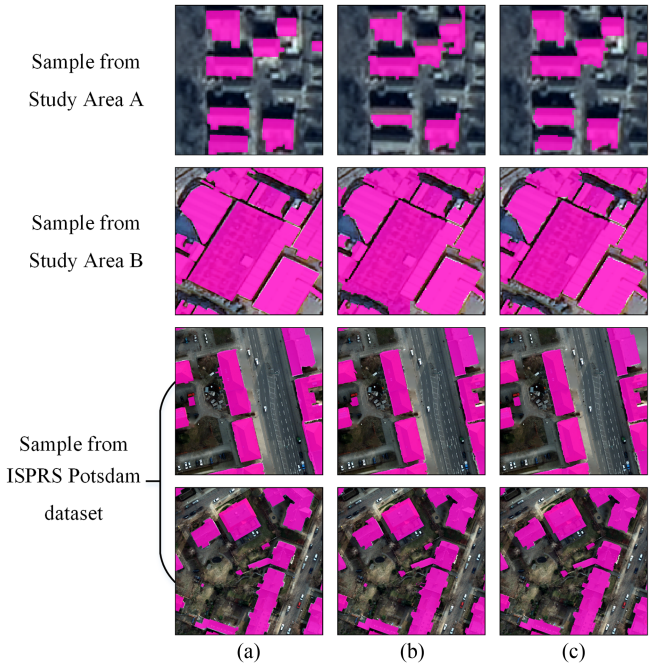


Fig. 12. Comparison of extraction results of different image scenes. (a) Manual annotation. (b) MRS-CNN. (c) DexiNed-MRS-CNN. (d) Proposed method. The pink area in the figure shows the buildings extracted by the above methods.

TABLE IV
COMPARISON OF ACCURACY EVALUATION RESULTS OF DIFFERENT IMAGE SCENES

| Study Area | Method | P(%) | R(%) | OA(%) | F1 |
|---|---|---|---|---|---|
| Study Area A | MRS-CNN [34] | 65.82 | 52.00 | 85.00 | 0.58 |
| | DexiNed-MRS-CNN | 67.03 | 61.39 | 86.40 | 0.64 |
| | **The proposed method** | **92.00** | **80.50** | **94.80** | **0.86** |
| Study Area B | MRS-CNN [34] | 83.48 | 81.88 | 83.90 | 0.82 |
| | DexiNed-MRS-CNN | 84.67 | 81.24 | 84.20 | 0.83 |
| | **The proposed method** | **90.49** | **82.45** | **88.10** | **0.86** |
| Study Area C | MRS-CNN [34] | 77.68 | 91.04 | 90.30 | 0.84 |
| | DexiNed-MRS-CNN | 86.96 | 93.19 | 94.30 | 0.90 |
| | **The proposed method** | **97.03** | **93.55** | **97.50** | **0.95** |
| ISPRS | MRS-CNN [34] | 93.70 | 96.20 | 97.30 | 0.95 |
| | DexiNed-MRS-CNN | 93.12 | **97.72** | 97.50 | 0.95 |
| | **The proposed method** | **95.51** | 96.96 | **98.00** | **0.96** |

the proposed method over the traditional methods when applied to complex scenes. However, the advancements MA-DexiNed in this study have stronger feature mining ability, which makes the network more suitable for complex image scenes than other networks. Nevertheless, the proposed method combines the line features obtained by MA-DexiNed and the region features obtained by MRS so that the confused segmentation objects that often occur when applied to complex image scenes can be precisely segmented. Therefore, it is appreciable that this study's proposed building extraction method is more suitable for complex image scenes.

### D. Limitations of the Proposed Method

This method also has certain limitations. For example, the MA-DexiNed can obtain more accurate and thinning edge features; however, it is still incapable of absolutely preventing missed detection and false detection. This is related to the network's performance and the small coverage area of the edge detection samples adopted in this method.

Simultaneously, although the proposed framework has a good extraction effect on building edges, the time cost of this method is high. Because the proposed framework cascades several different remote sensing image processing algorithms, including edge detection, postprocessing algorithms, image segmentation, and OCNN, some of which involve specifically complex processing or even professional software to complete, makes the end-to-end method integration is fraught with difficulties. In addition, the data used in the experiments are all off-nadir images, so there are many shadows and occlusions in the images.

## VI. CONCLUSION

With the emerging development of domestic HRS satellite technology, the interest in building extraction is growing. However, buildings in complex image scenes extracted by most existing methods are inaccurate. This study provides an approach that combines region-line feature fusion with OCNN to extract structures in complicated visual situations by cascading a deep learning edge detection network with OCNN. The experimental results demonstrate that the proposed method can extract accurate and complete edges of buildings. At the same time, the technique performs more prominently in complex image scenes. Therefore, it is more applicable to complex image scenes than simple ones.

The significant contributions of this study include the following.

1) A new edge detection network named MA-DexiNed for building edge extraction in complex image scenes is proposed. The experimental results of building edge extraction demonstrate that the network can obtain good detection results regarding building edge accuracy and small buildings detection.
2) A novel edge thinning connection algorithm according to the rule judgment is proposed, which experimentally verified the merits of thinning the edge probability maps and maintaining contiguous line features of buildings.
3) For the complex image scene building extraction task, a novel method is proposed to fuse supervised segmentation-based building edge line features with unsupervised segmentation-based building region features and

combine them with OCNN. This method improves the accuracy of image segmentation and provides a new idea for accurately extracting buildings from complex image scenes or in the case of imperfect samples.

However, the suggested technique in this study concentrates on the accuracy and completeness of building edges, and it does not consider the entire process's overhead regarding time and cost. Therefore, the essential research focus for the future is how to increase the method's automation and lower the time cost. To overcome this issue, alternative, well-designed approaches can be used instead of MRS to create an end-to-end building extraction network structure with a high degree of integration. Additionally, it should be noted that we only apply the method to the binary classification task of buildings. Due to the edge detection network, the current method can only enhance the edges of one type of object in one classification task. It is also challenging to break through this limitation to achieve fast multiclassification. However, it is possible to attempt to design multilayer edge detection networks to achieve hierarchical extraction of the edges of various objects to achieve the effect of enhancement of every edge in the whole image. This can provide novel ideas for land cover classification with high accuracy.

## REFERENCES


[1] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[2] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 808.

[3] Y. Qin, Z. Cao, H. Li, X. Wang, and W. Zhuo, "Building localization from forward-looking infrared images for UAV guidance," *Gyroscopy Navigation*, vol. 4, no. 4, pp. 188–197, 2013.

[4] Q. Hu, L. Zhen, Y. Mao, X. Zhou, and G. Zhou, "Automated building extraction using satellite remote sensing imagery," *Automat. Construction*, vol. 123, 2021, Art. no. 103509.

[5] Z. Shao, Z. Zhou, X. Huang, and Y. Zhang, "MRENet: Simultaneous extraction of road surface and road centerline in complex urban scenes from very high-resolution images," *Remote Sens.*, vol. 13, no. 2, 2021, Art. no. 239.

[6] Q. Zhang and J. Wang, "Detection of buildings from Landsat-7 ETM+ and SPOT panchromatic data in Beijing, China," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2002, vol. 5, pp. 2977–2979.

[7] B. Peng, D. Ren, C. Zheng, and A. Lu, "TRDet: Two-stage rotated detection of rural buildings in remote sensing images," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 522.

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

[9] T. Zuo, J. Feng, and X. Chen, "HF-FCN: Hierarchically fused fully convolutional network for robust building extraction," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 291–302.

[10] T. Lu, D. Ming, X. Lin, Z. Hong, X. Bai, and J. Fang, "Detecting building edges from high spatial resolution remote sensing imagery using richer convolution features network," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1496.

[11] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1480–1484.

[12] L. Xu, D. Ming, T. Du, Y. Chen, D. Dong, and C. Zhou, "Delineation of cultivated land parcels based on deep convolutional networks and geographical thematic scene division of remotely sensed images," *Comput. Electron. Agriculture*, vol. 192, 2022, Art. no. 106611.

[13] A. Jiao, M. He, and H. Luo, "Research on significant edge detection of infrared image based on deep learning," *Infrared Technol.*, vol. 41, no. 1, pp. 72–77, 2019.

[14] X. Lv, D. Ming, T. Lu, K. Zhou, M. Wang, and H. Bao, "A new method for region-based majority voting CNNs for very high resolution image classification," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1946.

[15] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

[16] G. Yang, Q. Zhang, and G. Zhang, "EANet: Edge-aware network for the extraction of buildings from aerial images," *Remote Sens.*, vol. 12, no. 13, 2020, Art. no. 2161.

[17] D. Li, G. Zhang, Z. Wu, and L. Yi, "An edge embedded marker-based watershed algorithm for high spatial resolution remote sensing image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2781–2787, Oct. 2010.

[18] J. Chen, J. Li, D. Pan, Q. Zhu, and Z. Mao, "Edge-guided multiscale segmentation of satellite multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4513–4520, Nov. 2012.

[19] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[20] I. Sobel, *Camera Models and Machine Perception*. Stanford, CA, USA: Stanford Univ., 1970.

[21] A. Kucharski and A. Fabijańska, "CNN-watershed: A watershed transform with predicted markers for corneal endothelium image segmentation," *Biomed. Signal Process. Control*, vol. 68, 2021, Art. no. 102805.

[22] M. Baatz and A. Schape, "Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation," in *Proc. Beiträge zum AGIT-Symp.*, 2000, pp. 12–23.

[23] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 1050.

[24] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[25] X. Jin and C. H. Davis, "Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information," *Eur. Assoc. Signal Process. J. Adv. Signal Process.*, vol. 2005, no. 14, pp. 1–11, 2005.

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

[27] X. Huang and L. Zhang, "An adaptive mean-shift analysis approach for object extraction and classification from urban hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4173–4185, Dec. 2008.

[28] P. S. Tiwari, H. Pande, and B. N. Nanda, "Building footprint extraction from IKONOS imagery based on multi-scale object oriented fuzzy classification for urban disaster management," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 34, pp. 1–7, 2006.

[29] Y. Bengio, *Learning Deep Architectures for AI*. Norwell, MA, USA: Now Publishers Inc, 2009.

[30] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015.

[31] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[32] F. Ji et al., "Aircraft detection in high spatial resolution remote sensing images combining multi-angle features driven and majority voting CNN," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2207.

[33] B. Zeng et al., "Top-down aircraft detection in large-scale scenes based on multi-source data and FEF-R-CNN," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1108–1130, 2022.

[34] S. Wei, T. Zhang, S. Ji, M. Luo, and J. Gong, "BuildMapper: A fully learnable framework for vectorized building contour extraction," *J. Photogrammetry Remote Sens.*, vol. 197, pp. 87–104, 2023.

[35] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5614812.

[36] Z. Liu, Q. Shi, and J. Ou, "LCS: A collaborative optimization framework of vector extraction and semantic segmentation for building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5632615.

[37] X. Lv, D. Ming, Y. Chen, and M. Wang, "Very high resolution remote sensing image classification with SEEDS-CNN and scale effect analysis for superpixel CNN classification," *Int. J. Remote Sens.*, vol. 40, no. 2, pp. 506–531, 2019.

[38] Y. Chen, D. Ming, and X. Lv, "Superpixel based land cover classification of VHR satellite image combining multi-scale CNN and scale parameter estimation," *Earth Sci. Inform.*, vol. 12, no. 3, pp. 341–363, 2019.
[39] K. Zhang et al., "Distance weight-graph attention model-based high-resolution remote sensing urban functional zone identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608518.
[40] W. Zhou, D. Ming, X. Lv, K. Zhou, H. Bao, and Z. Hong, "SO–CNN based urban functional zone fine division with VHR remote sensing image," *Remote Sens. Environ.*, vol. 236, 2020, Art. no. 111458.
[41] 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.
[42] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[43] 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.
[44] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[45] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[46] J. Liu, H. Huang, H. Sun, Z. Wu, and R. Luo, "LRAD-Net: An Improved lightweight network for building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 675–687, 2023.
[47] J. Hui, M. Du, X. Ye, Q. Qin, and J. Sui, "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 786–790, May 2019.
[48] Q. Li, Y. Shi, and X. X. Zhu, "Semi-supervised building footprint generation with feature and output consistency training," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623217.
[49] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10548–10559, 2021.
[50] L. Xia, X. Zhang, J. Zhang, H. Yang, and T. Chen, "Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2187.
[51] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.
[52] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[53] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu, "Statistical edge detection: Learning and evaluating edge cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 1, pp. 57–74, Jan. 2003.
[54] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.
[55] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3828–3837.
[56] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 193–202.
[57] X. S. Poma, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust cnn model for edge detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1923–1932.
[58] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. Assoc. Comput. Machinery*, vol. 60, no. 6, pp. 84–90, May 2017.
[59] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[60] R. Liu et al., "Multiscale road centerlines extraction from high-resolution aerial imagery," *Neurocomputing*, vol. 329, pp. 384–396, 2019.
[61] M. Rezaee, M. Mahdianpari, Y. Zhang, and B. Salehi, "Deep convolutional neural network for complex wetland classification using optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3030–3039, Sep. 2018.
[62] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

**Dehui Dong** is currently working toward the Ph.D. degree in the School of Information Engineering, China University of Geosciences (Beijing), Beijing, China.

His research interests include deep learning for high spatial resolution remote sensing images processing and applications.

**Dongping Ming** (Member, IEEE) received the B.E. degree in land administration and cadastral surveying from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1999, the M.E. degree in cartography and geographic information engineering from the Wuhan University, Wuhan, in 2002, and the Ph.D. degree in cartography and geographic information system from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, in 2006.

She is currently a Full Professor with the School of Information Engineering, China University of Geosciences, Beijing. Her research interests include high spatial resolution remote sensing image processing and analysis, remote sensing application of resources and environment monitoring, intelligent prevention and control of geo-hazards.

**Qihao Weng** (Fellow, IEEE) received the A.S. degree in geography from the Minjiang University, Fuzhou, China, in 1984, the M.S. degree in geography from the South China Normal University, Guangzhou, China, in 1990, the M.A. degree from The University of Arizona, Tucson, AZ, USA, in 1996, all in geography, and the Ph.D. degree in remote sensing and geographic information system from the University of Georgia, Athens, GA, USA, in 1999.

He is a Professor of Earth and environmental systems and the Director of the Center for Urban and Environmental Change, Indiana State University, Terre Haute, IN, USA. From 2008 to 2009, he visited the NASA Marshall Space Flight Center, Huntsville, AL, USA, as a Senior Research Fellow. He has authored 241 articles and 14 books, with over 19 900 citations and an H-index of 64. His research focuses on remote sensing analysis of urban ecological and environmental systems, land-use and land-cover changes, urbanization impacts, and environmental sustainability.

Dr. Weng is also an Elected Fellow of American Association for the Advancement of Science (AAAS) and American Society for Photogrammetry and Remote Sensing (ASPRS) and a member of International Society for Photogrammetry and Remote Sensing (ISPRS), American Geophysical Union (AGU), and American Association of Geographers (AAG). He was the recipient of distinguished career awards, including the NASA Senior Fellowship, the AAG Distinguished Scholarship Honors Award, the Taylor & Francis Lifetime Achievements Award, and the Japan Society for the Promotion of Science (Short-term S[E]) Fellowship. He has been the Organizer and Program Committee Chair of the biennial IEEE sponsored International Workshop on Earth Observation and Remote Sensing Applications (EORSA) conference series since 2008. He was the National Director of ASPRS from 2007 to 2010. He has been invited to give more than 110 talks by organizations and conferences worldwide. He also serves as the Editor-in-Chief of the *ISPRS Journal of Photogrammetry* and *Remote Sensing*. He is also a Series Editor of *Remote Sensing Applications Series* (Taylor & Francis) and *The Imaging Science Journal* series (Taylor & Francis).

**Yi Yang** received the B.E. degree in computer science and technology and M.E. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Hubei, China, in 2009 and 2012, respectively.

He is currently a Research Assistant with natural resources investigation and monitoring research center, Chinese Academy of Surveying and Mapping, Beijing, China. His research interests include remote sensing image processing and deep learning.

**Kun Fang** received the B.E. degree in geographic information system from the Central South University, Changsha, China, in 2004, and the Ph.D. degree in cartography and geographic information engineering from the China University of Geosciences, Beijing, China, in 2009.

He is currently a Lecturer with the School of Information Engineering, China University of Geosciences. His research interests include spatial database and spatio-temporal data analysis**.**

**Lu Xu** is currently working toward the Ph.D. degree in surveying and mapping from the School of Information Engineering, China University of Geosciences, Beijing, China.

His research interests include comprehending very high spatial resolution images. The specific research interests include Object-Based Image Analysis (OBIA) theory and image information extraction by deep learning methods.

**Tongyao Du** is currently working toward the master's degree in surveying and mapping with the China University of Geosciences (Beijing), Beijing, China.

Her research interests include high-resolution remote sensing imagery processing and applications by deep learning methods.

**Yu Zhang** is currently working toward the master's degree in surveying and mapping with the School of Information Engineering, China University of Geosciences (Beijing), Beijing, China.

Her research interests include intelligent processing and applications of high spatial resolution remote sensing images.

**Ran Liu** is currently working toward the master's degree in surveying and mapping with China University of Geosciences (Beijing), Beijing, China.

His research interests include image information extraction by deep learning and landslide susceptibility assessment.