Refined Extraction Of Building Outlines From High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering

Yakun Xie, Jun Zhu, Yungang Cao, Dejun Feng, Minjun Hu, Weilian Li, Yunhao Zhang, and Lin Fu

***Abstract*—The automatic extraction of building outlines from high-resolution images is an important and challenging task. Convolutional neural networks have shown excellent results compared with traditional building extraction methods because of their ability to extract high-level abstract features from images. However, it is difficult to fully utilize the multiple features of current building extraction methods; consequently, the resulting building boundaries are irregular. To overcome these limitations, we propose a method for extracting buildings from high-resolution images using a multifeature convolutional neural network (MFCNN) and morphological filtering. Our method consists of two steps. First, the MFCNN, which consists of residual connected unit, dilated perception unit, and pyramid aggregation unit, is used to achieve pixel-level segmentation of the buildings. Second, morphological filtering is used to optimize the building boundaries, improve the boundary regularity, and obtain refined building boundaries. The Massachusetts and Inria datasets are selected for experimental analysis. Under the same experimental conditions, the extraction results achieved with the proposed MFCNN are compared with the results of other deep learning models that have been commonly used in recent years: FCN-8s, SegNet, and U-Net. The results on both datasets reveal that the proposed model improves the F1-score by 3.31%–5.99%, increases the overall accuracy (OA) by 1.85%–3.07%, and increases the intersection over union (IOU) by 3.47%–8.82%. These findings illustrate that the proposed method is effective at extracting buildings from complex scenes.**

***Index Terms*—Building outline extraction, high-resolution images, morphological filtering, multifeature convolutional neural network (MFCNN).**

I. Introduction

As A critical component of basic urban geographic information, buildings play an important role in population estimation, change monitoring, urban planning, and smart city construction [1]–[3]. Consequently, the automatic extraction of building outlines from high-resolution images has always been a fundamental task in the field of remote sensing research [4].

In recent years, as the spatial resolution of remote sensing images has increased remarkably, remote sensing-based applications involving building extraction techniques have seen considerable development [5], [6]. Many building extraction algorithms have been proposed by scholars [7]–[12]. Depending on the data sources used, the existing methods can be divided into the following three categories.

1) Optical image methods [13]–[16].
2) Light detection and ranging point cloud methods [17], [18].
3) Combinations of optical image and point cloud methods [19]–[22].

Optical image methods rely on the spatial and spectral characteristics of optical images, such as textural features [14], geometric features [15], edge features [16], multispectral features [8], and shadow features [13]. LiDAR point cloud methods rely mainly on information extracted from point cloud data, such as elevation information, which can be clearly captured in point cloud data and used to identify buildings [17], [18]. Methods of the third type use multisensor data for building extraction. Such methods can achieve better results through the combined consideration of complementary features such as spectral information, spatial features, and elevation [19]. These combined approaches have achieved certain success in the extraction of building outlines; however, remote sensing images typically exhibit nonuniform regions, large intraclass variances, and low interclass variances, making it impossible to establish a suitable predefined model for object extraction. Moreover, the types of objects captured in such images are diverse and complex, and hence, traditional segmentation, classification, and edge extraction algorithms lack the ability to perform deep semantic feature extraction. Nevertheless, it is possible to automatically extract deep image features through machine learning methods by constructing deep neural networks.

The theory of deep learning was originally proposed in 2006 by Hinton *et al.* [23]. Deep learning constitutes the process of acquiring high-level abstract features from data by constructing mathematical models to achieve improvements in the classification accuracy and detection accuracy. In recent years, a

Manuscript received July 30, 2019; revised November 26, 2019 and February 27, 2020; accepted April 26, 2020. Date of publication April 30, 2020; date of current version May 15, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 41871289 and Grant 41771451, and in part by the Sichuan Youth Science and Technology Innovation Team under Grant 20CXTD0102. *(Corresponding author: Yungang Cao.)*

The authors are with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: yakunxie@my.swjtu.edu.cn; zhujun@swjtu.edu.cn; yungang@swjtu.cn; djfeng@swjtu.edu.cn; demon@my.swjtu.edu.cn; vgewilliam@my.swjtu.edu.cn; zhangyh0506@my.swjtu.edu.cn; vge_fulin@my.swjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.2991391

variety of neural network models have been proposed, including convolutional neural networks (CNNs) [24], recurrent neural networks [25], and deep belief networks [26]. These networks have been employed for a variety of high-performance computer vision tasks, such as image classification [24], natural language processing [27], speech recognition [28], remote sensing image processing [29], [30] and other applications [31], [32]. Among these networks, CNNs have achieved superior results in image classification tasks. Consequently, numerous scholars have developed many improved algorithms based on CNNs, such as FCN [33], SegNet [34], U-Net [35], Mask R-CNN [36], and DeepLab [37]–[39].

On the basis of this research, many neural network algorithms have been applied to remote sensing images for building extraction. Maggiori *et al.* used FCN, Skip, and multilayer perceptron (MLP) models for building extraction experiments on multiregion aerial images [40]. Zhao *et al.* used Mask R-CNN to extract and regularize the boundaries of buildings [41]. Both of the above-mentioned methods use existing basic models to extract buildings from images without considering the distinctive characteristics of those buildings in remote sensing images. In addition, Mhin used the patch CNN method for building extraction; principal component analysis (PCA) was employed to reduce the dimensionality of the original images, and the output was postprocessed by means of conditional random fields to improve the accuracy [42]. Furthermore, Alidoost *et al.* used a patch-based CNN architecture to extract both roads and buildings [43]. However, CNN models consider fixed-level features, limiting their recognition ability, and preventing the effective extraction of multiscale building features.

More recently, many neural network frameworks have been specifically adapted to the multiscale features of remote sensing images. For instance, Yuan designed a deep neural network with a simple structure for the extraction of buildings from aerial scenes and generated a large amount of labeled data using geographic information system building footprint data [44]. Deng *et al.* used a novel feature extraction method combined with a multiscale object proposal network and an accurate object detection network to construct multiscale features for building extraction [45]. Liu *et al.* extracted hierarchical building information through a multilevel building detection framework based on deep learning models consisting of the Gaussian pyramid technique and CNNs [46]. Li *et al.* developed a novel deep adversarial network for determining the high-order regularities of buildings; the network consists of both a generator and a discriminator, where the former is a deep CNN and the latter is an adversarial discriminator network [47]. Bittner *et al.* extracted building information from multisource remote sensing images using an end-to-end FCN that combines spectral and height information from different data sources and automatically generates a full-resolution binary build mask [1]. Lin *et al.* constructed a deep network architecture by combining residual blocks and dilated convolutions to extract buildings and improve the computational efficiency [48]. Majd *et al.* proposed a novel object-based deep CNN to solve the variation in scale of objects in very-high-resolution (VHR) images [49]. Liu *et al.* used a deep convolutional encoder–decoder with spatial pyramid pooling to reduce the loss of detailed information [50]. Although these works are all notable, the current feature utilization capability for remote sensing images is still insufficient, as described by each of the following shortcomings.

1) Imperfect image feature acquisition. In the encoding stage of a neural network, additional features can be acquired by increasing the depth of the network. However, increasing the network depth can result in a vanishing gradient or gradient divergence, which will reduce the overall performance of the network and lead to imperfect image feature acquisition.
2) Insufficient acquisition of morphological building features from remote sensing images. Buildings exhibit several obvious morphological features. In the encoding stage of a neural network, shallow features can be acquired by means of small local receptive fields (e.g., a 3×3 kernel). However, it is impossible to effectively learn deep morphological features, such as the obvious linear and right-angle features of buildings, from a high-resolution image.
3) Inadequate consideration of the multiscale features of remote sensing images. Each layer of a network contains unique information. In the decoding stage of a neural network, a single bottom-level feature is convolved upwards layer by layer, and the coding features cannot be effectively integrated with the decoding features. Consequently, the multiscale features of images cannot be comprehensively captured.
4) Irregularity of extracted building boundaries. The extracted boundaries of buildings typically exhibit many sawtooth features due to segmentation based on pixel-level semantics.

To overcome these limitations, this article proposes a building extraction framework based on a multifeature convolutional neural network (MFCNN) that considers the distinctive characteristics of remote sensing images. The network framework consists of three components. First, a refined residual connection unit (RCU) is used to improve the ability to learn deep and complex features during the encoding process. Then, adaptive dilated perception units (DPUs) are used to better map the morphological features of buildings. Finally, the loss of multiscale features is avoided by means of a pyramid aggregation unit (PAU). On this basis, to optimize the regularity of the building boundaries, a morphological filtering-based building regularization method is developed that uses morphological filtering to optimize the overall contour information to mitigate sawtooth phenomena along building boundaries. The results obtained in this way are more consistent with the real boundaries of buildings.

The rest of this article is organized as follows. The methods are presented in Section II, including a detailed description of the neural network framework and the regularization method. In Section III, the dataset images, parameter settings, and results are described in detail. The results of comparisons with other methods and a sensitivity analysis of our method are reported in Section IV. Finally, the conclusion and a discussion of future work are presented in Section V.

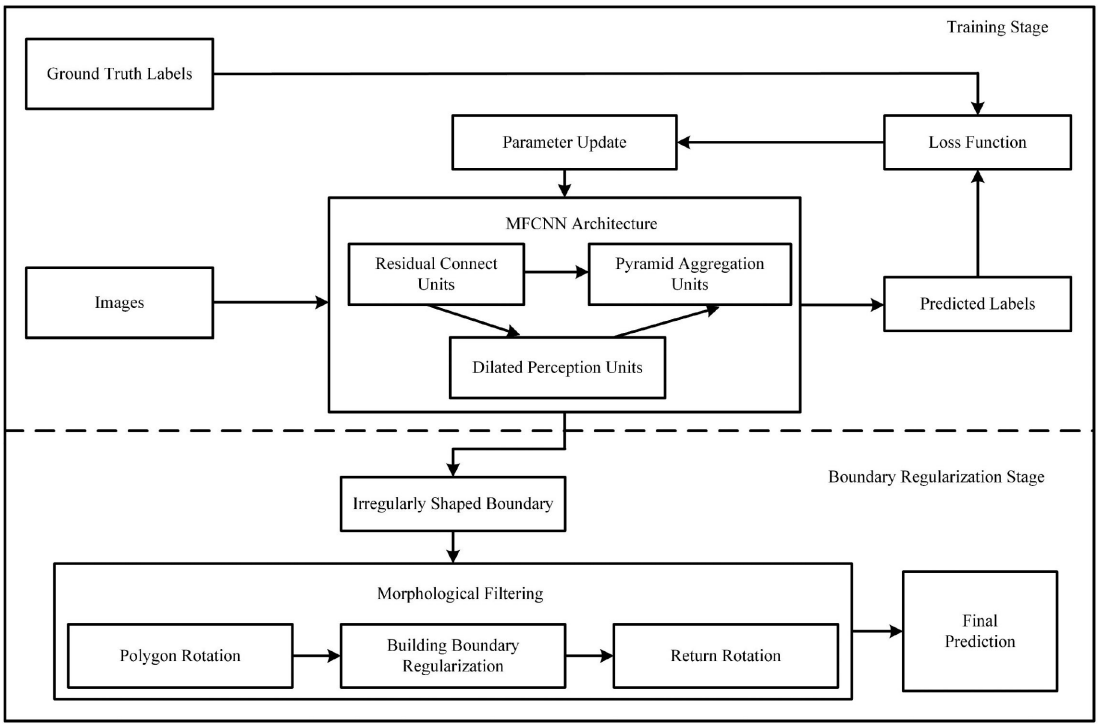

Fig. 1. Framework of our proposed approach: The training stage and the boundary regularization stage.

II. Methods for the Extraction of Building Outlines in Remote Sensing Imagery

In this article, a refined building extraction method for high-resolution remote sensing images based on the combination of an MFCNN and morphological filtering is proposed, as shown in Fig. 1. The proposed method is divided into two parts. The first part consists of the CNN framework for multifeature fusion; the MFCNN structure is symmetric to ensure that the generated outputs are of the same size as the inputs. The second part includes a contour refinement method based on morphological filtering that resolves the problematic irregular contours generated by pixel-level segmentation to achieve more refined architectural outlines.

A. Semantic Segmentation Model for the Extraction of Building Outlines

To overcome the limitations of the existing network frameworks for the extraction of building outlines from remote sensing images, this article proposes a building extraction framework based on an MFCNN, as shown in Fig. 2. The MFCNN is an end-to-end symmetric training structure consisting of an encoder network and a decoder network. The encoder portion consists of an RCU and multiple DPUs. In Fig. 2, each wide blue arrow represents an RCU, and the DPUs are represented by D1 and D2. The PAU is included in the decoder portion.

1) Residual Connected Unit: To extract rich features from remote sensing images, we use an RCU in the encoder portion of MFCNN, as shown in Fig. 3. The residual mapping structure connects the input layer to the next layer through a skip connection to allow the changes between layers to be considered during training. Compared with a traditional mapping structure, a residual mapping structure converges more easily and can help to avoid degradation in network performance due to an increase in the network depth [51]. To prevent overfitting, a batch normalization layer and a rectified linear unit layer are added along with the residual structure to establish a refined residual mapping [52], [53]. These refined residual units are used in the encoder to increase the nonlinear expression capability of the network and enable the network to obtain rich information from remote sensing images.

2) Dilated Perception Unit: To achieve both a large receptive field and a high spatial resolution, we build two DPUs by combining several dilated convolutions, as shown in Fig. 4. The DPUs are applied following the layers corresponding to the 1/4 and 1/8 scales in the encoding process to obtain morphological building features (e.g., linear and right-angle features) from a larger field of view. As the number of pooling layers increases, the spatial resolution of the image decreases. The shape characteristics of buildings also change with a change in the image resolution [54]. To adapt to the appearance of morphological building features at different resolutions, a group of adaptive dilated convolutions is constructed to better map these features. Each set of dilated convolutions consists of a 1×1 convolutional layer and four 3×3 dilated convolutional layers. The 1×1 convolutional layer can enhance the nonlinear expression ability of the network while reducing the feature dimensions [55]. Different rate values are set for feature maps at different levels. The 1/4-scale feature map size is 64, and the rate values of the corresponding dilated convolutions are set to 1, 6, 12, and 18 [see Fig. 4(a)]; the 1/8-scale feature map size is 32, and the rate values of the corresponding dilated convolutions are set to 1,

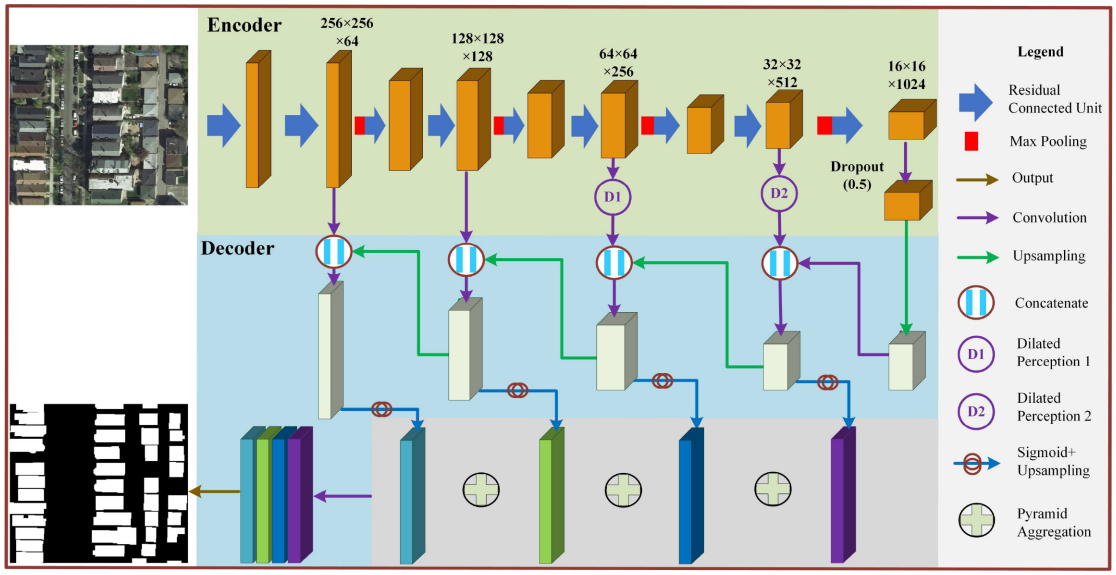

Fig. 2. Structure of the MFCNN.

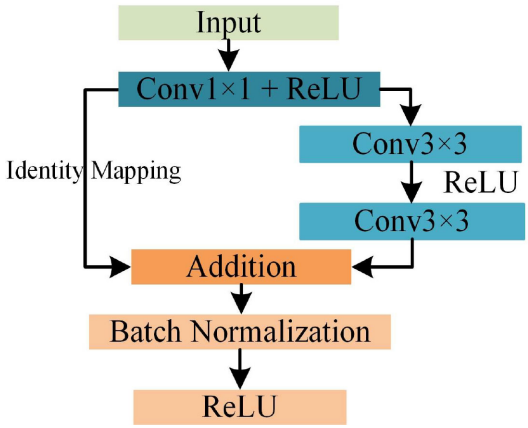

Fig. 3. Structure of a RCU.

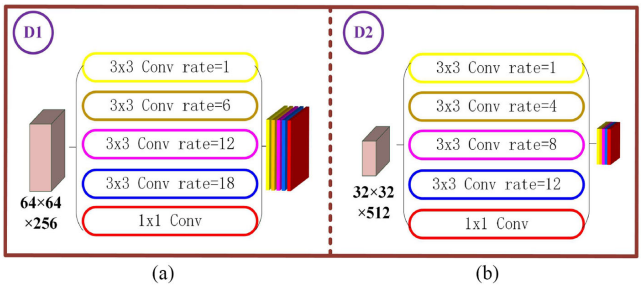

Fig. 4. Structure of the DPU. (a) Dilated perception unit 1 (D1 in Fig. 2). (b) Dilated perception unit 2 (D2 in Fig. 2).

4, 8, and 12 [see Fig. 4(b)]. Thus, adaptive dilated convolution dilation rates are established for differently sized feature maps. Thus, the morphological features of buildings in deep feature maps are obtained using receptive fields of different sizes, and these fields play an active role in the effective extraction of these features.

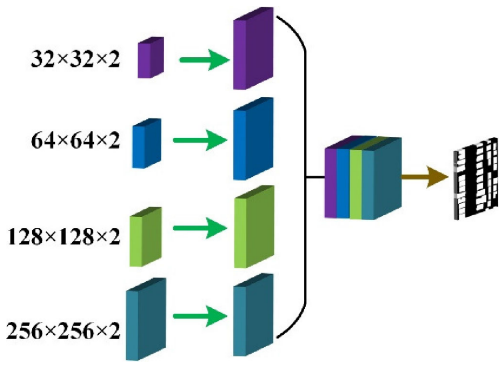

Fig. 5. Structure of the PAU.

3) Pyramid Aggregation Unit: To better acquire multiscale features, we establish a PAU, as shown in Fig. 5. The output of the first layer is upsampled to the original scale to predict the final output, and the outputs of the 1/2-, 1/4-, and 1/8-scale layers are also upsampled to the original scale for physical aggregation operations. In our method, bilinear sampling is adopted for upsampling operations. In this multiscale aggregation operation, all of the information from the different scales is utilized in both a separate and an integrated manner. Compared with a single multiscale weighted output, this physical combination strategy enables the aggregation of all multiscale features, thereby improving the optimization of the network [56]. Moreover, with this method, the output features depend only on the original network, and almost no additional calculation time is added.

B. Morphological Filtering Algorithm for Building Outline Optimization

When a neural network is used for pixel-level semantic segmentation, the output building boundaries are irregular. This article proposes a morphological filtering method for the refinement of building outlines that consists of three steps: polygon rotation,

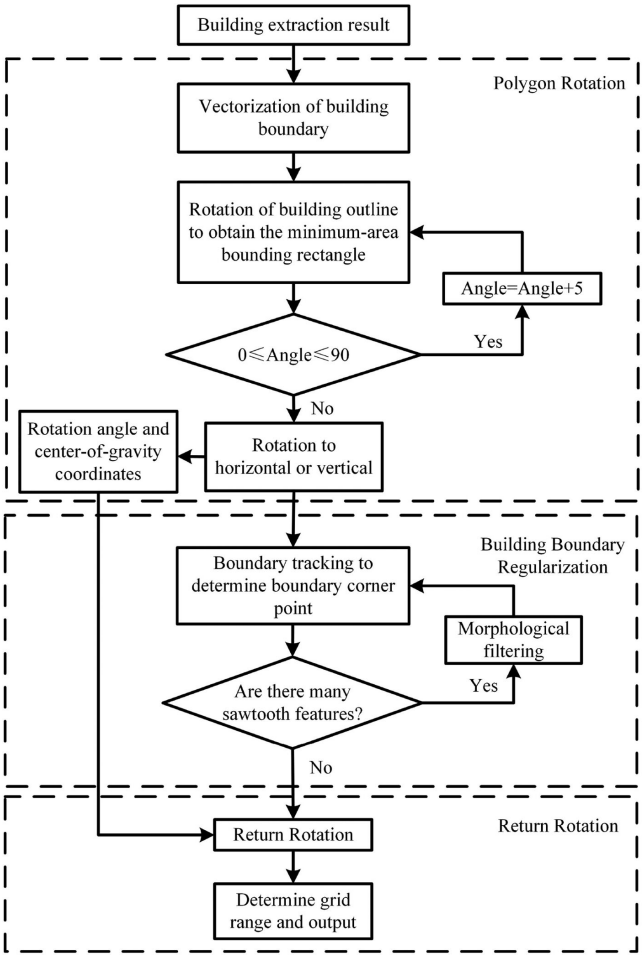

Fig. 6. Flow-chart of the building outline refinement algorithm.

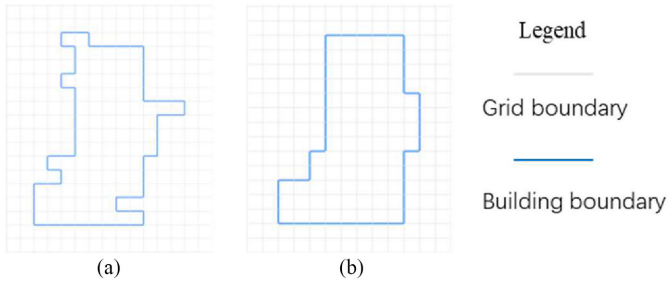

Fig. 7. Results of morphological filtering on a building boundary. (a) There are many sawtooth features in the originally extracted building boundary. (b) After morphological filtering, no sawtooth features remain.

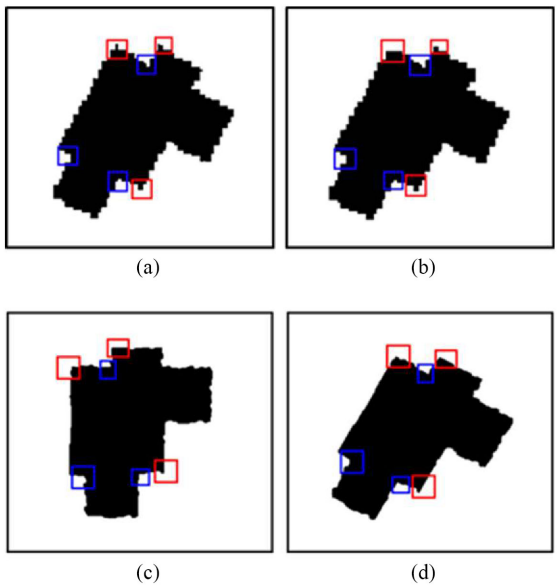

Fig. 8. Comparison among the grid fill results. (a) Initial raster result after extraction. (b) Original image after direct raster normalization. (c) Regularized result after rotation. (d) Result of rotation back to the original image angle after regularization.

building boundary regularization, and return rotation. In this method, a building boundary is optimized based on its overall contour information. The detailed flow chart of the algorithm is shown in Fig. 6.

1) Polygon Rotation: First, the binarized result is obtained for each building, and the boundary vector outline of the building is obtained via the boundary tracking algorithm [57]. Then, the barycentric coordinates of the resulting polygon are rotated through rotation angles of 0° to 90° in steps of 5°, as shown in (1). The corresponding rotation angle and minimum-area bounding rectangle of the polygon are obtained at each rotation. Finally, after the rotation process is complete, all values are compared to obtain the minimum area and the corresponding rotation angle. The angle corresponding to the minimum area is considered to be the horizontal or vertical direction of the polygon

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{1}$$

where (x, y) are the original coordinates, (x_1, y_1) are the coordinates after rotation, and θ is the angle of rotation.

2) Building Boundary Regularization: First, the outline of the building is rotated to the horizontal or vertical direction through polygon rotation. Then, a raster line is created at the pixel scale (gray line), and the number of corner points on the building boundary (blue line) and the building area are obtained. After analysis, we found that the number of corner points of buildings with different areas has different value intervals. Finally, the sawtooth features are determined by the building area and the number of corner points. If there are many sawtooth features [as shown in Fig. 7(a)], raster fill operations (open and close operations) are performed to obtain a regularized boundary [as shown in Fig. 7(b)].

3) Return Rotation: First, the building is rotated back to its original orientation based on the rotation angle and center-of-gravity coordinates obtained in the polygon rotation step, thus ensuring that the output will be consistent with the input. Then, the final raster boundary is determined via the ray method.

To illustrate the effect of the proposed method, a single building is selected to show the details of this procedure, as shown in Fig. 8. There are many sawtooth features in the originally extracted raster boundary of the building, as shown in Fig. 8(a). The result of applying morphological filtering directly to the original image is shown in Fig. 8(b). Although this operation has the

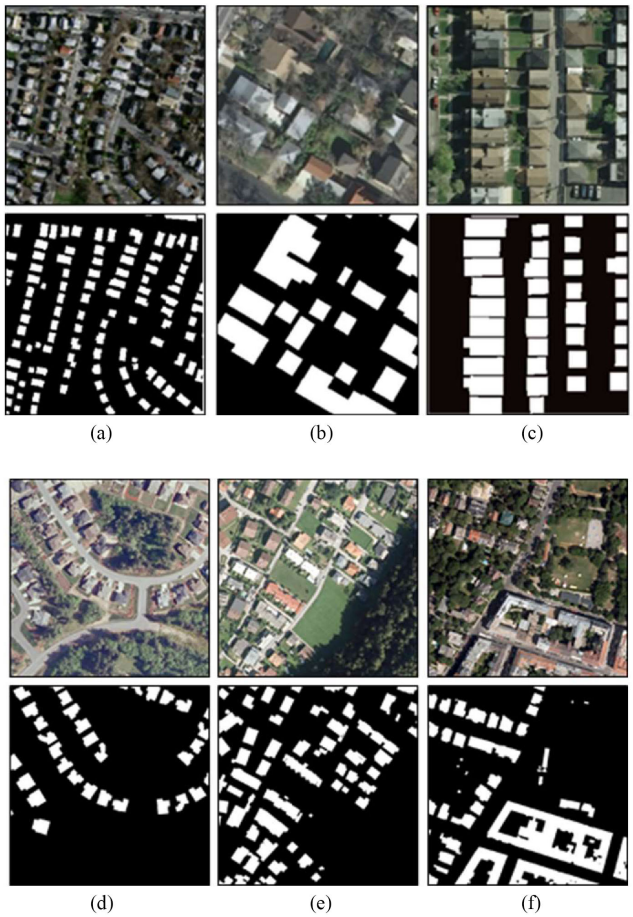

Fig. 9. Close-ups of the images from the datasets and their corresponding ground-truth building masks. (a) Massachusetts. (b) Austin. (c) Chicago. (d) Kitsap County. (e) West Tyrol. (f) Vienna.

effect of optimizing the boundary (note the regions highlighted with red and blue boxes), sawtooth phenomena are still evident. The result of morphological filtering after rotation is shown in Fig. 8(c), which illustrates that the noise has been essentially eliminated. Fig. 8(d) shows the result of rotation back to the original image angle after morphological filtering. This figure indicates that after the grid filling and regularization processes, the occurrence of redundant points (highlighted with red boxes) and hole noise (highlighted with blue boxes) is reduced. Thus, the proposed method can eliminate sawtooth phenomena along extracted building boundaries and yield finer building outlines.

III. Experiment Results

A. Dataset Descriptions

In the building extraction tests reported in this article, two standard datasets (including satellite imagery and aerial imagery) were employed to verify the effectiveness of the proposed method. The pixels in these images are categorized with two labels, namely, building and nonbuilding, as shown in Fig. 9.

1) Massachusetts: The Massachusetts building dataset was presented by Mnih [42]. This dataset contains 151 images of buildings throughout Boston, Massachusetts. The size of each image is 1500 × 1500 pixels, the spatial resolution is 1 m, and the images consist of three channels: red, green, and blue. The dataset was split into three groups of 137 training images, four validation images, and ten testing images; there were no overlapping areas. In addition, for ease of network training, each training and validation image was cropped into grid patches with dimensions of 256 × 256 pixels, and the data were augmented in our experiments. Finally, 7835 training samples and 862 test samples were generated.
2) Inria: The Inria dataset was presented by Maggiori *et al.* [40]. This dataset contains 360 aerial images, each of which covers a large area of 810 km^2 in ten different cities and includes different settlements and landscapes. The images have dimensions of 5000 × 5000 pixels with a spatial resolution of 0.3 m. The ground truth is provided only for the training set, which covers five cities (Austin, Chicago, Kitsap County, West Tyrol, and Vienna). For comparability, we split the dataset as described by Maggiori *et al.* [40] and created a validation set by excluding the first five tiles of each area from the training set (images 1–5 of each location for validation and images 6–36 for training) [40]; i.e., 155 images were retained for training, 25 images were set aside for validation, and the remaining 180 images were utilized for testing. As was done for the Massachusetts dataset, each training and validation image was cropped into grid patches with dimensions of 256 × 256 pixels, and the data were augmented in our experiments, yielding 93 000 and 15 000 image tiles for training and validation, respectively.

B. Experimental Configuration

1) Training Environment Description: The size of the training dataset for each experiment reported in this article was 3 × 256 × 256. All training and testing were implemented using TensorFlow and Keras on the Windows 10 platform with an Nvidia GeForce RTX 2080Ti 11G graphics card.

2) Hyperparameter Settings: All hyperparameters used in this experiment were optimal parameters selected by experts in remote sensing based on comparisons among the results of repeated trials. The number of epochs was set to 100, the batch size was set to 10, and the initial learning rate was set to 0.001. By monitoring the value of the loss function, the learning rate was reduced by 0.95 after every 5 consecutive epochs in which the performance did not improve. The MFCNN has four output levels, and the final loss function is defined as shown in the following equation:

$$\text{Loss} = \sum_{i=1}^{n} \lambda_i L_i \tag{2}$$

where $n = 4$, $\lambda = 1$, and each individual loss function component L is defined as a cross-entropy loss function [58], as shown in the following equation:

$$\text{L} = -\frac{1}{m} \sum_{i=1}^{m} g^i \ln p^i + \left(1 - g^i\right) \ln \left(1 - p^i\right) \tag{3}$$

where p^i is the predicted probability distribution for category i, g^i is the probability distribution of the corresponding ground-truth label, and m is the total number of training images.

To reduce overfitting and improve the generalization ability of the network, random cropping, random rotation, and fuzzy and random noise operations were performed on all data in the two datasets. Finally, the samples were balanced by the number of building pixels in both datasets.

3) Evaluation Metrics: To quantitatively evaluate the performance of the proposed method in extracting buildings from remote sensing images and to compare that performance with the results of other researchers, an accuracy evaluation is presented based on the IOU (4). The IOU, which can consider both incorrect detections and missed detections, has become the standard evaluation metric for semantic segmentation [33]. The OA (5) is also used in this article as a global accuracy evaluation metric. In addition, three common evaluation criteria, namely, the F1-score (6) and the precision and recall (7), are also evaluated. To facilitate an accurate analysis of the extraction results, the results are visually displayed as follows: as shown below, correctly classified building pixels (TP) are marked in green, missed building pixels (FN) are marked in red, the pixels incorrectly identified as building pixels (FP) are marked in blue, and correctly classified nonbuilding pixels (TN) are marked in white

$$\mathrm{IOU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{5}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{6}$$

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

where TP, FN, FP, and TN are the pixel classification results evaluated by comparing the extracted building pixels with the ground-truth points.

TP: True positive, i.e., the number of correctly extracted building pixels.
FN: False negative, i.e., the number of missed building pixels.
FP: False positive, i.e., the number of erroneously detected building pixels.
TN: True negative, i.e., the number of correctly extracted nonbuilding pixels.

C. Analysis of the Experimental Results

1) Introduction of the Baseline Models: We selected three baseline models for comparison: FCN-8s, SegNet, and U-Net. These baseline models and their parameter settings are introduced as follows.

1) FCN-8s: The FCN family of models was proposed by Long *et al.* [33]. Pixel-level semantic segmentation is realized by changing the last fully connected layer into a convolutional layer, and the images are restored to their original size through upsampling. The different FCN models are referred to as FCN-32s, FCN-16s, and FCN-8s based on their different numbers of pooling layers. We used the best-performing FCN-8s model for building extraction and initialized the network with VGG-16. The stochastic gradient descent (SGD) algorithm was applied to solve the optimization problem during training.
2) SegNet: SegNet is an end-to-end semantic segmentation method that was originally proposed by Badrinarayanan *et al.* [34]. This method is based on the VGG-16 framework; however, the fully connected layer is removed to construct a symmetric encoder-decoder network structure. The network performance is improved by performing upsampling operations followed by the downsampling of each layer in the coding network by the corresponding layer in the decoding network. Again, we used the SGD algorithm to solve the optimization problem during training.
3) U-Net: U-Net, proposed by Ronneberger *et al.* [35], is an improved symmetric network based on the FCN. This method overcomes the insufficient detail segmentation capability of the FCN by means of a splicing-based feature fusion method. In addition, U-Net realizes multiscale image feature recognition by partially fusing the extracted output features during the upsampling process. As before, we used the SGD algorithm to solve the optimization problem during training.

2) Massachusetts Building Dataset: Fig. 10 shows a qualitative comparison among the results obtained by the different methods on the Massachusetts test dataset. Two different images are selected for the comparison, and the results obtained by each method are initially compared at the pixel level. The results shown in Fig. 11 correspond to the regions within the red rectangles in Fig. 10. In general, the results of the proposed method are significantly better than those of the other three methods.

In the first row of Fig. 10, the building areas are small, and the roof texture information of the buildings is similar to the road texture information. In the second row of Fig. 10, the building areas are relatively large, and the roof texture information of the buildings is complex. The boundaries of the buildings in the two images are not only affected by the buildings' own shadows but also densely distributed. The pixel comparison results show that the boundaries are more clearly delineated in the MFCNN output than in the FCN-8s, SegNet, or U-Net output. In particular, as shown by the black boxes in the first row of Fig. 11, the basic information of small buildings can be accurately obtained by the proposed method even under complex conditions, while the other methods result in more missed and incorrect extractions at the boundaries. As highlighted by the boxes in the second row of Fig. 11, for densely distributed large buildings, the proposed method can also more effectively distinguish the boundaries and obtain more refined boundary representations. Finally, the pixel-level results show that the results of the proposed method contain less noise.

Table I shows the quantitative evaluation results for the proposed method and the other deep learning methods on the same dataset and under the same computer performance conditions. The IOU is used as a standard evaluation metric for semantic segmentation. Comparisons among the IOU values reveal that

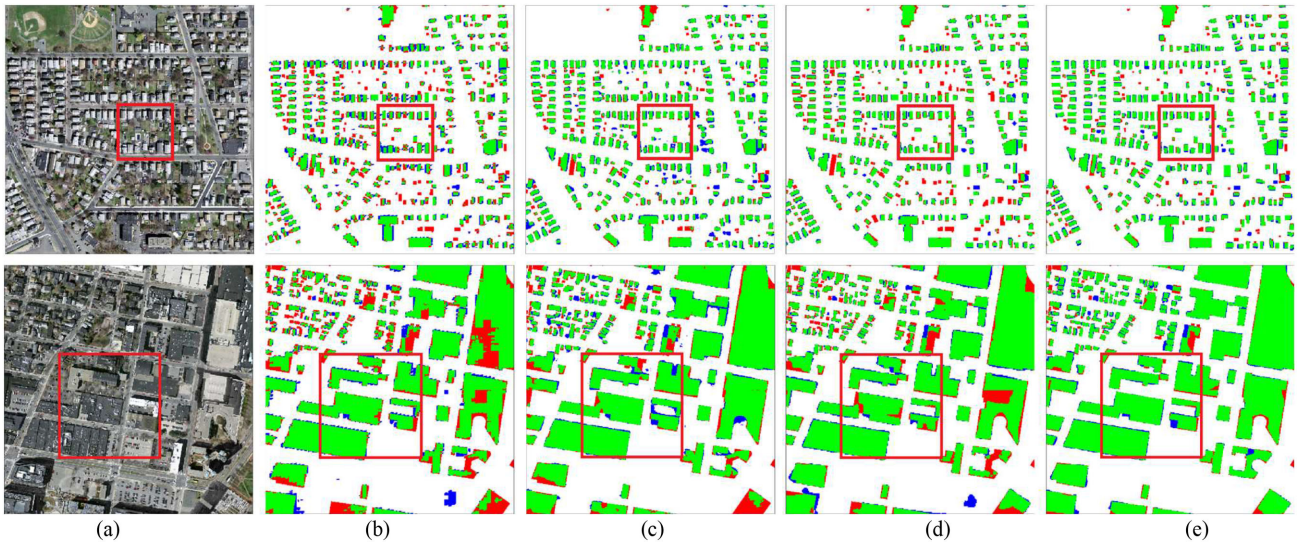

Fig. 10. Results of roof segmentation using four methods in different experimental areas of the Massachusetts building test dataset (zoomed-in views with more details are provided in Fig. 11). (a) Input image. (b) Output of FCN-8s. (c) Output of SegNet. (d) Output of U-Net. (e) Output of the MFCNN.

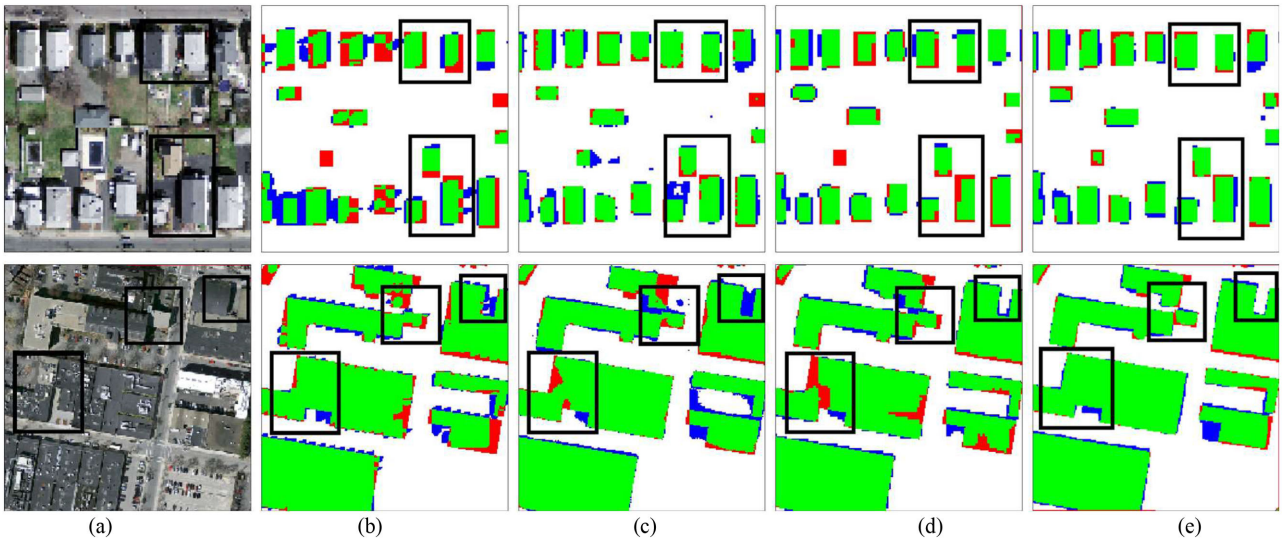

Fig. 11. Comparison among the building extraction results of different algorithms in local areas (the smaller areas in the red rectangular frames in Fig. 10). (a) Input image. (b) Output of FCN-8s. (c) Output of SegNet. (d) Output of U-Net. (e) Output of the MFCNN.

TABLE I
COMPARISON AMONG THE RESULTS OF DIFFERENT METHODS ON THE MASSACHUSETTS BUILDING TEST DATASET

Dataset	Method	OA (%)	P (%)	R (%)	F1 (%)	IOU (%)
Massachusetts	FCN-8s	92.65	81.72	78.51	80.39	69.97
	SegNet	93.16	82.12	80.17	81.14	70.03
	U-Net	93.56	83.11	83.14	83.07	71.38
	MFCNN	95.41	88.65	84.13	86.38	74.85

the proposed method achieves the best performance in extracting buildings with high complexity. Compared with the other algorithms, the IOU of the MFCNN is increased by 3.47%–4.88%. At the same time, the OA, precision, recall, and F1-score for the MFCNN are also significantly higher than those for the other networks, further proving the effectiveness of the algorithm proposed in this article.

3) Inria Building Dataset: A qualitative evaluation of the results of applying the proposed algorithm to the Inria dataset is shown in Fig. 12. Two images, one from an urban area and

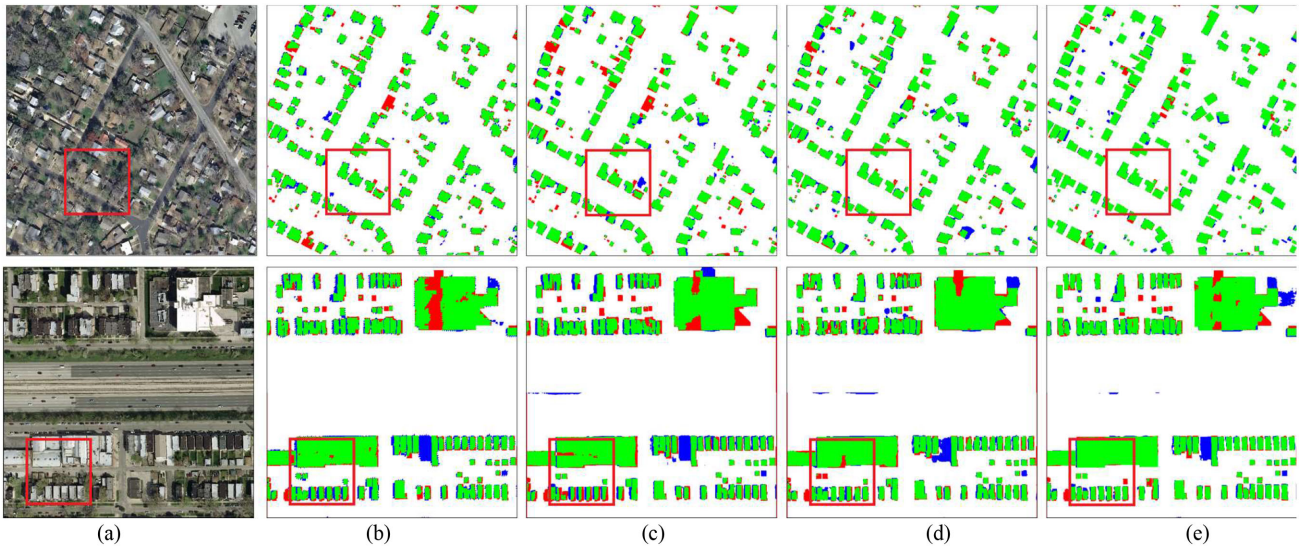

Fig. 12. Results of roof segmentation using four methods in different experimental areas of the Inria building dataset (zoomed-in views with more details are provided in Fig. 13). (a) Input image. (b) Output of FCN-8s. (c) Output of SegNet. (d) Output of U-Net. (e) Output of the MFCNN.

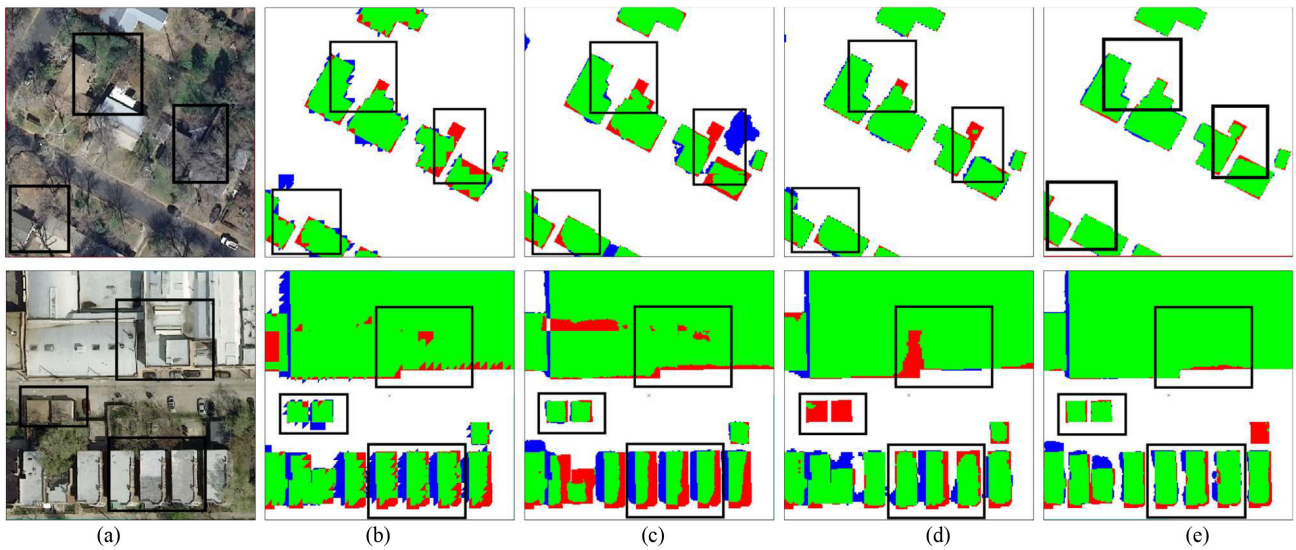

Fig. 13. Comparison among the building extraction results of different algorithms in local areas (the small areas in the red rectangular frames in Fig. 12). (a) Input image. (b) Output of FCN-8s. (c) Output of SegNet. (d) Output of U-Net. (e) Output of the MFCNN.

one from the outskirts of a city, are selected for comparison, and zoomed-in views of the regions within the red rectangles in Fig. 12 are presented in Fig. 13 to show more details.

The first row of Fig. 12 contains the results for the suburban image. The buildings in this figure are dense and heavily occluded by vegetation cover. Consequently, it is difficult to distinguish the boundary information. The second row in Fig. 12 shows the results for the urban image, which contains noise such as artificial nonbuilding features and shadows. Noise presents an important challenge in building extraction. The pixel-level comparison results in Fig. 12 demonstrate that the qualitative results of the proposed algorithm are significantly better than those of the other algorithms. Considering the details shown in Fig. 13, the black boxes in the first row of Fig. 13 indicate that the proposed algorithm can effectively obtain the fine boundaries of buildings in areas affected by vegetation. By contrast, in the results of the other methods, there are many cases of missed detection. The black boxes in the second row of Fig. 13 show that the extracted boundary information can also be clearly expressed in regions with large confounding entities such as shadows.

Table II shows the results of the accuracy evaluation for the different methods tested on the same dataset and under the same computer performance conditions. The reported results include the OA, precision, recall, F1-score, and IOU. The IOU of the MFCNN is increased by 8.82%, 8.25%, and 5.96% compared with those of the FCN-8s, SegNet, and U-Net single-layer outputs, respectively. This finding also confirms that the dilated convolution and pyramid aggregation strategies introduced in the proposed network help to improve the building extraction accuracy.

TABLE II
COMPARISON AMONG THE RESULTS OF DIFFERENT METHODS ON THE INRIA BUILDING DATASET

Dataset	Method	OA (%)	P (%)	R (%)	F1 (%)	IOU (%)
Inria	FCN-8s	93.75	83.46	82.41	82.89	70.53
	SegNet	93.83	85.47	81.96	83.70	71.10
	U-Net	94.20	87.16	82.14	84.52	73.39
	MFCNN	96.82	88.58	87.91	88.38	79.35

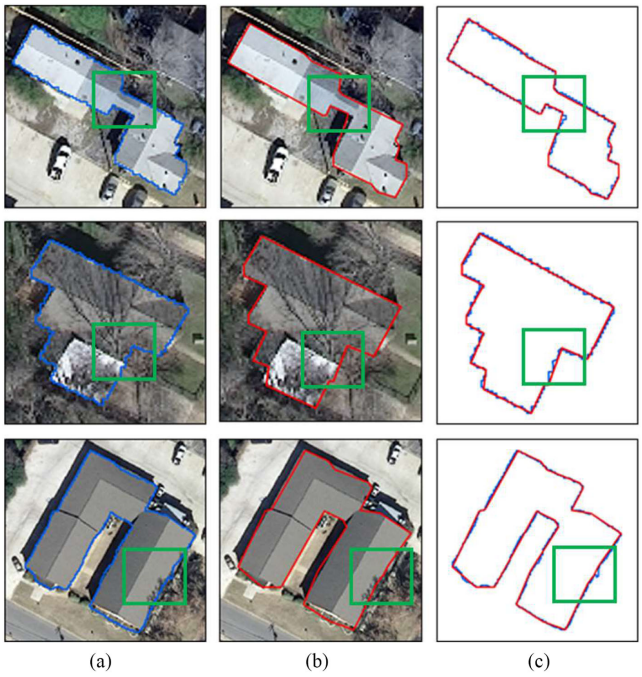

Fig. 14. Regularized building boundary output. (a) Results of vectorizing the original output. (b) Results of morphological filtering. (c) Result of overlapping building boundaries.

4) Refined Building Boundary Output: Morphological filtering is used as a postprocessing method to improve the extraction results. By considering the linear and right-angle features of buildings, the problem arising from polygon irregularities generated through segmentation is resolved to refine the extracted boundaries. The final building boundaries are obtained by vectorizing the final extraction results. In Fig. 14(a), the first column shows the results of vectorizing the original output. Many serrations can be observed on the boundaries that do not overlap with the original building boundaries. The results after the morphological filtering process are shown in the second column of Fig. 14(b). These boundaries are basically linear and coincide well with the true boundaries. To better show the differences in the results, the vectorized building boundaries in the first two columns are superimposed in the third column of Fig. 14(c). After postprocessing, the concave and convex parts of the building boundaries are clearly flatter. The proposed method effectively eliminates sawtooth phenomena along the building boundaries (as highlighted by the green frames in the figure), and the resulting shapes tend to be more regular.

To quantify the impact of morphological filtering regularization on the extraction of building boundaries, we further compare the raw output and the results after regularization in Table III. The root mean square (RMS) and Hausdorff distance are added to the original accuracy evaluation metrics [59]. The RMS is the root mean square of the closest distance from each point on the boundary to the minimum-area bounding rectangle and is used to express the degree of regularity of the building. The Hausdorff distance, which is often used as an evaluation metric for building regularization results, is used as the criterion to evaluate the shape similarity. An accuracy comparison reveals that the IOU achieved with the proposed method is basically equal to the IOU score produced by the original network framework, but compared with the original output, the RMS values for the Massachusetts and Inria datasets are reduced by 0.65 m and 0.32 m, respectively, after postprocessing, and the Hausdorff distance is reduced by 0.76 m and 0.28 m, respectively. The lower RMS after morphological filtering indicates that the building boundaries are more regular, and the decrease in the Hausdorff distance verifies that the extracted boundaries are more similar to the boundaries of the original buildings.

IV. DISCUSSION

A. Comparison of the Building Extraction Results With Recent Research

In this section, we summarize several studies conducted on the Massachusetts dataset and the Inria dataset in recent years and compare the results with those of our method. Accuracy comparisons for the Massachusetts dataset and the Inria dataset are shown in Tables IV and V, respectively. For the Massachusetts dataset, we focus on comparing the F1-score and OA. The F1-score, a weighted average of the precision and recall, is a statistical indicator used to measure the accuracy of a classification model. From the comparison, relative to the best results from the existing studies, the F1-score of the MFCNN is improved by 2.31%–2.88%, and the OA is improved by 0.81%–1.29%. Hence, the F1-score and OA of the proposed method are superior to those of the other methods, as shown in Table IV.

In contrast to the Massachusetts dataset, the Inria dataset has a resolution of 0.3 m and contains five different cities with more obvious spectral and spatial dissimilarities; thus, the results obtained on this dataset can serve as a good indicator for the generalizability of a method. We compare the accuracy of the MFCNN with the accuracies achieved in recent research on the same validation dataset, as shown in Table V. As seen from

TABLE III
ORIGINAL OUTPUT VERSUS REGULARIZED OUTPUT

Dataset	Method	OA (%)	F1 (%)	IOU (%)	RMS (m)	Hausdorff distance (m)
Massachusetts	MFCNN	95.41	86.38	74.85	1.93	2.52
	MFCNN+ Postprocessing	95.63	86.13	74.52	1.28	1.76
Inria	MFCNN	96.82	88.38	79.35	1.35	1.87
	MFCNN+ Postprocessing	96.39	88.46	79.43	1.03	1.59

TABLE IV
COMPARISON AMONG THE MASSACHUSETTS DATASET RESULTS

Method	Postprocessing	F1-score (%)	OA (%)
Aleshehhi et al. [60]	SLC		94.60
Cascaded multitask [61]		83.50	94.12
U-Net with Xception and multitask [61]		84.07	94.23
MFCNN		**86.38**	**95.41**

TABLE V
VALIDATION ACCURACY OF DIFFERENT MODELS ON THE INRIA AERIAL IMAGE LABELING DATASET

Methods	IOU (%)	OA (%)
FCN [40]	53.82	92.79
MLP [40]	64.67	94.42
SegNet (Single-Loss) [62]	72.57	95.66
SegNet with multitask loss [62]	73.00	95.73
2-levels U-Net+Aug. [63]	74.55	96.05
MFCNN	**79.35**	**96.82**

the results, the IOU is improved by 4.80%–25.53% compared with the existing studies, and the OA is improved by 0.77%–4.03%, further confirming that the accuracy of the proposed method is improved compared to that of the existing methods.

B. Effectiveness of the MFCNN

1) Introduction of the Model Without the RCU, DPU, or PAU: To effectively extract building information from remote sensing images, CNNs are still being explored as a popular class of deep learning algorithms. The network depth, the size of the field of view, and the multiscale nature of the output are all important in the building extraction process. Therefore, to improve the ability of our neural network to extract building features from images, we construct the neural network by combining three basic units, namely, the RCU, DPU, and PAU. To highlight the role of each type of basic unit, the influencing factors were analyzed by utilizing a control variable. The details of the modified network without the RCU, DPU, or PAU are introduced as follows.

1) Without the RCU: To investigate the role of the RCU, we deleted this unit and replaced it with a 3×3 convolution structure in the encoder portion while leaving the rest of the network unchanged.
2) Without the DPU: The DPU includes two sets of dilated convolutions, both of which are removed to examine the influence of the DPU on MFCNN, while the remainder of the network is unchanged.
3) Without the PAU: To compare the differences between the single output level and the four output levels, the PAU is replaced by a single-layer output structure, and the rest of the network is unchanged.

2) Building Extraction Results Without the RCU, DPU, or PAU: For comparison, three versions of the method proposed in this article from which the RCU, DPU, or PAU are excluded are tested under the same experimental conditions on both the Massachusetts dataset and the Inria dataset, and the results are compared in Table VI. Evidently, the RCU, DPU, and PAU are all beneficial to the results. The OA values on the Massachusetts dataset are increased by 1.47%–2.22%, and the IOU values are increased by 4.07%–5.20%. Furthermore, the OA values on the Inria dataset are increased by 2.99%–3.86%, and the IOU values are increased by 8.43%–9.12%. Thus, it is obvious that the proposed network architecture has a positive effect on the extraction of building outlines.

The degree of neural network learning determines the advantages and disadvantages of the final extraction results. The RCU can better learn deep and rich information from remote sensing images. For the network structure lacking the RCU, the OA values on the Massachusetts and Inria datasets are decreased by 1.47% and 3.86%, respectively, and the IOU values are correspondingly decreased by 4.07% and 9.12%. The acquisition of deeper information helps to improve the nonlinear representation capability of the network; hence, the extraction results obtained on the two datasets are improved.

The distinctive morphological characteristics of buildings play an important role in the ability to completely capture building boundaries. The DPU helps to obtain a larger field of view without losing resolution information, thus allowing the morphological features of buildings to be more effectively acquired. In remote sensing images, each building has its own size and characteristics, and a single field of view cannot capture all of its distinctive features. Hence, the construction of dilated convolution groups corresponding to the feature map allows the IOU to be increased by 2.76% and 6.60% on the Massachusetts and Inria datasets, respectively, thus helping to improve the building extraction performance.

The effective integration of multiscale features has always constituted a problem that needs to be solved for the extraction of ground objects from remote sensing images [56]. To avoid the loss of features caused by multilayer weighted outputs, the

TABLE VI
COMPARISON OF THE RESULTS WITHOUT THE RCU, DPU, OR PAU

Dataset	Method	OA (%)	P (%)	R (%)	F1 (%)	IOU (%)
Massachusetts	Without the RCU	93.94	82.73	78.73	80.52	70.78
	Without the DPU	94.28	87.90	81.31	84.36	72.09
	Without the PAU	93.19	84.46	77.12	80.14	69.65
	All	95.41	88.65	84.13	86.38	74.85
Inria	Without the RCU	92.96	83.33	81.87	82.47	70.23
	Without the DPU	94.05	84.99	83.78	84.41	72.75
	Without the PAU	93.83	85.47	80.96	83.17	70.92
	All	96.82	88.58	87.91	88.38	79.35

proposed network preserves all extracted features at different scales by means of physical aggregation. Compared with the single-layer output obtained without the PAU, the OA values are increased by 2.22% and 2.99% on the Massachusetts and Inria datasets, respectively, and the IOU values are correspondingly increased by 5.20% and 8.43%, thus showing that better output results are obtained.

C. Morphological Filtering

The large number of jagged boundary features generated through pixel-level semantic segmentation results in irregularly shaped boundaries. Therefore, to improve the regularity of the extracted boundaries, we improve the classification results by means of morphological filtering to obtain refined building boundaries. We fine-tune each boundary by means of simple rotation and morphological filtering operations to remove the large amount of aliasing caused by pixel-level segmentation. The results obtained after morphological filtering are more regular and more similar to the real building boundaries, as shown in Fig. 14. The effectiveness of the proposed method at refining the extracted boundaries through morphological filtering is demonstrated by the higher degree of regularity.

V. CONCLUSION AND FUTURE WORKS

Extracting buildings from high-resolution images has always been an important and challenging problem. In recent years, with the development of deep learning, CNNs have been effectively used for the extraction of building boundaries. In this article, we have proposed a building extraction framework for applications involving high-resolution remote sensing images with multiscale features. The effectiveness of the proposed algorithm framework has been validated on datasets containing images with different resolutions and complex scenes. Based on the characteristics of remote sensing images, a neural network architecture is constructed to include a refined residual structure, adaptive scaled dilated convolution groups, and a multiscale pyramid aggregation scheme. The proposed algorithm has been implemented to achieve the effective extraction of rich features from remote sensing images, effectively learn the morphological features of buildings, and accurately extract buildings of different scales. The extraction results show that the generated building boundaries have complete structures and a reduced level of noise, such as holes and other discontinuities. To solve the problem of sawtooth phenomena along the boundaries generated through pixel-level semantic segmentation, a method for building boundary regularization based on morphological filtering has been proposed. The morphological filtering process is performed after the boundary is rotated to a standard orientation, and the distinctive linear and right-angle features of buildings are effectively utilized. As seen from the accuracy of the results, the regularity of the boundaries is effectively improved. As shown in this article, our method can reduce the numbers of erroneously detected building pixels and missed detections. Then, through a separate postprocessing method, refined building contours can be effectively obtained that are better suited to practical engineering applications.

With the development of remote sensing technology, the number of high-resolution remote sensing images available will continue to increase, data acquisition will become easier, and the applications of such images will become more extensive. This article has proposed an effective method for improving the extraction of buildings from remote sensing images. However, the extraction of buildings with fuzzy boundaries and uncommon shapes is still difficult, and the extraction accuracy still needs to be improved. In future studies, we will further optimize deep neural networks to improve the efficiency and accuracy. At the same time, the generalizability of the network model to different data sources and different regions will be further studied to allow the building extraction results to be more effectively applied to practical engineering problems.

REFERENCES

[1] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.
[2] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks," 2016, *arXiv:1602.06564*.
[3] D. Lu, H. Tian, G. Zhou, and H. Ge, "Regional mapping of human settlements in southeastern China with multisensor remotely sensed data," *Remote Sens. Environ.*, vol. 112, pp. 3668–3679, 2008.
[4] H. Mayer, "Automatic object extraction from aerial imagery—A survey focusing on buildings," *Comput. Vis. Image Understanding*, vol. 74, pp. 138–149, 1999.
[5] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, pp. 1082–1086, 2019.
[6] N. Longbotham, C. Chaapel, L. Bleiler, C. Padwick, W. J. Emery, and F. Pacifici, "Very high resolution multiangle urban classification analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1155–1170, Apr. 2012.

[7] T. Kim and J. Muller, "Development of a graph-based approach for building detection," *Image Vis. Comput.*, vol. 17, pp. 3–14, 1999.

[8] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, pp. 236–248, 2007.

[9] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.

[10] X. Ru, H. Zhang, T. Wang, and L. Hui, "Using pan-sharpened high resolution satellite data to improve impervious surfaces estimation," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 57, pp. 177–189, 2017.

[11] X. Huang, W. Yuan, J. Li, and L. Zhang, "A new building extraction post-processing work for high spatial resolution remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, pp. 654–668. 2017.

[12] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from mutispectral GeoEye-1 imagery," *Photogramm. Eng. Remote Sens.*, vol 77, pp. 721–732. 2011.

[13] J. Peng and Y. Liu, "Model and context-driven building extraction in dense urban aerial images," *Int. J. Remote Sens.*, vol. 26, pp. 1289–1307, 2005.

[14] W. Su *et al.*, "Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery," *Int. J. Remote Sens.*, vol. 29, pp. 3105–3117, 2008.

[15] A. Huertas and R. Nevatia, "Detecting buildings in aerial images," *Vis. Graph. Image Process.*, vol. 41, pp. 131–152, 1988.

[16] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the Sobel operator," *IEEE J. Solid-State Circuits.*, vol. 23, no. 2, pp. 358–367, Apr. 1988.

[17] M. Awrangjeb and C. S. Fraser, "An automatic and threshold-free performance evaluation system for building extraction techniques from airborne LIDAR data," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4184–4198, Oct. 2014.

[18] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Contextual classification of lidar data and building object detection in urban areas," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 152–165, 2014.

[19] F. Karsli, M. Dihkan, H. Acar, and A. Ozturk, "Automatic building extraction from very high-resolution image and LiDAR data with SVM algorithm," *Arab. J. Geosci.*, vol. 9, 2016. [Online]. Available: https://link.springer.com/article/10.1007/s12517-016-2664-7

[20] Y. Yan, Z. Tan, N. Su, and C. Zhao, "Building extraction based on an optimized stacked sparse autoencoder of structure and training samples using LIDAR DSM and optical images," *Sensors*, vol. 17, 2017. [Online]. Available: https://doi.org/10.3390/s17091957

[21] S. Zabuawala, H. Nguyen, H. Wei, and J. Yadegar, "Fusion of LiDAR and aerial imagery for accurate building footprint extraction," *Image Process. Mach. Vis. Appl. II*, vol. 7251, 2009, Art. no. 72510Z.

[22] H. Zhang, L. Hui, L. Yu, Y. Zhang, and C. Fang, "Mapping urban impervious surface with dual-polarimetric SAR data: An improved method," *Landsc. Urban Plan.*, vol. 151, pp. 55–63, 2016.

[23] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[24] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[25] G. Alex, L. Marcus, F. Santiago, B. Roman, B. Horst, and S. Jürgen, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. no. 5, 855–868, May 2009.

[26] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[28] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2011.

[29] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.

[30] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[31] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. Conf. Comp. Vis. Pattern Recognit.*, Jun. 2014, pp. 2155–2162.

[32] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 640–651, 2014.

[34] 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.

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

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017, *arXiv:1703.06870*.

[37] 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.

[38] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[39] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[40] 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. Symp. Geosci. Remote Sens.*, 2017, pp. 3226–3229.

[41] K. Zhao, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. Comput. Vision Pattern Recognit. Workshops*, 2018, pp. 247–251.

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

[43] F. Alidoost and A. Hossein, "A CNN-based approach for automatic building detection and recognition of roof types using a single aerial image," *PFG-J. Photogramm. Remote Sens. Geoinf. Sci.*, vol. 86, pp. 235–248, 2018.

[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] Z. Deng, S. Hao, S. Zhou, J. Zhao, L. Lin, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens*, vol. 145, pp. 3–22, 2018.

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

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

[48] J. Lin, H. Song, and W. Jing, "ESFNet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019.

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

[50] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

[52] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[53] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

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

[55] W. Liu and J. Lee, "An efficient residual learning neural network for hyperspectral image superresolution," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1240–1253, Apr. 2019.

[56] 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, pp. 3308–3322, 2019.

[57] W. Wang, D. Jing, X. Li, H. Hu, W. Xu, and H. Guo, Y Ding, "A grid filling based rectangular building outlines regularization method," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 43, pp. 318–324, 2018.
[58] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[59] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
[60] R. Alshehhi, P. R. Marpu, L. W. Wei, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, 2017.
[61] 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.*, vol. 16, no. 5, pp. 786–790, May 2018.
[62] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," 2017, *arXiv:1709.05932*.
[63] A. Khalel and M. El-Saban, "Automatic pixelwise object labeling for aerial imagery using stacked u-nets," 2018, *arXiv:1803.04953*.

Yakun Xie received the B.S. degree in survey engineering from the School of Survey Engineering, Henan University of Urban Construction, Pingdingshan, China, in 2015, and the M.S. degree in geodesy and survey engineering from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2018, where he is currently working toward the Ph.D. degree.

His research interests include intelligent fire control, computer vision, and remote sensing image processing.

Jun Zhu received the M.S. degree in geodesy and survey engineering from Southwest Jiaotong University, Chengdu, China, in 2003, and the Ph.D. degree in cartography and geographic information systems from the Chinese Academy of Sciences, Beijing, China, in 2006.

From 2007 to 2008, he was a Postdoctoral Research Fellow with the Chinese University of Hong Kong, Shatin, Hong Kong. Currently, he is a Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University. His research interests include computer vision, 3-D GIS technology, and virtual geographic environments.

Yungang Cao received the M.S. degree in geodesy and survey engineering from Southwest Jiaotong University, Chengdu, China, in 2003, and the Ph.D. degree in cartography and geographic information systems from the Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently an Associate Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University. His research interests include computer vision, deep learning, and remote sensing image processing.

Dejun Feng received the M.S. and Ph.D. degrees in geodesy and survey engineering from Southwest Jiaotong University, Chengdu, China, in 2001 and 2004, respectively.

He is currently an Associate Professor with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University. His research interests include geographic information systems and remote sensing image processing.

Minjun Hu received the B.S. degree in surveying engineering from the College of Geomatics, Xi'an University of Science and Technology, Xi'an, China, in 2017. She is currently working toward the master's degree at Southwest Jiaotong University, Chengdu, China.

Her research interests include computer vision and remote sensing image processing.

Weilian Li received the B.S. degree in survey engineering from Tianjin Chengjian University, Tianjin, China, in 2015. He is currently working toward the Ph.D. degree at Southwest Jiaotong University, Chengdu, China.

His research interests include virtual geographic environments and disaster scene visualization.

Yunhao Zhang received the B.S. degree in remote sensing science and technology from the College of Resources and Environment, Chengdu University of Information Technology, Chengdu, China, in 2015. He is currently working toward the Ph.D. degree at Southwest Jiaotong University, Chengdu, China.

His research interests include 3-D GIS technology and virtual geographic environments.

Lin Fu received the B.S. degree in geographic information science from the School of Land and Resources, China West Normal University, NanChong, China, in 2017. He is currently working toward the Ph.D. degree at Southwest Jiaotong University, Chengdu, China.

His research interests include 3-D GIS technology, deep learning, and virtual geographic environments.