Building Footprint Extraction From Unmanned Aerial Vehicle Images Via PRU-Net: Application to Change Detection

Wei Liu, Jiawei Xu, Zihui Guo, Erzhu Li, Xing Li, Lianpeng Zhang, and Wensong Liu

***Abstract*—As the manual detection of building footprint is inefficient and labor-intensive, this study proposed a method of building footprint extraction and change detection based on deep convolutional neural networks. The study modified the existing U-Net model to develop the "PRU-Net" model. PRU-Net incorporates pyramid scene parsing (PSP) to allow multiscale scene parsing, a residual block (RB) in ResNet for feature extraction, and focal loss to address sample imbalance. Within the proposed method, building footprint extraction is conducted as follows: 1) unmanned aerial vehicle images are cropped, denoised, and semantically marked, and datasets are created (including training/validation and prediction datasets); 2) the training/validation and prediction datasets are input into the full convolutional neural network PRU-Net for model training/validation and prediction. Compared with the U-Net, PSP+U-Net (PU-Net), and U-Net++ models, PRU-Net offers improved footprint extraction of buildings with a range of sizes and shapes. The large-scale experimental results demonstrated the effectiveness of the PSP module for multiscale scene analysis and the RB module for feature extraction. After demonstrating the improvements in building extraction offered by PRU-Net, the building footprint results were further processed to generate a building change map.**

***Index Terms*—Building footprint change detection, deep convolutional neural network (DCNN), U-Net, unmanned aerial vehicle (UAV) image.**

I. Introduction

CHINA has undergone rapid urban expansion due to accelerated economic development over the past two decades. This urban expansion has resulted in a continuous increase in building footprints [1], [2]. The rapid changes in building footprints have caused many problems, including the deterioration of effective land use, the reduction of cultivated land, environmental pollution, and ecological destruction [3]. Therefore, accurate building footprint extraction and detection of the building footprint change is critical to the promotion of proper urban planning and achievement of sustainable economic and environmental development.

Presently, researchers rely on traditional image processing techniques to extract building footprint and detect building related changes in satellite images. These methods have gradually evolved from pixel-oriented to object-oriented techniques. Specifically, these methods and models involve the identification of uniform regions [4], watershed segmentation [5], morphological index [6], [7], clustering extraction of urban changes [8]–[10], bottom-up and top-down hybrid algorithms [11], and simple geometric structure methods [12]. Further, regarding multitarget change detection, Jabari *et al.* [13] proposed a change criterion that uses multivariate expansion to overcome nonlinear imaging condition differences and that utilizes multispectral properties for optical change detection. Zhang *et al.* [14] incorporated multiscale uncertainty analysis in a novel object-based change detection technique for unsupervised change detection in very high-resolution (VHR) images. However, these approaches have certain limitations. First, the registration of images generated in the same place but at different times must be highly accurate to allow for adequate detection performance. In practice, such high accuracy is often difficult to achieve, resulting in high omission and error rates [15]. Second, remote sensing data must meet rigorous requirements (e.g., similar resolutions and spectral features) and generally rely on artificially designed features. Moreover, most image analysis methods are only suitable for specific data [16]. Therefore, most existing models exhibit poor generalizability.

To overcome the limitations of traditional building footprint extraction methods, deep learning technology has developed rapidly in recent years. Deep convolutional classification networks usually comprise a fully connected layer that in turn requires input images of a fixed size. These network models have certain disadvantages, including their high storage overhead, low computational efficiency, and areal sensing restrictions. To achieve semantic segmentation, Long *et al.* [17] proposed the fully convolutional network (FCN); this approach allows the semantic segmentation of images of arbitrary sizes. However, the semantic segmentation achieved by FCNs often lacks adequate refinement [18]. Ronnerberger *et al.* [19] proposed the U-Net model, an extended FCN that employs a coder–decoder structure. Through a combination of low-level feature mappings, high-level complex features are constructed to achieve

Manuscript received September 28, 2020; revised November 8, 2020 and December 15, 2020; accepted January 13, 2021. Date of publication January 18, 2021; date of current version February 9, 2021. This work was supported in part by the Xuzhou Science and Technology Key R&D Program (Social Development) under Project KC20172, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and in part by the Grant from the State Key Laboratory of Resources and Environmental Information System. *(Corresponding author: Wei Liu.)*

The authors are with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China (e-mail: liuw@jsnu.eud.cn; xujiawei@jsnu.edu.cn; guozihui@jsnu.edu.cn; liezrs2018@jsnu.edu.cn; lixing@jsnu.edu.cn; zhanglp2000@126.com; liuwensongupc@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3052495

accurate positioning and address image segmentation. Owing to its ability to acquire full resolution maps, U-Net has become a general framework for some of the state-of-the-art semantic segmentation methods [20], CNNs and U-Net have been widely used in building footprint extraction [21]–[23]. Yang *et al.* [24] benefited from the scalability of CNNs and conducted a novel comparative analysis of four state-of-the-art CNNs for extracting building footprints throughout the United States. Majd *et al.* [25] proposed a novel object-based deep CNN (OCNN) framework for building extraction with VHR images. Saha *et al.* [26] exploited CNN features in a novel unsupervised context-sensitive framework-deep change vector analysis method for building change detection in multitemporal VHR images. Ji *et al.* [27] designed a Siamese U-Net with shared weights that significantly improves the segmentation accuracy. The aforementioned studies not only developed and validated new CNNs and U-Net models for building footprint extraction but also proposed the novel use of unmanned aerial vehicle (UAV) images to improve building footprint extraction and change detection.

However, the application of U-Net for building extraction remains hindered by certain limitations. First, U-Net is an FCN with good scalability, and its final layer features are attained by sampling, amplifying, and integrating the features of the previous layers. However, the skip-connection method it employs causes a semantic gap and does not allow for adequate feature expression [28]; thus, building footprint changes cannot be accurately estimated. To solve the semantic gap problem, Zhou *et al.* [29] proposed the application of the U-Net++ model; this model connects the encoder and decoder subnetworks through a series of nested and dense skip pathways. The redesigned skip pathways can reduce the semantic gap between the feature maps of the encoder and decoder subnetworks that often occur within existing models. Second, the features transmitted by U-Net may contain classification ambiguity or nonboundary related information [19] that can affect the classification results and building boundaries. It is easy to confuse the classification between artificial objects (e.g., roads and fences) and buildings [30]. Moreover, the tones and angles of building inclinations in UAV images vary according to specific UAVs and shooting angles. Some researchers have proposed that feature extraction in such models could be improved by inputting multiscale images and that building location accuracy could be improved by adding boundary information [20], [31]. However, the above strategies have two drawbacks. First, multiscale image input can result in repeated calculations and inefficiency. Second, the input of boundary information requires additional data and increased model complexity. Other researchers have contended that feature extraction could be improved by adjustments to the CNN itself, such as the incorporation of multiscale fusion and dilated convolution [32]. Although dilated convolution ensures a larger receptive field while parameters remain constant, it does not allow for the effective detection of certain small building footprints.

To address the above-mentioned problems, this study proposes a new building footprint extraction and change detection method. First, building footprints are efficiently and accurately extracted from UAV images using the PRU-Net model. Subsequently, change detection is performed to identify new and demolished building footprints. The proposed method improves on the performance of previous methods in the following four aspects.

1) Multiscale semantic parsing is performed using a pyramid scene parsing (PSP) module to allow improved footprint extraction of various building types.
2) A residual skip pathway is employed to address the semantic gap associated with the skip connection method.
3) A focal-loss (FL) function is utilized *in lieu* of the cross-entropy (CE) loss function to mitigate imbalanced building footprint extraction between positive and negative samples.
4) The results are processed and combined with GIS spatial analysis to extract new and demolished building footprints.

To test the performance of the PRU-Net model and its performance in building footprint detection, building footprint extraction experiments were performed over a 230 km^2 area, and building footprint change detection was tested over a 46 km^2 area. According to experimental results, the PRU-Net model exhibited strong generalizability, stable performance, and high robustness.

II. Materials and Methods

This study integrated deep convolutional neural networks (DCNNs) and high-resolution UAV image data to fully tap the value of UAV images and to extract building footprint and detect the changes in the images generated at different times using various shooting angles and definition. The overall framework is divided into three stages—data preprocessing; PRU-Net model training, validation, and prediction; and change detection postprocessing. This framework is shown in Fig. 1. First, the training/validation and prediction datasets are generated using high-resolution UAV images. Second, the training/validation dataset is input into the PRU-Net model for training and validation to obtain the optimized network model. The prediction dataset is input into the trained PRU-Net model to obtain the building footprint prediction results. Finally, the prediction results are processed via the merge, hole filling, projection/registration, and vectorization operations. Based on vectorization results, new and demolished building footprints are extracted via buffer, overlay, and negative intersection operations, successively.

To test the building footprint extraction capacity of the PRU-Net model, UAV images generated in May 2020 covering approximately 230 km^2 of Yizheng City were selected. As shown in Fig. 2(a), the resolution was 0.1 m, and districts 1, 2, and 3 (in red boxes) were considered in relatively more detailed building footprint extraction tests.

UAV images acquired during the two periods covering Jiangbei New District in Nanjing were selected to test the efficiency of the proposed method in detecting the changes in building footprint. As shown in Fig. 2(b), the resolution was 0.1 m, area covered measured approximately 46 km^2, and UAV images were generated during two periods: May and August 2020. Jiangbei New District is a new urban district that underwent

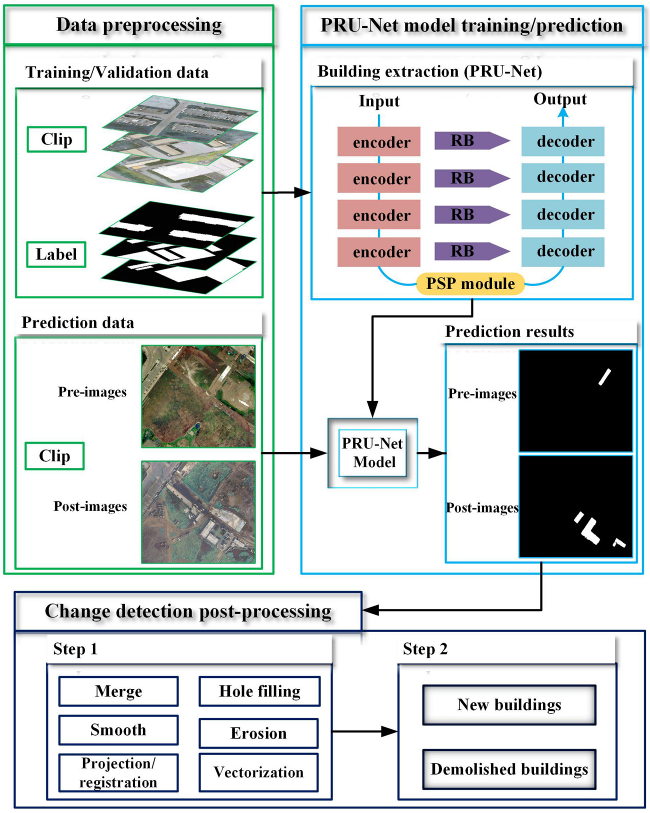


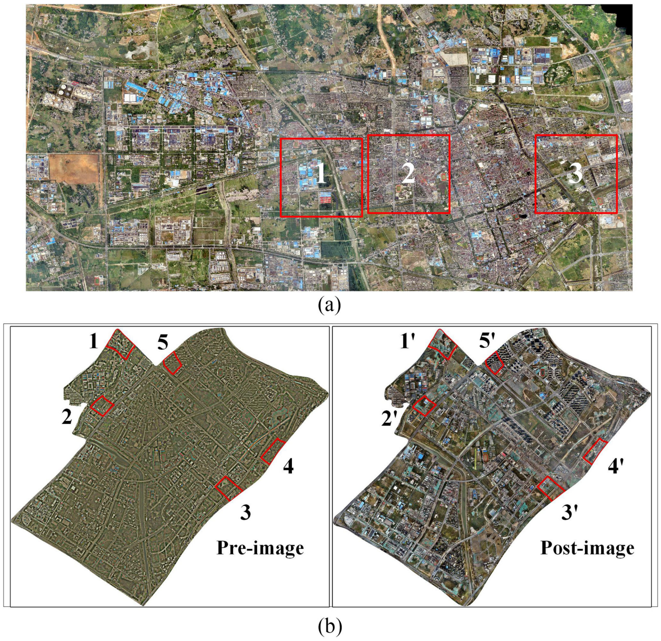
Fig. 1. Framework of proposed method.

drastic changes in building footprint during the study period. Therefore, local policymakers need to acquire information on building footprint change every three months to regulate land resources. Building footprint change results for districts 1, 2, 3, 4, and 5 were analyzed in detail.

A. Data Preprocessing

In this study, datasets were preprocessed as follows. The UAV images were linearly stretched by 2%, that is, the pixel value at which the histogram reached 2% of the total was used as the lower limit (MinValue), and the pixel value at which the histogram reached 98% of its highest value was used as the upper limit (MaxValue). If a pixel value was greater than the MinValue but less than the MaxValue, it was stretched from 0 to 255. If the pixel value was less than the MinValue, it was set as 0. If the pixel value was greater than the MaxValue, it was set as 255. This stretching operation was performed to overcome the considerable color differences between different splicing areas.

The UAV images were classified into three groups—a training set, a validation set, and a prediction set. A sliding window was used to maintain a 50% overlap rate for the prediction set, and each UAV image was divided into 256 × 256-pixel image blocks. Training and validation images were manually labeled using the Labelme tool, and the labels were subjected to one-hot processing to convert them into two-channel data, that is, building footprints were expressed as 255, and other features were expressed as 0.



(a)

(b)

Fig. 2. UAV datasets. (a) UAV image of building footprint extraction. (b) Two-phase UAV image of building footprint change detection.

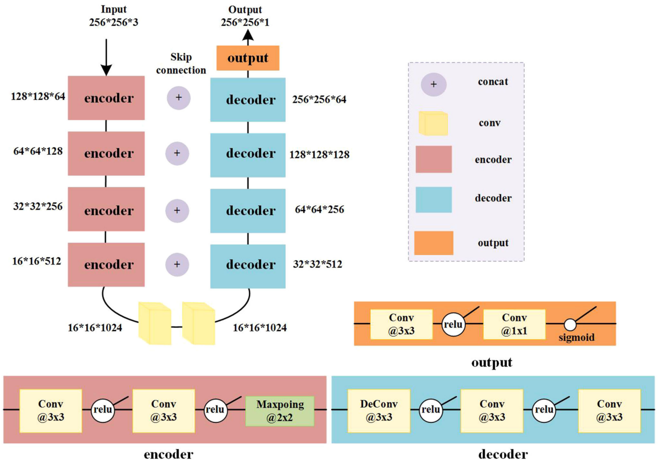


Fig. 3. U-Net network structure.

B. U-Net Network Model

U-Net is a network structure proposed in the IEEE International Symposium on Biomedical Imaging contest in 2015 [19]. Its U-shaped structure is realized through a contracting and expanding network (see Fig. 3).

The contracting network is mainly responsible for the down sampling and extraction of high-dimensional feature information. Each down sampling operation includes two 3 × 3 convolutional operations and one 2 × 2 pooling operation, wherein a rectified linear unit (ReLU) is used as the activation function. Each down sampling operation is performed to halve the image size and double the number of features. The expanding network is mainly responsible for up sampling. In each up-sampling operation, the output features are merged with the features of the mapped contracting network via skip connection to offset the

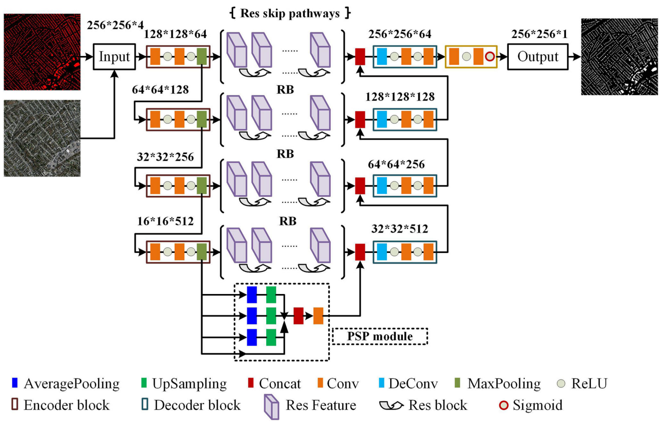


Fig. 4. PRU-Net network structure.

missing boundary information. Finally, the 1 × 1 convolutional operation is performed to map the previously extracted features into the proper category.

Compared with other networks, the advantages of the U-Net model include its simple structure, short training period, and smaller number of training parameters. However, the depth of U-Net is weak compared to that of Inception and ResNet.

C. PRU-Net Network Model

U-Net has demonstrated reliable performance in the segmentation of medical images. However, its feature fusion and extraction methods are not well diversified owing to its fully symmetrical decoding–coding method. In this study, a PSP module [33] was added to the top code block of U-Net, thereby expanding U-Net's multiscale feature aggregation ability and improving its ability to extract building footprints of different sizes. While the U-Net skip connection method has been observed to effectively restore image details, the direct fusion of its low- and high-layer features results in a semantic gap [28]. However, the use of residual connection can reduce semantic gaps. In this study, residual skip pathways are used instead of the typical U-Net skip connection. The integration of the two modules mentioned above and incorporation of the improved loss function in U-Net and the modified U-Net model proposed in this study is accordingly considered as the "PRU-Net." Its structure is shown in Fig. 4.

1) Structure of PRU-Net: As shown in Fig. 4, PRU-Net features a symmetrical structure with equally sized input and output images and 23 convolutional layers. Four down sampling and four up sampling operations are performed. There is no full connection layer, and the features generated by each layer contain enhanced semantic information. The left half of the PRU-Net network is structured as a typical U-Net, wherein high-layer features are extracted by gradually reducing the spatial dimensions of input data. As the network's core entails four groups of convolutional operations, two consecutive 3 × 3 convolutions (with the same number of convolution kernels) are used within each group. From the shallow to the deep group, the number of convolution kernels increases exponentially from 64 to 512. Spatial dimensionality reduction is performed through a pooling operation between every two groups, and some unimportant high-frequency information is filtered out. As a result, the spatial size of the feature map is reduced. After the convolutional operations, ReLU is used as the activation function to reduce training-stage gradient dispersion originating from network deepening. In addition, a batch norm (BN) layer was added to the downstream path to enable the normalization of the features of each network layer in each batch, thereby ensuring relatively stable distribution of each layer and improving the model's convergence rate and capacity [34]. To acquire the complex features of building footprints with different shapes and sizes, the PSP module is added to extract multiscale features.

The right half of the network consists of deconvolutional layers from bottom to top, and the input consists of two parts—deep abstract features obtained through deconvolution at the upper layer and features (including shallow local features) generated by the corresponding left half of the network. The two input features are integrated through residual skip pathways to gradually restore building footprint details and spatial dimensions. In addition, a dropout layer was added to the merger path, to randomly neutralize 50% of the weights of hidden nodes, thereby improving the generalizability of the network and preventing overfitting to a certain extent [35]. To eliminate the confusion of up sampling after feature fusion, a 3 × 3 convolution kernel is used to perform two convolution operations. After the above operations have been performed four times, the output images of the last layer can be restored to the size of the input images, and each pixel is classified and predicted.

2) Multiscale Scene Parsing: Inputting multiscale images is a viable solution to the considerable variation in building footprints with respect to material and scale. By inputting multiscale images, Liu *et al.* [31] rendered a network to adopt multiscale building footprint features. Ji *et al.* [36] used both original and down sampled images to influence the SiU-Net model into adopting multiscale building footprint features. The use of a multiscale feature extraction module (e.g., ASPP, SPP, and Vortex) is also a viable solution. The ASPP module introduces the dilated convolution of different ratios to segment multiscale features, thereby reducing the loss of information caused by pooling operations [37]. The SPP module designs pyramid pooling operations to obtain fixed-length feature dimensions and multiscale feature information [38], whereas the Vortex module further improves the ASPP module by adding an average pooling operation before the dilated convolution to obtain rich feature information [39]. These modules perform well in natural and remote sensing image segmentation. However, ASPP, SPP, and Vortex exhibit difficulty in the acquisition of global information around building footprints. Considering the improved spatial resolution of UAV images, building footprints in such images usually exhibit high intraclass and low interclass differences. In the process of building footprint extraction, the acquisition of building footprint features, information on the scene around the building footprints, and the aggregation of scenes in different regions is necessary. Thus, scenic information indicative of the relationship between different subregions can be retained and building footprints can be extracted more effectively.

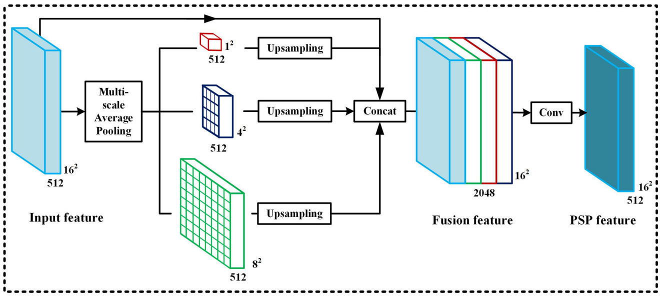


Fig. 5. PSP module.

PSP is a feature fusion method based on multiscale pooling [33]. Through average pooling, PSP aggregates multiscale semantic information from different regions to improve the network’s ability to obtain global information. Compared with dilated convolution, pooling operations involve far fewer calculations. To acquire global information on multiscale features of building footprints, this study introduced the PSP module after the fourth code block of PRU-Net. Compared with the PSP module in the study conducted by Zhao *et al.* [33], the PSP module proposed in this study directly performs deconvolution operations without a convolution dimensionality reduction after average pooling at each layer. In the process, it selects an appropriate layer and average-pooling factor for the UAV image set. As shown in Fig. 5, the PSP module uses 16 × 16, 8 × 8, and 2 × 2 average pooling for feature mapping to generate multiscale features, including 1 × 1, 4 × 4, and 8 × 8 features. Then, the multiscale features are up sampled and restored to match the size of the original input features, and fused features are generated through aggregation operations. Finally, PSP features are generated through 1 × 1 convolution of the fused features. The PSP module contains hierarchical global information on different scales, thereby enabling PRU-Net to fully acquire global information on multiscale building footprints.

3) Residual Skip Pathways: A unique contribution of U-Net was its introduction of skip connections; this innovation enabled the network to offset spatial information lost in pooling operations during coding and decoding [40]. However, the first skip connection of U-Net merges features after the first pooling operation and features before the last deconvolution operation. The former are low-level features, whereas the latter are advanced features that contain rich semantic information resulting in the semantic gap identified by Ibtehaz and Rahman [28] who thus proposed the use of residual connection instead of the original skip connections. In this study, residual connections are used as reference in the network, and the improved residual skip pathways were integrated into the PRU-Net network model. This residual structure may also alleviate potential degradation associated with deep network optimization [41].

As shown in Fig. 6(b), a residual skip pathway comprises *n* residual blocks (RBs) that are connected in a head-to-tail manner. Feature F_j^n is acquired after feature F_j^1 , acquired by the coder, and is extracted by *n* RBs. The deconvolved features are fused with F_j^n and are then jointly input into the decoder.

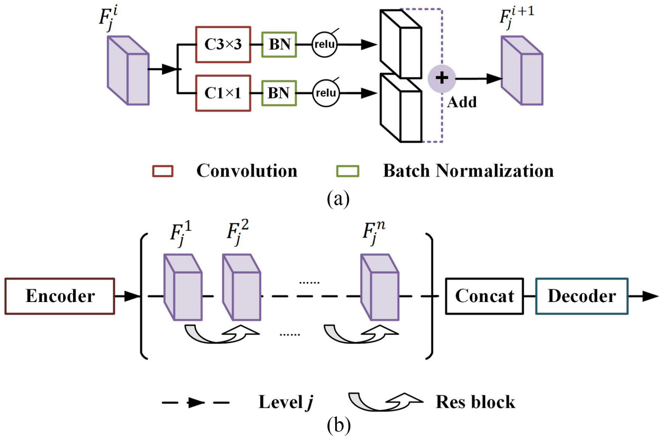


Fig. 6. Residual skip pathways. (a) RB. (b) Residual skip pathways consisting of *n* RBs.

The process of RB extraction is described as

$$F_j^{i+1} = R[N(F_j^i \otimes C_1]\oplus R[N\left(F_j^i \otimes C_3\right] \tag{1}$$

where F_j^i denotes the original features after the RB is input, and F_j^{i+1} denotes the features after extraction by the RB. In addition, ⊗ denotes the convolution operation, ⊗denotes the “add” operation, C_1 and C_3, respectively, denote the 1 × 1 and 3 × 3 convolution kernels, *N* denotes the BN layer, and *R* denotes ReLU. The overall design is referred to as RB, as shown in Fig. 6(a). The RB does not change the number of channels or the size of any feature. Considering the calculation efficiency, a varying number of RB is set in the residual skip pathway of each layer. Three main combination modes (2/2/2/2, 1/2/3/4, and 2/4/6/8) are employed to extract residual features. For example, 2/2/2/2 indicates that two RBs are used in the residual skip pathway of each layer, and this combination mode was found to yield optimum performance. Therefore, the 2/2/2/2 RB structure was used in the PRU-Net network.

4) Loss Function: Imbalanced data categories often affect model performance. According to the calculations, the pixels occupied by building footprints and backgrounds accounted for 9% and 91% of the dataset for Jiangbei New District and 13% and 87% of the dataset for Yizheng City, respectively. Therefore, it is necessary to address the imbalance between categories. Typical solutions to data imbalance issues include: 1) increasing the number of samples of low-data categories and 2) designing a cost function to promote network learning of positive examples. The first solution requires copious amounts of manual marking and is very time consuming. As the main objective of this study was to extract building footprint from UAV images (i.e., dichotomy problem), a cost function was considered. In semantic segmentation, CE loss is often used as the optimized cost function to address dichotomy problems, as expressed in

$$\mathrm{CE}\ (y, y') = \begin{cases} -\log\left(y'\right) & \text{if } y = 1 \\ -\log\left(1 - y'\right) & \text{if } y = 0 \end{cases} \tag{2}$$

where *y* is a one-hot vector (values: 0 and 1; if the category is the same as that in the label, the value is 1; otherwise, the value is 0), and *y'* denotes the probability of the prediction result with respect to the label. However, when CE loss is used to segment building footprints and backgrounds, and building pixels are far fewer than background pixels, the model can become severely biased toward backgrounds, resulting in poor building footprint extraction.

Weighted CE (WCE) can be used to address this problem. Specifically, the weight penalty factor *a* is applied to background samples to reduce the preference of the model for background samples and to improve building footprint extraction, as shown in

$$\text{WCE }(y,y') = \begin{cases} -a\log(y') & \text{if } y=1 \\ -(1-a)\log(1-y') & \text{if } y=0 \end{cases} \tag{3}$$

where *a* represents a balance between the number of building footprint samples and that of background samples. However, a large number of building footprint extraction targets is easy negative samples. Although the loss value of these samples is exceptionally low, the number of easy negative samples is relatively too large. Easy negative samples ultimately dominate the total loss value, resulting in low building footprint extraction accuracy.

As the use of easy negative samples scarcely improves model performance, the model should mainly focus on difficult negative samples. Lin *et al.* [42] proposed adding the γ penalty factor to improve FL, wherein the loss value of samples with a high degree of confidence is appropriately reduced. Thus, the model prioritizes difficult negative samples, as expressed in

$$\text{FL }(y,y') = \begin{cases} -(1-y')^{\gamma}\log(y') & \text{if } y=1 \\ -y'^{\gamma}\log(1-y') & \text{if } y=0. \end{cases} \tag{4}$$

Overall, (3) addresses the imbalance between positive and negative samples, and (4) addresses the imbalance between difficult and easy samples. Equation (5) combines elements of (4) and (5), thereby optimizing FL

$$\text{FL }(y,y') = \begin{cases} -a(1-y')^{\gamma}\log(y') & \text{if } y=1 \\ -(1-a)\,y'^{\gamma}\log(1-y') & \text{if } y=0. \end{cases} \tag{5}$$

Specifically, *a* solves the imbalance between positive and negative samples, and γ solves the imbalance between difficult and easy samples. The incorporation of FL as presented in (5) allows the proposed model to focus on difficult negative samples and improves the accuracy of building footprint extraction.

D. Change Detection Postprocessing

Change detection was performed in two steps. Step 1 addressed isolated points, planes, and holes in the building footprint extraction results yielded by the PRU-Net model. To optimize the results, such disturbances must be removed through hole filling, smoothing, and erosion. In addition, it was necessary to perform merge, projection, registration, and vectorization operations on the optimized results. Hole filling and smoothing were performed using the floodFill and Dilation functions, respectively, in OpenCV, whereas erosion was performed using the SieveFilter function in GDAL. Projection, registration, and vectorization operations were performed using specific tools in ArcGIS.

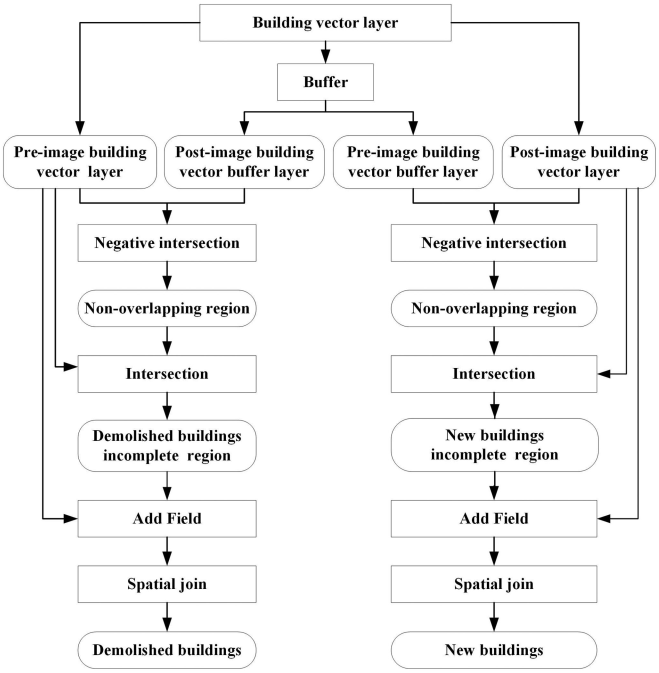


Fig. 7. Postprocessing of building footprint change detection.

As the UAV images were generated from different shooting angles during the two shooting periods, the building footprints captured in these images were displaced relative to each other. Thus, in step 2, it was necessary to perform overlay analysis by building a buffer area to detect building footprint changes. Specifically, buffer areas were built according to the image vectorization results (generated in step 1) for the two sets of UAV images. Based on the buffer areas, the operations specified in Fig. 7 were performed, which ultimately allowed for the acquisition of new and demolished building footprints.

1) Demolished Building Footprints: The building footprint vector layer was buffered, and the preimage and proimage building footprint vector buffer layers were generated, in succession. To extract nonoverlapping regions, a negative intersection operation was performed on the preimage building footprint vector layer and the proimage building footprint vector buffer layer. Intersection processing was performed on the nonoverlapping region layer and preimage vector layer to obtain demolished building footprint incomplete region. Field values were added to the demolished building footprint incomplete region, and the same field was added to the preimage vector layer (with different field values). The two phases of layers were connected to one view by the spatial join operation, and the demolished building footprints were selected by conditional filtering.

2) New Building Footprints: To extract nonoverlapping regions, a negative intersection operation was performed on proimage building footprint vector layer and preimage building footprint vector buffer layer. Intersection processing was performed on the nonoverlapping region layer and proimage vector layer to obtain new building footprint incomplete region. Field values were added to the new building footprint incomplete region, and the same field was added to the proimage vector layer (with

TABLE I
EXPERIMENT CONFIGURATION

CPU	GPU	Memory	System	Disk	TensorFlow	Keras	Python	CUDA
Intel Xeon E5-V4	P4000	16GB	Ubuntu 16.04	3T	2.0	2.1.5	3.70	9.0

TABLE II
HYPERPARAMETER CONFIGURATION

Hyper parameter	U-Net	PU-Net(Unet+PSP)	PRU-Net
Number of iterations	100	200	200
Learning rate	1e-4	1e-4	1e-4
Batch size	12	12	12
Image enhancement	Yes	Yes	Yes

different field values). The two phases of layers were connected to one view by spatial join operation, and the new building footprints were selected by conditional filtering.

E. Evaluation Metrics

To objectively evaluate the performance of the PRU-Net model, this study utilized five common assessment indicators: Precision, recall, F1 score, intersection over union (IoU), and kappa

$$\text{Precision} = TP/(TP + FP) \tag{6}$$

$$\text{Recall} = TP/(TP + FN) \tag{7}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{IoU} = \text{Area}\,(D \cap T)\,/\,(D \cup T) \tag{9}$$

$$\text{Kappa} = \frac{(TP + TN)\,/N - (TP + FP)\,(FN + TN)\,/N^2}{1 - (TP + FP)\,(FN + TN)\,/N^2}. \tag{10}$$

TP, *FP*, *FN,* and *TN* stand for true positives, false positives, false negatives, and true negatives, respectively. *D* is the detected building, *T* is the ground truth buildings, and *N* is the total number of samples.

III. RESULTS

A. Network Training and Optimization

The network model proposed in this study was based on the deep learning platform Tensorflow2.0 and Keras 2.1.5 coding. Two NVIDIA Quadro GPUs (memory: 16 GB) were used for model operation, as described in Table I. The batch size was set to 12, and the learning rate was set to 1e-4. Hyperparameters were all determined experimentally, as described in Table II. The Adam optimization function was used in this study owing to its high calculation efficiency, relatively low memory consumption, and superior performance in addressing nonstationary models. To improve the generalizability of the model and eliminate overfitting, the dataset was enhanced through random inversion, translation, and zooming and noise adjustments; dropout and BN layers were added to reduce the effect of overfitting. The dropout rate was set to 0.5. To prevent overfitting, training was sometimes terminated earlier than planned. Hence, the number of training iterations completed by the model may differ from the initial settings; however, this did not affect the comparative evaluation of network segmentation performance.

B. Building Footprint Extraction Results

A building footprint extraction experiment was performed using UAV images of Yizheng City to test the performance of different network models. Fig. 8 compares the building footprint extraction performance of U-Net, PU-Net, U-Net++, and PRU-Net. Green denotes TP, blue denotes FN, and red denotes FP. The first through sixth columns represent the original image, label, U-Net, PU-Net, U-Net++, and PRU-Net, respectively. U-Net had the highest detection error rate, manifested as a large number of red pixels. Compared with U-Net, PU-Net produced better results, but still exhibited a relatively higher FN, manifested as a large number of blue pixels. Compared with PU-Net, U-Net++ reduces the misjudgment of roads and water bodies as building footprints with high precision but a low recall rate. PRU-Net extracts the boundary of the building footprint clearly more in line with the actual shape of the building footprint. Moreover, it can show more detailed information of the building footprint, and the phenomena of speckles and holes were basically eliminated.

The first test image represented a suburban area and contained many factories and some dense building footprints. U-Net could not identify the large building footprints accurately. PU-Net offered better detection than U-Net, but has a high detection omission rate. U-Net++ is similar to PU-Net with respect to building footprints extraction. PRU-Net detected more large building footprints that better reflected the actual contours of building footprints; the recall and IoU could reach 0.9565 and 0.7238, respectively (see Table III). The second test image represented the city center. In this image, building footprint groups were very dense, and building footprint types are complex, thereby complicating the extraction process. U-Net, PU-Net, and PRU-Net effectively extracted dense building footprint groups. However, the overall identification accuracy of U-Net was low, for example, it misclassified some roads as building footprints, and its recall was merely 0.8535 (see Table III). The detection error rate of PU-Net and U-Net++ was marginally reduced, as indicated by its markedly fewer red pixels. PRU-Net again yielded the most optimal building footprint extraction, and its recall rate was as high as 0.9346 (see Table III). The third test image represented the new urban area. It contained regularly arranged high-rise building footprints and dense and scattered low building footprints. U-Net was prone to misclassify roads as building footprints and had a remarkably high omission rate that manifested as a few green pixels and many red pixels. PU-Net reduced the omission rate to a certain extent and yielded a 4.17%

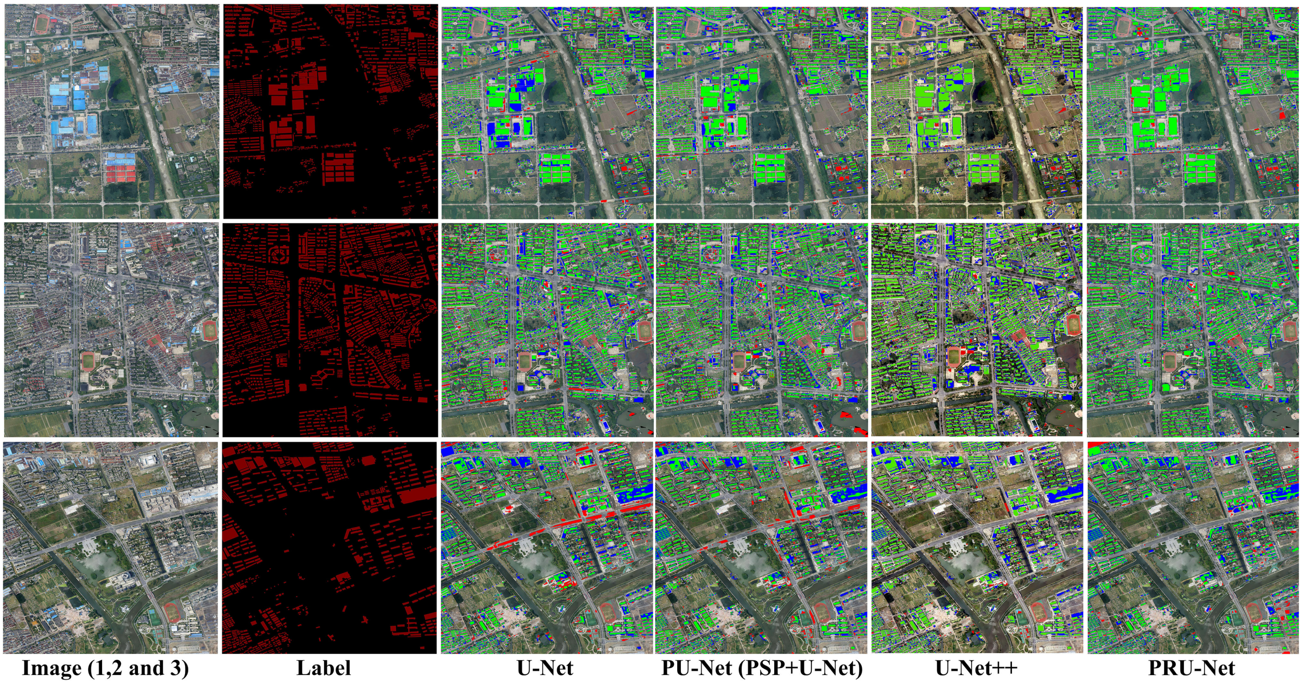


Fig. 8. Results of building footprint extraction (green: TP, blue: FN, red: FP).

TABLE III
COMPARISON OF BUILDING FOOTPRINT EXTRACTION ACCURACY

Image	Model	Precision	Recall	F1	Kappa	IoU
Image1	U-Net	0.8487	0.8640	0.8563	0.8015	0.5986
	PU-Net	0.8794	0.8994	0.8893	0.8357	0.6458
	U-Net++	0.8912	0.8765	0.8837	0.8379	0.6472
	PRU-Net	0.8886	0.9565	0.9213	0.8932	0.7238
Image 2	U-Net	0.8017	0.8535	0.8268	0.7843	0.5696
	PU-Net	0.8154	0.8931	0.8525	0.8013	0.5855
	U-Net++	0.8375	0.8642	0.8506	0.7993	0.5764
	PRU-Net	0.8404	0.9346	0.8850	0.8471	0.6724
Image 3	U-Net	0.8100	0.8732	0.8404	0.7912	0.6067
	PU-Net	0.8370	0.9096	0.8718	0.8237	0.6437
	U-Net++	0.8598	0.8764	0.8680	0.8132	0.6503
	PRU-Net	0.8569	0.9518	0.9019	0.8653	0.6984

increase in recall ratio. PRU-Net did not misclassify roads as building footprints and yielded optimum detection among the four models.

Table III quantitatively compares the performance of the four models. In suburbs, city centers, and new urban areas, U-Net yielded a kappa value of 0.8051, 0.7843, and 0.7912, respectively. The PU-Net and U-Net++ models exhibited average detection ability, thereby yielding average kappa accuracy. However, the PRU-Net can yield the highest kappa values, compared with the U-Net, exhibiting relative improvement in performance: 11.44%, 8.01%, and 9.37% improvement in suburbs, city centers, and new urban areas, respectively. In suburbs and new urban areas, the U-Net++ model yielded the highest precision. Meanwhile, the PRU-Net model yielded the highest and most stable recall rate, F1, kappa, and IoU, indicating that it achieves a better balance of precision and recall. Thus, the use of PRU-Net is clearly beneficial in the extraction of building footprints from UAV images.

C. Building Footprint Change Detection Results

UAV images covering a 46 km^2 area in Jiangbei New District of Nanjing City were selected to conduct a large-scale experiment of building footprint change detection. First, the trained PRU-Net network was used to extract building footprints (see Fig. 9). Then, based on the building footprint extraction results, change detection postprocessing was conducted to obtain information on new and demolished building footprints (see Fig. 10).

As shown in Fig. 9, definition and shooting angle of the UAV images obtained during the two periods differed, causing building footprint parallax distortion. In the preimages, spectral information was weak, and image definition was affected to some extent. Therefore, many holes and tiny spots existed in the building footprint extraction results, and the building footprint boundaries were not sufficiently smooth. These adverse effects were eliminated through hole filling, smoothing, and erosion, yielding a minor impact on building footprint change detection. In addition, the definition of preimages was relatively low; therefore, some building footprints in the preimages were not detected. However, such building footprints were detected in the higher quality postimages. For example, the building footprint in the blue box in the fourth row in Fig. 9 had a significant

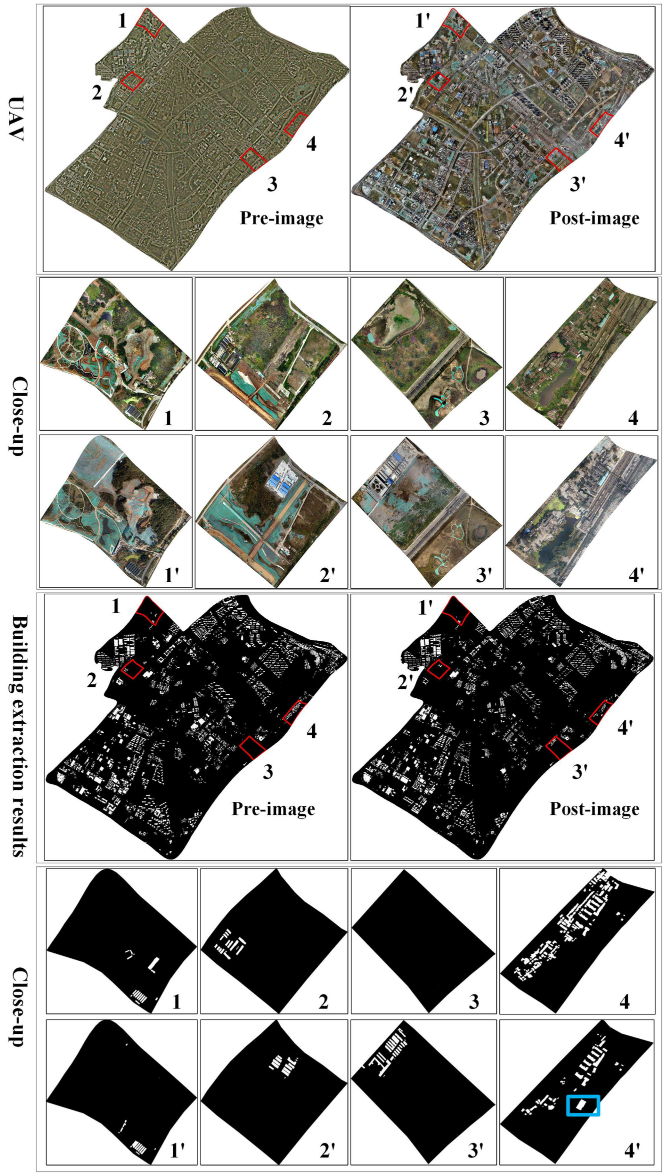


Fig. 9. Building footprint extraction results.

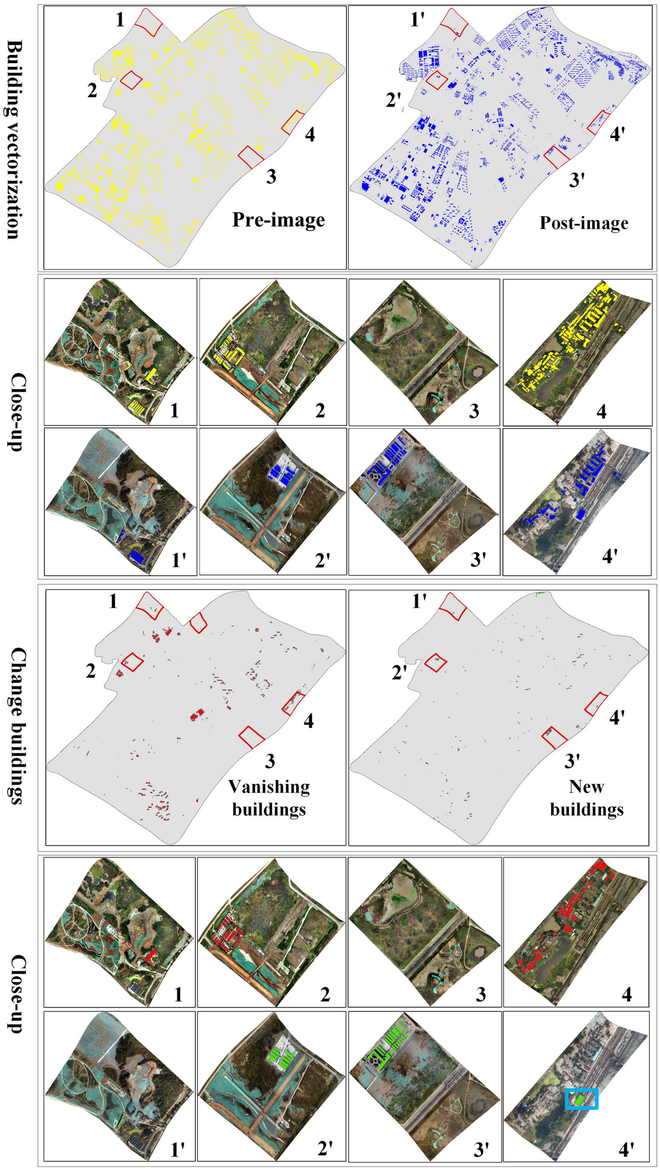


Fig. 10. Postprocessing results of building footprint change detection.

impact on the building footprint change detection results, as it was inaccurately classified as a new building footprint. However, overall, PRU-Net managed to produce good building footprint extraction results from the postimages, and the building footprint boundaries in the postimages were relatively clear, thereby facilitating building footprint change detection postprocessing.

In Fig. 10, the first line is building footprint vectorization results of overall layer, the yellow vector patches are preimage building footprint extraction results, and the blue vector patches represent postimage building footprint extraction results. In the second line, the yellow vector patches represent the close-up of the preimage building footprint extraction results, and the blue vector patches represent the close-up of the postimage building footprint extraction results. The third line is building footprint change detection results of overall layer; red vector patches were demolished building footprints, and blue vector patches are new building footprints. In the fourth line, red vector patches represent demolished building footprints close-up, and blue vector patches represent new building footprints close-up. From the detection of the building footprint in the blue box in the fourth row of Fig. 10, the building footprint was evidently regarded as a new building footprint due to inaccurate interpretation results that had a relatively greater impact on the detection results.

Owing to the differences in the shooting angle of the UVA images, building footprints can sometimes appear to be parallax distortion. During building footprint change detection postprocessing, the choice of buffer radius is particularly important, which considerably affects the results. The buffer radius should be determined strictly according to the maximum parallax distortion of the unchanging building footprint before and after the time phase. In this study, the experimental results showed that a buffer radius of 15 m was suitable for the target area. The

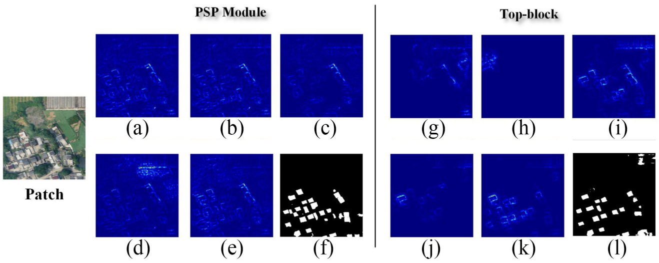


Fig. 11. Visualization of building footprint extraction feature.

TABLE IV
COMPARISON OF DIFFERENT MODELS (THE FIRST ROW PRESENTS RESULTS OF THE BENCHMARK U- NET MODEL; THE SECOND ROW PRESENTS RESULTS WITH THE ADDED PSP MODULE; THE THIRD ROW PRESENTS RESULTS WITH THE ADDITION OF AN RB; THE FOURTH ROW PRESENTS RESULTS WITH THE ADDITION OF AN FL MODULE)

Model	Precision	Recall	F1	Kappa	IoU
Baseline	0.8207	0.8636	0.8416	0.7952	0.5857
+PSP	0.8472	0.8982	0.8720	0.8186	0.6201
+PSP+RB	0.8556	0.9374	0.8946	0.8514	0.6806
+PSP+ RB +FL	0.8629	0.9501	0.9044	0.8645	0.6951

impact of building footprint parallax distortion will be discussed in further detail below.

IV. Discussion

A. Visual Analysis of Multiscale Scene Parsing

To explore the feature segmentation ability of the multiscale scene analysis module, visual analysis of the network features was performed. Fig. 11 shows the multiscale scene-parsing module and the top convolution output (i.e., top block) after the fourth down sampling by the original U-Net. Fig. 11(a)–(c) shows the convolution output associated with three levels (red, blue, and green) in Fig. 5, and Fig. 11(d) shows a feature output channel of the module. Fig. 11(g)–(j) presents four feature graphs of the two convolutional layers of the top block. Compared to the original U-Net module, the PSP module extracted more detailed building footprint features. Fig. 11(e) and (k) shows the output of the last convolutional layer of the PSP and U-Net modules, respectively. After the introduction of the PSP module, the network also maintained high feature output when the final convolution was activated. Fig. 11(f) and (i) shows the final prediction results of the two networks. The network incorporating the PSP module acquired additional building footprint features and yielded more accurate predictions than the original U-Net network.

B. Impact of Network Module on Building Footprint Extraction

Table IV presents the test results concerning multiscale scene parsing, residual skip pathways, and different loss function modules. As described in Table IV, the PSP module in the second model allowed for multiscale feature aggregation and thus could capture the building footprint information that was more in line with real labels. In addition, it improved the IoU by 5.87%. The introduction of the RB module in the third model offset the semantic gap arising from the skip connection between the bottom and top layers of the network. In addition, the precision rate, recall ratio, F1, and IoU were improved to a certain degree. Specifically, the precision rate was improved by 4.25%, the recall ratio by 8.55%, F1 by 6.30%, and IoU drastically by 16.2%. Following the addition of the PSP and RB modules, the possibility of misclassifying roads and water bodies as building footprints was significantly reduced, and the detection rate of large building footprints was also significantly improved, thereby improving the IoU of building footprints significantly. FL fine-tuning training was incorporated into the fourth model to improve the model's ability to detect difficult negative samples, especially building footprint boundaries. Under the fourth model configuration, precision rate, recall ratio, F1, and IoU reached their peaks. The incorporation of FL effected a continuous 2.13% improvement in IoU.

C. Impact of Parallax Distortion on Building Footprint Change Detection

With respect to building footprint change detection, UAV images significantly differ from VHR image sets. Usually, VHR images shooting height can reach several hundred kilometers. For the VHR images, the effect of parallax distortion on building footprints caused by tones and angles can be disregarded because of the high orbit at which these images were captured. However, UAV flying height typically does not exceed 1000 m, so the tones and angles of building footprint inclination in UAV images vary drastically according to specific UAVs and shooting angles, especially with respect to high-rise buildings [see Fig. 12(a) and (b)]. This can greatly influence the detection of building footprint changes in two-stage UAV images. The building footprint extraction and vectorization results shown in Fig. 12(c)–(f) indicate that parallax distortion had a minor impact on the detection of low building footprints but a significant impact on the detection of high-rise building footprints. To eliminate the impact of parallax distortion on high-rise building footprint change detection, it was necessary to perform buffer processing on building footprints, and it was particularly important to select a reasonable buffer radius. Fig. 12(g) and (h) shows the building footprint change detection results when the buffer distance was set to 5 m. For buildings with a parallax distortion that exceeded 5 m, the results contained several errors. However, due to the short buffer distance, slight building footprint changes in the small building footprints surrounding high rises were generally well detected [as shown in the blue boxes in Fig. 12(g) and (h)]. Fig. 12(i) and (j) shows the change detection results when the buffer distance was 10 m. A moderate number of errors was observed for building footprints with a parallax distortion that exceeded 10 m, and the changes in some small buildings close to the high-rise building within 10 m were not detected [e.g., the small new building footprints in the blue box in Fig. 12(j) were not detected successfully]. Fig. 12(k) and (l) shows the change detection results when the buffer distance is 15 m. While some

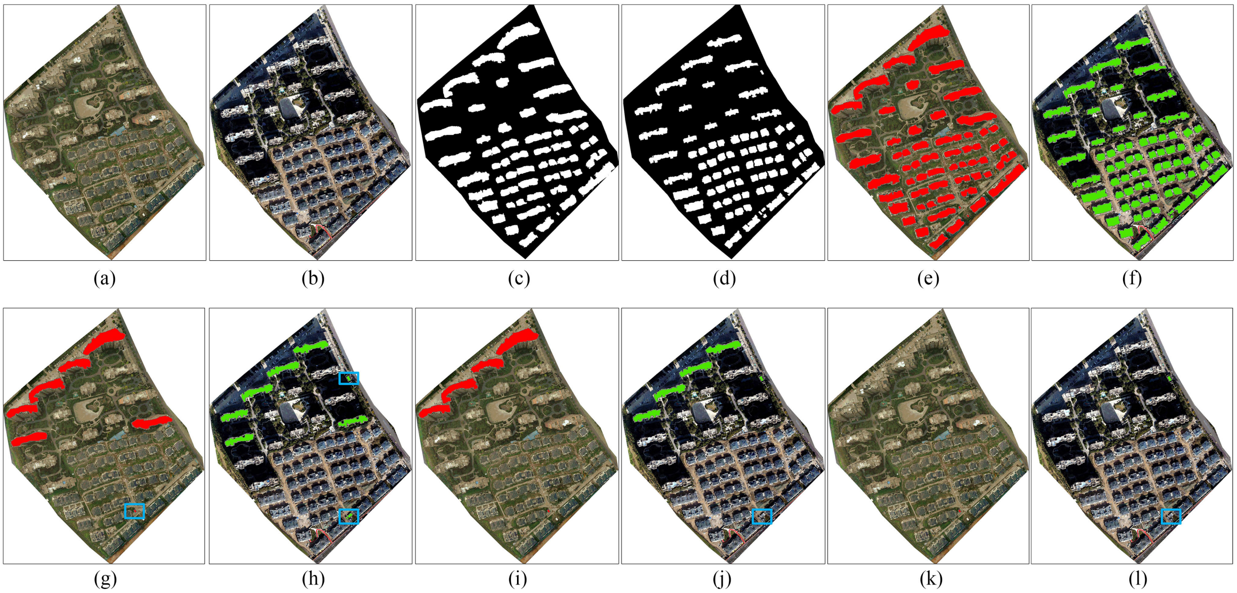


Fig. 12. (a) and (b) Two-phase UAV images. (c) and (d) PRU-Net extraction results. (e) and (f) Vectorized results. Building footprint changes results under the conditions of (g) and (h) 5 m; (i) and (j) 10 m; and (k) and (l) 15 m buffer distances, respectively (red represents the demolished building footprint and green represents the new building footprint).

TABLE V
IMPACT OF PARALLAX DISTORTION DISTANCE ON DETECTION RESULTS OF BUILDING FOOTPRINT CHANGE

Distance (m)	Precision	Recall	F1
5	0.8182	0.9825	0.8929
10	0.8486	0.9632	0.9023
15	0.9618	0.9517	0.9567
20	0.9642	0.9135	0.9382

changes in small buildings next to high rises were not detected [blue box in Fig. 12(l)], overall errors in building footprint change detection were reduced effectively.

Table V shows the impact of parallax distortion distance on the results of change detection of building footprint. The highest recall was 0.9825 when the distance was set to 5 m. However, the parallax distortion distance of many high-rise buildings exceeds 5 m, yielding a precision of only 0.8182 in building change detection. When the distance of parallax distortion is set to 20 m, the highest precision can be obtained, but recall is reduced to 0.9135. When the parallax distortion distance is set to 15 m, a more balanced precision and recall can be obtained, and the F1 value is 0.9567, which is the highest. When the parallax distortion distance changes from 10 to 15 m, the precision drastically changes from 0.8486 to 0.9618, indicating that the parallax distortion distance of high-rise buildings in this study region ranged between 10 and 15 m. Based on these results, the recommended parallax distortion distance was set at 15 m.

V. CONCLUSION

This study proposed a building footprint extraction and change detection method. The proposed fully convolutional PRU-Net network was used to extract building footprints from UAV images with varying definition and shooting angle. Building footprint change detection postprocessing was used to address the significant parallax distortion changes in building footprints within UAV images acquired during two different periods. A large-scale experiment confirmed that the proposed method can automatically perform accurate building footprint extraction from multiperiod, multiangle, and multiresolution UAV images, and the results can be applied to quickly identify building footprint changes. The proposed method can effectively detect the new and demolished building footprints in UAV images. However, the current method requires postprocessing, and an end-to-end building footprint change detection network model could be established in the future to realize building footprint change detection in one step. Overall, the results suggest that the proposed method is efficient and practical for extensive use, including in newly developed urban areas that are experiencing rapid change and thus must be routinely re-evaluated by policymakers.

REFERENCES

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

[2] Y. Shi, Q. Li, and X. X. Zhu, “Building footprint generation using improved generative adversarial networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 16, no. 4, pp. 603–607, Apr. 2019.

[3] P. Xiao, X. Zhang, D. Wang, M. Yuan, X. Feng, and M. Kelly, “Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition,” *ISPRS J. Photogramm. Remote Sens.*, vol. 119, no. 9, pp. 402–414, 2016.

[4] C. Beumier and M. Idrissa, "Building change detection from uniform regions," in *Iberoamerican Congress on Pattern Recognition.*, Springer, Berlin, Heidelberg, 2012, pp. 648–655.

[5] M. Turker and E. Sumer, "Building-based damage detection due to earthquake using the watershed segmentation of the post-event aerial images," *Int. J. Remote Sens.*, vol. 29, no. 11/12, pp. 3073–3089, 2008.

[6] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[7] S. Liu, Q. Du, X. Tong, A. Samat, L. Bruzzone, and F. Bovolo, "Multiscale morphological compressed change vector analysis for unsupervised multiple change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4124–4137, Sep. 2017.

[8] W. Li, Y. Wu, and Z. Hu, "Urban change detection under large view and illumination variations," *Acta Automatica Sinica*, vol. 35, no. 5, pp. 449–461, 2009.

[9] T. Leichtle, C. Gei, M. Wurm, T. Lakes, and H. Taubenbck, "Unsupervised change detection in VHR remote sensing imagery – An object-based clustering approach in a dynamic urban environment," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 54, pp. 15–27, 2017.

[10] Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "An approach to multiple change detection in VHR optical images based on iterative clustering and adaptive thresholding," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1334–1338, Aug. 2019.

[11] B. Liu, K. Tang, and J. Liang, "A bottom-up/top-down hybrid algorithm for model-based building detection in single very high resolution SAR image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 6, pp. 926–930, Jun. 2017.

[12] P. Lukashevich, B. Zalessky, and A. Belotserkovsky, "Building detection on aerial and space images," in *Proc. Int. Conf. Inf. Digit. Technol.*, 2017.

[13] S. Jabari, M. Rezaee, F. Fathollahi, and Y. Zhang, "Multispectral change detection using multivariate Kullback-Leibler distance," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 163–177, 2019.

[14] Y. Zhang, D. Peng, and X. Huang, "Object-based change detection for VHR images based on multiscale uncertainty analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 13–17, Jan. 2018.

[15] L. Gu, S. Xu, and L. Zhu, "Detection of building changes in remote sensing images via flows-UNet," *Acta Autom. Sinica*, vol. 46, no. 6, pp. 1291–1300, 2020.

[16] H. Su, Y. Yu, Q. Du, and P. Du, "Ensemble learning for hyperspectral image classification using tangent collaborative representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3778–3790, Jun. 2020.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[18] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks," 2016, *arXiv: 1602.06564*.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Medical Image Comput. Computer-Assisted Intervention*, Cham, Switzerland: Spinger, 2015, pp. 234–241.

[20] 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," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, 2018.

[21] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla 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.

[22] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 91–105, 2019.

[23] W. Liu, M. Y. Yang, M. Xie, Z. Guo, and D. Wang, "Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2912.

[24] L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the united states," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[25] 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.

[26] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[27] H. Ji, X. Li, X. Wei, W. Liu, and L. Wang, "Mapping 10-m resolution rural settlements using multi-source remote sensing datasets with the Google Earth Engine platform," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2832.

[28] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[29] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learn. Medical Image Analysis Multimodal Learn. Clinical Decision Support*, Cham, Switzerland: Springer, 2018, pp. 3–11.

[30] R. Jaturapitpornchai, M. Matsuoka, N. Kanemoto, S. Kuzuoka, R. Ito, and R. Nakamura, "Newly built construction detection in SAR images using deep learning," *Remote Sens.*, vol. 11, no. 12, 2019, Art. no. 1444.

[31] Y. Liu, Z. Zhang, R. Zhong, D. Chen, and L. Sun, "Multilevel building detection framework in remote sensing images based on convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3688–3700, Oct. 2018.

[32] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2016, *arXiv:1511.07122*.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Honolulu, HI, USA, pp. 6230–6239, 2017.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

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

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2014.

[39] C. W. Xie, H. Y. Zhou, and J. Wu, "Vortex pooling: Improving context representation in semantic segmentation," 2018, *arXiv:1804.06242*.

[40] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learn. Data Labeling Medical Appl.*, Cham, Switzerland: Springer, 2016, pp. 179–187.

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

[42] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

Wei Liu received the M.S. degree in cartography and geographic information engineering from the China University of Mining and Technology, Xuzhou, China, in 2007, and the Ph.D. degree in cartography and geographic information engineering from the China University of Mining and Technology, Xuzhou, China, in 2010.

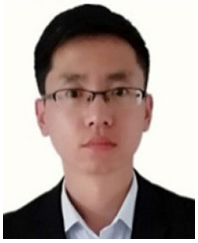
He is currently an Associate Professor with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include spatial data quality checking, high-resolution remote sensing image processing, and GIS development and applications.

Jiawei Xu received the B.S. degree in geographic information system from Nanjing Xiaozhuang University, Nanjing, China, in 2019. He is currently working toward the B.S. degree in the cartography and geographic information engineering with Jiangsu Normal University, Xuzhou, China.

His research interests include high-resolution remote sensing image processing and GIS development and applications.

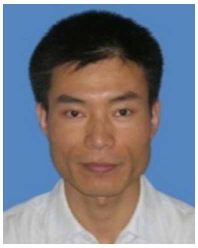
Zihui Guo received the B.S. degree in geodesy and survey engineering, in 2018 from Jiangsu Normal University, Xuzhou, China, where he is currently working toward the B.S. degree in the cartography and geographic information engineering.

His research interests include spatial data quality checking and high-resolution remote sensing image processing.



Erzhu Li received the M.S. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2014, and the Ph.D. degree in cartography and geographic information system from Nanjing University, Nanjing, China, in 2017.

He is currently a Lecturer and Researcher with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include high-resolution image processing and computer vision in urban remote sensing applications.

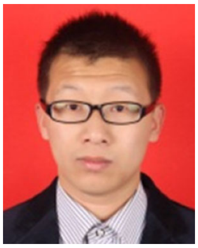


Xing Li received the M.S. degree in cartography and geographic information engineering from the Shandong University of Science and Technology, Qingdao, China, in 2004, and the Ph.D. degree in photogrammetry and remote sensing from East China Normal University, Shanghai, China, in 2009.

He is currently an Associate Professor with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include high-resolution image processing and computer vision in urban remote sensing applications.

Lianpeng Zhang received the M.S. degree in geodesy and survey engineering from the Shandong University of Science and Technology, Taian, China, in 1989, and the Ph.D. degree in photogrammetry and remote sensing from the Shandong University of Science and Technology, Qingdao, China, in 2003.

He is currently a Professor with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include high-resolution image processing and computer vision in urban remote sensing applications.



Wensong Liu received the M.S. degree in cartography and geographic information engineering from the China University of Petroleum, Qingdao, China, in 2016, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019.

He is currently a Lecturer and Researcher with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include PolSAR images processing and change detection.