ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

The Application of Deep Convolution Neural Network to Building Extraction in Remote Sensing Images

Weiyang[1, a] and Xiaofang Liu[1]

[1]College of Computer Science, Sichuan University of Science Engineering, Zigong, 643000, China.

[a]Email:18064064@qq.com

Abstract

The use of high-resolution remote sensing images to quickly and accurately detect urban building information is the current research focus. In this paper, aiming at the problems of small target loss, rough edge and poor semantic segmentation in the traditional algorithm of extracting buildings from high-resolution remote sensing images, an improved deep convolutional neural network based on U-Net is proposed to realize the end-to-end semantic segmentation at the pixel level. The model fusion strategy was adopted to improve the segmentation accuracy, and the mIoU in the data set reached 70.4%.

Keywords

Remote sensing image, deep learning full convolutional neural network, semantic segmentation, model fusion.

1. INTRODUCTION

As the main research content of remote sensing image application field, remote sensing image target recognition has important theoretical research significance and extensive application value. Taking high-resolution remote sensing images of urban areas as an example, some of the targets identified are artificial features such as buildings, roads and Bridges. Effective identification and accurate positioning of these targets has always been an urgent problem. Buildings are an important part of urban basic geographic information, so it is necessary to study the extraction method of buildings with high precision and high efficiency. With the development of deep learning theory, deep neural network model has been widely used in different industries and achieved good performance in computer vision tasks[1-2].People have made exploration in the combination of deep learning and remote sensing image application, verified the feasibility of using deep neural network, especially convolutional neural network, to process remote sensing data, and proposed a deep neural network model suitable for building extraction of high-resolution remote sensing image. With the rapid development of computer hardware, the deep learning method has shown strong performance in the classification and detection tasks in the field of image processing[3-4],in which convolutional neural network (CNN) is the representative and performs well in the image-level regression and classification tasks. However, building extraction is a semantic segmentation task at the pixel level, and the use of CNN will cause the memory overhead to rise sharply, the computational efficiency to be low, and the perception area to be limited[5].Accordingly, the Fully Convolutional Networks (FCN)[6]remove the Fully Convolutional layer in traditional CNN, and deconvolution the end feature map to generate segmentation results consistent with the resolution of the input image, thus realizing the classification at the pixel level and applying it to the target extraction in the field of remote sensing[7-9].However, as the depth of the network

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

increases, the feature dimension increases and the detail information is lost, which will lead to the insufficient accuracy of the extraction results. Subsequent examples include SegNet[10],DeconvNet [11],and U-Net[12].On the basis of FCN,U-Net convolutional neural network adopts symmetrical structure design, fusing features of low dimension and high dimension, and achieving higher extraction accuracy in the field of medical image segmentation. Liuzhe [13]et al. designed a CT image liver segmentation method based on the improved U-Net and Morphsnakes algorithm,and the segmentation results were enhanced. Based on the model, GuillaumeChhor[14]et al. used Adam Optimizer to replace the stochastic gradient descent algorithm, added batch standardized acceleration training, and used loss based on Dice coefficient to realize the extraction of building materials in remote sensing images. Wu guangming et al.[15]proposed an improved network with double constraints, which optimized the parameter updating process and improved the extraction accuracy of buildings. In the existing methods, neural network is used to perform semantic segmentation of remote sensing images. The segmentation results have some problems such as smooth edges and loss of detail information, which affect the accuracy to some extent.

Although these methods are better at processing natural images, they are not suitable for remote sensing images. Because high-resolution remote sensing images contain millions of building targets per image, and in the image the building structure is fine, these are different from natural images. However, we can realize the semantic segmentation of natural images, and combine the advantages of different network models to study a robust method for building segmentation of remote sensing images for complex scenes.

Although the results of building extraction based on the fusion of high-resolution remote sensing image and deep neural network are good, the research on improving the structure of deep neural network to improve the precision of building extraction still has a lot of room for development and further research is needed.

Based on the above analysis, this paper proposes an end-to-end deep product neural network model based on U-Net for building segmentation.The cross entropy loss function and joint loss function are redesigned to improve the accuracy of building segmentation.At the same time, the building segmentation method of remote sensing image of the two models is fused to extract higher precision.

2. RESEARCH METHOD

2.1. U-Net (End-to-End Full Convolution Network Model)

In order to represent the model in theory, S represents the remote sensing image, and M represents the GT (ground truth) image. When the pixels of the (i, j) position of the remote sensing image s are represented as buildings, $M(i,j) = 1$of the GT image at the same position corresponding to it. The goal of the model is to learn$p(M(i,j)|S)$from the sample data.

In high-resolution remote sensing images, an image often contains thousands of pixels, which makes the image take up a lot of storage space. Due to the limitation of computer hardware, only a whole image can be predicted by block, and then the predicted image can be restored by splicing. The direct prediction splicing will have obvious splicing marks, which will affect the segmentation accuracy of buildings. The way to solve this problem is to make predictions on overlapping image blocks. In this paper, a clipping layer is added to the model to solve this problem. The end-to-end model is based on this idea. That is:

$$p(N(M(i,j),\omega_m)|N(S(i,j),\omega_m)) \quad (1)$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

Where, $N(I(i,j),\omega)$represents an image block with (i, j) as the center and the size of$\omega * \omega$on the image I. Through the model prediction, the prediction with the size of$\omega_m * \omega_m(\omega_m < \omega_s)$is obtained. In addition, this solution can reduce the calculation time of the model.

2.2. Improved U-Net Architecture

The complexity of the basic U-Net architecture is not high, so using the basic model with low complexity can achieve high precision and achieve a balance between precision and complexity. The network architecture of U-Net, SeeFig. 1.

The remote sensing image data set used in this paper has complex background, rich feature types and wide spectrum. The basic U-Net network framework cannot effectively extract the pixel features of complex remote sensing images, so the depth of U-Net is Deep-UNet (D-Net) to extract more complex spectral features. The network architecture of D-Net , See Fig. 2.

The left half of the network is the lower sampling part, which is constructed according to VGG16, which is a typical structure in the convolutional neural network. By gradually reducing the spatial dimension of input data, high-dimensional features are extracted.The core is 5 groups of conv and MaxPooling, in which the first and second groups adopt two times of 3*3 convolution operation, with the number of convolution cores 64 and 128 respectively, and the third, fourth and fifth groups adopt three times of 3*3 convolution operation, with the number of convolution cores 256, 512 and 512 respectively. The BN(Batch Normalization) layer is added after each convolution operation to normalize the features of each layer of the network layer, so as to make the feature distribution of each layer more uniform and improve the fault tolerance of the model while speeding up the convergence speed of the model.

Left half and right half a center of symmetry, it is composed of a series of sampling layer, its core is with the sample for 5 groups Upsampling and conv, each group of input in addition to a layer of convUpsampling get deep abstraction features, and the corresponding sampling output of shallow layer under the local characteristics of the deep and shallow characteristics by skip-connection fusion, to restore the characteristic figure details and ensure its corresponding space information dimension unchanged.

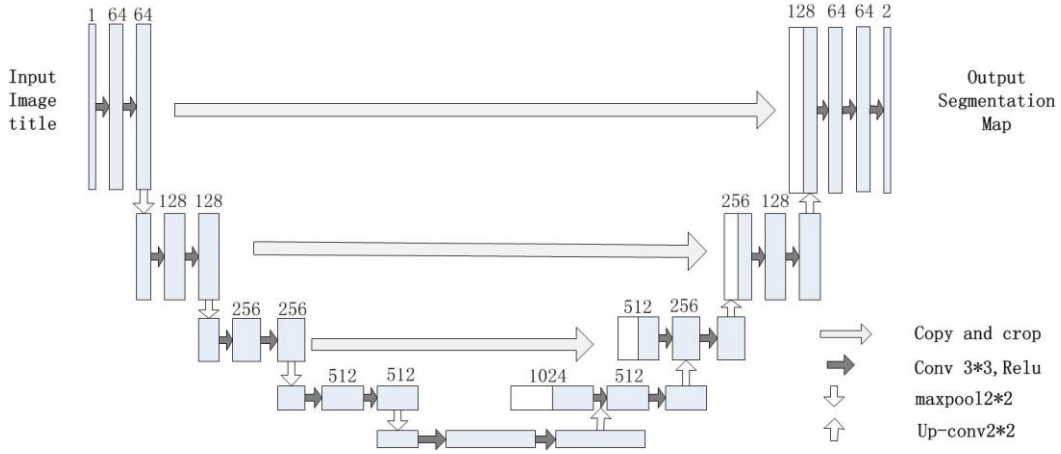


Figure 1. U-Net architecture

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

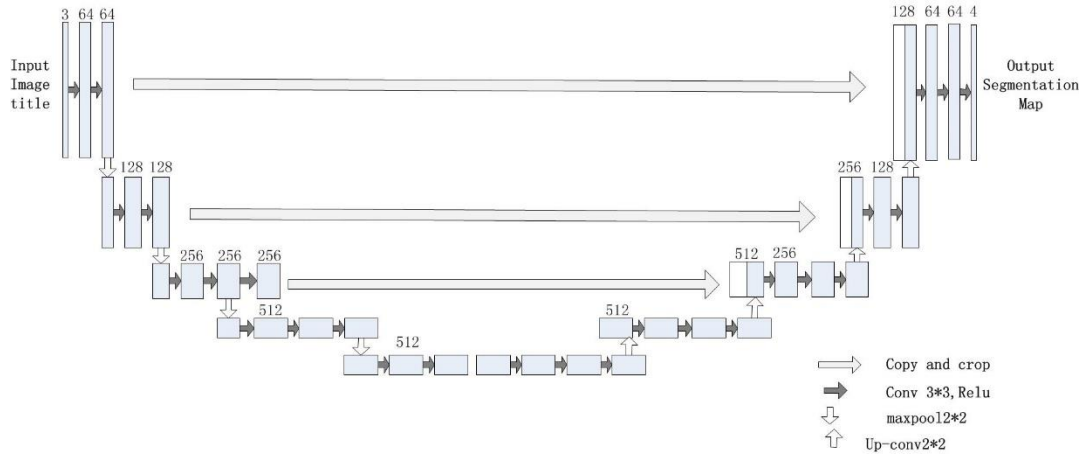


Figure 2. D-Net architecture

Specifically, the categories are expressed as$\delta = \{0,1 \ldots M\}(M=3)$, Set the input RGB remote sensing image and its gt to have N in total, and the set form is expressed as$\{(X_i, G_i)|X_i \in \delta^{H\times W\times 3}, G_i \in \delta^{H\times W} i = 0,1 \ldots, N\}$, Where$X_i$represents the input three channel RGB remote sensing image,$G_i$represents gt corresponding to RGB, H and W represent the height and width of remote sensing image respectively. The parameters of the network layer are$W = [w^{(1)}, w^{(2)}, \ldots w^{(L)}]$, Where l is the number of network layers, and each layer of the network is defined as$t^L(x, W^L)$, Then the model definition is as follows (2):

$$f(x, W) = t^L(t^{(L-1)}\left(\ldots t^{(2)}\left(t^{(1)}\left(x, w^{(1)}\right); w^{(2)}\right) \ldots; w^{(L-1)}\right); w^{(L)}) \tag{2}$$

Where the O-th componentf(x, w)is expressed as the score that pixel x belongs to category o. At the end of the network layer, softmax function is used to calculate the probability that the pixels of the input image belong to a certain category. The definition is as follows (3):

$$\rho(o|x, W) = \frac{\exp(f_o(x,W))}{\sum_{M=0}^{M}\exp(f_m(x,W))} \tag{3}$$

Finally, the multi classification function of visual loss is used to calculate the difference between the predicted value and the real value, and the most effective parameter W^*, which is defined as follows (4):

$$L\big(Y, \rho(Y|X)\big) = -log\rho(X|Y) = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{HW} X_{ij} log\rho(g_{ij}|X_{ij}, W) \tag{4}$$

2.3. Loss Function and Model Fusion

The loss function of the two models above adopts the Binary Cross Entropy function corresponding to the Sigmoid function, and converts the output probability value from the maximized probability to the minimized information Entropy through the negative logarithm. Its expression is as follows:

$$H = -\frac{1}{m}\sum_{i=1}^{m}\left[ylog\hat{y} + (1-y)\log\left(1-\hat{y}\right)\right] \tag{5}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

Where y represents the true value of the pixel value and $\hat{y}$ represents the predicted value of the pixel value. In addition, in order to make the task of building segmentation more accurate, mIOU formula is added to the binary cross entropy loss function to form the joint loss function. The loss function of the final model is as follows:

$$L = H - logJ_m \tag{6}$$

$$J_m\left(y, \hat{y}\right) = \frac{1}{m}\sum_{i=1}^{m} \frac{y_i \cdot \hat{y}_i}{y_i + \hat{y}_i - y_i \cdot \hat{y}_i} \tag{7}$$

Inspired by the model fusion method, the prediction results of the two models are fused by weighting, that is, to retain the segmentation extraction of large buildings by codec segmentation network, and to retain the accurate segmentation of U-Net segmentation network on small buildings. Combining Formula 1, the fusion formula of model prediction is as follows:

$$P = \alpha * P_e + \beta * P_u, \alpha + \beta = 1 \tag{8}$$

Among them,$P_e$is the prediction result of codec network; P_u is the prediction result of D-Net network. The model fusion process,See Fig.3. In the input image, take the given point as the center, extract the image blocks with the scale size of 112^2px and 128^2px, respectively, and send them to the two models for prediction. After the model operation, the prediction results with the size of 80^2px are obtained. Finally, the final segmentation mask image is obtained by weighting the prediction results according to the above formula. Experiments show that whenα = 0.45, β = 0.55, the fusion model can get the best results.

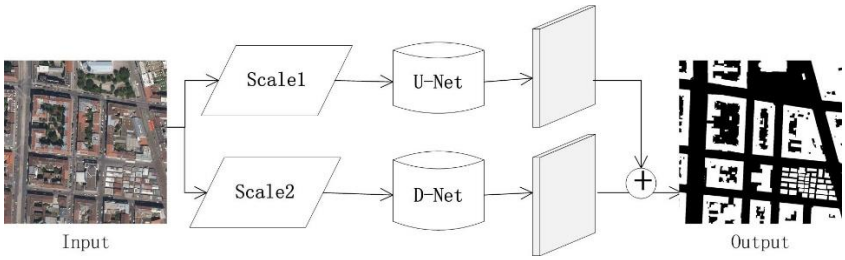


Figure 3. Model fusion

3. EXPERIMENT

3.1. Experimental Data

The paper selects the INRIA aerial image data set [22] released by the National Institute of information and automation of France in 2018 as the research data. The spatial resolution of the dataset is 0.3m, and there are 180 images covering 5 cities, each city contains 36 high-resolution remote sensing images. Due to the need to distinguish individual buildings, this paper selects 36 images over Austin area in the United States and different urban residents in Australia, and uses the method of random division, 31 of them are used for training, 5 images are used for testing.The pixel points of each image are 5000 × 5000, and the coverage is about 2.25 km^2. Considering the performance of the computer, the original image is cut to 128 × 128 pixels with 144 pixels as the step, and the samples without buildings are eliminated. 4526 training samples and 762 test samples are obtained. In order to generate the sample tag

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

corresponding to the instance segmentation, this paper uses the function in scikit image to generate the corresponding instance tag from the two class true value image, and assign different tag values to different buildings. The sample situation, See Fig. 4.

Figure 4. (a)original(b) gt (ground truth)

3.2. Evaluation Criteria

In order to evaluate the segmentation effect of this dataset, three commonly used evaluation indicators were used in this paper, namely mean Intersection Over Union (mIOU), Overall Accuracy (OA) and F1 Score.Suppose there are k different categories, let n_{ij} represent the number of pixels with actual category I but predicted result J, $t_i = \sum_j n_{ij}$ represent the total number of pixels with category i, $p_i = \sum_i p_{ij}$represent the total number of pixels with predicted result i.The calculation formulas are as follows:

$$mIOU = \frac{1}{k} \times \sum_i \frac{n_{ij}}{t_i + p_i - n_{ij}} \tag{9}$$

$$O\,A = \frac{\sum_i n_{ij}}{\sum_i t_i} \tag{10}$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

3.3. Network Model Construction

The two proposed models were studied using 180 high-resolution images in the Inria dataset and their corresponding semantic annotation maps.First, 144 image blocks with the size of 128 ×128 pixels were randomly cut out of 180 images, and the horizontal, vertical or mirror image of the images were flipped randomly. The image patches with the dimension of 144×128×128 were constituted as the input of the encoding and decoding segmentation network. In the process of one iteration, 5000 batches of the above dimensional image patches were used to learn network parameters by batch gradient descent, with a total of 25 iterations. For the U-Net segmentation network model, image patches with dimensions of 64×128×128 were used as model input during 25 iterations, and 8000 batches of image patches were used for training in each iteration. In order to eliminate other interference factors, all experiments in this paper adopt the same optimization algorithm and data amplification method. The training data set was flipped horizontally and mirrored with p=0.5.At the same time, 90° ,180° and 270° were used to rotate the training data with uniform probability. In the experiment, all models adopted Adam optimization algorithm, with a learning rate of 0.0004, a batch size of 2, training times of 200 and k iterations. The program in the experiment USES the Keras deep learning framework, with a graphics card of NVIDIA TITAN Xp and a CPU of Intel i7-8700k.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

3.4. Experimental Results and Analysis

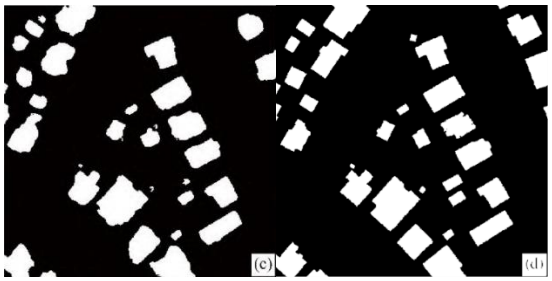
In the experiment, U-Net and the improved d-net based on U-Net proposed in this paper were trained respectively, and the performance of extracting buildings by two network dichotomies was compared. The results, SeeTab. 1.The performance of d-net proposed in this paper is better than that of U-Net, and the accuracy is improved by about 3.2%.

Table 1. Extraction accuracy of buildings with different network structures

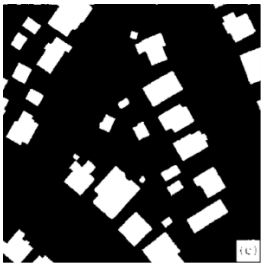
methods	Acc / %	F1 / %	MIoU / %
U-Net	90.38	83.32	63.92
D-net	93.64	84.86	68.34

After the model training, the sample results of the model in three different regions of the test set are given. The prediction results of the two models and the prediction results of model fusion are shown in figure 5.From top to bottom, the figure shows the results of RGB remote sensing image, Ground Truth, U-Net, D-Net network and model fusion. It can be seen that the U-Net network can better identify and extract the outline of large buildings, while the d-net network can better extract the edge details of buildings and the outline of small buildings because of the combination of high and low level feature maps. Model fusion combines the advantages of both. At the same time, the results showed false positive (identifying non-buildings as buildings) and false negative (not identifying buildings) for small areas.

(a)RGB remote sensing image (b)Ground Truth



(c)U-Net results (d) D-Net results



(e) Fusion results

Fig 5. Prediction results of the model on the test set

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

Results of the three models on the test set, SeeTab. 2.Table 2showing mIoU, Acc, and F1 by region. As can be seen from the table, the overall result of the fusion model on the test set is improved by nearly 2% compared with d-net, reaching 70%, that is, the area of the predicted area covers 70% of the area of the real area

Table 2. Evaluation results of the model on the test set

Model		Bellingh	Blooming	Innsbr	SanFran	EastTy	Over
U-Net	mIoU	67.96	49.80	67.88	58.79	75.23	63.9
	Acc/	92.87	91.44	90.20	87.77	89.63	90.3
	F1/%	82.70	83.10	86.70	76.90	87.20	83.3
D-Net	mIoU	69.78	57.67	69.93	68.76	75.55	68.3
	Acc/	96.43	93.09	94.44	89.03	95.21	93.6
	F1/%	86.20	86.70	84.20	78.90	88.30	84.8
Fusion model	mIoU	71.67	62.01	70.27	70.65	77.78	70.4
	Acc/	97.67	96.32	96.53	90.29	97.92	95.7
	F1/%	89.00	88.70	88.50	88.80	87.90	88.5

The paper, a building semantic segmentation model based on high-resolution remote sensing image is designed and implemented. The experimental results show that the image patch size of the input codec segmentation network is larger and contains more context details, so that the codec network can better extract the outline of large buildings. The D-Net network, which combines the image details at the high level, is more sensitive to the features of small buildings. The model fusion combines the advantages of the two, and can obtain more satisfactory semantic segmentation results in the building segmentation test set.

4. CONCLUSIONS

In this paper, U-Net convolutional neural network is applied to extract buildings from remote sensing images, and the accurate extraction of buildings in the target area is realized. In addition, when extracting buildings in complex urban areas, environmental information is easily confused with building information, resulting in poor extraction results. In view of the above problems, this paper proposes a feature enhanced d-net neural network, which enhances the low-dimensional detail information in the network transmission process and improves the model's ability to obtain the building details. Through training and testing on remote sensing image data sets covering 340 square kilometers, and comparing with other methods, experiments show that the fusion model proposed in this paper can achieve an average of 70.47% on mIoU, 95.74% on accuracy and 88.58% on F1 coefficient, which is better than U-Net and d-net methods. In addition, due to the existence of factors such as solar shadow, shielding and differences in the characteristics of buildings, the integrity of the extraction of buildings will be affected to some extent. It is not comprehensive to consider only the color or brightness characteristics of the pixel itself and its local areas. Therefore, in the future work, the shadow and occlusion of buildings in the image will be brought into the research scope to improve the extraction effect of buildings.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202003_6(3).0017

REFERENCES

[1] Szegedy C, Liu W, Jia Y,et al. Going deeper with convolutions // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 1–9

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. South Lake Tahoe, 2012: 1097–1105

[3] Russakovsky O, Deng J,Su H,et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[4] Li C X, Cao L, Zhang Y L,et al.Knowledge-based deep reinforcement learning: a review[J]. Systems Engineering and Electronics, 2017, 39(11): 217-227.

[5] LECUN Y L, BOTTOU L,BENGIO Y, et al.Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[6] LONG J, SHELHAMER E,DARRELL T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.

[7] Maggiori E,Tarabalka Y,Charpiat G, et al.Convolutional neural networks for large-scale remote-sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(2): 645-657.

[8] Marmanis D,Wegner J D,Galliani S, et al.Semantic segmentation of aerial images with an ensemble of CNNs[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016, 3: 473-480.

[9] Song Q S, Zhang C, Chen Y, et al. Road segementation using full convolutional neural networks with conditional random fiels[J]. Journal of Tsinghua University (Science and Technology),2018(8):725-731.

[10] Badrinarayanan V, Kendall A, Cipolla R.Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[11] Noh H, Hong S, Han B.Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision, Dec. 7-13, 2015, Santiago, Chile. New York: IEEE, 2015, 178: 1520-1528.

[12] Ronneberger O, Fischer P, Brox T.U-Net: Convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W, et al. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015.Cham: Springer, 234-241.

[13] Liu Z, Zhang X L, Song Y Q, et al.Liver segementation with improved U-Net and Morphsnakesalgorithm[J]. Journal of Image and Graphics, 2018,23(8):1254-1262.

[14] Guillaume C, Cristian B A, Ianis B L.Satellite image segmentation for building detection using U-Net. [EB/OL]. (2017). [2019]. https:// www.semanticscholar.org/paper/Satellite-Image-Segmentation-for-Building-Detection-Chhor -Aramburu/abb13964a435e1ac0c77b7dd68095e9da81b90aa.

[15] Wu G M,ChenQ,RyosukeS,et al.High precision building detection from aerial imagery using a U-Net like convolutional architecture[J]. ActaGeodaetica et CartographicaSinica, 2018,47(6):864-872.