ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1

Research on Building Extraction from Remote Sensing Images based on Deep Learning

Guokuan He [a, b], Yuhong Wang [a, *], Ping Nie [c]

[a] School of Surveying and Information Engineering, West Yunnan University of Applied Sciences, Dali 671006, Yunnan, China,

[b] China First Metallurgical Group CO., LTD, Wuhan 430070, Hubei, China,

[c] Guangxi Transportation Design Group Co., Ltd, Nanning 530029, Guangxi, China

* email: 511611478@qq.com

Abstract: In this study, the LoveDa dataset is used to carry out experiments. In the data preprocessing stage, the sample size of the dataset is effectively expanded through data enhancement and normalization operations such as rotating, flipping, etc., so that the model can learn more images of buildings with different shapes and angles during the training process. Subsequently, the U-Net network model is trained using the processed dataset, paying close attention to the model's accuracy, F1 value and other key indicators throughout the process, flexibly adjusting the model parameters and hyper-parameters according to the changes in the indicators, and preserving the model weights after the training is completed; finally, the model is tested through comparison experiments with other methods in order to evaluate its performance. The experimental results show that this method extracts buildings with an accuracy as high as 0.9428, a recall of 0.8104, and an F1 value of 0.8227, which improves the accuracy by 20% compared with the traditional method, and the recall and F1 value are also significantly improved, which proves that U-Net has a significant advantage in the extraction of buildings from remote sensing images, and powerfully supports the urban planning and management work.

Keywords: Deep learning; buildings; remote sensing imagery; U-net network model; feature extraction.

1. Introduction

1.1 Research Background and Significance

Building extraction is an important application field in remote sensing technology, and remote sensing technology to extract the feature information of target features aims to identify, locate and extract the geospatial information of buildings in remote sensing images. As one of the main components of the urban surface, the automatic extraction of buildings has a wide range of application prospects in the fields of urban planning, urban management, disaster monitoring, etc.

Deep Learning (DL)is a new research direction in the field of computer vision, mainly by learning the intrinsic laws and representation levels of sample data [1], so that the computer can have the human-like analyzing and learning ability. 2014 by Jonathan Long et al. was the first to propose the application of convolutional neural networks to image segmentation in the image, which breaks through the restrictions limited by the size of the original image [5], and for the first time to complete the end-to-end learning approach and for the pixel-level images to begin to try to predict that the full convolutional neural network (Fully Connected Network (FCN), FCN's most important feature is that the fully connected layer is replaced by a convolutional layer, thus occupying one of the classical models in image segmentation tasks.

1.2 Current Status of Domestic and International Research

1.2.1 Building Extraction Based on Deep Learning

In recent years, the rapid development of the field of computer vision, deep learning in remote sensing image processing has made substantial progress, especially the research and application of

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1
convolutional neural network (CNN) and deep learning network structure (such as SegNet, U-Net, HRNet, Mask R-CNN, etc.). Foreign scholars started to explore the application of deep learning in remote sensing image processing earlier, and they developed innovative building extraction algorithms, which can not only accurately identify the building area, but also finely extract the contour, which lays a solid foundation for applications such as building information extraction and urban planning.

1.3 Research Content and Technical Route

1.3.1 Research Content

The main research content of this thesis is the study of remote sensing image building extraction based on deep learning, which accomplishes end-to-end feature extraction of remote sensing image buildings by using the architecture of deep learning (DL).To address these challenges, this paper proposes a building extraction method, U-Net network model, which needs to be comprehensively considered and improved in three aspects: data preprocessing, network model training, and model testing.

1.3.2 Technical Approach

U-Net network model is a kind of convolutional neural network (CNN) architecture commonly used in image segmentation, through the encoder-decoder structure, to realize the extraction of remote sensing image features and spatial details of the information retained. the U-Net network model technical route, as shown in Figure 1-1.

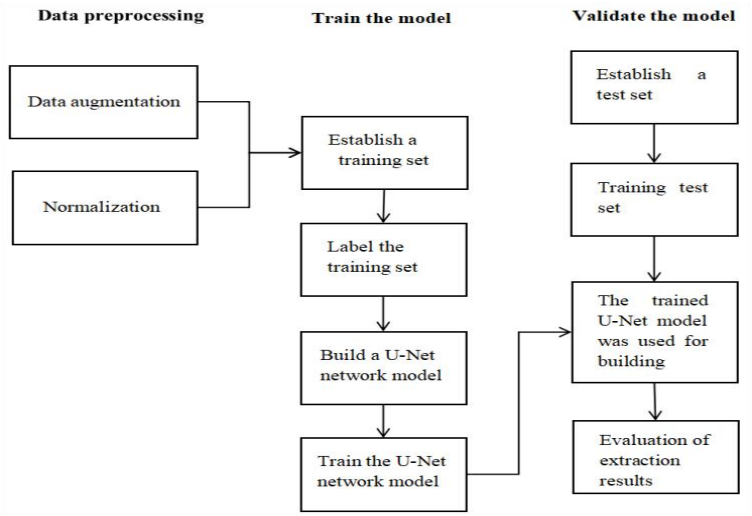


Figure 1-1. Technical roadmap flowchart

2. Data Processing and Experiments

2.1 Data Processing

2.1.1 Data Set

This paper adopts the open source LoveDA building dataset released in NeurIPS 2021 as the data source for this experiment, download address: https://pan.baidu.com/share/init?surl=YrU1Y4Y0dS0f_OOHXpzspQ, the dataset comes from Google Earth, the collection Time is July 2016, the whole image covers an area of about 536.15 square kilometers, the number of remote sensing images has 5987 0.3m high-resolution images, 600 images each of Nanjing, Changzhou and Wuhan are selected to do the experiment, and other information is shown in Table 2-1.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1

Table 2-1. Details of the LoveDA dataset

Key	Value
Coverage Area	Nanjing, Changzhou, Wuhan
Number of Remote Sensing Images	1800
Resolution	0.3m
Single Image Size	1024*1024
Original Image Bit Depth	8-bit
Label Image Bit Depth	8-bit
Number of Original Image Channels	Three Channels
Number of Label Image Channels	Two Channels

2.1.2 Data Preprocessing

A series of preprocessing measures are adopted before inputting the original image data into the model for training, including data enhancement (rotating, flipping) and normalization (unifying the pixel values into the same interval), which mainly aim at eliminating the unfavorable factors affecting the classification of features, optimizing the quality of the images, and enhancing the key features of the images, which not only improve the generalization ability of the model, but also enhance the accuracy and stability of the building extraction.

Data enhancement is a technique that generates new data by performing a series of geometric transformations on the original data, as shown in Figure 2-1.

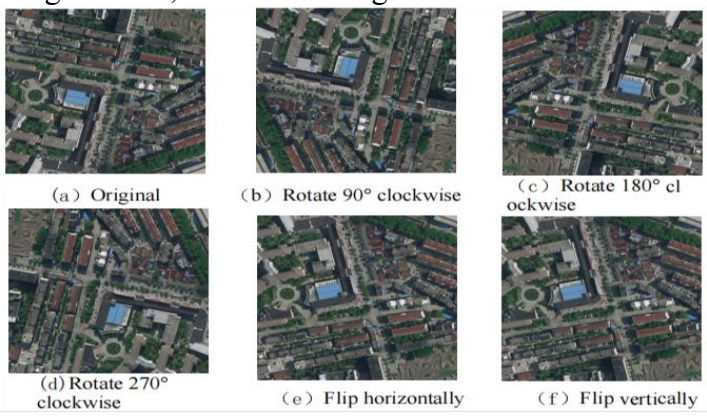


Figure 2-1. Schematic diagram of data augmentation effect

There are various methods of data normalization, the most commonly used are linear normalization (Min-Max Normalization) and standardization (Standardization), the main purpose of which is to normalize the dimensional features to the same interval of values, so as to improve the convergence speed of the model and the training efficiency. One of the linear normalization (Min-Max Normalization) is to center the data according to the minimum value, and then according to the polar deviation (maximum value - minimum value) for scaling, and finally the data will converge to the range of [0,1], the normalization formula is:

$$x = \frac{x - min(x)}{max(x) - min(x)} \tag{2.1}$$

Standardization, on the other hand, is the process of scaling data so that it falls into a small specific interval, transforming it into a distribution with a mean of 0 and a standard deviation of 1. The standardization formula is:

$$x = \frac{x - \mu}{\sigma} \tag{2.2}$$

Which is the original data, is the mean of the data set, is the standard deviation of the data set, the choice of which normalization method depends on the specific characteristics of the data and the needs of the model, will not change the distribution of the data, but only a linear dimensionless processing of the data.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1

2.1.3 Creating a Label Library

LabelMe is a graphical image annotation tool written in the Python programming language that enables human-computer interaction and launches the software from the command line to accomplish label production. It allows the user to create labels for tasks such as image segmentation, target detection, etc. through a simple graphical interface with more features including: annotation of input image data in the form of polygons, rectangles, circles, polylines and points. For example, Figure 2-2 shows some of the building labeled images after the original image is labeled by LabelMe.



Figure 2-2. Partial original images and labeled images

2.2 Experiments

2.2.1 Model Construction

The whole network architecture presents a symmetrical shape, like the letter "U" as shown in Fig. 2-3, the encoder part forms the left side of the U, the decoder part forms the right side of the U, and the middle Skip-connect connects the two sides, which is in the shape of a U, and thus the name U-Net is given to it.

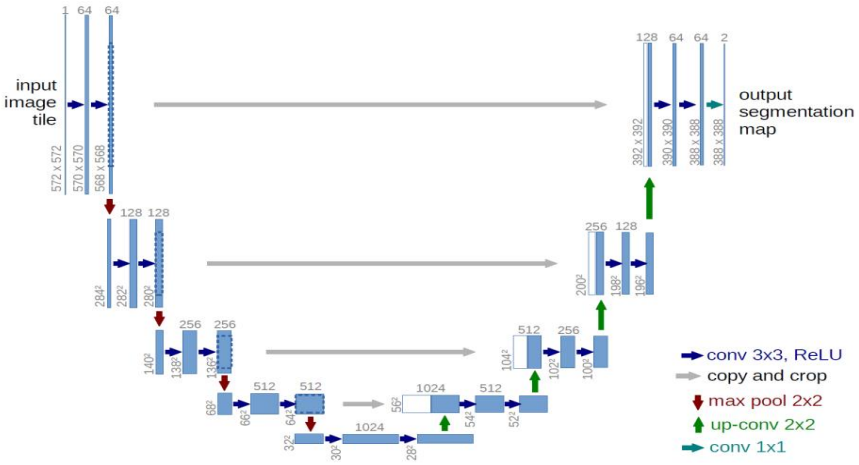


Figure 2-3. Schematic diagram of U-Net network model architecture

2.2.2 Experimental Environment

The experimental environment selected for this paper is based on the deep learning framework, the programming language is Python, the program editor is PyTorch, the operating system is Windows, the graphics card used for training is NVIDIA GeForce GTX 1650 Ti, and the memory size of the graphics card is 8G.

2.2.3 Model Parameter Configuration

The U-Net network model training process is heavily based on PyTorch as a framework, and the model training uses the back-propagation algorithm to update the weights and biases of the model. During the training of the model, the cross-entropy loss function is used to measure the difference between the probability distribution of the model output and the true labels, and the optimizer iteratively updates the parameters to minimize the loss, and the parameter configurations are shown in Table 2-2 in order to achieve more accurate predictions.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1

Table 2-2. Model training parameters

Training Parameters	Parameters
Initial Learning Rate	0.001
Batch	200
Batch Size	16
Convolution Kernel	3 * 3
Optimizer	Adam
Activation Function	ReLU

2.2.4 Evaluation Metrics

Based on the building extraction results of the U-Net network model for remote sensing images, the trained network is evaluated using the validation set, and the metrics, such as Recall, Precision, F1-Measure, Overall Accuracy, OA, and Kappa Coefficient), etc. to evaluate the performance and generalization ability of the model. The formula for calculating the specific indexes is as follows:

Recall is the proportion of correctly predicted as buildings to the proportion of total buildings, and the calculation formula is:

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

Precision is for the prediction outcome, which indicates the proportion of all true positive samples that are predicted as positive by the model, and is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

F1-Measure is a weighted harmonic average of Recall and Precision, which combines the performance of these two metrics and is calculated as:

$$F1_Measure = 2 * \frac{Precision * Recall}{(precision + Recall)} \tag{2.5}$$

Where, TP denotes the number of features where the network model predicts the result to be a building and actually is also a building, FP denotes the number of features where the network model predicts a building but is actually a background, and FN denotes the number of features where the prediction is a background but is actually a building.

Overall Accuracy (OA) is a key indicator of the performance of the classification algorithm, which evaluates the overall classification effect of the model by comparing the ratio of the number of correctly classified pixels to the total number of pixels to be classified, and is calculated as:

$$OA = \frac{a + e + i}{a + b + c + d + e + f + g + h + i} \tag{2.6}$$

Kappa Coefficient (Kappa Coefficient) is a statistical tool for evaluating the performance of a classifier to exclude the effect of random consistency by comparing the actual and expected accuracies, with a range of values from -1 to 1, but usually between 0 and 1, and is computed as:

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{2.7}$$

$$P_e = \frac{(a+d+g)*(a+b+c)*(b+e+h)*(d+e+f)*(c+f+i)*(g+h+i)}{(a+b+c+d+e+f+g+h+i)^2} \tag{2.8}$$

Depending on the magnitude of the Kappa coefficient, the consistency level of the classifier can be categorized into five classes: from 0.00 to 0.20 as very low, from 0.21 to 0.40 as average, from 0.41 to 0.60 as moderate, from 0.61 to 0.80 as high, and from 0.81 to 1 as almost perfect.

3. Experimental results and analysis

3.1 Experimental Results

This study conducted in-depth training experiments on the U-Net network model based on the LoveDA dataset and obtained significant results. Analyzing the experimental data, it is found that as

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1
the number of iterations of the U-Net network model increases, its performance tends to stabilize and the accuracy rate shows a steady upward trend, and the accuracy index is shown in Table 3-1. Its model extraction precision is 94.28%, which means that the model can accurately recognize and extract the key information in the image; the recall rate is 81.04%, which indicates that it can effectively cover most of the real instances when detecting the target object; and the F1 value is 82.27%, which shows that the model has achieved a good balance between precision and recall, which can accurately recognize the target object and reduce the number of misclassification and missed detections.

Table 3-1. Evaluation table of method accuracy index in this paper

Extraction Method	Precision	Recall	F1-Measure
U-Net	0.9428	0.8104	0.8227
Changzhou	0.9314	0.8009	0.8161
Nanjing	0.9467	0.8092	0.8219
Wuhan	0.9503	0.8211	0.8301

0The buildings of remote sensing images are extracted based on the U-Net network model, as shown in Figure 3-1, and the results show that the U-Net network model has good adaptability and generalization ability on LoveDA dataset, in which most of the features of the buildings can be recognized and extracted, which is a good interpretation of the superiority of the deep learning technology.

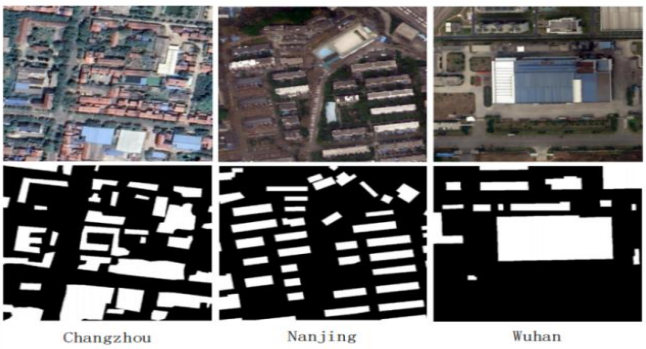


Figure 3-1. Schematic diagram of the extraction effect of the method in this article

The evaluation indexes of other methods for extracting buildings are shown in Table 3-2, and a conclusion can be drawn from the indexes: the overall classification accuracy of Support Vector Machine (SVM) with Kappa coefficients is better than that of the neural network extraction method, but the overall accuracy is still to be improved compared with the method in this paper.

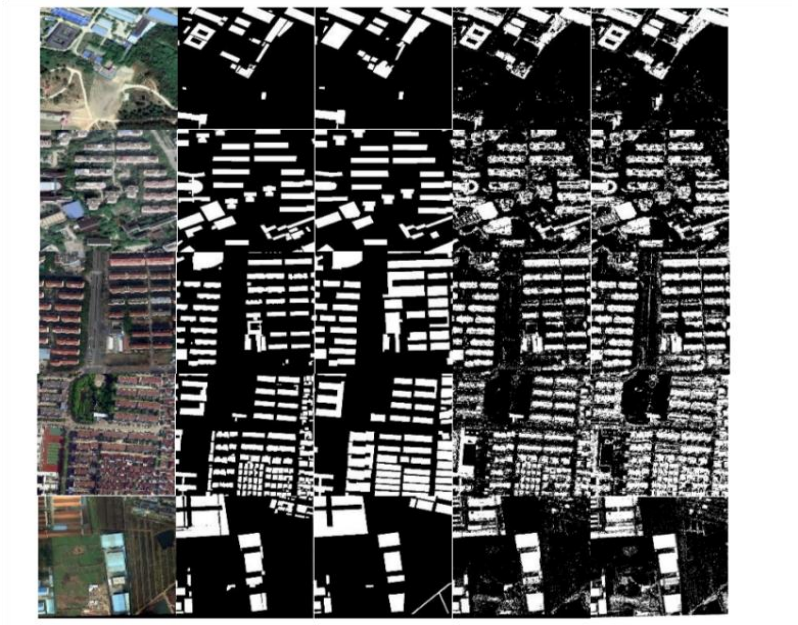
Table 3-2. Evaluation of accuracy indexes of other extraction methods

Extraction Method	Overall Accuracy	Kappa Coefficient
Neural Network Method	74.6903%	0.6965
Changzhou	74.6589%	0.6642
Nanjing	74.7010%	0.7107
Wuhan	74.7110%	0.7146
Support Vector Machine	86.7111%	0.7885
Changzhou	86.7106%	0.7861
Nanjing	86.7108%	0.7895
Wuhan	86.7119%	0.7899

3.2 Comparative Analysis of Extraction Results of Different Methods

As seen in Figure 3-2, the method of this paper is compared with the other two methods, and it is concluded that the different methods are able to respond to the scope of the building outline, but the extraction effect of neural network and support vector machine (SVM) on the building can only

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1
detect the approximate location and rough boundaries, and can not be processed in a refined manner, and there is redundancy and lack of data, i.e., there is a misclassification of the building and omission of detection. The U-Net network model proposed in this paper extracts the contour of the building better, although there are also a few misclassifications and omissions, but compared with the other two methods, the extraction effect is obviously larger and effective. The corresponding conclusion is that the U-Net network model has a better effect on the building extraction of remote sensing images.



（a）Original image （b）Real labels （c）Methods of this article （d）Neural networks （e）SVM

Figure 3-2. Different methods for extracting building comparison maps

4. Summary

By using the open-source LoveDA building dataset as the data for this experiment, a study was conducted on the extraction of buildings in remote sensing images. This paper is based on the U-Net semantic segmentation network model of deep learning technology to realize end-to-end image segmentation technology, and the buildings in remote sensing images are effectively extracted. The research conclusions of this thesis are as follows:

(1) The U-Net network model is able to accurately classify the features in remote sensing images and effectively extract the features of the target features through the learning of a large number of sample labels, thus realizing the recognition and extraction of buildings.

(2) When dealing with large-scale remote sensing image segmentation tasks, the U-Net model shows significant performance advantages. Specifically its encoder is able to deeply excavate the advanced and abstract features in remote sensing images; In addition, the jump connection mechanism in the model can effectively fuse the low-level features in the encoder part with the high-level features in the decoder part, thus significantly improving the accuracy of the segmentation results.

(3) The results show that the model achieves excellent performance in the building extraction task, with its accuracy, recall and F1 value reaching 0.9428, 0.8104 and 0.8227, respectively. Compared with the traditional building extraction methods, the U-Net network model not only improves the completeness of the extracted building contours, but also improves the accuracy by more than 20%, and also improves the accuracy, recall and F1 value. rate and F1 value are also improved. These experimental data fully demonstrate the superiority and effectiveness of the U-Net network model in building extraction from remote sensing images.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.6.1.P1

References

[1] Hai Songhao. Rapid Detection and Recognition of Traffic Signs Based on Convolutional Neural Networks [D]. Harbin University of Science and Technology, 2021. DOI: 10.27063/d.cnki.ghlgu.2020.001110.

[2] Fully Convolutional Networks for Semantic Segmentation.[J]. Shelhamer Evan;;Long Jonathan;;Darrell Trevor. IEEE transactions on pattern analysis and machine intelligence,2017

[3] Lv Daoshuang, Lin Na, Feng Lirong, et al. Building Extraction from High-Resolution Remote Sensing Images Based on an Improved U-Net Network [J]. Geospatial Information, 2021, 19(01): 30-34+4.

[4] Liu Yifan. Research on the Method of Building Extraction from High-Resolution Remote Sensing Images Based on Convolutional Neural Networks [D]. China University of Mining and Technology, 2020. DOI: 10.27623/d.cnki.gzkyu.2020.000296.

[5] Zhang Han. Research on Image Semantic Segmentation Methods Based on Attention Mechanism [D]. Anhui University, 2021. DOI: 10.26917/d.cnki.ganhu.2021.000304.

[6] Hao Xuzheng. Research on Pedestrian Detection and Recognition Based on Convolutional Neural Networks [D]. Tianjin Polytechnic University, 2018.

[7] "Deep Learning for Urban Building Detection Using Remote Sensing Data" by Wang et al. (2019).[8] Zhang Jiaxin. Research on ResNet Image Classification Model Based on Tensor Composition Attention [D]. Liaoning Technical University, 2022. DOI: 10.27210/d.cnki.glnju.2022.000331.

[8] Zhang Jiaxin. Research on ResNet Image Classification Model Based on Tensor Composition Attention [D]. Liaoning Technical University, 2022. DOI: 10.27210/d.cnki.glnju.2022.000331.

[9] Chen Yunying. Research on Building Extraction from High-Resolution Remote Sensing Images Based on LTHNet [D]. Jiangxi University of Science and Technology, 2023. DOI: 10.27176/d.cnki.gnfyc.2023.000750.

[10] Zheng Yi. Research on Cloud Detection Algorithm Based on Multi-scale Semantic Segmentation of Satellite Remote Sensing Images [D]. Harbin Institute of Technology, 2022. DOI: 10.27061/d.cnki.ghgdu.2022.001109.

[11] Huang Xin. Research on Scene Semantic Segmentation Methods Based on Attention Mechanism [D]. University of Electronic Science and Technology of China, 2021. DOI: 10.27005/d.cnki.gdzku.2021.001680.

[12] Yuan Peng, Wang Ke, Xiao Jian. Cultivated Land Extraction from High-Resolution Images Based on RMAU-Net Network Model [J]. Hubei Agricultural Sciences, 2023, 62(08): 182-188+196. DOI: 10.14088/j.cnki.issn0439-8114.2023.08.029.

[13] Hinton, Geoffrey E, Osindero, Simon, Teh, Yee-Whye. A fast learning algorithm for deep belief nets.[J].Neural computation,2006,18(7)

[14] "Deep Learning for Urban Building Detection Using Remote Sensing Data" by Wang et al. (2019).

[15] Liang Hongyu. Research on Image Semantic Segmentation Methods Based on Contextual Feature Fusion [D]. Harbin University of Commerce, 2022. DOI: 10.27787/d.cnki.ghrbs.2022.000314.

[16] "A Deep Learning Approach to Urban Building Extraction from Remotely Sensed Data" by Wang et al. (2017).