Complex & Intelligent Systems (2022) 8:1847–1855
https://doi.org/10.1007/s40747-021-00322-z

ORIGINAL ARTICLE

Remote sensing image building detection method based on Mask R-CNN

Qinzhe Han[1] · Qian Yin[1] · Xin Zheng[1] · Ziyi Chen[1]

Received: 3 December 2020 / Accepted: 1 March 2021 / Published online: 27 March 2021

Abstract
Quickly and conveniently identifying buildings in disaster areas plays an important role in disaster assessment. To achieve the technical requirements of flood disaster relief projects, this paper proposes a building extraction method for use with remote sensing images that combines traditional digital image processing methods and convolution neural networks. First, the threshold segmentation method is used to select and construct a training dataset. Then, a variety of preprocessing methods are used to enhance the selected dataset. Finally, the improved Mask R-CNN algorithm is used to detect buildings in the images. Experiments show that compared to the R-CNN algorithm, the proposed method improves detection accuracy and reduces the computational time.

Keywords Target detection · Remote sensing image · Deep learning · Mask R-CNN

Introduction

Buildings are an important class of artificial objects in high-resolution remote sensing images. Detecting the number and shape of buildings in remote sensing images plays an important role in decision-making issues such as land planning, mapping, and post-disaster reconstruction. Particularly in earthquake prevention and disaster mitigation, the number of remaining buildings and the degree of damage are important indicators. An indicator of the number of remaining buildings can be linked to multiple indicators, such as the number of deaths, the number of affected households, and the degree of property loss. The degree of damage is also of great significance for post-disaster reconstruction. Therefore, for earthquake prevention and disaster reduction, the recognition of buildings in remote sensing images of disaster areas is of great significance.

With the recent and rapid development of satellite remote sensing technology, remote sensing images are becoming more accurate. In a high-precision image, the difference in the texture features inside a building is becoming increasingly obvious; thus, it is more difficult to identify building features using artificial selection. The traditional classification method based on the texture features of ground objects is becoming increasingly feasible in high-resolution remote sensing images. Additionally, because the edges of buildings are often adjacent to other features, misclassification will directly lead to a loss in consistency between the final extraction results of buildings and the original buildings. Therefore, further research on building detection methods is of great significance to improve the accuracy of building extraction results.

The primary difficulties of automatic building identification are as follows:

1. As a data source, high-resolution remote sensing images are different due to a series of imaging conditions, such as scale and spectral range, and the universality of this method is poor;
2. There is a large different in the structures of buildings, and the same type of buildings (e.g., residential) will

✉ Qian Yin
zhengxin@bnu.edu.cn

✉ Xin Zheng
yinqian@bnu.edu.cn

Qinzhe Han
201821210036@bnu.edu.cn

Ziyi Chen
chenziyi@bnu.edu.cn

[1] Image Processing and Pattern Recognition Laboratory, School of Artificial Intelligence, Beijing Normal University, Beijing, China

exhibit different texture characteristics and geometric characteristics in different areas and environments;

3. Buildings tend to obscure each other visually, and shadows and other interference factors in images also yield unexpected errors in the building modeling process.

Therefore, since the advent of remote sensing technology, many scientists and researchers have been trying to improve and explore new methods to improve the accuracy and speed of remote sensing image automatic classification algorithms. The traditional building detection methods typically include extracting buildings by combining the spectral and shape features of buildings using multiscale feature fusion, object-oriented extraction, etc. Traditional target detection algorithms have achieved good results in certain datasets; however, due to the marked changes in the texture features of remote sensing images in real geographical environments, and the characteristics of buildings in different regions also changing greatly, there are few general solutions.

Recently, with the rapid development of deep learning and artificial intelligence, deep convolution neural networks have made great progress in the field of semantic segmentation and target recognition. Compared to the traditional feature classification algorithm, the deep neural convolution network does not require manually designed features; however, through the deep learning method, the computer can automatically learn and capture features for the current data set. Therefore, deep learning methods tend to produce more stable results for remote sensing image recognition with different accuracies, stronger universality and faster calculation speeds.

Related work

The primary idea of the traditional remote sensing image classification method is to use spectral information, texture information, spatial correlation and other information of the target object to determine the class attributes of ground objects. During image target recognition, a series of factors, such as spectral information, texture information, and spatial shape information, are considered comprehensively, and the object-oriented idea is used to achieve the purpose of classification and target extraction. Wu et al. primarily used the traditional remote sensing image feature extraction method and subtly combined the spectral characteristics and shape features of buildings to extract buildings from remote sensing images [1]. Huang et al. further used the multi-feature fusion method to extract features of buildings in remote sensing images and finally chose to use the SVM classifier to classify ground objects [2]. Fang et al. used the spatial position relationship between shadows and buildings in densely built areas, and used the graph cutting algorithm to accurately plot the outline of buildings [3]. Among foreign researchers, Wang et al. used the object-oriented method and used the pixel-based maximum likelihood classification method to process the image and achieved good results [4].

Although traditional building extraction methods perform well with certain data sets [5, 6], they have poor universality, and their results are different between two different data sets. Traditional classification methods typically do not make full use of the texture, shape and other features of buildings, and the ability to model the spatial relationship between buildings and their backgrounds is insufficient; thus, mistakes and omissions are common. Concurrently, considering the different characteristics of different buildings under the category of buildings and even the distinct differences between regions, the stability of the traditional extraction algorithm is poor, and generalizability is typically poor; thus, the traditional extraction method cannot meet the requirements of the proposed experiment.

The proposal of deep learning in 2006 is a milestone in the field of artificial intelligence. The feature extraction of buildings in remote sensing images often produces different optimal feature combinations with different regions and houses. The deep learning method does not require the algorithm writer to extract features, but they must provide a large-scale remote sensing image data set, which can complete the model construction, and then use the model to extract buildings. Compared to traditional methods, the entire process achieves a faster operating speed, better accuracy, and better building contour extraction without post-processing. Therefore, most current building extraction algorithms use deep learning methods [7].

In current image classification and retrieval, deep convolution neural networks are the most commonly used machine learning models. During target recognition and extraction, a convolutional neural network can automatically learn target features from data and fuse different levels of features. Therefore, the final model generated by the convolutional neural network has a strong ability to learn features. Additionally, convolutional neural networks can also directly describe the end-to-end automatic feature extraction process, which prevents a shortage of manually designed features and improves the target detection speed.

In the field of target detection, region-based CNN (R-CNN) can be used to extract the convolution features of regions and make localization and classification by region more robust, leading to better segmentation. Girshick used the idea of He Kaiming's SPP_Net [8] for reference and improved the R-CNN by adopting the method of physical-like sampling mapping, which markedly improved the model's detection speed; this method is called Fast R-CNN [9]. In 2015, Girshick cooperated with He Kaiming and others to continue to improve Fast R-CNN [10] by inserting a regional recommendation network to share all convolution layers with fast R-CNN, which is called Faster R-CNN. The Mask R-CNN [11] proposed in the same year is extended based on Faster R-CNN, and a new branch for predicting the object mask is added to the bounding box recognition branch in parallel. Based on this ingenious design, the target in an image can be detected, and a high-quality segmentation result can also be produced for each target.

With the rapid development of deep learning models, the field of remote sensing image detection and plotting has also improved. Liu et al. combined a traditional CNN network and an FCN network to detect targets with small areas and fragmented distributions in remote sensing images with high accuracy [12]. Fan et al. modified the design of the pooling layer in the convolutional neural network, reducing the loss of characteristic image information [13]. Dong et al. focused on real-time detection and the difficulty of detecting small buildings, and improved the YOLOv3 algorithm to strengthen the detection ability [14].

Based on these analyses, this paper designs a Mask R-CNN network-based full convolution neural network algorithm of the R-CNN network that is based on the FCN structure. First, the convolution neural network is used to extract the building features of remote sensing images, and then deconvolution is used to reconstruct the features. Combined with the real requirements of the disaster reduction process, the number and damage of buildings are evaluated. Concurrently, the building object detection image and semantic segmentation results of a single building are obtained. Finally, a building target detection experiment is performed in a real disaster area, and a high detection accuracy is achieved.

Research method

Preprocessing

The essence of deep learning is to establish a multilevel machine learning model and learn a large amount of training data to obtain better image features, and then improve the accuracy of classification or prediction. The quality of the image itself is important for the training of the neural network, which is related to the quality of the entire network.

The experimental objects of this experiment are primarily rural areas in China with frequent occurrences of mountain floods and other geological disasters. Buildings in this study are often distributed in blocks along the river, the distance between each building is small, the aggregation is strong, and the mutual shielding is serious. Therefore, it is necessary to preprocess remote sensing images. Next, the preprocessing of remote sensing images is introduced in three parts.

Threshold segmentation and fuzzy clustering are used to screen the training data

Because no comprehensive remote sensing image building dataset currently exists in China, we must use algorithms to filter data to select a suitable training dataset. Typically, gray information is one of the most important features in digital images. Therefore, in traditional target detection experiments, threshold segmentation is the simplest and most basic experimental method. The remote sensing images used in this experiment primarily belong to flood-prone areas, and the primary bodies are villages built near water in mountainous areas. In this type of remote sensing image, there are many types of ground objects, such as farmland, vegetation, roads, buildings, and water bodies. Due to the different materials used in different types of objects, different gray values often appear in the image. Based on the peak value of the gray histogram, we can roughly distinguish different types of ground object features.

The idea of the threshold segmentation algorithm is simple. Typically, after the digital image is processed for gray values, a certain threshold is calculated based on the gray histogram, and then the gray level is redivided based on this threshold. There are many different variations in the real application of this method. For example, in the design of how to distinguish the gray threshold, we can learn from the concept of potential energy in physics and consider positive and negative samples as positive and negative charges; we, thus, construct an appropriate potential function to fit the peak value of the gray level in the histogram and use the intersection point of two adjacent peaks as the threshold to segment the histogram. This method can improve the accuracy of the experimental results, although the mathematical expression is complex.

When the threshold segmentation method is used to screen the images in the data set, the building texture of certain images appears adequate to the naked eye; however, the threshold segmentation algorithm cannot distinguish buildings accurately. After many experiments, regions of a remote sensing image that are identified as “buildings” are often multiple buildings with different textures. The threshold segmentation algorithm will cause certain buildings to be undetectable. Therefore, the fuzzy clustering method [15] is also used in this section to screen the buildings.

Fuzzy clustering is a common clustering method that aims to determine how many types of samples should be classified in advance and then iteratively performs this classification based on the optimal principle until the classification is reasonable. This method is not commonly applied to remote sensing images because fuzzy clustering is computationally intensive compared to the traditional K-means clustering method. In this experiment, fuzzy clustering is only used as a supplement when threshold segmentation does not perform well. In most images, the texture features of buildings are not different.

With certain images that are easily recognized by the naked eye, but the threshold segmentation results are not ideal, the experiment will assume that there are two buildings with different characteristics in the image and then use the fuzzy clustering algorithm to cluster them again. If the clustering result is improved compared to the threshold segmentation, then the clustering result is added to the training data set. In this part of the experiment, the effectiveness and integrity of the training data space can be guaranteed.

Removing shadows

In high-precision remote sensing images, shadows are common. As one of the important factors that affect the amount of image data information, shadows are an important feature variable in images. However, during this target recognition experiment, considering that the buildings in rural areas are typically gray, the shadow areas in remote sensing images will tend to be confused with buildings with a low gray value, which increases the error rate of target recognition.

To reduce errors and improve accuracy, we use a shadow removal method to transform the image covered by shadows with less information into a high-quality image with a large amount of information, which is of great significance for the proposed deep learning method.

Shadows are caused by light being blocked; thus, the brightness value of shadows primarily comes from reflection or scattering of other geographical factors. In a high-precision remote sensing image, the brightness value of shadows is lower than that of other areas; however, the shadow area has no effect on the texture features of the surface objects, and its color is typically uniform and regular.

Based on the characteristics of these shadow areas, the retinex algorithm in the HSV color gamut space is used to remove shadows in this experiment [16].

Image enhancement by adding noise and flipping

Typically, the size of the training data set is an important parameter in image classification and target detection. If too few training data are available, network model training cannot be completed and leads to results not fitting and poor accuracy. In this experiment, because the training dataset primarily comes from the real remote sensing images of flood disaster areas and requires manual annotation, the dataset is smaller than other target detection datasets. Therefore, image data enhancement is typically a necessary method to improve the quality of the training dataset.

How to increase the size of the training data set and train the convolutional network model with robustness through limited or small amounts of data has been investigated by many researchers. Typically, the primary methods of data enhancement include flipping, rotation, translation, blurring, and increasing noise. In this experiment, considering the characteristics of the building data, we primarily use two methods to enhance the data of the manually marked image, which are fuzzy and noise increasing.

We first define the function library imported into OpenCV and then define various data enhancement method functions. Finally, we traverse the training data set and enhance the data using the pseudorandom number allocation method. After data enhancement, the training data set mitigates the problem of insufficient training data sets.

Analysis and improvement of deep learning network structure

After preprocessing, the training dataset for the neural network is basically complete. This chapter discusses the structure of the neural network used in the experiment. First, the R-CNN is used to complete the target detection task, and the final output includes the location and category information of the target object. Second, based on the requirements of building area measurement in this experiment, the fully connected layer in the R-CNN is modified to a convolutional layer, which makes the entire network a fully convolutional neural network. The fully convolutional neural network exhibits a fast operation speed and classifies the types of ground objects in remote sensing images using semantics, which is more suitable for the requirements of this experiment. The network models for testing include the SegNet model and U_net model.

Finally, Mask R-CNN is used to detect buildings in the remote sensing images. Mask R-CNN combines the advantages of these two classification models and can accurately output the location information of the target object while extracting the target object's shape as a mask. This experiment uses the advantages of a fully convolutional neural network, optimizes certain network structures in Mask R-CNN, and achieves good results.

R-CNN network model

As one of the most effective methods in the field of target detection, R-CNN series methods recommend candidate regions and then use a convolution network to classify candidate regions. Considering the slow computation speed of R-CNNs, this paper uses the Fast R-CNN model in its experiment. The Fast R-CNN algorithm structure is simple, continuously extracts the feature map through a series of convolution and pooling operations, and finally uses the fully connected layer for target detection and judgment. The process of target detection includes the following three key modules: region recommendation, feature extraction and region classification. First, region classification is used to generate approximately 2000 region recommendations for the input image to form the candidate detection set. Second, feature extraction is used to extract the fixed length 4096 feature vectors from each region recommendation. Last, the third module classifies regions to score and filter each recommended region.

Fast R-CNN can meet the target detection requirements of simple images with a size of 400 * 400, and the accuracy of the experimental results is also high. However, in this experiment, because a remote sensing image is large size and contains rich detail, Fast R-CNN computes slowly and achieves lower detection precision due to too many regions being recommended.

From the perspective of improving computation speed, a fully convolutional neural network should yield good performance.

Fully convolutional network model

In the CNN algorithm, continuous convolution and pooling operations for feature extraction are performed on the image. After the image features are extracted, there are typically two network design ideas: the R-CNN mentioned above, which uses a fully connected neural network for logistic regression classification; and a convolutional network, which upsamples the feature image to achieve feature reconstruction. Compared to R-CNN, which is primarily used in the field of target detection, the fully convolutional neural network is more commonly used in the field of semantic segmentation. After a target feature is extracted, the full convolutional network is used for feature reconstruction, which improves calculation efficiency and eliminates the shortcomings of constraining local features.

The SegNet [17] model is a common fully convolutional neural network that is comprised of a group of symmetric encoders and decoders. Its structure is simple, and model detection is fast. During encoding, SegNet uses convolution to extract features and increases the receptive field via pooling, which is similar to the feature extraction stage in R-CNN. During decoding, SegNet reproduces the features after image classification through deconvolution until the original size of the image is restored. R-CNN typically stops at image classification. In this experiment, it is necessary to extract the building shape from the building in the remote sensing image; thus, it is necessary to use deconvolution for semantic segmentation.

U_net [18] is also a common full convolution network that is typically used in medical images. Medical images are typically large and, thus, similar in size to the remote sensing images in this experiment. In general, it is not possible to input the original image into the network directly. In U_net, a sliding window is used to scan the original image, and the slices of the original image are used for training or testing. During slicing, the image will be expanded to ensure the accurate prediction of the boundary image block.

Concurrently, the U_net model is also a network with a simple structure. The fully symmetrical encoder and decoder structure is similar to the SegNet model. During semantic segmentation using deconvolution, the U_net model must be matched with the feature map generated during convolution. In the target detection task, the feature map is only used to make the model perform reinforcement learning, and the high-level feature maps are often not related to the low-level feature maps. This characteristic must be improved.

Finally, because the fully convolutional neural network ultimately does not use a fully connected layer, it can generate image segmentation maps of any size, which facilitates the investigation of the shape of buildings at multiple scales. Concurrently, the number of parameters for the convolution calculation is much lower than that of the fully connected layer. Therefore, using this fully convolution model can improve the computational efficiency of the network and make the experimental results richer, more flexible and more suitable for the requirements of this experiment.

Analysis and improvement of R-CNN network structure

From the many available deep learning models, the Mask R-CNN model is selected as the backbone in this experiment. Mask R-CNN is based on Faster R-CNN and adds a new branch to the bounding box recognition branch to predict the object mask. In the pooling layer, the model uses align pooling to replace the maximum pooling, which mitigates the misalignment between the ROI and the extracted features during the pooling process and improves the accuracy of the extracted feature contours to the height of the instance segmentation. This design can detect a target in the image and provide a high-quality segmentation result for each target.

Although Mask R-CNN can extract target shapes that cannot be resolved by R-CNN, its network structure still uses R-CNN as the backbone. Combined with the deconvolution operation in the decoding part of the SegNet model, we know that modifying the original fully connected network to a fully convolutional network can improve the model calculation speed and make better use of the previous extraction characteristic map. In this improved model, before box regression and classification, we use a convolutional layer with a convolution depth of 1024 instead of a fully connected layer with 1024 neurons in the R-CNN.

In the process of testing the U_net model, the image features extracted by shallow convolution operations are primarily information such as texture, pixels, and position relationships; the features extracted by deep convolution operations are often more abstract and high-level semantic features. In this experiment, buildings are often densely arranged in remote sensing images. When using a deep convolutional neural network to detect a target image with many small targets, image features will, thus, be lost as the network propagates.

Therefore, in the improved Mask R-CNN model, the U_net model is used for reference during the fusion of feature maps at different scales. To calculate image features with different depths, this experiment uses ResNet101 for feature extraction. Considering that the number of feature maps should be controlled within a reasonable range, this experiment modified the residual neural network based on the original Res101 model. It is assumed that each residual convolution module contains an input layer, three convolution layers, two activation layers, and an output layer. Compared to the original feature extraction network, although the calculation speed has been reduced, gradient dispersion and gradient explosion are prevented, and concurrently, we can obtain a reasonable number of feature images.

Then, when performing feature fusion in the RPN network, the bottom layer feature image must pass the 1 * 1 convolution to obtain the same number of channels as the previous layer feature image; high-level feature images must be upsampled to obtain the same length and width as the next-level feature images. Thus, the two layers are added together to obtain a new fused feature image. Figure 1 shows this process.

In this experiment, the algorithm flow of building recognition using the improved Mask R-CNN is as follows:

1. Select the training dataset that has been preprocessed above, and divide it into 20% as the backup of the verification set.
2. Input the new dataset to the improved residual neural network to obtain the corresponding feature map, and perform feature fusion on it;
3. Set a predetermined number of ROIs for each anchor point in the trained feature map to obtain multiple candidate ROIs;
4. The candidate ROIs are sent to the RPN network for binary classification and pooled;
5. These ROIs are then classified via logistic regression and MASK generation.

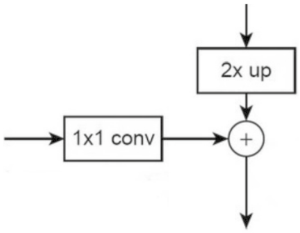

Fig. 1 Fusion of feature maps

Experimental results and analysis

Preparation of experimental data

Because no complete remote sensing image data set about buildings exists in China, this paper chooses to use a manual annotation method to construct a data set for use in this study. The source of the data set includes three parts:

1. DOTA [19] data set;

   The DOTA dataset contains 2806 aerial images from different sensors and platforms, covering various scales, positions and shapes. The DOTA dataset contains more types of target objects than other datasets. Although it does not include the target type of the building, the relatively large image scale and rich target types markedly reduce the difficulty of building identification.

   Images in the DOTA dataset have a high resolution and can be used as the primary body of the training set data to extract building features. A total of 70 images are used in this part, and a total of approximately 2000 buildings are marked.
2. Real remote sensing images of Hebei Province, China;

   The DOTA dataset is primarily composed of remote sensing images of rural areas in Canada. Although there are many green areas, these regions are visually different from mountain villages in China. To enhance the universality of the training data, this paper adds real remote sensing images of flood disaster areas and performs more model training.

   This portion of the remote sensing images was downloaded from Google Earth with a pixel level of 20 and an image resolution of 0.12 m. A total of 20 images are used in this part, and approximately 700 buildings are marked.
3. Data provided by CCF big data competition (high-definition remote sensing image of a city in southern China in 2015);

   This dataset is relatively small and contains 5 large-scale RGB remote sensing images with labels; the image size range is approximately 4000 * 4000 pixels. There are four types of objects marked in these images: vegetation, buildings, water bodies, roads and others. Because real data sets are often large, the experimental data in this part can be used to supplement the model.

   Based on the previous preprocessing methods, we separately process these three parts of the data sets using threshold segmentation and then select the remote sensing images with moderate building density. Then, based on the real situation of the remote sensing images, we perform fuzzy and other data enhancement processing. After preprocessing, we perform manual annotations.

The manual annotation tool selected in this paper is the VGG Image Annotator (VIA). After exporting labeled buildings to JSON format, they can be input into Mask R-CNN for experimentation.

Results and analysis

The experimental environment used in this study is a Win10 system with Python3; Keras, which is used as the deep learning tool; and an NVIDIA GTX 1660TI video card. To verify the effectiveness of the experimental algorithm, the network model mentioned in the previous chapter is selected for comparison. Considering that the difficulty of this experiment is that all buildings cannot be identified easily, this article uses the common kappa coefficient and precision rate to evaluate accuracy. Considering the differences in focus between the semantic segmentation method and the target detection method, the kappa coefficient is used to evaluate the accuracy of the SegNet and U_net models, and the precision rate is used for the accuracy evaluation of the R-CNN series of algorithms, where the larger the value, the higher the accuracy. Relevant formulae are as follows:

$$\text{Kappa coefficient} = \frac{P_0 - P_\mathrm{e}}{1 - P_\mathrm{e}}, \tag{1}$$

$$P_0 = \frac{\mathrm{TP} + \mathrm{TN}}{N{*}N}, \tag{2}$$

$$P_\mathrm{e} = \frac{(\mathrm{TP} + \mathrm{FP}){*}(\mathrm{TP} + \mathrm{FN}) + (\mathrm{TN} + \mathrm{FN}){*}(\mathrm{TN} + \mathrm{FP})}{N{*}N}, \tag{3}$$

$$\text{Accuracy rate} = \frac{\text{Number of predicted buildings}}{\text{Total number of buildings}}, \tag{4}$$

where N is the image size; TP is the number of pixels describing buildings and predicted to describe buildings; TN is the number of pixels of other features, and the detection result is also the number of pixels of other features; FP is the number of pixels that indicate other features are predicted to be buildings; and FN is the number of pixels that indicate that a building is predicted to be other features.

Because there are 4000 * 4000 remote sensing images in the training and verification data sets, we write the interface for image segmentation and fusion. The primary process includes the following steps:

1. Take the DOTA remote sensing image and the real remote sensing image of the disaster area from the data set, and calculate their average size;
2. The size of the segmented image is set equal to 1;
3. The length and width of the remote sensing image are divided into 256; if not, fill in 0, and then segment.
4. The large image is segmented using the designed image size as the step size, and the small images are sent to the model for prediction.
5. The predicted small image is spliced into a new image and restored to the size of the original image.

In addition to the interface for image size, this experiment also made modifications that are more suitable for remote sensing image detection. In the design of the initial parameters of the model, the experiment uses the method of parameter migration and uses the weights trained with the COCO dataset as the pretraining weights of the algorithm model in this paper. Considering that the number of negative samples in remote sensing images is often much greater than that of positive samples, the candidate regions whose intersection ratio with the labeled frame is greater than 0.6 are considered positive samples when setting the candidate regions, and the setting between 0.2 and 0.6 is background; those below 0.2 do not participate in training.

In model training, network parameters must be set, including the learning rate, batch size, and activation function. In this experiment, the initial learning rate is designed to be 0.1, and the batch size is designed to be 15. Then, the learning rate optimization strategy is used in the experiment, a higher learning rate is used at the beginning of training, and the learning rate gradually decreases with the training. Because there is only one GPU in the experimental equipment, the GPU only processes one image at a time.

Figures 2 and 3 show the experimental results of the Mask R-CNN algorithm and ordinary R-CNN algorithm (Table 1).

As shown in the tables, the proposed method finishes in less computation time and improves the detection accuracy compared to the traditional R-CNN algorithm. In addition, in the comparison of Kappa coefficients, the kappa coefficient value of SegNet and U_net models is about 0.75, while the coefficient value of the experimental method of the paper is about 0.9. It can be seen that the prediction of this paper is more accurate. Figures 4 and 5 show a comparison of the Mask R-CNN algorithm before and after the improvement.

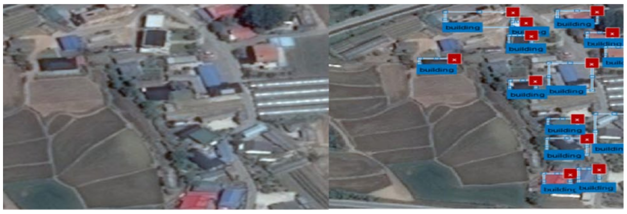

Fig. 2 Original picture and R-CNN test results

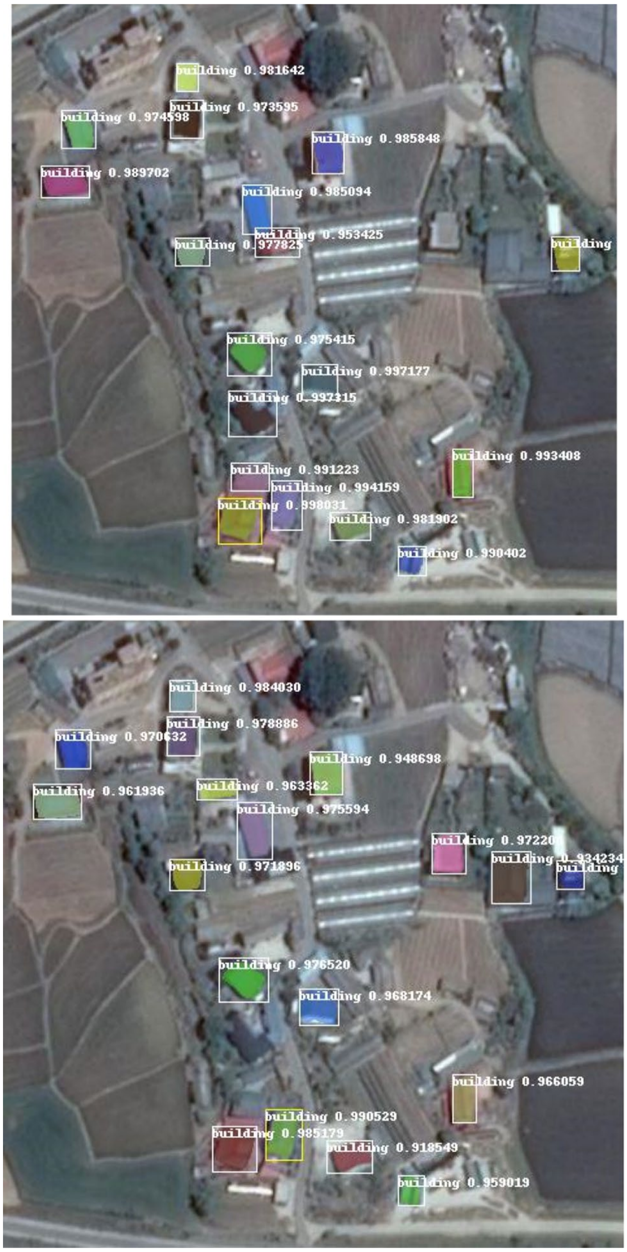

Fig. 3 Mask R-CNN test results and improved Mask R-CNN test results

Table 1 Algorithm accuracy rate

	R-CNN	Mask R-CNN	Improved Mask R-CNN
DOTA Dataset	77.7%	72.7%	**81.8%**
Remote sensing image of disaster area	59.2%	57.7%	**72.2%**
Model training time	10 h	Nearly 10 h	Nearly 3 h

bold numbers represent the experimental results of the improved Mask R-CNN

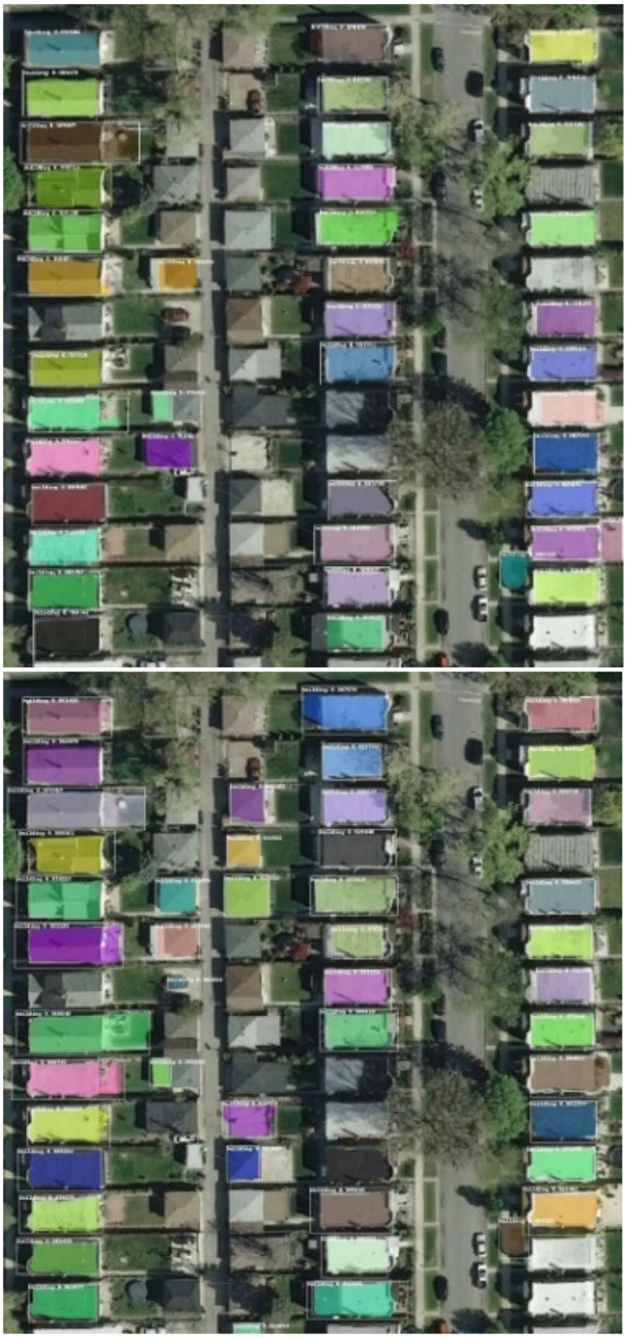

Fig. 4 Comparison of DOTA test data

Conclusion

In this paper, the advantages of the traditional unsupervised object segmentation method are used to construct a data set. Combined with the improved Mask R-CNN algorithm of the residual network, a building detection method for remote sensing images of flood disaster areas is designed. Via experimental detection, as described in the previous chapter, the proposed method improves model accuracy

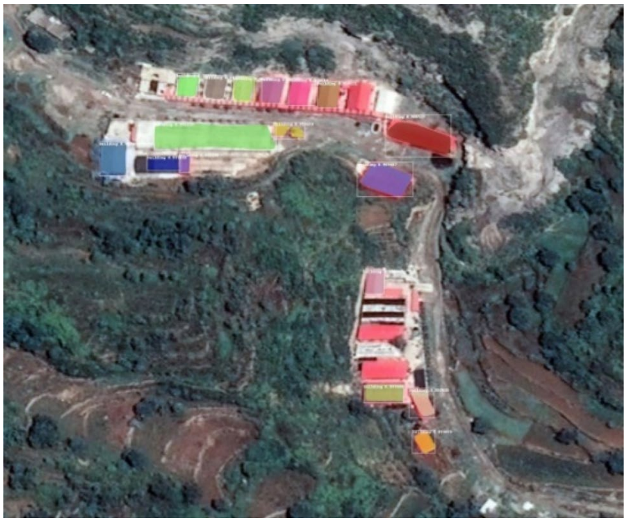

Fig. 5 Disaster area remote sensing image

and markedly reduces computation time, highlighting its improved performance.

In this experiment, algorithm updating is used to reduce computation time as much as possible without reducing the precision of the results. In later experiments, we plan to improve model accuracy and design an algorithm with the goal of training the model as accurately as possible.

Acknowledgements The research work described in this paper was fully supported by the National Key R&D program of China (2017YFC1502505) and the Joint Research Fund in Astronomy (U2031136) under cooperative agreement between the NSFC and CAS. Qian Yin and Xin Zheng are the authors to whom all correspondence should be addressed.

Declaration

Conflict of interest Corresponding authors declare on behalf of all authors that there is no conflict of interest. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Wu W et al (2012) Building extraction from high resolution remote sensing imagery based on spatial-spectral method. Geomat Inf Sci Wuhan Univ 7:800–805
2. Huang X et al (2007) Classification of high spatial resolution remotely sensed imagery based upon fusion of muitiscale features and SVM. J Remote Sens 11:48–54
3. Xin F, Shanxiong C (2019) High-resolution remote sensing image building extraction in dense urban areas. Bull Surv Mapp
4. Xu-dong W, Jian-ming G, Bai-jun J et al (2008) Mixed-pixel classification of remote sensing images of cellular automata. J Surv Mapp 37(1):42–48
5. Bateson CA, Asner GP, Vessman CA (2000) Endmember bundless: a new approach to incorporating endmember variability into spectral mixture analysis. IEEE Trans Geosci Remote Sens
6. Wang Q, Tenhuen JD (2004) Vegetation mapping with multitemporal NDVI in north Eastern China Transect. Int J Appl Obs Geoinf 6:17–31
7. Sun ZJ, Xue L, Xu YM et al (2012) Overview of deep learning. Appl Res Comput 29(8):2806–2810
8. He K, Zhang X, Ren S et al (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
9. Girshick R (2015) Fast R-CNN. IEEE international conference on computer vision, pp 1440–1448
10. Ren S, He K, Girshick R et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 91–99
11. Kaiming H, Georgia G, Piotr D, Ross G, Mask R-CNN (2017) The IEEE international conference on computer vision (ICCV), pp 2961–2969
12. Yulan L, Xiaoxia H, Hongyang Li et al (2019) Extraction of informal solid waste in towns and villages based on convolutional neural network and conditional random field method. J Geoinf Sci 21(2):259–268
13. Rongshuang F, Yang C, Qiheng X et al (2019) High-resolution remote sensing image building extraction method based on deep learning. J Surv Mapp 48(1):38–45
14. Dong B, Xiong FH et al (2020) Research on remote sensing building detection based on improved Yolo v3 algorithm. Comput Eng Appl 56(18):209–213
15. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geoences 10(2–3):191–203
16. Jobson DJ (2004) Retinex processing for automatic image enhancement. J Electron Imaging 13(1):100–110
17. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
18. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation[C]//NAVAB N. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 234–241
19. Xia GS, Bai X, Ding J et al (2017) DOTA: a large-scale dataset for object detection in aerial images

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.