EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Encoder-decoder structure based on conditional random field for building extraction in remote sensing images

Yian Xu[1,*]

[1]Department of Architectural Engineering, Anyang Vocational and Technical College, Anyang 455000 China

Abstract

The application of building extraction involves a wide range of fields, including urban planning, land use analysis and change detection. It is difficult to determine whether each pixel is a building or not because of the large difference within the building category. Therefore, automatic building extraction from aerial images is still a challenging research topic. Although deep convolutional networks have many advantages, the networks used for image-level classification cannot be directly used for pixel-level building extraction tasks. This is caused by successive steps larger than one in the pooling or convolution layer. These operations will reduce the spatial resolution of feature maps. Therefore, the spatial resolution of the output feature map is no longer consistent with that of the input, which cannot meet the task requirements of pixel-level building extraction. In this paper, we propose a encoder-decoder structure based on conditional random field for building extraction in remote sensing images. The problem of boundary information lost by unitary potential energy in traditional conditional random field is solved through multi-scale building information. It also preserves the local structure information. The network consists of two parts: encoder sub-network and decoder sub-network. The encoder sub-network compresses the spatial resolution of the input image to complete the feature extraction. The decoder sub-network improves the spatial resolution from features and completes building extraction. Experimental results show that the proposed framework is superior to other comparison methods in terms of the accuracy on open data sets, and can extract building information in complex scenes well.

Keywords: building extraction, encoder-decoder structure, conditional random field, feature extraction.

Received on 19 November 2021, accepted on 30 November 2021, published on 07 December 2021

doi: 10.4108/eai.7-12-2021.172362

*Corresponding author. Email: 910675024@qq.com

1. Introduction

With the rapid development of urban construction, buildings, as the most basic part of a city, have become one of the most variable types of artificial objects in basic geographic data [1,2]. Timely and accurate extraction of remote sensing information is of great significance to urban planning, disaster management, digital city, geographic database update and other fields [3,4].

The basic theoretical research on building extraction using remote sensing images began in the 1980s. In the decades of development, edge extraction, image segmentation, Digital Surface Model (DSM), deep learning and other methods have been applied to building

information extraction, and some research achievements have been achieved [5,6].

While high resolution remote sensing image contains a lot of details, it also has some noise problems, which makes the effect of building extraction not good. Due to the large spectral difference in buildings and the noise problem in high resolution remote sensing images, the traditional pixel-based methods are prone to the discontinuous "salt-and-pepper phenomenon" in building extraction. That is, the same ground objects are not divided into one category, leading to the phenomenon of uniform land fragmentation. The method based on segmentation adopts object-oriented thought, overcomes the "salt and pepper phenomenon" and other shortcomings of the method based on pixel, so more and more scholars pay attention to it. Chao et al. [7] proposed a multi-scale segmentation method based on IKONOS image using object-oriented strategy. Wegne et al. [8] proposed a combination of region segmentation and Markov random field algorithm for remote sensing image scene modeling and building extraction. However, in the segmentation based building extraction method, it is difficult to accurately describe the spatial location, size and other information of the segmented building, and it is difficult to obtain the real building features from the image due to the reasons of tree occlusion and shadow occlusion. In order to improve the accuracy of building extraction, some scholars try to introduce auxiliary information into the extraction method.

Gao et al. [9] proposed a method to automatically extract building samples based on building shadows and accurately verify buildings. This method had the highest extraction accuracy in suburban buildings. LiDAR point cloud combining with texture features and Markov random fields can effectively extract building information in a variety of complex environments, and has been widely used [10-12]. Building extraction combined with auxiliary information can effectively improve the extraction accuracy, but it requires a high cost to obtain high-quality LiDAR point cloud data. In addition, building extraction methods combining spatial information, geometry, texture and other features of high-resolution remote sensing images are gradually developed. Karantzalos et al. [13] introduced a priori shape model of typical buildings related to geometric attributes, and achieved a good extraction effect, but it was not applicable to urban areas with dense buildings. AkCay et al. [14] combined the PLSA (Probabilistic Latent Semantic Analysis) model and morphological analysis to identify ground objects in high-resolution remote sensing images, thus avoiding the difficulty of establishing regular geometric shapes. But the outline of the extracted building was irregular. Li et al. [15] introduced Conditional Random Fields (CRF) into building extraction, combining pixel-level information and segmental level to identify roofs, which could effectively deal with roofs of complex shapes [16]. CRF is developed on the basis of Markov Random Fields, eliminating the strict independence assumption of Markov random airport. Its good global nature can well connect local features and realize the organic integration of bottom-up and top-down target semantics. However, because of the lack of large-scale spatial interaction information modeling ability of conditional random field, it is easy to produce different degrees of ground object smoothing problem [17,18].

In recent years, the development of deep learning has greatly promoted the progress of building extraction. Chen et al. [19] designed a 27-layer deep convolutional neural network with convolution and deconvolution in view of the characteristics of building shape rules, diverse appearance and complex distribution, and realized the prime-level extraction of buildings in high-resolution images. Xu et al. [20] proposed a new neural network framework RESU-NET, which could extract buildings in urban areas with very high resolution remote sensing images. Huang et al. [21] proposed an end-to-end trainable gated residual refinement network, which was based on Fully Convolutional network (FCN). It used excellent feature learning and end-to-end pixel labeling capabilities of FCN combining high-resolution aerial imagery and LiDAR point clouds for building extraction. Complex neural network can extract buildings with high accuracy, but it often needs a lot of computing time, and the existing deep learning methods are not good at extracting geometric structure integrity, building edge and other details.

Conditional random field can make up for the deficiency of deep learning in building extraction by modeling image context information by using spatial neighborhood information of both labeled and observed images. Shrestha et al. [22] improved the performance of FCN by introducing Exponential Linear Unit (ELU) and combining it with CRF to make full use of image spatial neighborhood information and enhance building boundaries. Sun et al. [23] designed a multi-task network that enabled FCN to generate both mask and edge information, and used the conditional random field model to refine the results of FCN, effectively improving both time and space efficiency. Li et al. [24] proposed the Feature Pairwise Conditional Random Field (FPCRF) framework, which used Convolutional Neural Networks (CNN) as a feature extractor to achieve fine-grained segmentation of graph models. However, in the combination of deep learning and conditional random field model, the existing deep learning models used for building extraction often extract building features from a single scale, and tend to ignore small scale building objects. However, the traditional image spatial information of conditional random field modeling can

only use the neighborhood information of pixels, but not the large-scale spatial information. In addition, the pixel-based processing method is easy to cause the loss of some building details and the phenomenon of salt-and-pepper noise in the building.

To solve the above problems, this paper proposes a new encoder-decoder structure based on conditional random field for building extraction in remote sensing images. In this paper, the cost term of local category labeling is introduced in traditional conditional random fields, so that the binary potential energy can effectively reflect the linear combination of the spatial relations of adjacent pixels and the information of local region category labeling, and keep the details of regular buildings. In addition, in order to alleviate the influence of large spectral differences within buildings, deep features with higher abstraction degree and stronger representation ability are extracted through autonomous learning. End-to-end building extraction in a single network can be achieved using encoder-decoder network.

2. Global-local detail aware conditional random field

In this paper, a global local detail aware conditional random field framework is proposed for building extraction from high resolution remote sensing images. Building extraction process can be divided into three stages: (1) using global and local integration D-LinkNet to model the unitary potential energy of conditional random field. The classification diagram output by global and local integration D-LinkNet is used as unary input of conditional random fields [25]. (2) Based on the classification graph, the connected region marker algorithm is used to obtain the segmentation priors. In addition to using the spatial context information of the image, the cost term of local category labeling is introduced. When the uncertainty of image labeling is strong, the category labeling can be obtained by referring to the label information of the neighborhood of the pixel. (3) α -expansion inference algorithm based on graph cut method is used to deduce the model and obtain the final building mark.

x is labeled image and y is observed image. When the observation field y is given and the random variable x_i obeys the Markov random field, the model constitutes a conditional random field.

The relationship between image marker sequence x and observation sequence y is modeled by unitary potential energy, and the probability of pixel acquiring building marker or non-building marker is calculated based on pixel characteristics. The unary potential energy $\varphi_i(x_i, y)$ can be defined as:

$$\varphi_i(x_i, y) = -\ln(P(x_i = b_k \mid f_i(y))) \quad (1)$$

Where x_i is the mark of pixel i, $i \in V = \{1,2,\cdots,N\}$. N is the pixel number in the image. The tag set $B = \{(k = 0,1) \mid b_k\}$. k=0 and k=1 indicate that the pixel is marked as building or non-building. f is the feature mapping function, which corresponds the image block to the feature vector. $f_i(y)$ represents the feature of pixel i.

In the early CRF-based segmentation methods [26], the appropriate features are firstly extracted and selected from the input image, and then the Structured Support Vector Machine (SSVM) [27] or other classifiers are used to learn the coefficients of CRF for segmentation. These methods require a large amount of calculation, and the existing potential energy function cannot fully consider the characteristics of high-resolution images, and lacks the ability of large-scale spatial interaction information modeling, which is easy to produce different degrees of ground object smoothing results.

Zhou et al. [28] proposed a new network D-LinkNet (LinkNet with Pretrained Encoder and Dilated Convolution) based on the LinkNet (Exploiting Encoder Representations for Efficient Semantic Segmentation) network [29]. D-linkNet has few parameters, high computational efficiency, and can make use of multi-scale building features of image. However, the extensive use of dilated convolution weakens the spatial connection between adjacent neurons and makes it difficult for the following network layer to extract local spatial information in the image. Therefore, this paper proposes a global and local integrated D-LinkNet to learn image features and calculate the tag b_k obtained in pixel x_i based on its feature vector.

In order to realize pixel-level building extraction from aerial images, this paper designs a convolutional neural network with encoder-decoder structure based on VGG-16 network, as shown in figure 1. Firstly, the spatial resolution of the network structure feature map gradually decreases, which is similar to the coding process. In order to meet the requirements of pixel-level building extraction, this paper uses continuous transposed convolution to restore the spatial resolution of feature images, which is similar to the decoding process.

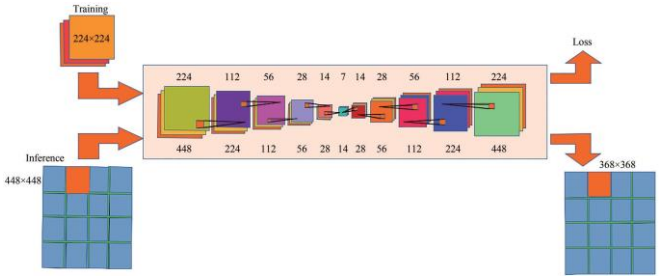


Figure 1. Visualization of the network structure

In order to achieve pixel-level building extraction from aerial images, this paper designs the network structure based on VGG-16 network. However, due to two factors, the full connection layer in VGG-16 is discarded. On the one hand, after discarding the full connection layer, the number of parameters to be learned is reduced from 138M to 14.7M, which reduces the network training pressure caused by too many learning parameters. On the other hand, it can preserve more spatial information in the feature map. Table 1 lists the network details used in this paper, where K represents the kernel size of the convolutional/deconvolution layer or pooling layer, P represents the size of the edge filling in the process of convolution, S represents the stride length, and N represents the number of convolutional filters.

Table 1. Details of the network structure

Layer	Parameter	Size
Input	[illegible]	224×224
Conv+BN+ReLU	K=3,P=1,S=1,N=64	224×224
Conv+BN+ReLU	K=3,P=1,S=1,N=64	224×224
Max-pooling	K=2, S=2	112×112
Conv+BN+ReLU	K=3,P=1,S=1,N=128	112×112
Conv+BN+ReLU	K=3,P=1,S=1,N=128	112×112
Max-pooling	K=2, S=2	56×56
Conv+BN+ReLU	K=3,P=1,S=1,N=256	56×56
Conv+BN+ReLU	K=3,P=1,S=1,N=256	56×56
Conv+BN+ReLU	K=3,P=1,S=1,N=256	56×56
Max-pooling	K=2, S=2	28×28
Conv+BN+ReLU	K=3,P=1,S=1,N=256	28×28
Conv+BN+ReLU	K=3,P=1,S=1,N=256	28×28
Conv+BN+ReLU	K=3,P=1,S=1,N=256	28×28
Max-pooling	K=2, S=2	14×14
Conv+BN+ReLU	K=3,P=1,S=1,N=512	14×14
Conv+BN+ReLU	K=3,P=1,S=1,N=512	14×14
Conv+BN+ReLU	K=3,P=1,S=1,N=512	14×14
Max-pooling	K=2, S=2	7×7
Deconv+BN+ReLU	K=4,P=1,S=1,N=512	14×14
Deconv+BN+ReLU	K=3,P=1,S=1,N=512	14×14
Deconv+BN+ReLU	K=3,P=1,S=1,N=512	14×14
Deconv+BN+ReLU	K=4,P=1,S=1,N=512	28×28
Deconv+BN+ReLU	K=3,P=1,S=1,N=512	28×28
Deconv+BN+ReLU	K=3,P=1,S=1,N=512	28×28
Deconv+BN+ReLU	K=4,P=1,S=1,N=256	56×56
Deconv+BN+ReLU	K=3,P=1,S=1,N=256	56×56
Deconv+BN+ReLU	K=3,P=1,S=1,N=256	56×56
Deconv+BN+ReLU	K=4,P=1,S=2,N=128	112×112
Deconv+BN+ReLU	K=3,P=1,S=1,N=128	112×112
Deconv+BN+ReLU	K=3,P=1,S=1,N=64	224×224
Deconv+BN+ReLU	K=3,P=1,S=1,N=64	224×224
Conv	K=1,P=0,S=1,N=2	224×224
Softmax	----	224×224

2.1. Construction of global local Multiparallel expansion convolution module

It is difficult to extract sufficient features from a single sensory field due to differences in building scales. To solve this problem, global local integration D-LinkNet constructs a global local multi-parallel expansion convolution module, which extracts building features by first increasing the expansion rate and then decreasing the expansion rate. In addition, the expansion convolution module in this paper is divided into expansion convolution of cascaded mode and parallel mode. The parallel expansion convolution effectively compensates for the loss of local structure information in the process of

feature extraction. As shown in figure 2, from the bottom, when the expansion of the stack of convolution of inflation is set to 1, 2, 4, 8, 3, 1, each layer of the receptive field size using different inflation rates in each branch parallel building feature extracting is 1, 3, 7, 15, 31, 37, 37, respectively. Finally, all branches feature extraction characteristics through the combined operation will get the final result. Since each path has a different receptive field, the network can integrate building features of different scales.

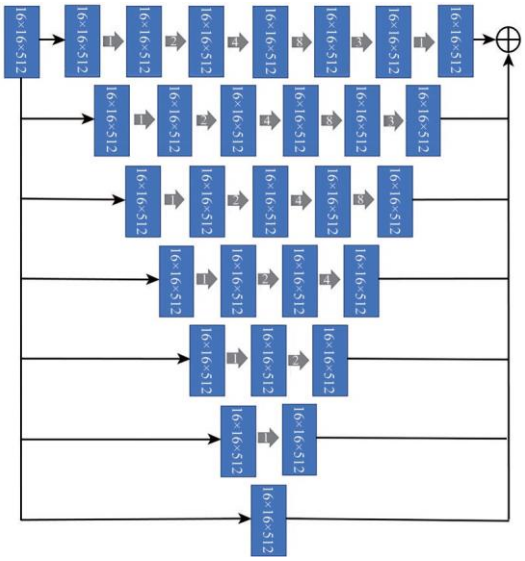


Figure 2. The structure diagram of global and local multiple parallel dilated convolution module

The binary potential energy $\varphi_i(x_i, y_i, y)$ is used to construct the relationship between the current node and its neighborhood node, which makes the spatial interaction of local pixels can be considered in building extraction.

2.2. Local category marking costs

In this paper, a linear combination of neighborhood smooth term and local class labeled cost term is introduced to model binary potential energy. In the process of classification iteration, this framework can fully consider the mark of each pixel and make use of the details. The definition is as follows:

$$\varphi_i(x_i, y_i, y) = g_{ij}(y) + \theta \cdot \Theta_B(x_i, y_i \mid y) \quad (2)$$

When $x_i = x_j$, $\varphi_i(x_i, y_i, y)$ =0. Where $g_{ij}(y)$ represents the spatial smoothing term of modeling adjacent pixels. $\Theta_B(x_i, x_j \mid y)$ is a local category tag cost term with size $|B| \times |B|$, representing the penalty between adjacent pixels x_i and corresponding tags of x_j . θ is the parameter that controls the influence of the cost of class marking on the binary potential energy.

Function $g_{ij}(y)$ models the spatial interaction relationship between neighborhood pixel i and j , which can be expressed as:

$$g_{ij}(y) = dist(i, j)^{-1} \exp(-\beta \| y_i - y_j \|^2) \quad (3)$$

Where (i, j) represents the coordinate pairs of neighborhood pixels. The function $dist(i, j)$ is its corresponding Euclidean distance. y_i and y_j represent the spectral vectors at position i and j . Parameter β is set as the reciprocal of two times of the mean square deviation of spectral vector differences of all adjacent pixels in the image.

The cost of local category marking item $\Theta_L(x_i, x_j \mid y)$ uses observational image data to model the spatial relations of category markers x_i and x_j in each neighborhood, which are defined as follows:

$$\Theta_L(x_i, x_j \mid y) = \frac{\min(P(x_i \mid f_i(y)), P(x_j \mid f_j(y)))}{\max(P(x_i \mid f_i(y)), P(x_j \mid f_j(y)))} \quad (4)$$

Since the category probability $P(x_i \mid f_i(y))$ can be obtained from the global local integration D-LinkNet, the local category labeling cost term is the interaction effect of the current labeling of adjacent positions i and j based on the observed image data y . Therefore, the introduction of the cost term of local category marking enables the framework to effectively consider all pixel marks in the iterative process and keep the details of the building.

2.3. Segmentation priori (SP)

When the number of samples, accuracy and richness are limited, the binary potential energy of traditional conditional random fields is difficult to distinguish ground object details and image noise accurately, and the basic analysis unit of object-oriented method is segmentation object. Therefore, a larger scale of spatial information can be used to effectively mitigate the effect of spectral differences within the building and background region classes [30].

The segmentation prior is based on the classification graph using the connected region marking algorithm. In this paper, we use the classical eight neighborhood connected region labeling algorithm of set data structure to get segmentation objects. Then, based on the original building extraction graph, the mark of each segmentation

object can be obtained through the maximum voting strategy. The segmentation prior can be defined as:

$$P(x_i = b_{seg}) = \max(P(x_i = b_k)), k \in |B| \quad (5)$$

Where b_{seg} represents the object mark category of the segmentation region where the pixel is located.

2.4. α -expansion model inference based on graph cut method

In the global local detail sensing conditional random field model, according to the characteristics of high resolution image, the potential function is extracted and constructed for buildings. After the completion of model construction, it is necessary to predict the optimal building extraction effect of the test image through model reasoning, that is, to obtain the optimal tag configuration for the pixel, which is defined relative to the cost function. In order to obtain the optimal markers, researchers proposed many inference methods, such as Iterative Conditional Modes (ICM) and graph-cuts algorithm. However, ICM is sensitive to the selection of initial value and is prone to fall into local minimum value. In this paper, graph-cuts based α -expansion algorithm is used for inference [31].

The α -expansion algorithm sets a local search strategy for energy functions with metric attributes. This strategy can solve the problem that the algorithm easily falls into the local minimum solution when the moving space is small. According to the local search strategy, the α -expansion algorithm iterates continuously through graph-cuts algorithm within the cycle, and each iteration calculates the global minimum of the class II marking problem.

3. Experiment and result analysis

The WHU building dataset consists of an aerial dataset and a satellite dataset [32]. The aerial image data was taken from New Zealand, covering an area of about 450km^2, with a ground resolution 0.3m and containing more than 187000 well-marked buildings. The images in the area are cropped to 8189 images with 512×512 pixels. The samples are divided into a training set, a validation set and a testing set, consisting of 4736 images, 1036 images and 2416 images, respectively. The east Asia satellite image data consists of six adjacent satellite images with a ground resolution of 2.7m and covers an area of 550km^2. The entire image is seamlessly cropped into 17388 images with 512×512 pixels, comprising 29085 buildings, with samples divided into a training set and a test set. 13662 images are used for training and 3726 for testing.

3.1. Experiment setting

Experiments are performed on the WHU building data set, both aeronautical and satellite. A D-LinkNet convolution conditional smoothing classifier based on convolutional CRFs and DPSCRF (Detail Preserving Smoothing Classifier) Conditional Random Fields) and Fully Connected/Dense CRF are analyzed as comparative experiments. The experimental design and model parameters on both datasets are consistent. DPSCRF uses support vector machine (SVM) to construct unitary potential energy, models the linear combination of pixel space neighborhood relation and local category marking cost term of binary potential energy, and adopts object-oriented thought to merge and segment priors. FullCRF constructs a fully connected CRF model of image complete pixel set. ConvCRF uses ResNet to construct unary potential energy on the basis of FullCRF, and adds conditional independence assumption to CRF reasoning, which can formalize most of the reasoning into convolution.

Firstly, SVM is used to classify high resolution remote sensing images. The region of interest was selected and the images were classified by SVM classifier based on the color features of the images. The classification results were used to construct the unitary potential energy of DPSCRF. The unitary potential energy of ConvCRF is constructed using ResNet, and a pre-trained network model is used in the experiment. In the CRF reasoning part, the filter size h is set to 3. For the model in this paper, all images in the training set are first used to train unitary potential energy, and then iterative reasoning is carried out uniformly for the model during testing. Based on D-LinkNet34, global and local integration D-LinkNet is proposed to replace the unitary potential energy of CRF. In the experiment, the expansion rates of stacked dilatancy convolution are set as 1, 2, 4, 8, 3, 1. Finally, 7 branches including the original feature graph are fused.

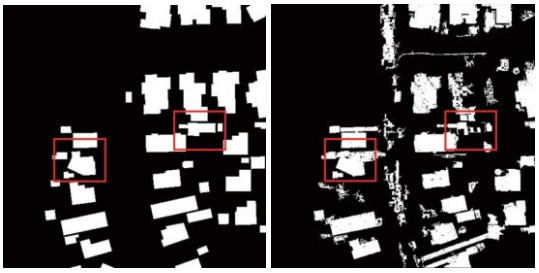
In this paper, three evaluation criteria are selected for building extraction, which are Accuracy (the proportion of the number of correctly predicted pixels in the total number of pixels), Precision (the proportion of correctly predicted pixels within the category), and Intersection-over-Union (IoU: the ratio of intersection and union between the predicted results and the real value of a certain category within the model).

3.2. Results analysis

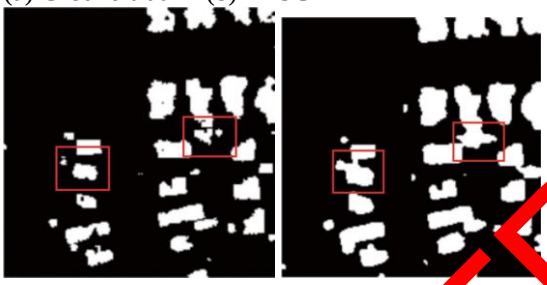
The original image is shown in figure 3 (a), and the corresponding manually marked building area is shown in figure 3 (b). Figure 4 shows the results of building extraction by using different models.

(a)Aerial image (b) labeled image

Figure 3. Original images and labels



(a) Ground truth (b) DPSCRF



(c)FullCRF (d) ConvCRF

(e)D-LinkNet (f) Proposed method

Figure 4. Building extraction results from aerial image

As can be seen from figure 4, DPSCRF has poor extraction effect with more discrete pixels and rough boundary. FullCRF has an average effect. Although there is no obvious salt-and-pepper noise, boundary ambiguity exists.

ConvCRF algorithm has a good effect on building extraction, there is almost no noise, and the boundary ambiguity problem is greatly improved. However, for small buildings, the extracted boundary is rough. In contrast, D-LinkNet and the proposed method can extract clear building boundaries without noise. In addition, the global local detail sensing conditional random field model removes the small building blocks incorrectly extracted by D-LinkNet and can effectively maintain details.

Accuracy, precision and IoU are calculated to extract building performance for quantitative evaluation model, as shown in table 2.

Table 2. Quantitative evaluation of building extraction from WHU aerial dataset/%

Method	Accuracy	Precision	IoU/%
DPSCRF	91.46	80.01	67.75
FullCRF	95.35	81.05	80.02
ConvCRF	96.81	94.99	87.32
D-LinkNet	98.52	93.73	90.59
Proposed	98.56	94.98	91.73

As can be seen from Table 2, the overall accuracy of DPSCRF and FullCRF is poor. The IoU calculation result of ConvCRF reached 87.32%, and other indexes are also relatively high. The accuracy of D-LinkNet and IoU are both higher than that of ConvCRF, while the accuracy of D-LinkNet is slightly lower than that of ConvCRF. However, the accuracy evaluation results are significantly higher than DPSCRF and FullCRF, indicating that D-LinkNet can be well applied in the field of building extraction. As for the proposed model in this paper, it can be seen that it is higher than the other four models in terms of accuracy, precision and IoU, and also significantly improved compared with D-LinkNet.

In order to further verify the validity of the model, this paper compares some cutting-edge methods in the field of building extraction, including Spatial Residual Inception Convolutional Neural Network (SRI-NET) and Deep Encoding Network (DE-NET) and EU-Net (Efficient Fully Convolutional Network). The quantitative evaluation results are shown in table 3.

Table 3. Accuracy comparison of the state-of-the-art methods for building extraction

Method	IoU/%
SRI-Net	89.09

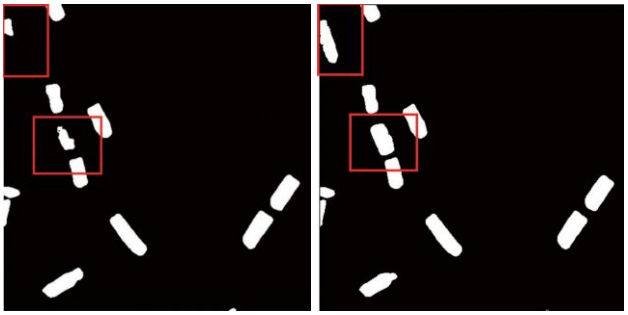
DE-Net	90.13
EU-Net	90.57
Proposed	91.73

We also conduct experiments on WHU satellite dataset. The original image is shown in figure 5(a), the corresponding artificially marked building area is shown in figure 5(b), and the rest is the result of building extraction by using different models.

As shown in figure 5, the architectural objects extracted by FullCRF, ConvCRF and D-Linknet are incomplete. FullCRF has the worst overall effect of building extraction. In addition to many missed inspection problems, there are also irregular building boundaries and false inspection problems. The omission problem of ConvCRF is not as serious as that of FullCRF, but there are obvious false detections in ConvCRF. Compared with FullCRF and ConvCRF, D-LinkNet significantly improved the irregular boundary problem in building extraction. But there are still some phenomena of missing and mis-detection. As can be seen from the figure, the buildings extracted by the method in this paper have no obvious problems of misdetection and missing detection, and the boundary of building objects is relatively regular, which has a good effect of maintaining details [33-35].

(a)raw image (b) ground truth

(c)FullCRF (d)ConvCRF



(e)D-LinkNet (f) Proposed

Fig.5 Original satellite images and building extraction results

This paper compares some cutting-edge methods in the field of building extraction, such as SR-FCN (Scale Robust Convolutional Neural Network), and obtains the quantitative evaluation results of building extraction as shown in table 4. As can be seen from Table 4, the presented method in this paper has obvious advantages over its comparison method in terms of accuracy, precision and intersection ratio. Among them, the extraction accuracy of all models based on WHU satellite data set is similar. The IoU of ConvCRF is 78.83%, which is obviously higher than FullCRF and SR-FCN, and other indexes are also relatively high, indicating that the ConvCRF model can adapt to the application of building extraction. The IoU value of D-LinkNet is nearly 10% higher than that of ConvCRF, and the accuracy is also higher, indicating that D-LinkNet has more accurate extraction ability than ConvCRF. The proposed method is superior to the other four methods in each detection index and has good building extraction ability.

Table 4. Quantitative evaluation of building extraction from WHU satellite dataset/%

Method	Accuracy	Precision	IoU/%
FullCRF	98.20	51.99	49.33
SR-FCN	91.37	79.01	64.01
ConvCRF	99.27	86.26	78.83
D-LinkNet	99.46	90.14	88.11
Proposed	99.54	91.82	89.83

Unlike aerial data sets, WHU satellite data sets are sparsely distributed and have lower image resolution than aerial data sets. The above factors bring some difficulties to the model extraction of buildings in this paper.

Therefore, the Precision and IoU of satellite data set test are all lower than those of aerial data set. From the perspective of visual effect, the model in this paper has a greater improvement in satellite data set than D-LinkNet.

4. Conclusion

This paper puts forward a encoder-decoder structure based on conditional random field for building extraction in remote sensing images. The following conclusions are drawn from the tests using aerial and satellite images from the WHU building dataset.

(1) D-LinkNet is used to model the unitary potential energy of the conditional random field model, which can effectively combine the features of buildings of different scales and make the structure of acquired building objects more complete.

(2) By adding segmentation priors into the construction of binary potential energy, the problems of noise and spectral difference in the image can be effectively dealt with, and the building classification map with clean background can be obtained.

(3) The introduction of the cost term of local category marker meets the high requirements of building extraction task for building detail information extraction, and can capture the details that are difficult to be identified by the network. The model is tested on both aeronautical and satellite datasets, and the IoU index is up to 91.72% and 89.83%, respectively. The framework can, therefore, be adapted to both aeronautical and satellite data sets.

In the existing experiments, the extraction accuracy of this framework is high, but its training and testing data are only limited to small-scale images. In the future, the application of large-scale high-resolution remote sensing image in building extraction will be further studied, and a more complete building information will be extracted by combining multi-source geographic information data.

Acknowledgements.

The author would like to thank the reviewers for their excellent comments.

References

[1] Wang J, Qin Q, Xin Y, et al. A Survey of Building Extraction Methods from Optical High Resolution Remote Sensing Imagery[J]. Remote Sensing Technology & Application, 31(4), 653-662, 2016.

[2] Li Yan, Xiaowei Wang, and Shoulin Yin. Campus Garbage Image Classification Algorithm Based on New Attention Mechanism [J]. International Journal of Electronics and Information Engineering. Vol. 13, No. 4, pp. 131-141, 2021.

[3] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM, 60(6), 84-90, 2017.

[4] Smirnov E A, Timoshenko D M, Andrianov S N. Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks[J]. AASRI Procedia, 6, 89-94, 2014.

[5] S. Yin and H. Li. Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582.

[6] Shahid Karim, Ye Zhang, Shoulin Yin, Irfana Bibi. A Brief Review and Challenges of Object Detection in Optical Remote Sensing Imagery [J]. Multiagent and Grid Systems. 16(3), 227-243, 2020

[7] Chao T, Yihua T, Xiaojie G, et al. Object-oriented Method of Hierarchical Urban Building Extraction from High-resolution Remote-Sensing Imagery[J]. Acta Geodaetica et Cartographica Sinica, 2016, 39(1):39-45.

[8] J. D. Wegner, U. Soergel and B. Rosenhahn, "Segment-based building detection with conditional random fields," 2011 Joint Urban Remote Sensing Event, 2011, pp. 205-208, doi: 10.1109/JURSE.2011.5764756.

[9] X. Gao, M. Wang, Y. Yang and G. Li, "Building Extraction From RGB VHR Images Using Shifted Shadow Algorithm," in IEEE Access, vol. 6, pp. 22034-22045, 2018, doi: 10.1109/ACCESS.2018.2819705.

[10] Syed G, Mohammad A, Lu G. An Automatic Building Extraction and Regularisation Technique Using LiDAR Point Cloud Data and Orthoimage[J]. Remote Sensing, 2016, 8(3):27.

[11] Yin, S., Li, H. & Teng, L. Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images [J]. Sensing and Imaging, vol. 21, 2020. https://doi.org/10.1007/s11220-020-00314-2

[12] Broersen T, Peters R, Ledoux H. Automatic identification of watercourses in flat and engineered landscapes by computing the skeleton of a LiDAR point cloud[J]. Computers & Geosciences, 2017, 106(sep.):171-180.

[13] K. Karantzalos and N. Paragios, "Recognition-Driven Two-Dimensional Competing Priors Toward Automatic and Accurate Building Detection," in IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 1, pp. 133-144, Jan. 2009, doi: 10.1109/TGRS.2008.2002027.

[14] H. G. AkÇay and S. Aksoy, "Automatic Detection of Geospatial Objects Using Multiple Hierarchical Segmentations," in IEEE Transactions on Geoscience and Remote Sensing, vol. 46, no. 7, pp. 2097-2111, July 2008, doi: 10.1109/TGRS.2008.916644.

[15] E. Li, J. Femiani, S. Xu, X. Zhang and P. Wonka, "Robust Rooftop Extraction From Visible Band Images Using Higher Order CRF," in IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 8, pp. 4483-4495, Aug. 2015, doi: 10.1109/TGRS.2015.2400462.

[16] Lafferty J D. Conditional random fields: probabilistic models for segmenting and labeling sequence data[J]. /Proceedings of the Eighteenth International Conference on Machine Learning (ICML). San Francisco: Morgan Kaufmann Publishers: 282-289, 2001.

[17] J. Zhao, Y. Zhong and L. Zhang, "Detail-Preserving Smoothing Classifier Based on Conditional Random Fields for High Spatial Resolution Remote Sensing Imagery," in IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 5, pp. 2440-2452, May 2015, doi: 10.1109/TGRS.2014.2360100.

[18] Xiaowei Wang, Shoulin Yin, Ke Sun, et al. GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition [J]. Journal of Applied Science and Engineering, vol. 23, no. 3, pp. 555-561, 2020.

[19] K. Chen, K. Fu, X. Gao, M. Yan, X. Sun and H. Zhang, "Building extraction from remote sensing images with deep learning in a supervised manner," 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017, pp. 1672-1675, doi: 10.1109/IGARSS.2017.8127295.

[20] Yongyang X, Liang W, Zhong X, et al. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters[J]. Remote Sensing, 2018, 10(1):144.

[21] Huang J, Zhang X, Xin Q, et al. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 151(MAY):91-105.

[22] Shrestha S, Vanneschi L. Improved Fully Convolutional Network with Conditional Random Fields for Building Extraction[J]. Remote Sensing, 2018, 10(7).

[23] Sun J, Li W, Zhang Y, et al. Building segmentation of remote sensing images using deep neural networks and domain transform CRF[C]// Image and Signal Processing for Remote Sensing XXV. 2019.

[24] 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)," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 7502-7519, Nov. 2020, doi: 10.1109/TGRS.2020.2973720.

[25] Zhu Q Q, Li Z, Zhang Y N, et al. Global-Local-Aware conditional random fields based building extraction for high spatial resolution remote sensing images. National Remote Sensing Bulletin, 25(7): 1422-1433, 2021.

[26] Fayao Liu, Guosheng Lin, Chunhua Shen. CRF learning with CNN features for image segmentation[J]. Pattern Recognition, 2015, 48(10):2983-2992.

[27] Szummer M., Kohli P., Hoiem D. (2008) Learning CRFs Using Graph Cuts. In ECCV 2008. Lecture Notes in Computer Science, vol 5303. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88688-4_43

[28] L. Zhou, C. Zhang and M. Wu, "D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 192-1924, doi: 10.1109/CVPRW.2018.00034.

[29] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4, doi: 10.1109/VCIP.2017.8305148.

[30] Xiaowei Wang, Shoulin Yin, Desheng Liu, et al. Accurate playground localisation based on multi-feature extraction and cascade classifier in optical remote sensing images [J]. International Journal of Image and Data Fusion, vol. 11, no. 3. pp. 233-249, 2020.

[31] Y. Boykov, O. Veksler and R. Zabih, "Fast approximate energy minimization via graph cuts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1222-1239, Nov. 2001, doi: 10.1109/34.969114.

[32] Ji S, Wei S, Lu M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery[J]. International Journal of Remote Sensing, 2018:1-15.

[33] Shoulin Yin, Hang Li, Lin Teng, et al. An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images [J]. International Journal of Image and Data Fusion, vol. 11, no. 2, pp. 201-214, 2020. DOI: 10.1080/19479832.2020.1727573.

[34] Shoulin Yin, Ye Zhang and Shahid Karim. Region search based on hybrid convolutional neural network in optical remote sensing images[J]. International Journal of Distributed Sensor Networks, Vol. 15, No. 5, 2019.

[35] Shoulin Yin, Ye Zhang, Shahid Karim. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model[J]. IEEE Access. volume 6, pp: 26069 - 26080, 2018.