EXTRACTION OF BUILDINGS FROM HIGH-RESOLUTION REMOTE SENSING IMAGES BASED ON IMPROVED U-Net

Wenxi HUANG[1,*], Liufeng TAO[2], Xue LI[3], Xiaoyi HU[4]

Aiming at traditional classification methods in high-resolution remote sensing images, a modified semantic segmentation model U-Net was developed to more efficiently and accurately extract buildings. First, Google image was applied to create sample data sets which were used to improve U-Net network model. ResNet with different depths was employed as the backbone network to extract semantic information from images. Furthermore, attention mechanism module was added to refine the extracted feature map and improve the classification performance of surface features. The experimental results showed that, compared with Support Vector Machine (SVM) and SegNet, improved U-Net model (Attention Res-UNet) performed better in prediction performance and evaluation metrics, with mean values in accuracy, recall, F1 value, and Intersection over Union (IoU) reached 92.4%, 87.9%, 91.5%, and 89.9%, respectively. The improved U-Net model has a prediction performance that is closer to manual annotation, and can efficiently recognize and extract remote sensing image information and obtained high-precision extraction results. This method had certain application significance for surface features extraction.

Keywords: building extraction, U-Net, ResNet, attention mechanism.

1. Introduction

Quick and accurate detection and extraction of buildings from remote sensing images are of great significance in various applications including change detection, urban planning, and natural disaster prevention [1,2]. Continuous development of remote sensing technology has significantly improved the spatial resolution, spectral resolution, and temporal resolution of remote sensing images. High-resolution remote sensing images have become an important source of various land targets, including building data. However, in remote sensing images, terrain information, geometric structures, and texture features are more abundant and refined, which makes building extraction technology more urgently required [3,4].

* Corresponding author
[1] Institute of Seismology, China Earthquake Administration, Wuhan 430071, China, e-mail: hwxcug@163.com
[2] School of Computer Science, China University of Geosciences, Wuhan 430074, China
[3] Institute of Seismology, China Earthquake Administration, Wuhan 430071, China
[4] Institute of Seismology, China Earthquake Administration, Wuhan 430071, China

Therefore, achieving higher precision in automatic extraction of surface features has been a great hot topic in recent years.

Currently, traditional building extraction methods mainly include supervised classification [5], knowledge constraints [6], and template matching [7]. Among them, supervised classification applies manual methods to design building features, then trains classifiers such as random forest (RF) [8], support vector machine (SVM) [9], and ISO clustering [10], and finally employs the trained classifier to classify experimental data for building extraction. Knowledge constraints, such as rectangular constraints [11], constraints on geometric and radiometric differences [12], and diamond constraints [13], are applied for the identification and extraction of surface features under specific conditions in order to obtain better extraction results. Template matching is the process of determining building parameter templates and describing building information, and then, finding the most suitable algorithm based on correlation to obtain optimal extraction results [14].

Recently, with rapid development of artificial intelligence and significant improvement of computer processing capabilities, as well as continuous updating of massive learning sample data sets, deep learning techniques represented by Convolutional Neural Networks (CNN) [15] have been extensively employed in surface feature extraction from various types of satellite imagery [16]. Deep learning avoids a great amount of manual operations and automatically learns semantic features of building objects from image data set with large numbers of training samples to obtain accurate extraction results. This method is more applicable than traditional methods. However, batch processing based on CNN models may face various challenges such as excessive memory requirements, low computational efficiency, and limited perceptual regions [17]. To solve these problems, some scholars proposed Fully Convolutional Networks (FCN), which not only improved the ability to extract spectral and spatial features, but also removed fully connected layers, enhancing image segmentation efficiency and decreasing computational complexity. However, when extracting images using FCN model, the resolution of feature images is continuously decreased during forward propagation process. These segmentation results generated only by up-sampling end features had lower edge accuracy and lost some detailed information. In 2015, the U-Net model was first proposed in the field of medical image segmentation. This model is based on FCN model and integrated low- and high-dimensional features from traditional networks through symmetrical structures, improving image segmentation performance. In 2019, Tang et al. [18] used an improved U-Net network based on multi-scale feature structures to automatically detect the distribution of lung nodules, and the results showed that the model could more accurately locate lung lesion areas. Liu et al. [19] introduced a building extraction method on the basis of feature compression-activated U-Net model, which restored

spatial information at corresponding scales and identified buildings with different shapes better.

Furthermore, today, attention mechanisms are at research forefront in computer vision and natural language processing. Its basic idea is to teach models to focus their attention on important information and ignore unimportant ones. Consequently, it is extensively applied in such tasks as object detection and image segmentation.

This research applied Google high-resolution remote sensing images and constructed its own data set. At the same time, it improved U-Net network by using ResNet at different depths as its backbone network to enhance the performance of its classification. After the skip connections of U-Net, an attention module was added to refine its extracted feature maps. Finally, the optimal model is selected by comparing the extraction effect and extraction indicators of the building.

2. Building extraction based on improved U-Net model

2.1 U-Net network and its improvements

Fig. 1 shows the structure of U-Net network. As was seen from the figure, it has two main characteristics: first, a U-shaped structure, and second, a skip connection.

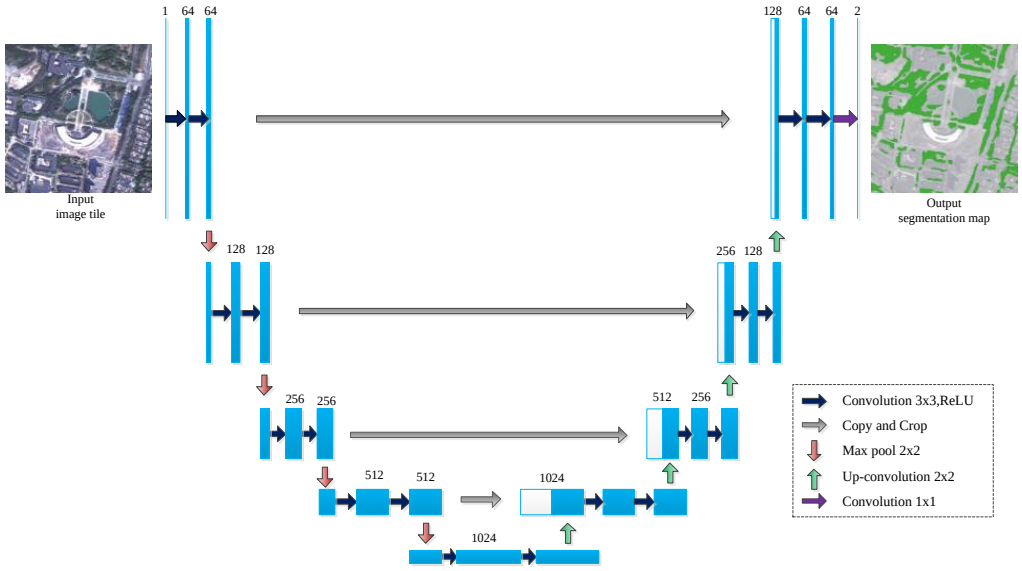

Fig. 1. The conventional network structure of traditional U-Net model

The U-Net network extends the idea of integrating low dimensional and high-dimensional features of fully convolutional neural networks, the left encoder part of U-Net performed four identical down-sampling operations through

convolution and pooling to extract advanced feature information. Each down-sampling convolution reduced image size to half of its original size and doubled the number of feature channels. The decoder on the right side used transposed convolution for 4 up-sampling to restore resolution. Each up-sampling and deconvolution doubled image size, the number of feature channels has been reduced to half of its original value. In order to decrease spatial information loss due to down-sampling process, skip connections were introduced and concatenate operation was applied to fuse more low-level semantic information in the feature maps recovered from up-sampling, refining segmentation results.

Although one of the major advantages U-Net is its simple structure, the backbone network applied for down-sampling in U-Net is stacked with ordinary Conv + BN + ReLu (CBR) modules, which may lack depth compared to network models such as ResNet and Visual Geometry Group (VGG). ResNet proposed the idea of residual learning, which to some extent solved the problems of information loss and significance loss, as well as gradient vanishing or exploding in too deep networks. At the same time, compared to VGG networks, ResNet had fewer parameters and could improve accuracy by considerably increasing the depth. The extraction effect is more excellent, while also preventing the classification accuracy from reaching saturation and the extraction performance from rapidly declining. Therefore, this article improved traditional U-Net by replacing U-Net backbone network with ResNet with stronger learning ability and attempted to experiment with ResNets of different layers (Res50, Res101 and Res152) to select the optimal combination model.

2.2 Attention module

The attention module applied in this research belonged to channel attention mechanism in soft attention mechanism [16]. It further enhanced the features among different categories by mutual correlation among the same class of target features to enhance the accuracy of classification. However, its greatest advantage was that it could integrate existing networks and improve network performance at low costs. Fig. 2 shows attention mechanism structure, which was divided into three parts.

1) Sequencing, also known as feature compression, converted a 2D channel into a real number, which in a sense represented channel importance.

2) Excitation, referring to the application of parameters to generate weights for each feature channel, could learn and explicitly model correlations among feature channels.

3) Feature recalibration, the function of which can be simply summarized as enhancing important features, weakening unimportant features, and making the extracted features more directional. Therefore, adding attention mechanism could

refine the extracted feature maps and improve the classification performance of surface features.

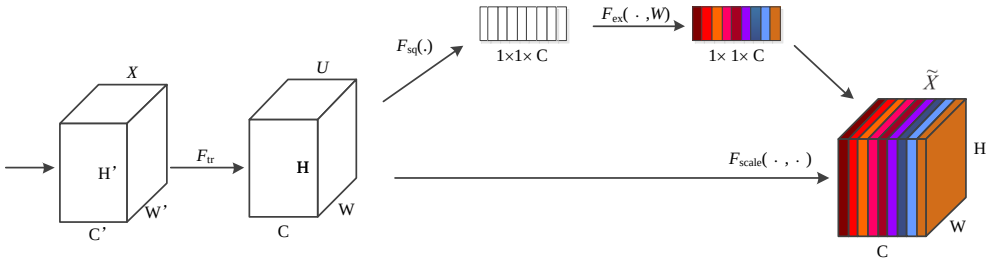

Fig. 2. Schematic diagram of channel attention module

2.3 Attention Res-UNet model

After the addition of attention module, Res-UNet was developed, as shown in Fig. 3. The sampling operation process of the red box encoder was as follows: input of cropped RGB three-channel sample image with 256×256 size into the feature extraction module of ResNet network, application of its convolutional and residual structures to meet the requirements of feature extraction, discarding the last three layers employed including average pooling layer, fully connected layer, and activation function layer for classification results, and finally obtaining a feature vector with 8×8 size and 2048 depth.

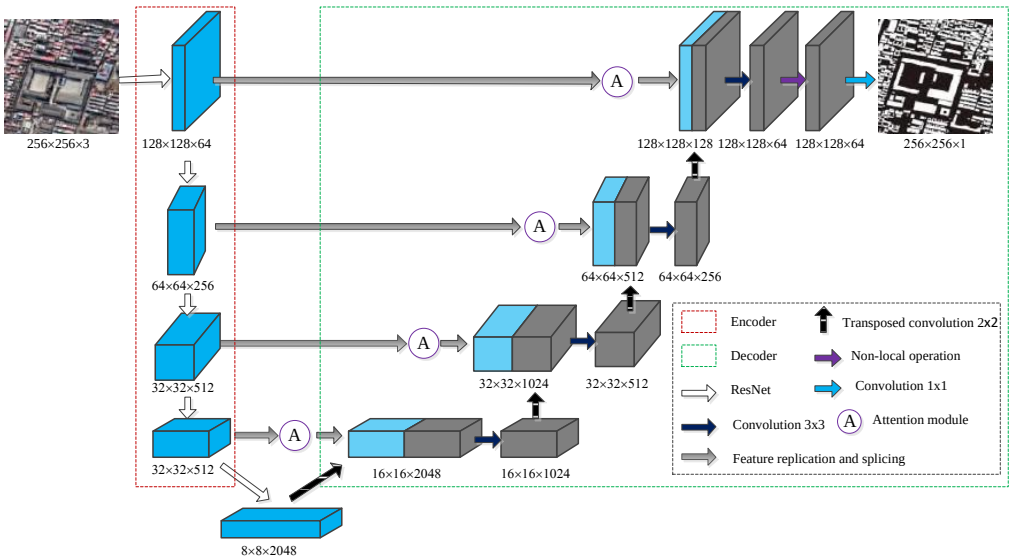

Fig. 3. Attention Res-UNet network architecture diagram

After the down-sampling was completed, the up-sampling process of green box decoder began. A transposed convolution layer with stride of 2 and convolution kernel size of 2×2 was sequentially set. After passing through this convolution

kernel, the size of feature layer could be gradually increased while its depth could be reduced. Based on U-Net model structure, concatenate method was applied to connect the information of equally sized feature maps in down-sampling and up-sampling lines. After concatenation, the feature map which integrated deep and shallow feature information entered attention module, which could further balance shallow features and make the fused information more flexible in adapting to land feature extraction, improving directionality. After passing attention module, a 3×3 convolutional layer, batch normalization layer, and ReLu were applied to correct linear unit operation, and this process was repeated twice. After repeating up-sampling operations according to the above rules, a feature layer of 128×128×64 was obtained. However, since convolutional layer could only preliminarily combine image information in receptive field, a nonlocal operation layer was added after this feature layer to provide accurate global information for each pixel. This solved the problem of large-scale recognition errors in image recognition of green areas under difficult classification conditions. Non-local operations only introduce global information and do not change feature layer size and depth. Finally, a 1×1 convolutional layer as well as up-sampling operation was applied to obtain buildings recognition results.

2.4 Experimental environment and parameter configuration

This experiment adopted the open-source deep learning framework Tensorflow_gpu 2.0.0, developed in Python. And all software and hardware operated in Windows 10 operating system. The computer was equipped with an Intel (R) Xeon (R) Silver 4210R CPU, with a main frequency of 2.4 GHz and 256 GB of memory. The experiment adopted dual GPU mode, with a 24 GB NVIDIA Tesla K80 GPU model graphics card. It was driven by Compute Unified Device Architecture (CUDA) Toolkit 10.0, corresponding to CUDA Deep Neural Network (cuDNN) V7.5.0. Batch processing was adopted for training, with batch size of 4 and training frequency of 100. The number of images employed for learning forward and backward propagations was set to 8 to ensure that the model fully learned the features of house samples.

3. Building extraction experiment

3.1 Research area and data

Taking Weining County, Guizhou Province (103° 36' E ~ 104° 45' E, 26° 36' N ~ 27° 26' N) as research area (Fig. 4), There are many ethnic minorities living in Weining County, with a variety of building types, including various classical ethnic characteristic buildings and modern buildings. The experimental data included Google satellite images as well as building vector data. Among them,

Google satellite images recorded in 2021 were adopted as high spatial resolution remote sensing image data, which included three bands (RGB) and ground resolution of 5 meters. The number of building vectors was obtained from the third national land use survey database in 2020 (http: //www.mnr. gov. cn/), which could effectively ensure the accuracy, objectivity, and timeliness of building vector data. ArcGIS Pro software was applied for converting building vector data into label images required for deep learning. Fig. 5 illustrates partial images and corresponding building labels in the building data set.

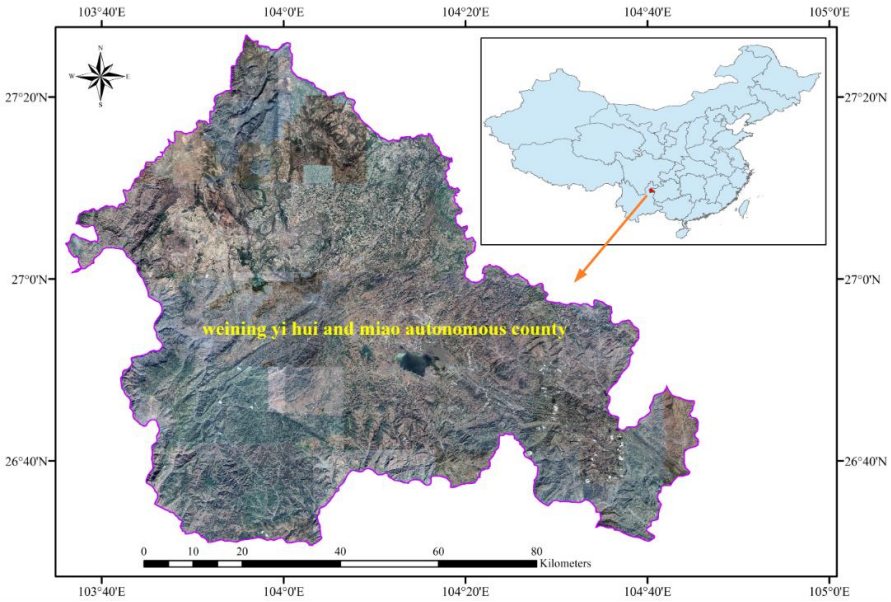

Fig. 4. The remote sensing image data of the study aera

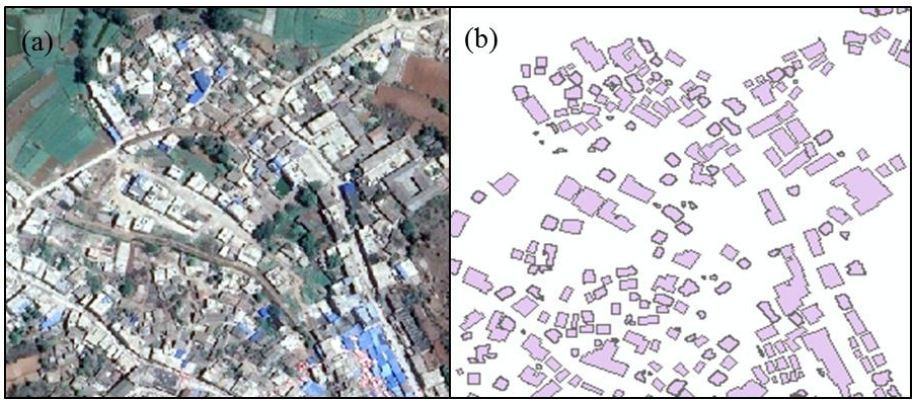

Fig. 5. Partial building images (a) and corresponding label data set (b)

3.2 Data and preprocessing

To increase the size of sample and correctly input it into the model, it was necessary to preprocess the original experimental data to obtain training and validation data set. Sample preprocessing mainly included data format conversion, seamless segmentation, sample screening, training and validation data set production, etc. Firstly, building vector data was converted into raster annotated data required for U-Net model training. Secondly, remote sensing image data and building grid data were synchronously sliced into same size (256×256), so that remote sensing image slice data corresponded one-to-one with building grid slice data. Thirdly, the proportion of building area in the grid slice data of each building was calculated and the slice data with smaller building area proportions were removed. Finally, sample data set was divided into training and validation data in certain proportions to facilitate the input of preprocessed sample data into U-Net model.

3.3 Experimental results and analysis

3.3.1 Analysis of building extraction results

A comparative experiment was conducted on building extraction in the same research area using improved U-Net, SVM, and SegNet models respectively. The extraction results (Fig. 6) revealed that the extraction results of SVM and SegNet models presented obvious errors and omissions. Other impermeable surfaces with the same material such as roads, parking lots, and exposed bedrock could easily be extracted as buildings. Building boundaries were not clearly extracted and there were several salt and pepper classification phenomena in the extraction results of SVM and SegNet models. Man-made surface features extraction results based on improved U-Net model were closest to real ground conditions, which better distinguished between buildings and backgrounds, obtained richer and more obvious boundary information, improved the ability of target boundary localization, and supplemented semantic information to a certain extent. Overall, the accuracy of building extraction obtained from this method was higher.

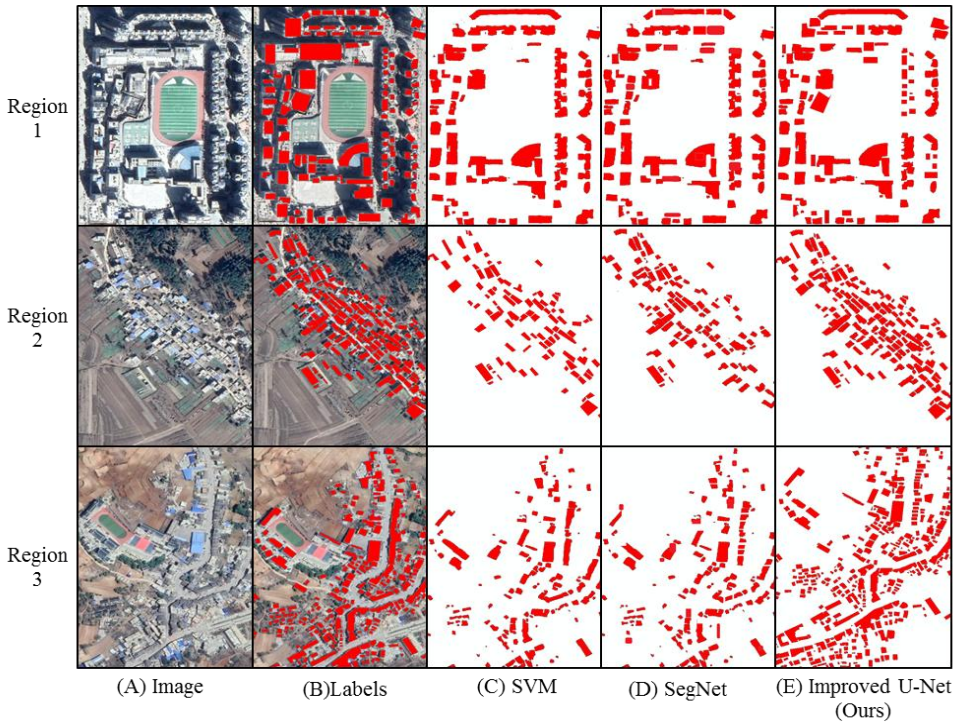

Fig. 6. Visual comparison of building extraction with three models

From the local building extraction results in three regions (Fig. 7), that the following conclusions were made: (1) for buildings with single surface features and shadows on the building itself (Fig. 7, Region 1), all three models correctly recognized buildings. However, SVM model was greatly affected by shadows and the shape segmentation effect of buildings was poor. SegNet model improved semantic segmentation performance and boundary localization ability to a certain extent, accurately and completely extracted buildings in images, and extracted more detailed information compared to SVM model. (2) For buildings with relatively simple surrounding environments and slightly complex structures (Fig. 7, Region 2), the boundary segmentation results of SegNet model were rough and SVM model had difficulty in correctly recognizing buildings, resulting in obvious missed detections. However, improved U-Net model accurately and clearly extracted buildings and had advantages in handling details. (3) For buildings with complex structures and surrounding environments (Fig. 7, Region 3), SegNet model had some errors in details and the model lost some boundary information. SVM model was not able to well distinguish between buildings and backgrounds, which resulted in obvious false positives. However, improved U-Net model could detect the precise contours of buildings, enhancing target localization ability and semantic segmentation accuracy in complex scenes.

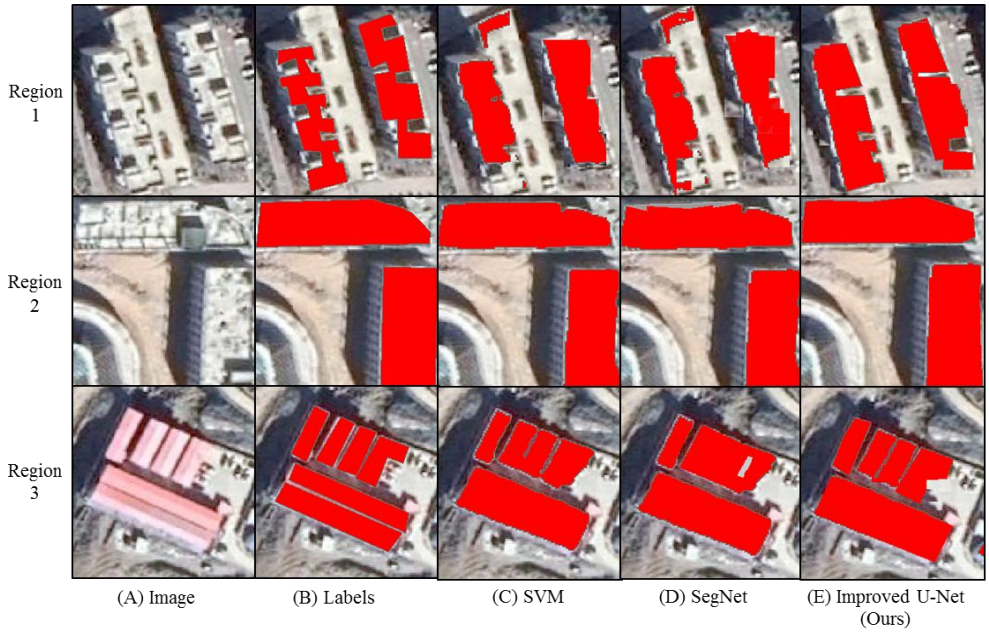

Fig. 7. Comparison of the extraction performance of different buildings with three models

Detailed comparison of building boundary extraction efficiency (Fig. 8) showed that improved U-Net model had higher extraction accuracy compared to the other two models: (1) Building boundaries extracted using SegNet and SVM models were more blurred and rough. Due to the effects of shadows around the buildings and materials of some impermeable surfaces, some semantic information of building boundaries was lost. Especially, extraction results of SVM model had many errors and omissions, which could not well locate building edges. (2) Compared with the other two models, improved U-Net model extracted smoother building boundary segmentations, containing more comprehensive semantic information and more accurate spatial position information, and had more advantages in detail processing.

Fig. 8. Details of building boundary division extracted with four models

The overall extraction results and detailed comparison results of the building indicated that after introducing attention mechanism, the model can effectively enhance the adaptability of U-Net on different terrain features, and effectively reduce misclassification and missed classification caused by small targets. The results suggested to some extent that the attention mechanism can effectively improve the specificity in extracting target feature information.

At the same time, compared to ordinary U-Net networks, our model replaces the backbone network with ResNet, and as the hierarchy of the backbone network deepens, we found that the phenomenon of misclassification and missed segmentation gradually decreases, the coherence gradually increases, the block boundaries gradually match the actual boundaries, and the number of point and block like fragmented patches also gradually decreases. Especially, the U-Net model using Res152 network is closest to manual annotation and has the best overall visual perception. This indicated that replacing the backbone network with different levels of ResNet can improve the segmentation accuracy of the network and make the segmentation results more accurate.

3.3.2 Precision comparative analysis

Precision, Recall, F1 value, and Intersection over Union of the target object extracting results were calculated for the three models to compare the accuracies of the extracted results. According to the obtained results presented Fig. 9, that the following conclusions were made: (1) in the three experimental areas shown in Fig. 7, improved U-Net model achieved the mean values of 92.4%, 87.9%, 91.5%, and 89.9%, for accuracy, recall, F1 value, and Intersection over Union, respectively, which were better than the other two models. (2) In recall tests, SVM model had the lowest recall rate because it had the highest number of missed pixels in building extraction results. (3) Improved U-Net model had the lowest number of missed pixels, resulting in the highest recall rate, especially in region 2.

Accuracy evaluation results of the three models further indicated that when extracting target features based on shallow machine learning algorithms (such as SVM), only the spectral information of pixels in the image and relationships among pixels were applied for image classification. However, within limited computing units, it was difficult to accurately express real target scene in the face of large-scale high-resolution satellite image data combined with complex and diverse land features. Extraction accuracy of deep learning methods (such as SegNet model) was higher than that of SVM model which also demonstrated good performance in extracting buildings from high-resolution remote sensing images. However, improved U-Net model had the advantages of improving target boundary localization and semantic segmentation capabilities. By processing of decoding layers, the spatial and edge information of building boundary segmentation was

improved and more detailed information of buildings were extracted from images, making segmentation results clearer and more complete. Therefore, using improved U-Net model, both the speed and efficiency of building extraction from high-resolution images were improved.

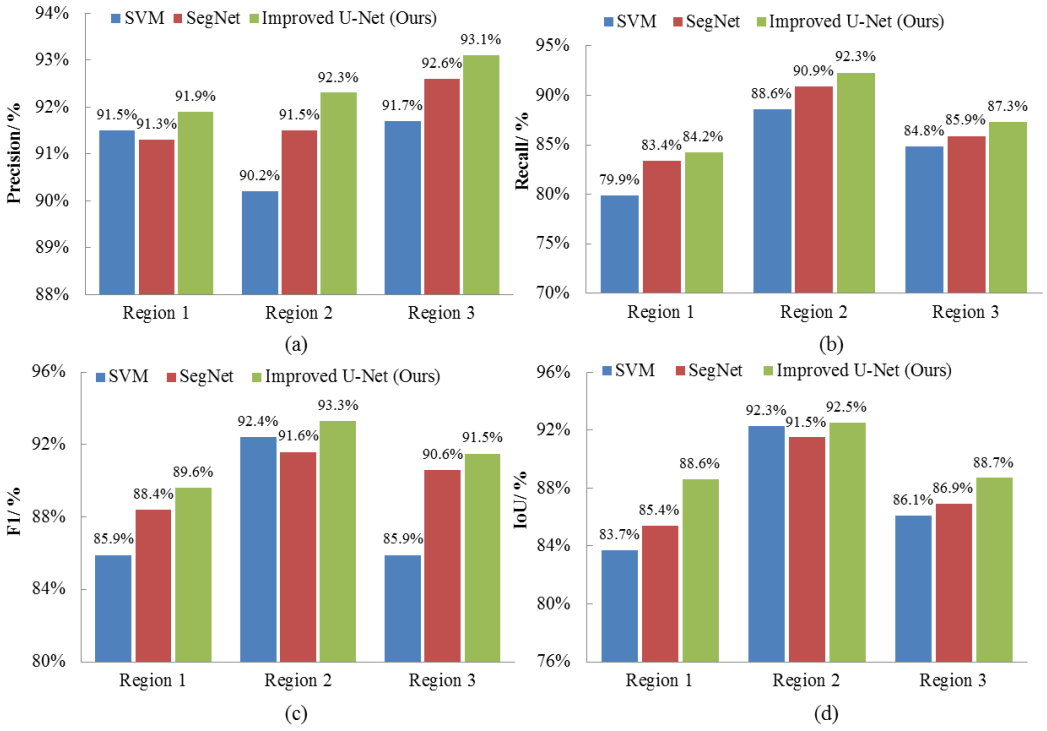

Fig. 9. Accuracy evaluation result of the three models ((a): Comparison Diagram of Prediction Accuracy of Three Models; (b): Comparison diagram of recall rate of three models; (c): Comparison diagram of F1-scores of three models; (d): Intersection over Union of three models.)

4. Conclusions and discussion

This research used Google satellite imagery and building vector data as data source, and through image preprocessing, Sample making tool of ArcGIS Pro was applied to create a self-made remote sensing image data set. Furthermore, this research improved the semantic segmentation model U-Net by using ResNet residual network as decoder backbone network for feature extraction, deepening network layers while preventing model degradation and gradient vanishing. Also, an attention module was added after U-Net skip connections to refine the extracted feature maps and improve the classification performance of the model. Image semantic segmentation model was applied for building extraction, which could achieve pixel level classification. For Google high-resolution images, it was able to recognize and extract buildings with minimum size of 0.8 m × 0.8 m. Finally, after

comparing the results obtained from three models, it was found that improved U-Net model presented the most accurate segmentation results, with fewer fragmented blocks and misclassification situations, almost close to manual annotation. Furthermore, the Mean precision and Mean Intersection over Union (MIoU) of 3 regions were the highest in evaluation indicators with 92.4 and 89.9, respectively, proving that this model was more suitable for the recognition and extraction of buildings in high-resolution remote sensing images.

This research used a self-built high-definition data set with limited data volume and manually annotated label images had visual interpretation errors and omissions, which had certain impacts on network training and prediction results. In future research, we will try to employ larger and better annotated samples to improve the quality of high-definition remote sensing image samples and segmentation efficiency of the model.

Acknowledgements

This work was supported by The Spark Program of Earthquake Technology of CEA (No. XH23028YA).

R E F E R E N C E S

[1] *S. Daranagama and A. Witayangkurn,* Automatic building detection with polygonizing and attribute extraction from high-resolution images, ISPRS International Journal of Geo-Information, Vol. **10**, Iss. 9, pp. 606, 2021.

[2] *H. Eric, M. Areeba, and Z. Kezhong,* PM2.5 pollution and endoplasmic reticulum stress response, Environmental Disease, Vol. **6**, pp. 111-115, 2021.

[3] *G. Liasis, and S. Stavrou,* Building extraction in satellite images using active contours and colour features. International Journal of Remote Sensing, Vol. **37**, Iss. 5, pp. 1127-1153, 2016.

[4] *E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez,* Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. **55**, Iss. 2, pp. 645-657, 2016.

[5] *X. Peng, R. Zhong, Z. Li, and Q. Li,* Optical remote sensing image change detection based on attention mechanism and image difference. IEEE Transactions on Geoscience and Remote Sensing, Vol. **59**, Iss. 9, pp. 7296-7307, 2020.

[6] *B. Tejeswari, S. K. Sharma, M. Kumar, and K. Gupta,* Building footprint extraction from space-borne imagery using deep neural networks. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. **43**, pp. 641-647, 2022.

[7] *X. Li, J. Deng, and Y. Fang,* Few-shot object detection on remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, Vol. **60**, pp. 1-14, 2021.

[8] *Y. Chen, M. Luo, and J. Peng,* Land use classification of industrial and mining reclamation area based on grid search random forest algorithm. Transactions of the Chinese Society of Agricultural Engineering, Vol. **33**, Iss. 14, pp. 250-257, 2017.

[9] *A. W. Putra, W. Putra, B. S. Mulyanto, and A. O. Gaffar,* A performance of combined methods of VCG and 16BCD for feature extraction on HSV. International Journal of Image, Graphics and Signal Processing, Vol. **13**, Iss. 3, pp. 13-32, 2021.

[10] *G. Tari, L. Jessen, P. Kennelly, A. Salman, T. Rainer, and P. Hagedorn,* Surface mapping of

the Milh Kharwah salt diapir to better understand the subsurface petroleum system in the Sab'atayn Basin, onshore Yemen. Arabian Journal of Geosciences, Vol. **11**, Iss. 15, pp. 428, 2018.

[11] *A. Di Pilato, N. Taggio, A. Pompili, M. Iacobellis, A. Di Florio, D. Passarelli, and S. Samarelli,* Deep learning approaches to Earth Observation change detection. Remote Sensing, Vol. **13**, Iss. 20, pp. 4083, 2021.

[12] *A. Smith, and R. Sarlo,* Automated extraction of structural beam lines and connections from point clouds of steel buildings. Computer-Aided Civil and Infrastructure Engineering, Vol. **37**, Iss. 1, pp. 110-125, 2022.

[13] *A. Ashtekar, and J. Lewandoski,* Differential geometry on the space of connections via graphs and projective limits. Journal of Geometry and Physics, Vol. **17**, Iss. 3, pp. 191-230, 1995.

[14] *S. Lhome, C. He, C. Weber, and D. Morin,* A new approach to building identification from very-high-spatial-resolution images. International Journal of Remote Sensing, Vol. **30**, Iss. 5, 1341-1354, 2009.

[15] *Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner,* Gradient-based learning applied to document recognition. Proceedings of the IEEE, Vol. **86**, Iss. 11, pp. 2278-2324, 1998.

[16] *S. Saha, Y. T. Correa, F. Bovolo, and L. Bruzzone,* Unsupervised deep learning based change detection in Sentinel-2 images. In 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), IEEE, pp. 1-4, 2019.

[17] *Z. Guo, X. Shao, Y. Xu, H. Miyazaki, W. Ohira, and R. Shibasaki,* Identification of village building via Google Earth images and supervised machine learning methods. Remote Sensing, Vol. **8**, Iss. 4, 271, 2016.

[18] *S. Tang, M. Yang, and J. Bai,* Detection of pulmonary nodules based on a multiscale feature 3D U-Net convolutional neural network of transfer learning. PLoS One, Vol. **15**, Iss. 8, pp. e0235672, 2020.

[19] *F. Liu, J. Luo, B. Huan, H. Yang, X. Hu, and N. Xian,* Building extraction based on SE-UNet. Journal of Geo-information Science, Vol. **21**, Iss. 11, pp. 1779-1789, 2019.