Research on High Altitude Remote Sensing Building Segmentation Based on Improved U-Net Algorithm

SHI Mengyuan, GAO Junchai

(*College of Electronic and Information Engineering*, *Xi'an Technological University*, *Xi'an* 710021, *China*)

Abstract: Building extraction from high resolution remote sensing image is a key technology of digital city construction[14]. In order to solve the problems of low efficiency and low precision of traditional remote sensing image segmentation, an improved U-Net network structure is adopted in this paper. Firstly, in order to extract efficient building characteristic information, FPN structure was introduced to improve the ability of integrating multi-scale information in U-Net model; Secondly, to solve the problem that feature information weakens with the deepening of network depth, an efficient residual block network is introduced; Finally, In order to better distinguish the target area and background area in the image and improve the precision of building target edge detection, the cross entropy loss and Dice loss were linearly combined and weighted. Experimental results show that the algorithm can improve the image segmentation effect and improve the image accuracy by 18%.

Keywords: Remote Sensing Image, FPN, Efficient Residual Block Networks, Loss Function

1 Introduction

High-resolution remote sensing image is an important national land resource, and it is of great significance to use it to extract buildings for economic forecast, urban digital construction and national defense construction[1, 9].

Extracting buildings from remote sensing images is essentially assigning semantic markers to each pixel[4]. In the traditional feature extraction of remote sensing image, the feature information of the image is generally obtained by artificial or machine learning methods, and then segmented and extracted[6, 8].

Cao[3] etc., they are problems in high-resolution remote sensing images, such as the large number of ground objects and the complex feature information, the segmentation edge is not clear and the object details are lost. In the pre-processing stage before image segmentation, the initial segmented image is obtained by using superpixel segmentation technology. In the process of region merging, objects are merged based on the heterogeneity among objects and the homogeneity within objects, combining with spectral, texture and shape features. The final image segmentation result is obtained by adjusting the global segmentation parameters to adjust the merging scale. Sun[4] etc., encoder-decoder frameworks are popular in semantic image segmentation, however, encoder-decoder models face two main problems. The one is structural stereotype which is receptive fields imbalance rooted in this kind of frameworks. The other is insufficient learning that deeper neural networks tend to encounter the notorious problem of vanishing gradients, suppress the adverse consequences of structural stereotype as far as possible. To alleviate the problem of insufficient learning, we propose a novel residual architecture for encoder-decoder models. Zhou[10] etc., since the traditional deep convolutional neural net-

work segmentation of high-resolution remote sensing images requires manual design of network architecture, it is excessively dependent on expert experience, time-consuming and laborious, and has poor network generalization ability. The resource balance item is added to the network architecture parameters to improve the stability of the search algorithm and reduce the update imbalance and discretization error generated in the pruning process. Secondly, some channels are selected for mixed operation of search space to save computing resources, improve search efficiency and alleviate network overfitting. Finally, according to the features of high resolution remote sensing image, such as complex ground objects, discrete distribution and wide space range, the Gumbel-Softmax Trick method is introduced to sample from discontinuous probability distribution to improve sampling efficiency.

U-Net is one of the most widely used convolutional neural networks. It obtains the eigenvalues of images through continuous convolution and pooling, and then restores the images by deconvolution[7]. At the same time, U-Net model combines the characteristics of coding and decoding as well as jumping network, and there is a mapping relationship between the extended and reduced network[4]. In the process of expansion, the corresponding shrink layer features are combined to compensate for the lost boundary information, so as to improve the prediction accuracy of the boundary information[17].

In this paper, buildings in high resolution remote sensing images are studied, Due to the diversity of building types, complex texture, irregular spatial distribution and other particularities, there are some problems such as missed detection, mistaken segmentation, fuzzy edge and incomplete segmentation of buildings. In this paper, a new U-NET network model is proposed[18], which adds FPN structure of multi-scale information to the network structure, to enrich the required feature information of each pixel in classification[11]. In a meanwhile, at the bottom of the network, using the efficient residual block network model can effectively reduce the attenuation of feature information caused by the deepening of the network[14]. Finally, two linear weighted combinations based on cross entropy loss and Dice loss are used to optimize the boundary detection of buildings[12]. Experimental results show that the improved method can improve the precision of building boundary segmentation and achieve better segmentation results.

2 Research Techniques and Methods

2.1 The U-Net Model

Firstly, the U-Net network technology is introduced into the cell segmentation of medical image, and good segmentation results are obtained. The main feature of the U-Net network is the use of a skip connection structure similar to that shown in Fig.1. In the coding part, its feature graph is superimposed on the corresponding network structure, and then the inverse convolution is carried out. Then, the feature information obtained from the encoder is directly imported into the corresponding decoder by using the jump connection layer, so that the system can get higher-level information in the process of up-sampling and deconvolution, such as contour information, location information, so as to provide more scale information for the subsequent image segmentation[17].

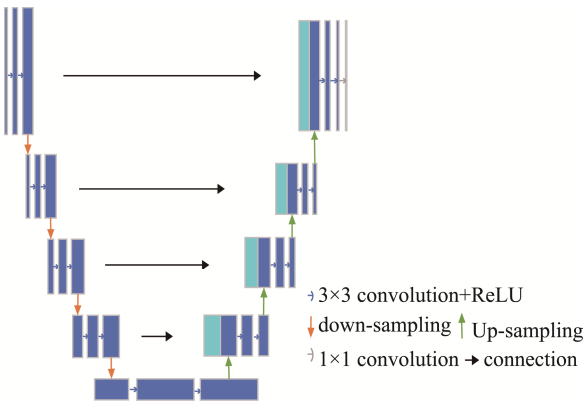


Fig.1 U-Net Network

Based on the U-Net model, this paper improves and optimizes it for the following reasons:

(1) The U-Net model is simple in construction and easy to improve, perfect and expand. Many scholars have made various modifications on the U-Net model

in the past, which facilitates the improvement and use of the algorithm in this paper. Moreover, U-Net has the advantages of fewer training parameters and short training cycle[6].

(2) U-Net model is mainly used for image segmentation. Through U-Net structure, images with small data set can be enriched and trained to achieve good segmentation effect. Since the data used in this method are all buildings based on satellite images, there is little data available for public use, the U-Net has very low requirements for training samples. In a sense, it can effectively overcome the problem of unsatisfactory model training effect caused by limited data.

2.2 Improved U-Net Model

Based on the structure of U-Net convolutional neural network, the linear combination method of FPN structure, efficient residual block and loss function is introduced. The improved U-Net network model is used to segment remote sensing building images. Its network structure is shown in Fig.2.

In this network model, FPN structure is added in the coding stage of network structure to obtain more scale characteristic information of remote sensing image data, and the upsampling of decoder is allowed to access the input information of the whole coding; Secondly, this paper proposes an efficient residual convolution block construction based on the empty convolution method, which aims to expand the perception domain, fully reflect multidimensional features, ensure the existence of gradient, and suppress gradient diffusion to some extent, so as to accelerate the convergence of the network [7]; In addition, cross entropy loss and Dice loss function are added to the loss function to improve the precision of building target edge detection.

2.2.1 FPN Structure

FPN structure, also known as feature pyramid, is widely used in target detection. FPN structure can effectively integrate multi-scale semantic information. FPN is mainly composed of three parts: bottom up, top down and horizontal connection [11], its structure is shown in Fig.3.

On the far left is the bottom-up process, that is, feature information of each scale is extracted from the convolutional neural network. In this process, the top-down method is adopted to sample the extracted feature data, so as to ensure that the upper convolutional features obtained have features similar to the bottom features[11]. On this basis, the information fusion features of multiple levels can be obtained by superposing the corresponding elements of the high-level feature map and the low-level feature map transmitted horizontally. Due to the small amount of convolution that the low feature layer undergoes, it contains a large number of texture features, while the high feature layer has better semantic characteristics on

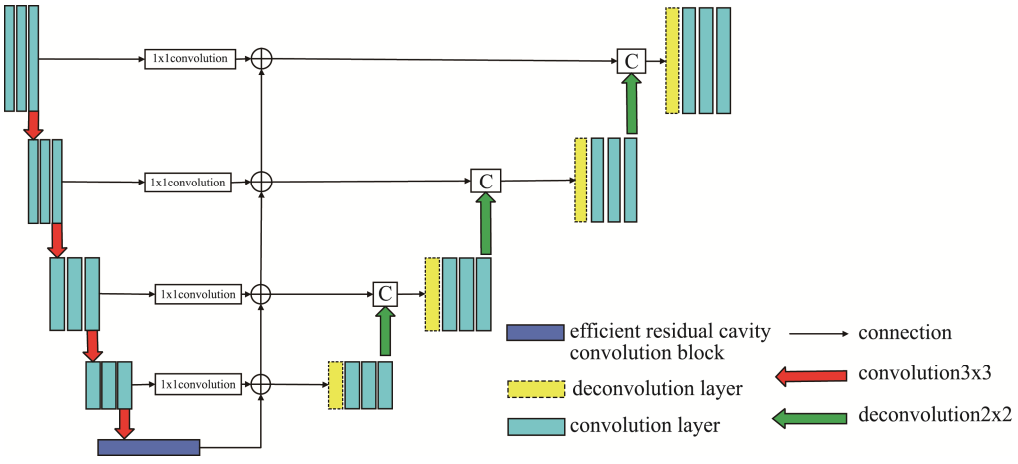


Fig.2 Improved U-NET Network Structure Diagram

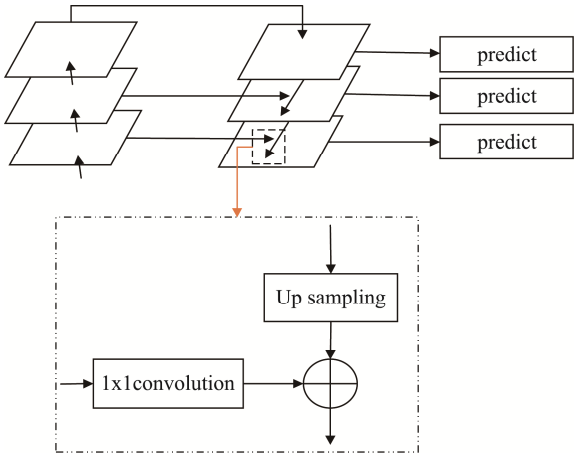


Fig.3 FPA Network Structure

multiple convolutional filtering operations. On this way, the feature of the upper layer is extracted to the same size as the feature of the lower layer, and the phase sum operation is carried out, and the texture of the bottom layer is used to supplement the feature of the upper layer, so as to enrich the characteristics of the fused image[16].

In addition, the upsampling of FPN adopts quadratic linear interpolation, because quadratic linear interpolation is a linear operation, without changing the original shallow level information, only linear interpolation method to increase the size of the feature graph, and the corresponding addition of corresponding elements ensures that the characteristics transmitted to each decoder layer is the addition and fusion. Compared with the conventional U-Net deconvolution operation, it can effectively prevent the data loss caused by deconvolution, retain more complete characteristic data information, and ensure the multi-dimensional data information of all levels contained in the decoder. Through the analysis of FPN and U-Net, we can see that the cross connection between them can be converted to each other. This characteristic is beneficial to the improvement of U-Net model by FPN structure. On the basis of making full use of U-Net network architecture, U-Net gives full play to all kinds of scale information.

2.2.2 Efficient Residual Network

In residual networks, residual learning modules are introduced. It consists of two parts, namely direct mapping and residual[14], its structure is shown in Fig.4.

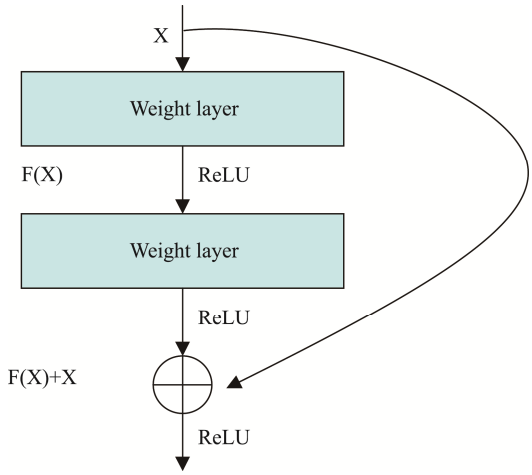


Fig.4 Residual Network

It can be seen from Fig.4 that the residual network has two layers. Such as the type (1):

$$F = W_2\sigma(x_l W_l) \tag{1}$$

σ represents the nonlinear activation function ReLU.

Then through the "quick way" and the second activation function ReLU, the output y is obtained. Such as the type (2):

$$y = F(x,\{W_i\}) + X \tag{2}$$

Thus, it can be seen that the difference between residual network and ordinary convolutional network is that the general convolutional network has problems such as incomplete information and long time consuming in the process of information transmission. However, residual network can realize object segmentation by learning residual information in the case of deep learning difficulty.

In this paper, by introducing three residual modules in the fifth layer of the network, on the one hand, the network information is not smooth, and on the other hand, the problem of poor segmentation effect caused by network degradation is solved.

2.2.3 Loss Function

The linear combination of cross entropy and Dice

loss can effectively overcome the problem that the target features of remote sensing image data set are not significant, improve the data imbalance and improve the precision of target segmentation.

Considering the loss degree of the whole image, more attention is paid to the loss change of the target building, so the segmentation error based on the feature region can be effectively eliminated[17].

The loss rate of Dice's loss function is L_D, the expression is shown in (3),

$$L_D = -\frac{\sum_{n=1}^{N}(p_n \times r_n) + \varepsilon \sum_{n=1}^{N}(1-p_n)(1-r_n) + \varepsilon}{\sum_{n=1}^{N}(p_n \times r_n) + \varepsilon \sum_{n=1}^{N}(2-p_n 1-r_n) + \varepsilon} \quad (3)$$

The loss rate of BCE loss function is L_B, the expression is shown in (4),

$$L_B = -\frac{1}{n}\sum_{n=1}^{N}(y_n \times \ln x_n + (1-y_n) \times \ln(1-x_n)) \quad (4)$$

The loss rate of linear combination loss function is L_{DB}, the expression is shown in (5),

$$L_{DB} = 0.5L_B + L_D \quad (5)$$

N indicates the number of all pixels of the segmented image; P represents the number of foreground pixels in the training set; R represents the actual number of foreground pixels in the training set; N indicates the number of training round; x_n is the independent variable of loss function training; y_n is the gradient factor of gradient change.

3 Experimental Results and Analysis

3.1 The Data Set

In order to verify the effectiveness of the improved network structure, the high resolution remote sensing image buildings published by the Institute of National Institute for Information and Automation in France are used as data sets to carry out validation experiments. The data coverage of this dataset is up to 810km^2, which the image types are aviation forward full color image and a resolution of 0.3m, covering many different terrains, for example, buildings in dense residential areas and some mountainous areas basically include all types of common buildings.

To maximize the calculation of the experimental platform operation performance, this study will be the original image is cut into 512 x 512 pixel resolution of sample images, a total of 387 pieces, and according to the proportion, reviews the research data and combining with the actual situation to make minor adjustments, which speak about according to the proportion of the data, then 336 training sets of remote sensing buildings covering 810 km^2 with a resolution of 512 x 512 and 51 corresponding test sets are obtained.

3.2 Experimental Environment

In this paper, Python programming language was adopted under PyCharm, Pytorch was used as the network development structure, and an improved U-Net model was used as the segmentation network to achieve segmentation of image data sets of different buildings. Finally, a conclusion was obtained through simulation experiment.

3.3 Model Training and Evaluation Indicators

The total set parameters of this experiment are as follows: 200 training rounds, batch size set as 16, commonly used Adam optimizer, learning rate 0.001, its results is shown in Fig.5[20].

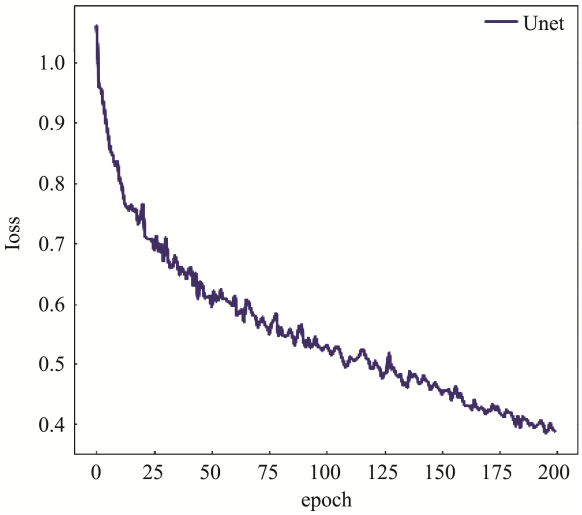


Fig.5 Loss Diagram

When performing semantic segmentation tasks, accuracy, Recall and F1-SOCre are generally used to evaluate the effect of the model indirectly. Each indicator is expressed as follows[22]:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{7}$$

$$F_1 = \frac{2RP}{R + P} \tag{8}$$

Here, TP is the positive sample of the correct prediction, FP is the positive sample of the wrong prediction, FN is the negative sample of the prediction error[18][19]. F1-socre is a measure of accuracy and recall rate, and is the harmonic average between the two side。

In this paper, in order to compare the detection effects of different algorithms on data sets, overall Pixel Accuracy (PA) and intersection ratio (IMoU) were selected to directly evaluate the algorithms. Where, the PA expression is shown in Equation (9):

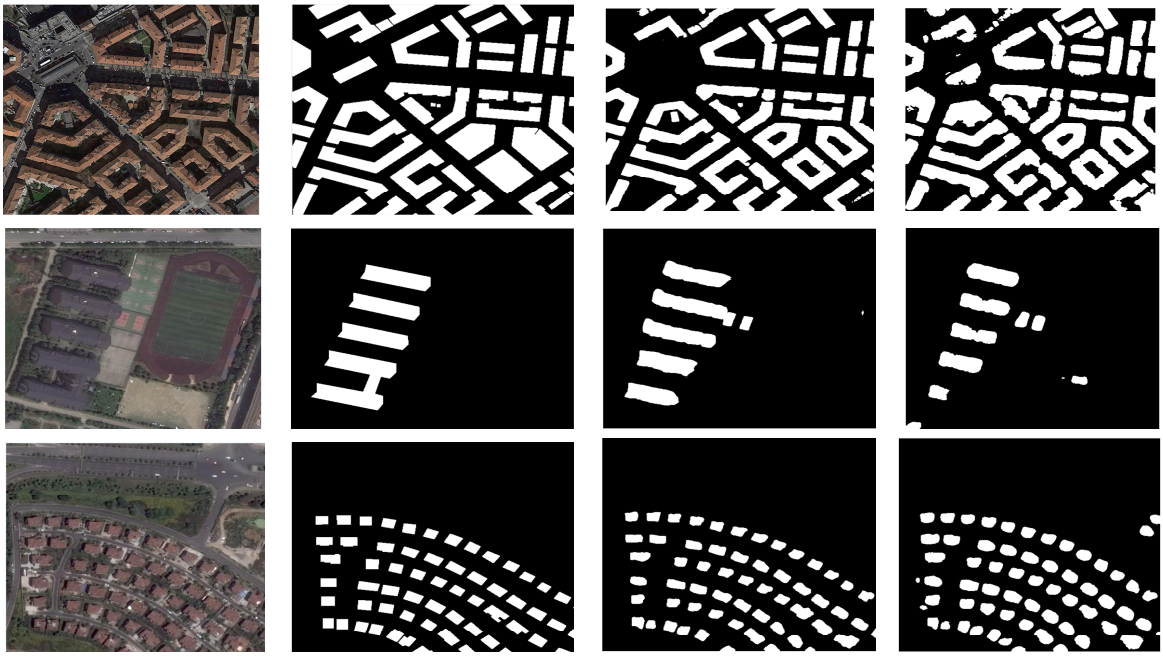
$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

As a standard measure of semantic segmentation, IMoU is an IoU based on classification and weighted average of the IoU to obtain a hole-based IMoU, the IMoU expression is shown in Equation (10):

$$IMoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN + FP + TP} \tag{10}$$

3.4 Analysis of Experimental Results

In this paper, by comparing the accuracy of U-Net and the improved algorithm in building segmentation of remote sensing image, the experimental result graph can be obtained as shown in Fig.6.



(a) The Original Image (b) Mask (c) Improved U-Net (d) U-Net

Fig.6 Experimental Results

Table 1 Evaluation Indexes of Different Algorithms

The Evaluation Index	IMoU (Building and Background)	IMoU (Building)	Time/s
U-Net	80.58	0.76	0.298
Improved U-Net	80.82	0.94	0.295

4 Conclusion

In this paper, an improved U-Net algorithm is proposed to solve the problems of low efficiency and insufficient segmentation of existing remote sensing building images. Firstly, by modifying the backbone network structure, FPN is used as the backbone network to improve the extraction and integration of building characteristic information; then, efficient residual block network is used to solve the problem of feature information weakening caused by network depth; finally, the fusion loss function is used to achieve the detection accuracy of building edge. From the experiment, it can be known that the improved network is more accurate for edge detection of building objects, and it can be seen from the comparison in Fig.5 (c) and (d) that the segmentation precision of the cutting network before and after the improvement on building edge contour has a great difference, and the precision of the improved network on building contour segmentation is as high as 80.82% (including the background part). Compared with the traditional unimproved segmentation network, the accuracy of the improved algorithm is improved by 18%, and the running speed of the improved algorithm is also improved. Therefore, the method proposed in this paper has certain reference value for remote sensing image segmentation.

References

[1] Wang C, Shi A Y, Gu A H, Xiong X B, Liu Q. A city high-resolution remote sensing image segmentation method combined with shadow compensation [J]. Journal of Electronic Measurement and Instrumentation, 2017, 31(10): 1687-1692.

[2] Xiang Z J, Can F S, Chu H, Huang L. High resolution remote sensing image segmentation algorithm based on superpixel [J]. Computer Engineering and Design, 2020, 41(05): 1379-1384.

[3] Cao X, Song C G, Zhang J, Liu C. Remote Sensing Image Segmentation based on Generative Adversarial Network with Wasserstein divergence[C]//. Conference Proceeding of 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2020)., 2020: 342-347.

[4] Sun Y, Tian Y, Xu Y P. Problems of Encoder-Decoder Frameworks for High-Resolution Remote Sensing Image Segmentation: Structural Stereotype and Insufficient Learning[J]. Neurocomputing, 2018, 330.

[5] Zhu H M, Tan R, Han L T, Fan H F, Wang Z J, Du B W, Liu S C, Liu Q. DSSM: A Deep Neural Network with Spectrum Separable Module for Multi-Spectral Remote Sensing Image Segmentation[J]. Remote Sensing, 2022, 14(4).

[6] Zhang R N, Yu L, Tian S W, Lv Y L. Unsupervised remote sensing image segmentation based on a dual autoencoder[J]. Journal of Applied Remote Sensing, 2019, 13(03).

[7] Chen G S, Li C, Wei W, Jing W P, Marcin Woźniak, Tomas Blažauskas, Robertas Damaševičius. Fully Convolutional Neural Network with Augmented Atrous Spatial Pyramid Pool and Fully Connected Fusion Path for High Resolution Remote Sensing Image Segmentation[J]. Applied Sciences, 2019, 9(9).

[8] Wang Y Q, Wang C X. High resolution remote sensing image segmentation based on multi-features fusion[J]. Engineering Review: Međunarodni časopis namijenjen publiciranju originalnih istraživanja s aspekta analize konstrukcija, materijala i novih tehnologija u području strojarstva, brodogradnje, temeljnih tehničkih znanosti, elektrotehnike, računarstva i građevinarstva, 2017.

[9] Zhao Q, Zhang S, Huang S L. Multi-scale and Multi-feature High Resolution Remote Sensing Image Segmentation[J]. International Journal of Applied Mathematics and Statistics™, 2013, 51(22).

[10] Zhou P, Yang J. Remote sensing image segmentation method based on neural network architecture search is adopted [J]. Journal of Xidian University, 2021, 48(05): 47-57+77.

[11] Wang X, Yu M, Ren H E. Remote sensing image semantic segmentation based on UNET and FPN [J]. Liquid crystal and Display, 2021, 36(03): 475-483.

[12] Gong Y J, Huang M, Huang X L. Flame image segmentation method based on improved Resnet-UNET [J]. Journal of Beijing Information Science and Technology University (Natural Science Edition), 2021, 36(05): 39-44.

[13] Li L X, Yuan Y, Wen S H. Building segmentation and extraction from high resolution remote sensing image

based on BAU-NET [J]. Journal of Yanshan University, 2021, 45(04): 335-342.

[14] Wang Y, Yang Y, Wang B S, Wang T, Pu X H, Wang C Y. Deep residual neural network for building segmentation in high resolution remote sensing images [J]. Remote sensing technology and application, 2019, 34(04): 736-747.

[15] Li Y, Li Y J, Liu J C, Fan H, Wang Q L. Research on image segmentation of steel surface defects based on improved RES-UNET network [J/OL]. Journal of Electronics and Information technology: 1-8.

[16] Gu S J, Pu X Z, Jing J W, Liu S. Muzzle flame segmentation method based on improved UNet network [J]. Foreign electronic measurement technology, 2021, 40(04): 16-21.

[17] Huang S P, Liu H N, Zhou K S, Liu J Y. Zebra crossing segmentation based on improved UNet [J]. Intelligent computers and applications, 2020, 10(11): 61-64+69.

[18] Xiang Y, Zhao Y D, Dong J H. Mining area change detection based on improved UNet twin network [J]. Journal of coal, 2019, 44(12): 3773-3780.

[19] Fu L H, Zhao Y, Jiang H X, Zhao R, Wu H X, Yan S X. Semi-supervised video object segmentation based on foreground perception and visual attention [J]. Electronic journals, 2022, 50(01): 195-206.

[20] Wang X B, Cao S P, Zhao H C, Liu P F, Tai B C. Bilateral feature aggregation and attention mechanism point cloud semantic segmentation [J]. Chinese Journal of Scientific Instrument, 2021, 42(12): 175-183.

[21] Wang Z D, Guo C. A semantic segmentation optimization ORB_SLAM2 in dynamic scenarios [J]. Journal of Dalian Maritime University, 2018, 44 (04): 121-126.

Author Biographies

SHI Mengyuan is currently a M.Sc. candidate, majoring in communication and information system.
E-mail: 951051762@qq.com

GAO Junchai received a Ph.D. degree of Engineering. She is currently an associate professor. Her main research interest includes computer vision.
E-mail: 527667658@qq.com