J. Meas. Sci. Instrum., 2024, 15(3): 318-327. http://jmsi.publish.founderss.cn
DOI: 10.62756/jmsi.1674-8042.2024033 jmsi@nuc.edu.cn

Automatic road extraction framework based on codec network

WANG Lin[1], SHEN Yu[1*], ZHANG Hongguo[1], LIANG Dong[2], NIU Dongxing[2]

1. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. China Railway Scientific Research Institute Co., Ltd., Chengdu 610036, China

***Corresponding author:** SHEN Yu (17516297579@163.com)

Received: February 10, 2023 **Revised:** April 13, 2023 **Accepted:** April 25, 2023

Abstract: Road extraction based on deep learning is one of hot spots of semantic segmentation in the past decade. In this work, we proposed a framework based on codec network for automatic road extraction from remote sensing images. Firstly, a pre-trained ResNet34 was migrated to U-Net and its encoding structure was replaced to deepen the number of network layers, which reduces the error rate of road segmentation and the loss of details. Secondly, dilated convolution was used to connect the encoder and the decoder of network to expand the receptive field and retain more low-dimensional information of the image. Afterwards, the channel attention mechanism was used to select the information of the feature image obtained by up-sampling of the encoder, the weights of target features were optimized to enhance the features of target region and suppress the features of background and noise regions, and thus the feature extraction effect of the remote sensing image with complex background was optimized. Finally, an adaptive sigmoid loss function was proposed, which optimizes the imbalance between the road and the background, and makes the model reach the optimal solution. Experimental results show that compared with several semantic segmentation networks, the proposed method can greatly reduce the error rate of road segmentation and effectively improve the accuracy of road extraction from remote sensing images.

Key words: remote sensing image; road extraction; ResNet34; U-Net; channel attention mechanism; sigmoid loss function

0 Introduction

Using the image features contained in the target remote sensing image to realize automatic information extraction is a general trend of information intelligence in the 21st century[1]. Since deep learning has strong independent learning capabilities and excellent automatic extraction capabilities, it has been widely used and rapidly developed in the field of remote sensing image extraction[2,3].

Convolutional neural network (CNN) has a good performance in image-level segmentation task[4]. However, it cannot meet the requirement of segmentation accuracy of remote sensing images. Accordingly, on the basis of CNN, Long et al.[5] proposed a fully convolutional network (FCN) to complete a pixel-level image segmentation. According to different objects and focuses of the problems, some semantic segmentation networks based on FCN have been proposed successively, such as SegNet[6,7], Deeplab[8,9], U-Net[10,11] and ResNet[12,13]. Owing to its excellent performance in medical image segmentation, U-Net has attracted the attention of many scientific research teams. For example, Liu et al. [14] integrated Morphsnakes algorithm into the classic U-Net and proposed an improved U-Net for CT image segmentation, which improves the segmentation effect of image edge. Jin et al.[15] proposed a double U-Net remote sensing image extraction model and optimized some parameters. By using the double U-Net joint training method, the feature fitting ability of the network was improved, and the segmentation accuracy of remote sensing images was also improved effectively. ResU-Net was proposed by integrating the residual modules[16]. R2U-Net was obtained by combining cyclic convolution with ResU-Net[17]. Attention U-Net added attention mechanism into the up-sampling and down-sampling process of U-Net[18].

Since actual road is complex, the remote sensing image to be processed has high dimension and strong background interference. The shallow network used by classic U-Net is difficult to establish the mapping between the remote sensing image and the segmentation result, which leads to insufficient accuracy the extraction results[19]. According to the characteristics of remote sensing images of the road, we proposed an RAU-Net model, which replaces the encoder of U-Net with

ResNet34, extracts the feature information of different receptive fields, and integrates them. At the encoding stage, the fused feature information was extracted, which provides more contextual semantic information for up-sampling road image restoration, enhances the low-dimensional detail information in the network propagation, and improves the segmentation accuracy. At the decoding stage, the channel attention mechanism was integrated, the information of feature map was selected and weighted, and the original feature was recalibrated in the channel dimension, which effectively avoids the loss of details in the segmentation process, and reduces the error rate of target image extraction from the road. Finally, we used an adaptive sigmoid loss function to obtain clear boundary with high confidence for improving segmentation accuracy, as well as a nonlinear activation function, exponential linear unit (ELU), to classify pixel points one by one for improving pixel-level classification effect.

1 Basic networks

1.1 U-Net

U-Net is a classical network model which combines low-dimensional features and high-dimensional features. It is composed of feature encoder and feature decoder. The network structure of U-Net is shown in Fig.1.

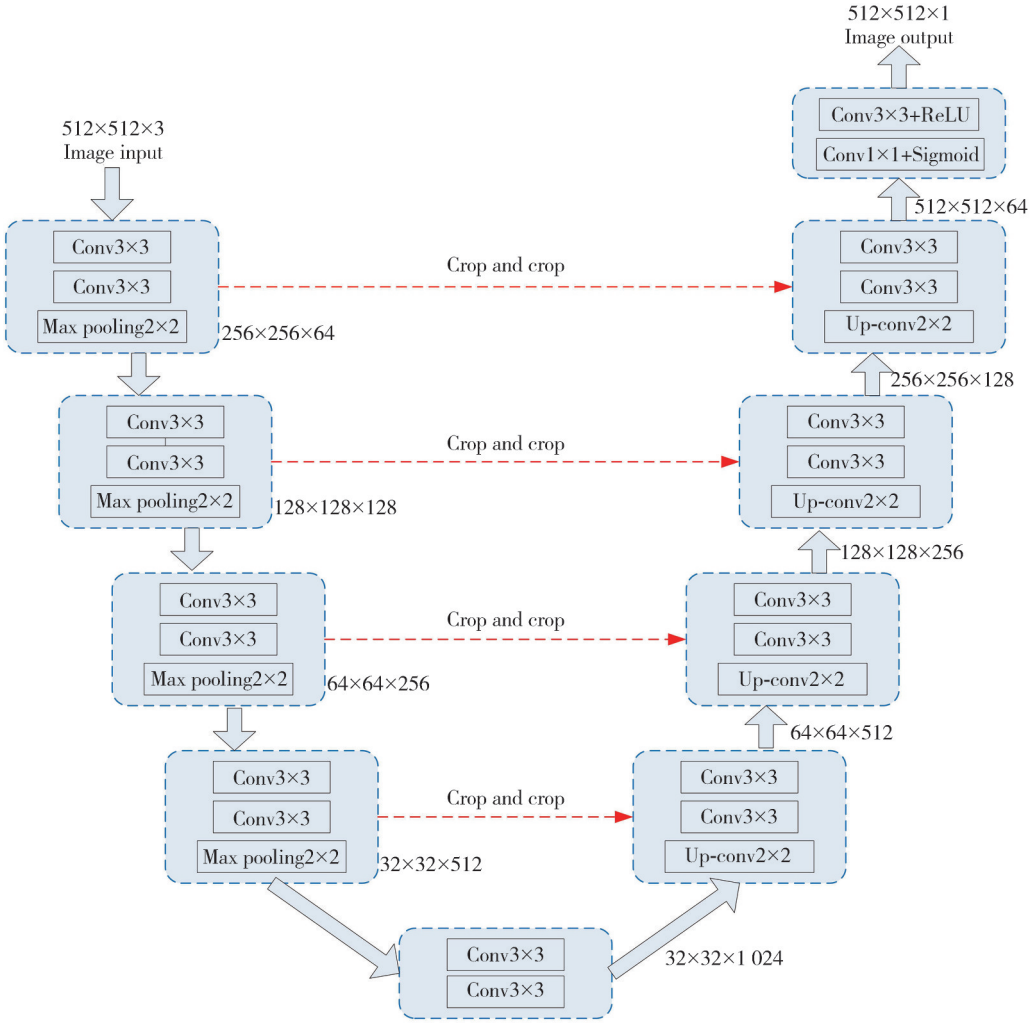

Fig. 1 U-Net feature extraction model

The front end of the network is the feature encoder, which is constructed based on VGG16[20,21], and high-dimensional image features with a reduction of image size are obtained via convolution and pooling operations. The rear end of the network is the feature decoder, which contains two 3×3 convolutions and one 2×2 deconvolution corresponding to the front end. In addition to inputting the deep abstract features obtained by up-sampling in the previous layer, it also inputs the shallow local features of the corresponding down-sampling output. The deep features are fused with the shallow features, and the detailed information of low-dimensional features at each level is introduced to ensure that the spatial dimension of information remains

unchanged.

1.2 ResNet34

On the premise that deep network can converge, with the increase of network depth, the accuracy rate will tend to be saturate or evenly decline, which has become a barrier to the development of deep neural network. ResNet34 is also inspired by the VGG network and the specific structure is shown in Fig.2.

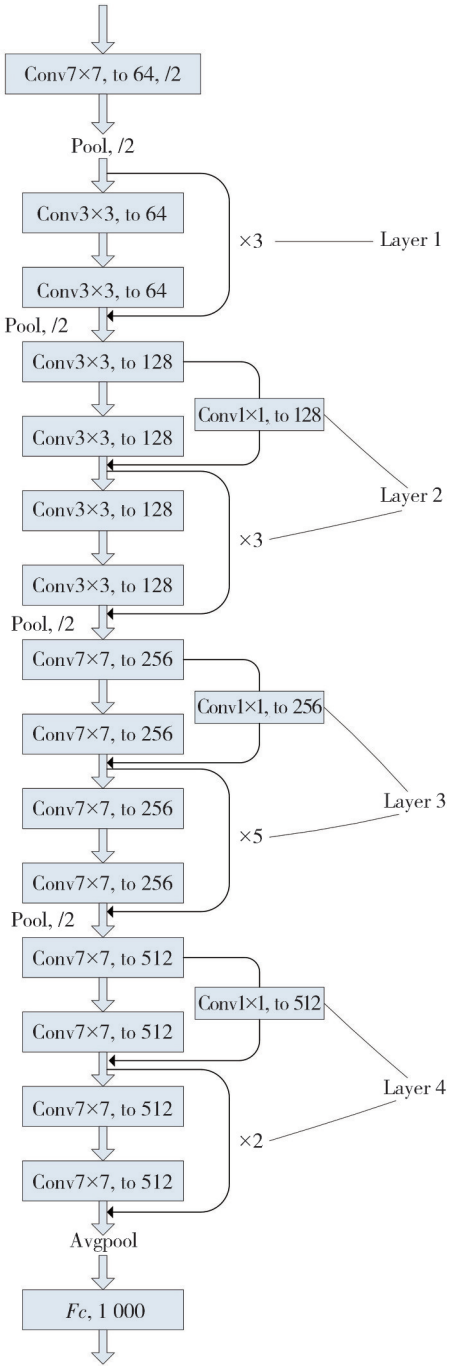

Fig. 2 ResNet34 feature extraction model

The ResNet34 mainly adopts a 3×3 convolution with step size of 2 for down-sampling, and the redundant items of the network are mapped identically by jumping connection, which makes the weight convergence of deep network more effective. The network ends with a global average pooling layer and a 1 000-dimensional fully connected layer with softmax, and the total number of weighted layers is 34. The whole network follows two simple design rules: Firstly, for the same output feature map size, each layer adopts the same convolutional layer; Secondly, if the size of the feature map is halved, the number of convolutional layers is doubled to keep the time complexity of each layer. The residual structure of ResNet34 can accelerate the training of deep neural network and improve the accuracy of the model.

1.3 Sequeeze-and-excitation block

The channel attention mechanism, which was proposed by Hu et al. in CVPR in 2018[22], is also called sequeeze-and-excitation block (SE block), as shown in Fig.3.

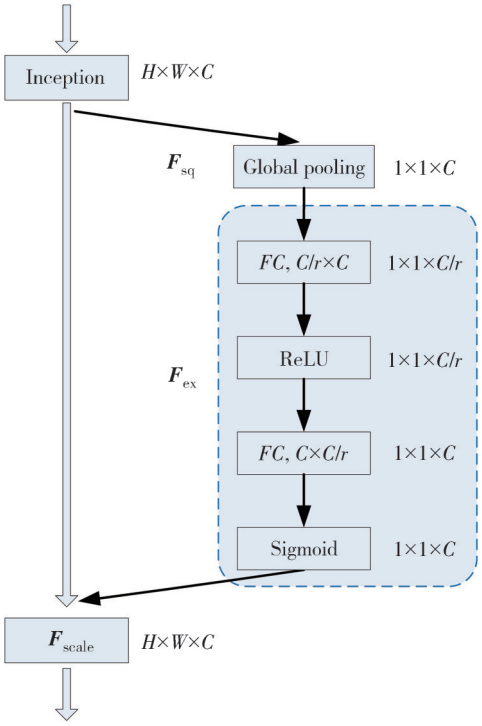

Fig. 3 SE block

It is a process of recalibrating the feature map. SE block is not a complete network structure but a substructure that can be embedded into other classification or detection models. The core idea is to learn feature weights based on the loss via the network, make the network pay attention to important channel features, filter irrelevant channel features, and recalibrate the features. Generally speaking, it is to make the network focus on a certain aspect of the image like human vision, and reduce the attention of other unconcerned information. In addition, exploring the areas with the most information in the image and more important channel features can improve the effect of sentiment

classification to a certain extent.

It can be seen from Fig. 3 that SE block is mainly composed of global average pooling (Squeeze), adaptive recalibration (Excitation), and feature fusion (Scale). The specific process is as follows. Firstly, the feature map is pooled by a 2×2 global average, and each 2-dimensional feature channel is compressed and mapped into a feature map with a dimension of $1\times1\times C$ consistent with the input structure to obtain global information. Then, the feature map passes through a fully connected layer and ReLU function, the dimension becomes $1\times1\times C/r$, where r is a scaling parameter and can be set at 16 or other values. The use of this parameter is to reduce the number of channels and thus to reduce the amount of calculation. Afterwards, the dimension is restored to $1\times1\times C$, and the weights of the two fully connected layers are reciprocal, which can integrate local information with category discrimination in the pooling layer. Next, the sigmoid function is used to get C weights of the feature map, where C represents the number of channels and is obtained by learning the previous fully connected layer and nonlinear layers. Finally, the probability map of feature excitation output is weighted by multiplication to the previous feature map by channel, and the original feature is recalibrated in the channel dimension to enrich the extracted information.

2 Proposed RAU-Net

2.1 Overall network architecture

Since the encoding structure of U-Net is actually a process of feature extraction and is highly similar to VGG16, the pre-trained ResNet34 is used as the encoding structure of U-Net that can extract features better via migration learning. At the decoding stage, RAU-Net with SE block is proposed. The overall network architecture is shown in Fig.4.

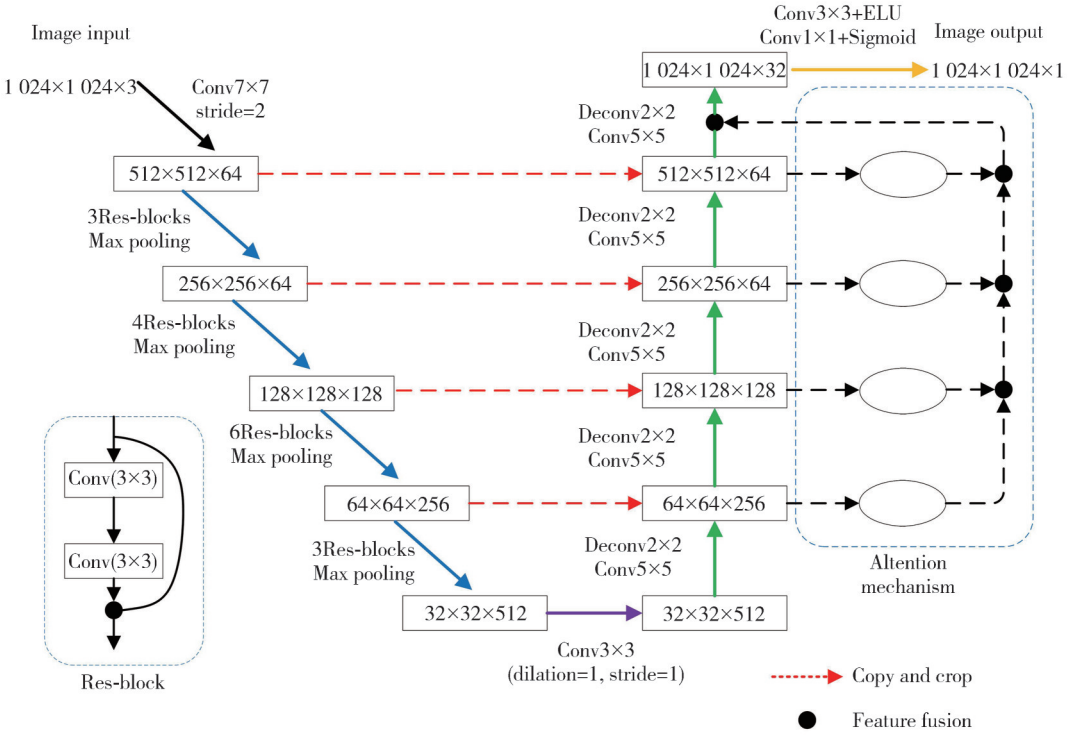

Fig. 4 RAU-Net network architecture

The convolution mode of RAU-Net adopts the same mode as that of U-Net, which makes the convolution operation start in case that the center of the convolution kernel coincides with the corners of the feature image. This mode keeps the size of feature map unchanged in the process of forward propagation, and makes it unnecessary for parameter adjustment to accurately calculate the change in image size.

In the RAU-Net architecture, each gray rectangular block represents a multi-channel feature map. The left part is the encoder, ResNet34, which is mainly composed of four encoding layers. The numbers of feature channels of four encoding output images are 64, 128, 256 and 512, respectively, which are correspondingly expanded to 8 times before encoding. After each convolution, pooling is performed to reduce the image size and improve the visual perception field. From low-level to high-level, the feature map is reduced

from 1 024×1 024 pixels to 32×32 pixels. The right part is the decoder of RAU-Net. The settings are the same as that of U-Net, but the original two 3×3 convolutional layers are replaced with one 5×5 convolutional layer, which reduces the complexity of neural network while maintaining segmentation accuracy. The encoder and the decoder are connected by dilated convolution, which adjusts the receptive field of feature points without reducing the resolution of feature map and with detailed spatial information. Except for the last convolution layer activated with sigmoid, there is an ELU activation behind each convolution layer.

RAU-Net adds channel attention mechanism in the process of feature encoding, selects information from high-level feature map, and obtains new feature map through weighting. It enhances emotional semantic information, retains richer features and more discriminative detailed information, and realizes the recalibration of features. The integration of channel attention mechanism enables the network to imitate the characteristics of human vision's selective attention to things. It improves the acquisition of pixel features of the road target images, and enables the network to recognize the target information on the road in advance, so as to achieve the purpose of quickly and correctly extracting the required information.

2.2 Activation function

ReLU function is often used as activation function in semantic segmentation networks, and it is defined as

$$f(x)=\max(0,x). \tag{1}$$

ReLU function almost exponentially increases in the process of forward use and greatly saves the time of calculation. However, if the input value is negative, the gradient will drop to zero and the neuron will be inactivated, and thus activation function cannot play the role.

In this study, ELU function is used to replace ReLU function, and it is defined as

$$\begin{cases} f_1(x)=\alpha(\mathrm{e}^x-1), & x\in(-\infty,0], \\ f_2(x)=x, & x\in(0,\infty), \end{cases} \tag{2}$$

where α is an adjustable parameter, which controls the average of the input activation values at 0 when the negative part of ELU function is saturated. Eq. (2) is an improvement of Eq. (1) by adding nonlinear factors and integrating the characteristics of sigmoid and ReLU, which can alleviate the gradient disappearance, improve the robustness of noise, and solve the problem of neurons inactivation of negative input[23]. With the addition of nonlinear factors, the model constructed by ELU function has a higher accuracy in the process of road target extraction from remote sensing images.

2.3 Adaptive loss function

Road target extraction is a binary classification problem. During the training process, there is an imbalance in the proportion of road and background. For example, the proportion of background is close to 90%, while the road only accounts for 10%. The traditional loss function uses an average weighting method to process each pixel. In this case, training efficiency is low, because the negative samples account for most of all training samples and do not contribute useful learning signals.

To solve this imbalance problem, the standard binary classification cross entropy loss function[24,25], sigmoid cross entropy loss (SCE Loss), is improved by redistributing the weights of positive and negative samples, and an adaptive sigmoid loss function is proposed. Its value is a measure of the level of system chaos, which is linearly related to the variability of the system. It represents the difference between the predicted probability distribution and the true probability distribution, and the smaller the difference, the better the constructed model. When the difference is 0, the predicted value is completely consistent with the true value. Based on this, the neural network will constantly adjust the weight parameters, making the value gradually tend to zero.

The sigmoid function converts the output into a probability map, which is defined as

$$p(x)=\frac{1}{1+\mathrm{e}^{-x}}, \tag{3}$$

where $p(x)$ is the probability of the pixel in the positive sample, and it is always a number between 0 and 1.

The adaptive sigmoid loss function for road detection is defined as

$$\begin{aligned} ASLoss(y_S,f_S(x))=&\alpha_{S_n}\sum_{i=1}^{M}\sum_{j=1}^{N_0}-(1-y_i^j)\log(1-\\ &p_1(x_i^j))+\alpha_{S_p}\sum_{i=1}^{M}\sum_{j=1}^{N_1}-y_i^j\log p_1(x_i^j), \end{aligned} \tag{4}$$

where x_i^j is the jth pixel of the ith image; y_i^j and $f_S(x)$ represent the corresponding label and the output of the last group, respectively; M is the minimum batch size; N_0 and N_1 represent the number of background pixels and the number of road pixels in each image, respectively; $p_1(x_i^j)$ represents the probability of the road pixels; α_{S_n}

and α_{S_p} represent the proportion of road and background in the training samples, respectively, and they are calculated by

$$\begin{cases} \alpha_{S_n} = \dfrac{\alpha_{S_1}}{\alpha_{S_1} + \alpha_{S_0}}, \\ \alpha_{S_p} = \dfrac{\alpha_{S_0}}{\alpha_{S_1} + \alpha_{S_0}}, \end{cases} \tag{5}$$

where α_{S_0} and α_{S_1} represent the number of background pixels and road pixels in all training samples, respectively. In this way, the weights can be automatically adjusted according to the ratio of positive samples and negative samples before training, which makes the loss contribution rates of the two types are more appropriate.

3 Experimental results and analysis

3.1 Experimental setup and preprocessing

The experimental hardware configuration was I9-9900F CPU, NVIDIA GeForceGTX2080Ti (32 GB) GPU, the operating system was 64-bit Ubuntu, and the deep learning framework was Tensonflow. In the experiment, the proposed RAU-Net model was trained by batch processing on DeepGlobe road extraction data set[26], of which 80% of the pictures were randomly selected as training set, and other pictures as verification set and test set. Each batch there were 160 images randomly selected to input the model for training. The learning rate was initially set to be 2×10^{-4}, minus 5 for 3 times, and then we observe the slow decrease of training loss. At the preprocessing stage, the images from the original data set were divided with an effective image with 1 024×1 024 pixels. In addition, by manual screening, the images with less than 20% of road in the divided remote sensing images were removed from the corresponding ground truth image. The roads were marked as the foreground, and other targets were marked as the background.

3.2 Evaluation index

In this study, four indexes, *accuracy*, *recall*, intersection over union (*IoU*) and *kappa* coefficient, were used to quantitatively evaluate extraction results, and they are calculated by

$$IoU = \frac{L_P}{L_P + E_P + E_N}, \tag{6}$$

where *IoU* reflects the coincidence degree between the predicted image and the ground truth image; L_P is the number of total pixels that are correctly recognized as the road by the model; E_P is the number of the pixels that mistakenly regard the background as the road. E_N is the number of the pixels that mistakenly regard the road as the background.

$$accuracy = \frac{L_P}{L_P + E_P}, \tag{7}$$

where *accuracy* represents the proportion of pixels with correct prediction in the total pixels.

$$recall = \frac{L_P}{L_P + E_N}, \tag{8}$$

where *recall* indicates how many pixels in the image are correctly predicted.

$$kappa = \frac{I_{CC} - I_{RE}}{1 - I_{RE}}, \tag{9}$$

where *kappa* stands for the consistency between the extracted result and the true value, I_{CC} is the overall recognition accuracy, and I_{RE} is the accidental consistency.

$$\begin{aligned} I_{CC} &= \frac{L_P + L_N}{M}, \\ I_{RE} &= \frac{(L_P + E_P)M_C + (E_N + L_N)M_U}{M^2}, \end{aligned} \tag{10}$$

where L_N is the number of total pixels correctly identified as the road by the model; M is the number of the total pixels recognized by the model; M_C represents the number of real road pixels in the sample, and M_U represents the number of pixels in the sample that are the real background.

3.3 Analysis of experimental results

In the experiment, the proposed network model RAU-Net corresponding to the balance between *accuracy* and *recall* was selected as the final parameter model, and it was compared with FCN8s, ResNet34, DeeplabV2, U-Net and Attention U-Net for evaluation and analysis. The MSRA initialization method[27] was used for parameter initialization, which only considers the number of inputs, and the weight initialization obeys a Gaussian distribution with the mean value of 0 and the variance of $2/n$ (where n is the number of inputs). The experimental evaluation indexes are shown in Table 1. The comparison histograms of evaluation indexes are shown in Figs.5—8.

Table 1 Evaluation indicators of networks

Method	*IoU*	*accuracy*	*recall*	*kappa*
FCN8s	0.457	0.655	0.692	0.736
ResNet34	0.482	0.784	0.671	0.745
U-Net	0.563	0.816	0.714	0.814
DeeplabV2	0.604	0.822	0.812	0.835
AttentionU	0.609	0.818	0.805	0.826
RAU-Net	**0.623**	**0.849**	**0.823**	**0.863**

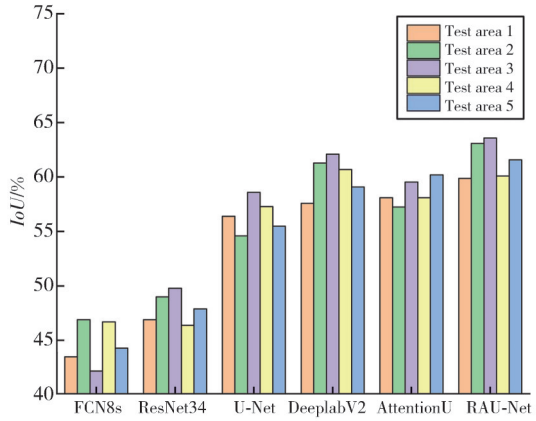

Fig. 5 Comparison of *IoU* results

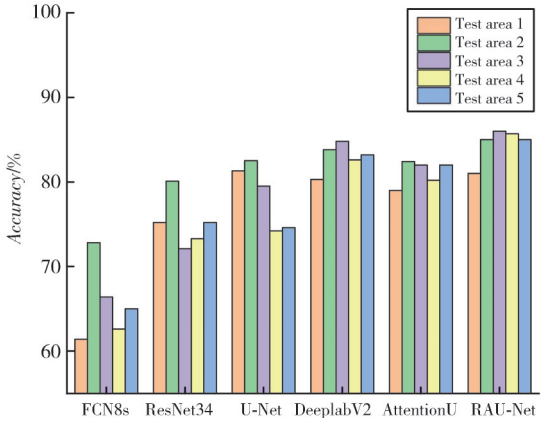

Fig. 6 Comparison of *accuracy* results

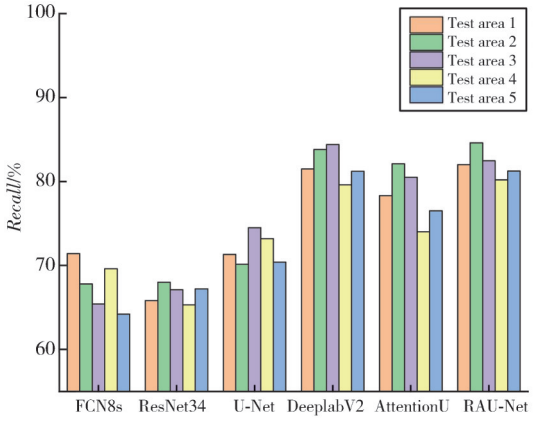

Fig. 7 Comparison of *recall* results

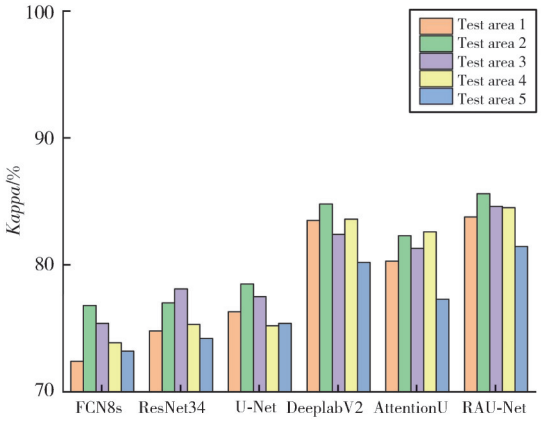

Fig. 8 Comparison of *kappa* results

It can be seen from Table 1 and Figs. 5 — 8 that ResNet34 has a significant improvement in accuracy compared with FCN8s. Compared with ResNet34, U-Net has improved in four evaluation indicators, and the *kappa* coefficient has improved significantly, but the improvement in accuracy is not obvious. Since DeeplabV2 introduces dilated convolution, Attention U-Net introduces attention mechanism, both of which are better than U-Net. Compared with other semantic segmentation networks, the proposed RAU-Net greatly optimizes the accuracy and has advantages in all evaluation indicators, which objectively proves the effectiveness of this method.

Figs. 9 — 13 are the segmentation effect diagrams by the networks in different situations.

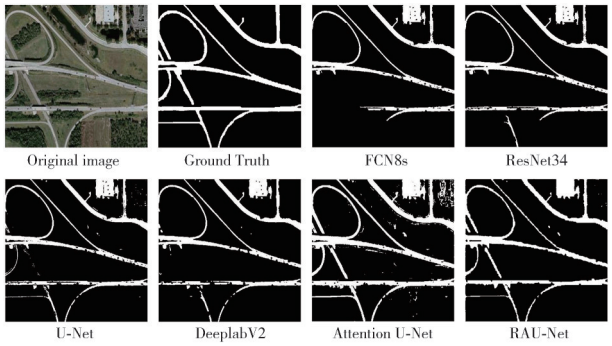

Fig. 9 Comparison results of road extraction for different network models in area 1

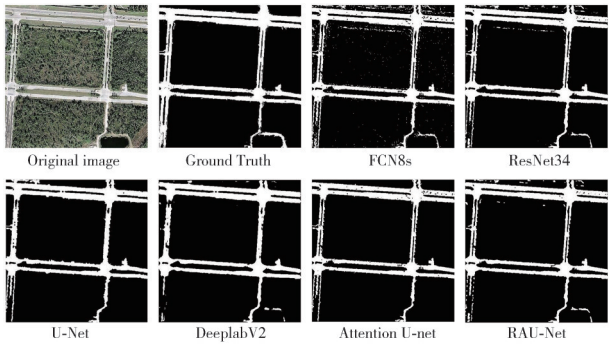

Fig. 10 Comparison results of road extraction for different network models in area 2

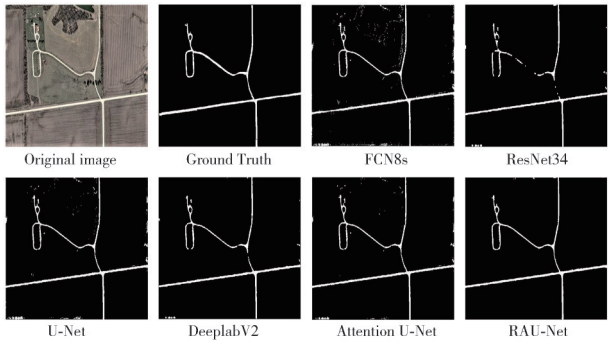

Fig. 11 Comparison results of road extraction for different network models in area 3

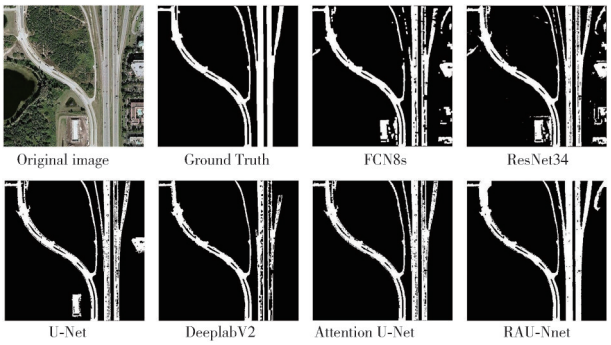

Fig. 12 Comparison results of road extraction for different network models in area 4

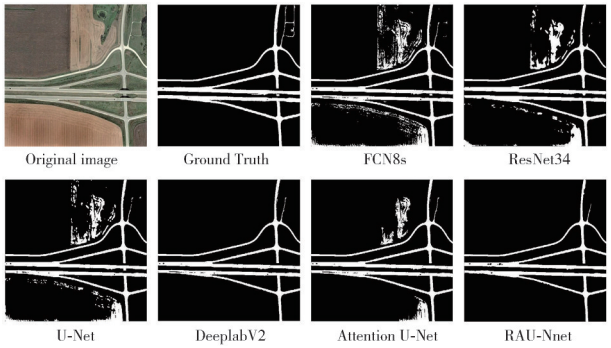

Fig. 13 Comparison results of road extraction for different network models in area 5

It can be seen that the results extracted by FCN8s have many isolated points, and road extraction is incomplete with fractures and holes and easily affected by nearby buildings. ResNet34 reduces the probability of fracture in road extraction, but adhesion still exists in complex and dense areas, which affects extraction effect. U-Net can effectively improve the adhesion phenomenon of road environment, and the extraction results are relatively complete, but there are still false extraction and missing extraction.

DeeplabV2 improves the segmentation accuracy of road extraction, and road segmentation is more accurate, but the extraction results still have fractures and burrs. Attention U-Net can effectively decreases the error of target extraction by U-Net. Comparative analysis shows that the proposed RAU-Net not only accurately and clearly separates the road from the background in complex environment, which ensures the integrity and delicacy of the extracted road, but also solves the problems of holes and fractures, which alleviates missing extraction. Therefore, the extracted road results have the highest similarity with the ground truth map.

In addition, in the same experimental environment and at the training stage, six semantic segmentation network models took 8.67 h, 10.56 h, 10.48 h, 11.23 h, 12.36 h and 11.73 h, respectively. It can be seen that the proposed RAU-Net improves the accuracy of road extraction, and does not significantly increase the time cost of calculation, which further proves the timeliness of the proposed method.

In order to verify the advantages of AS loss in segmenting high-resolution road remote sensing images, it is compared with SCE loss and the commonly used Dice coefficient loss function[28] (Dice loss) for classification problems. When the network is training, the variation curves of segmentation accuracy and training loss with the number of iterations are shown in Figs. 14 and 15, respectively. It can be seen that the segmentation accuracy of RAU-Net+AS loss is higher than that of the other two network models. With the increase of the number of iterations, the training loss of RAU-Net+AS tends to be flat and converges first, the second is RAU-Net+SCE loss, and the last is RAU-Net+Dice loss. It proves the effectiveness of the new loss function proposed theoretically.

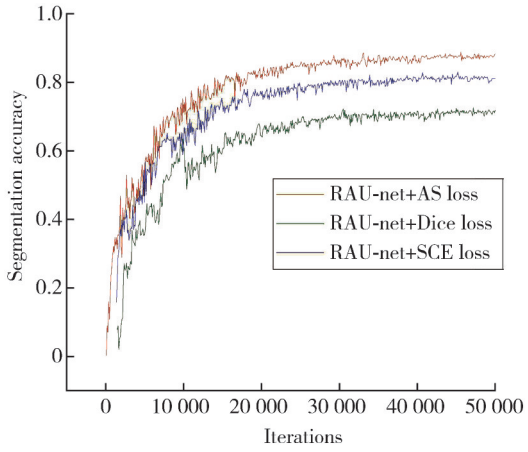

Fig. 14 Relationship between segmentation accuracy and iterations

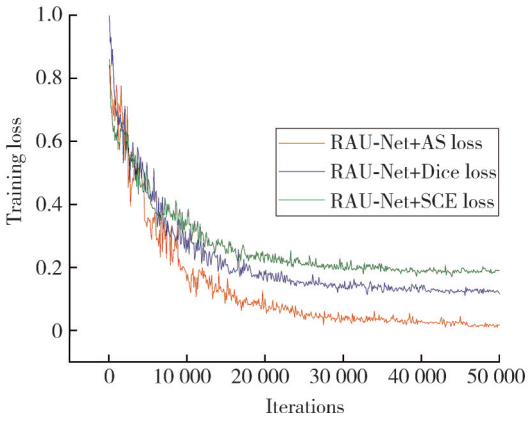

Fig. 15 Relationship between training loss and iteration times

4 Conclusions

In this work, we proposed a road extraction model RAU-Net for remote sensing images by changing the encoding structure of the U-Net. The residual module

was used to deepen the number of network layers so as to enhance the low dimensional detail information in network propagation and improve the segmentation accuracy. The channel attention mechanism was introduced into the decoder to recalibrate the features and increase the extracted detail information. The adaptive loss function was used to obtain clear boundary with high confidence and improve segmentation accuracy. Experiments show that the proposed RAU-Net performs better than FCN8S, ResNet34, DeeplabV2, U-Net and Attention U-Net in the four indexes of *IoU*, *accuracy*, *recall* and *kappa* coefficient, and the improvements are obvious. Because the characteristics of road are somewhat similar to those of rail transit, and a dense road environment still has a certain impact on the accuracy and completeness of the road extraction, the proposed model needs to be further optimized.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 61864025), 2021 Longyuan Youth Innovation and Entrepreneurship Talent (Team), Young Doctoral Fund of Higher Education Institutions of Gansu Province (No. 2021QB-49), Employment and Entrepreneurship Improvement Project of University Students of Gansu Province (No.2021-C-123), Intelligent Tunnel Supervision Robot Research Project (China Railway Scientific Research Institute (Scientific Research) (No.2020-KJ016-Z016-A2), Lanzhou Jiaotong University Youth Foundation (No. 2015005), Gansu Higher Education Research Project (No. 2016A-018), Gansu Dunhuang Cultural Relics Protection Research Center Open Project (No.GDW2021YB15).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] LU X Y, ZHONG Y F, ZHENG Z, et al. Multi-scale and multi-task deep learning framework for automatic road extraction. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(11): 9362-9377.

[2] HAN J W, ZHANG D W, CHENG G, et al. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(6): 3325-3337.

[3] ROMERO A, GATTA C, CAMPS-VALLS G. Unsupervised deep feature extraction for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(3): 1349-1362.

[4] GUO Z L, SHAO X W, XU Y W, et al. Identification of village building via google earth images and supervised machine learning methods. Remote Sensing, 2016, 8(4): 271.

[5] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation//The IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. New York: 2015: 3431-3440.

[6] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[7] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[9] PAPANDREOU G, CHEN L C, MURPHY K P, et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1742-1750.

[10] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation //Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[11] GAO X R, CAI Y H, QIU C Y, et al. Retinal blood vessel segmentation based on the Gaussian matched filter and U-Net. //2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Shanghai, China. New York: IEEE, 2017: 1-5.

[12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[13] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks//European Conference on Computer Vision, October 11-14, 2016, Amsterdam, The Netherlands. Cham: Springer, 2016: 630-645.

[14] LIU Z, ZHANG X L, SONG Y Q, et al. Liver segmentation with improved U-Net and Morphsnakes algorithm. Journal of Image and Graphics, 2018, 23(8):

1254-1262.
[15] JIN F, WANG L F, LIU Z, et al. Double U-Net remote sensing image road extraction method. Journal of Geomatics Science and Technology, 2019, 36(4): 377-381.
[16] SHANKARANARAYANA S M, RAM K, MITRA K, et al. Joint optic disc and cup segmentation using fully convolutional and adversarial networks//International Workshop on Ophthalmic Medical Image Analysis, International Workshop on Fetal and Infant Image Analysis. September 14, 2017, Québec City, QC, Canada. Cham: Springer, 2017: 168-176.
[17] ALOM M Z, YAKOPCIC C, TAHA T M, et al. Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net)// IEEE National Aerospace and Electronics Conference, July 23-26, 2018, Dayton, OH, USA. New York: IEEE, 2018: 228-233.
[18] OKTAY O, SCHLEMPER J, LE FOLGOC L, et al. Attention U-Net: learning where to look for the pancreas. 2018: 1804. 03999. http://arxiv. org/abs/1804. 03999v3.
[19] ZHANG Z X, LIU Q J, WANG Y H. Road extraction by deep residual U-Net. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.
[20] LIU Z H, WU J Z, FU L S, et al. Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion. IEEE Access, 2019, 8: 2327-2336.
[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. ArXiv e-Prints, 2014: arXiv: 1409. 1556.
[22] HU J, SHEN L, SUN G. Squeeze-and-excitation networks// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.
[23] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (ELUs). 2015: 1511. 07289. http://arxiv. org/abs/1511. 07289v5.
[24] DE BOER P T, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134(1): 19-67.
[25] CRESWELL A, ARULKUMARAN K, BHARATH A A. On denoising autoencoders trained to minimise binary cross-entropy. 2017: 1708. 08487. http://arxiv. org/abs/1708. 08487v2.
[26] DEMIR I, KOPERSKI K, LINDENBAUM D, et al. DeepGlobe 2018: a challenge to parse the earth through satellite images. 2018: 1805. 06561. http://arxiv. org/abs/1805. 06561v1.
[27] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1026-1034. ,
[28] MILLETARI F, NAVAB N, AHMADI S A. V-net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision, October 25-28, 2016, Stanford, CA, USA. New York: IEEE, 2016: 565-571.

基于编解码网络的自动道路提取框架

王　霖[1], 沈　瑜[1*], 张泓国[1], 梁　栋[2], 牛东兴[2]

1. 兰州交通大学 电子与信息工程学院, 甘肃 兰州 730070;
2. 中铁科学研究院有限公司, 四川 成都 610036

摘　要: 基于深度学习的道路提取是近十年来语义分割的热点之一。本研究提出一种基于编解码网络的遥感图像道路自动提取框架。首先,将预训练好的ResNet34网络迁移到U-Net网络并替换其编码结构,加深网络层数,降低道路分割的错误率和细节丢失。其次,利用通道注意力机制,对编码器上采样得到的特征图进行信息选择,将目标特征进行权重优化,在增强图像目标区域特征的同时抑制背景及噪声区域,优化背景复杂度高的遥感图像特征提取效果。最后,提出一种自适应sigmoid损失函数,解决道路和背景占比不平衡的问题,使模型在客观程度上达到最优解。实验结果表明,与其他几种语义分割网络方法相比。本研究所提的方法有效降低了道路提取的错误率,提高了道路图像的分割精度。

关键词: 遥感图像; 道路提取; ResNet34; U-Net; 通道注意力机制; sigmoid损失函数

引用格式: WANG Lin, SHEN Yu, ZHANG Hongguo, et al. Automatic road extraction framework based on codec network. Journal of Measurement Science and Instrumentation, 2024, 15(3): 318-327.