An Island Remote Sensing Image Segmentation Algorithm Based on A Fusion Network with Attention Mechanism

Tianyuan Chen¹², Hongfei Wang¹*, Hao Liu³⁴ and Peng Wu¹

¹Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China
²University of Chinese Academy of Sciences, Beijing 100049, China
³State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
⁴School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author’s e-mail: whf@csu.ac.cn

Abstract. With the increasing importance of islands in many fields, it has become the focus of research to obtain information from island remote sensing images efficiently by using image semantic segmentation algorithm. In recent years, deep learning methods based on convolutional neural network have been widely used in image segmentation. However, in view of the problems that remote sensing images contain richer ratio information and complex background interference, we propose an island remote sensing image segmentation algorithm based on a fusion network with attention mechanism, called AFU-Net. The network is built on the basis of FC_U-Net [1]. An attention mechanism is added to pre-weight the shallow features before deep-shallow layer feature fusion, in order to enhance the response capability of the target features, suppress the background interference and improve the segmentation accuracy of the network. The testing and comparative experiments on NWPU-RESISC45 dataset show that the quantitative metrics and visual effects of AFU-Net are greatly improved compared to FC_U-Net, and are also superior to other three state-of-the-art methods, U-Net, FCN, SegNet, which indicates the effectiveness of our method.

1. Introduction

Marine resources play a more and more important role in today’s world. As the main carrier for the utilization and development of marine, the status of island is also increasingly important. A large area of feature information contained in remote sensing images provides a data basis for the study of island [2]. Image semantic segmentation refers to the classification of each pixel in the image. Obtaining information from remote sensing images efficiently by using image semantic segmentation algorithm has increasingly become the focus of scholars’ research.

In recent years, scholars at home and abroad have done a lot of research in the field of image semantic segmentation and put forward many segmentation methods. In particular, the methods based on DCNN (deep convolutional neural network) [3] have been rapidly developed and widely applied in the field of computer vision. In 2015, Long et al. [4] proposed FCN (Fully Convolutional Network), which replaced the full connection layer of CNN with the convolutional layer, realized the image semantic segmentation in the form of end-to-end for the first time, and used skip structure to connect the shallow detail features
and the deep semantic features. In 2015, Ronneberger et al. [5] proposed U-Net, which transferred all features from the encoder to the corresponding decoder and connected them to obtain detail and semantic features, leading to very good performance in biomedical segmentation applications. In 2015, Badrinarayanan et al. [6] proposed SegNet, in which decoder used the pooling indexes provided by the corresponding encoder to perform nonlinear upsampling operations, saving storage space and time.

These deep learning segmentation methods have achieved good results in natural scene image segmentation. In 2016, Zhang L et al. [7] applied deep learning methods in remote sensing image segmentation, used convolutional networks to extract the spectral and texture features of remote sensing images, which can extract more features of remote sensing images. In 2018, Zhang et al. [8] adopted an encoder-decoder structure network, optimized the network with residual blocks, fused detailed information during the upsampling process, and achieved good results in road remote sensing image segmentation. In 2019, Xu Yue et al. [9] proposed an image segmentation algorithm based on an improved U-Net network architecture and a fully connected conditional random field, capable of extracting remote image features with high background complexity.

In order to improve the segmentation performance, many scholars applied the attention mechanism in the segmentation network, modeled the target feature information, highlighted the target features, suppressed the irrelevant features, so as to improve the segmentation accuracy. In 2016, Wang et al. [10] used the residual connection to implement the attention mechanism, in which the residual branch used the fully convolutional network to obtain the channel weights, then multiplied another branch’s feature maps pixel by pixel to realize the feature map weighted operation. In 2018, Oktay, O et al. [11] proposed U-Net with a soft attention structure, which monitored low-layer features through high-layer features to implement the attention mechanism, and adjusted activation values through learning parameters automatically. Good results have been achieved in medical image segmentation. In 2018, Huang et al. [12] proposed a simplified location attention mechanism to calculate the correlation between each pixel in the cross-shaped area. After repeating the above operation and using the softmax function to obtain the weights of all pixels, the feature maps are weighted pixel by pixel.

The above research has improved the image segmentation accuracy in complex scenes to a certain extent, but there are still deficiencies. For the problem that the remote sensing images contain richer scale information and more complex background interference, we propose a remote sensing image segmentation algorithm based on a fusion network with attention mechanism, called AFU-Net. The network is based on FC_U-Net network [1]. FC_U-Net is a fusion network, which improves U-Net and FCN network respectively to form two new modules, and then the semantic features and detail features from the two modules are fused by a fusion module. In this paper, the attention mechanism is added to the module corresponding to U-Net. Before the deep and shallow layer features are fused, the attention information extracted from the deep layer feature maps is multiplied by the shallow layer feature maps to guide the shallow-deep layer feature fusion. After that, high-resolution feature maps are obtained, which are richer than those obtained by direct fusion. The attention mechanism can weight the feature information, enhance the feature information of the target, and reduce the interference of useless information, thereby improving the segmentation performance of the model.

2. Fusion Model with Attention Mechanism

2.1. The network of AFU-Net

AFU-Net network is proposed on the basis of FC_U-Net network. The structure comparison of the two networks is shown in Figure 1.
The FC_U-Net model [1] was proposed by us in 2020. The model includes three modules, a module based on U-Net network (Module A), a module based on FCN network (Module B), and another module to fuse two modules (Module C).
Module A uses an encoder-decoder structure. The network can be classified into two parts: One part is the encoder. Through the convolutional layers, the normalization layers and the max-pooling layers, the input image is encoded in order to extract the semantic features of the image; The other part is the decoder, which decodes the encoded image through the deconvolutional layer, and finally obtains a preliminary semantic image of the same size as the input image. Between the encoder and decoder, the image semantic features from different layers are fused through the skip connections.

There is some improvement of this network compared with U-Net.

1) It deepens the network layers (from A16 to A19 and from A22 to A24), so as to extract feature information more fully.

2) It adds BN (Batch Normalization) layers (A2, A6, A10, A14, A18) to normalize the features from the previous layers, which make the features more evenly distributed and improve the fault tolerance of the model while accelerating the convergence speed of the model.

3) It adds the Dropout function with rate=0.5 after the last encoding layer (A21), which can reduce network complexity, reduce overfitting, and enhance network generalization ability.

Finally, Module A outputs a wealth of image semantic features with a small amount of detail features.

The module B only includes three types of layers, 3*3 convolutional layers, 5*5 convolutional layers and cropping layer. After the input image passes through 6 3*3 convolutional layers, 6 5*5 convolutional layers and a cropping layer, detailed feature maps of the same size as the input image are obtained.

Compared with FCN, Module B removes the pooling layers to prevent information loss and save image details. In addition, in order to solve the problem that the decrease of the receptive field [13] caused by the removal of pooling layers, the network deepens the convolutional layers to expand the receptive field, reduce background interference, and improve the prediction accuracy.

Finally, Module B outputs a wealth of detail features with less background interference.

2.2. Attention mechanism

(a) The location of Attention Gate (AG) in the network
The Attention Gate module is mainly inspired by the Attention U-Net proposed by Oktay, O. et al.\textsuperscript{[11]}. The module adds attention information to the skip structure of Module A. Module A consists of two parts: encoder and decoder. The encoder gradually reduces the spatial resolution of the output features through convolutional and pooling operations, while the decoder gradually repairs the details and spatial resolution through upsampling operations. There are some skip connections between the encoder and decoder, which fuse the features from shallow and deep layers to help the decoder repair the details better.

However, Module A directly uses features from shallow layers in the decoding stage, which will cause a large amount of background interference, resulting in a decrease in segmentation accuracy. For this problem, the attention mechanism is introduced, the attention information from the deep layer feature maps is used to guide the extraction of the shallow layer feature information so as to filter the feature information, suppress the features of the non-target area and pay more attention to the features of the target area.

The location of AG in the network is shown in Figure 2(a). Before the deep-shallow layer feature fusion, the module AG is used to readjust the output features of the encoder. This module generates a gating signal to control the importance of features at different spatial locations, as shown by the red circle in the figure above.

The internal structure of AG is shown in Figure 2(b), where $x_1$ represents input feature information from the encoder, $g$ represents the strobe signal information from the decoder, and attention coefficient $\alpha$ is obtained after a series of operations on $g$ and $x_1$, input feature $x_1$ is multiplied by attention coefficient $\alpha$ to get output feature information containing attention information.

The calculation process of the attention coefficient $\alpha$ is that, firstly, for the two input features with different sizes and depths, $g$ and $x_1$, $1\times1$ convolution layers $W_g$ and $W_x$ are used to change the dimension to the same channel to facilitate feature fusion.

$$f_1 = W_g \times g$$

$$f_2 = W_x \times x_1$$

The feature maps of the same channel are added to perform feature fusion.

$$f = f_1 + f_2$$

The fused feature maps pass through a linear rectification activation function (ReLU) to increase the sparsity of the neural network, enhance the generalization performance of the network, and improve the calculation efficiency.

$$f_3 = \sigma_1 \times f$$

The activated feature maps pass through a $1\times1$ convolution layer to increase the nonlinearity of the network and deepen the depth of the network.

$$f_4 = \psi \times f_3$$

Then set a Sigmoid activation function ($\sigma_2$) to map features between 0 and 1.

$$f_5 = \sigma_2 \times f_4$$

Finally, the resampler resamples the feature maps to the original $x_1$ size to obtain the attention coefficient $\alpha$. 
The whole process can be expressed as:

$$\alpha = \sigma_2(\psi(\sigma_1(W_g^g + W_c^c x^l)))$$  \hspace{1cm} (7)$$

Multiply the attention coefficient \(\alpha\) by the input features \(x^l\) to obtain the output features.

The proposed attention module uses the deep feature information as a weight to pre-weight the shallow features, enhance the target features’ responsiveness and suppress the background feature information, thereby reducing unwanted information interference, filtering out feature information and improving the segmentation accuracy of the network.

3. Experiment and Analysis

3.1. Dataset preparation

Remote sensing image data of this experiment is derived from NWPU-RESISC45\(^{[14]}\). This dataset is a publicly available benchmark for Remote Sensing Image Scene Classification (RESISC), created by Northwestern Polytechnical University (NWPU). It contains 31,500 images, covering 45 scene classes with 700 images in each class. These 45 scene classes include airplane, airport, baseball diamond and so on. We select island as the dataset of this experiment, with a total of 700 images whose pixel size is 256*256. The images are marked by ourselves. All images are randomly divided into a training dataset, a validation dataset and a testing dataset according to the ratio of 8:1:1. The training dataset includes 560 images, the validation dataset includes 70 images, and the testing dataset includes 70 images.

3.2. Experimental environment

In this experiment, the operating system is Ubuntu 18.04, CPU is E5-2630, memory is 128G, GPU is NVIDIA TITAN Xp, GPU acceleration library is CUDA10.0. The deep learning framework uses Keras2.2.4 as the frontend and TensorFlow1.13.1 as the backend.

3.3. Experimental parameter settings

1) Learning-rate: 0.001
2) Weight decay coefficient: 0.0001;
3) Batch size: 6;
4) Epoch: 200;
5) Optimization method: Adam;
6) Loss function: cross-entropy loss function.

3.4. Evaluation metrics

In this experiment, the mean accuracy\(^{[15]}\), pixel accuracy, mean region intersection over union (mean IU), frequency weighted IU (fw IU), dice coefficient (DIC), and matthew correlation coefficient (MCC) are statistically calculated, whose computation equations are

$$\text{mean accuracy} = \frac{\sum_{c=1}^{N} n_{cc} / \sum_{c=1}^{N} n_{cc}}{N}$$  \hspace{1cm} (8)$$

$$\text{pixel accuracy} = \frac{\sum_{c=1}^{N} n_{cc}}{\sum_{c=1}^{N} n_{cc}}$$  \hspace{1cm} (9)$$

$$\text{mean IU} = \frac{\sum_{c=1}^{N} n_{cc} / (\sum_{c=1}^{N} n_{cc} + \sum_{c=1}^{N} n_{cc} - n_{cc})}{N}$$  \hspace{1cm} (10)$$

$$\text{fw IU} = \frac{\sum_{c=1}^{N} n_{cc} / (\sum_{c=1}^{N} n_{cc} + \sum_{c=1}^{N} n_{cc} - n_{cc})}{\sum \sum_{c=1}^{N} n_{ic}}$$  \hspace{1cm} (11)$$

$$\text{DIC} = \frac{2 \times TP + FP + FN}{2 \times TP + TP \times TN - FP \times FN}$$  \hspace{1cm} (12)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$  \hspace{1cm} (13)$$

The \(n_{cc}\) is the number of class \(i\) segmented to be class \(c\), and \(i, c \in \{\text{island, none island}\}\). The \(N\) represents the number of object category, and is set as 2 in this study for island and non-island object. The TP, FP, TN, and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively.
3.5. Experimental results and analysis
AFU-Net network is based on FC_U-Net network. Therefore, in order to verify the effectiveness of AFU-Net, we first selected FC_U-Net as a comparative experiment. At the same time, we respectively selected three state-of-the-art methods, U-Net, FCN, SegNet. The same dataset, experimental parameters and evaluation standard with our experiment were used to ensure the reliability of the comparison results.

The experimental results are shown in figure 3. From left to right, each column is the original images, ground truth, AFU-Net detection results, FC_U-Net detection results, U-Net detection results, FCN detection results and SegNet detection results.

By comparing all the test results, we can find that SegNet has the most wrong and missed islands. For example, the small island in the upper right corner of the first image has few island boundary, some parts inside the island on the left half of the second image are incorrectly classified as the background, and several small islands in the third image are almost all missed, and only a few parts are correctly classified as islands. This may be due to the fact that SegNet [16] applies a little detail information from pooling indexes in segmentation, thus introducing a lot of background interference. FCN can segment most small islands, but the boundaries of them are relatively rough. For example, the three small islands in the upper right corner of the first image are all glued together without segmenting their respective boundaries. The second and third images also have problems such as island adhesion and unclear boundary segmentation. This may be because [17] FCN uses the feature information obtained from the upsampling operation, resulting in loss of details and rough boundaries. U-Net can also segment most small islands, but there is more serious noise on the boundary segmentation. Especially with the second image as the representative, there is a lot of noise in the middle island’s boundary, and the upper part of the right island is missed. This is because the semantic features of [18] islands are combined with some shallow details, which bring a lot of interference. FC_U-Net performs better than the above three networks. Not only can it segment most small islands, but the island boundary segmentation is also more accurate. This is because the deepening of the Module A network structure brings more semantic information, so the island can be well segmented. Module B uses deep convolutional layers to enrich the detail information, thereby making the island boundary segmentation more accurate. However, there is still some background interference. For example, some islands in the first and second images are missed, because the decoder of Module A directly introduces shallow details. Thus, introducing background interference information. The AFU-Net proposed in this paper solves this problem very well. It adds an attention module to the skip structure between the encoder and decoder, uses the attention
information from the deep network feature maps to guide the extraction of shallow network feature information, so as to filter the feature information, suppress the features of the non-target area and pay more attention to the features of the target area.

Therefore, in the segmentation results of AFU-Net, not only can most of the islands be segmented, but also the segmentation boundary is more accurate, and the background interference is also greatly reduced.

In order to further verify the proposed viewpoint, we quantitatively evaluated the detection effects of each network, and the results are shown in table 1. The results show that AFU-Net performs better than FC_U-Net, U-Net, FCN, SegNet in the mean accuracy, pixel accuracy, mean region intersection over union (mean IU), frequency weighted IU (fw IU), dice coefficient (DIC), and matthew correlation coefficient (MCC), which indicates the superiority of the network design.

Table 1. Quantitative evaluation results of AFU-Net, FC_U-Net, U-Net, FCN, SegNet.

| Method    | Mean Accuracy | Pixel Accuracy | Mean IU | Frequency Weighted IU | Mean DIC | MCC  |
|-----------|---------------|----------------|---------|-----------------------|----------|------|
| AFU-Net   | 0.8928        | 0.9132         | 0.8150  | 0.8314                | 0.9070   | 0.7971|
| FC_U-Net  | 0.8853        | 0.9045         | 0.8089  | 0.8255                | 0.8933   | 0.7887|
| U-Net     | 0.8153        | 0.8061         | 0.6650  | 0.6821                | 0.7973   | 0.6074|
| FCN       | 0.8221        | 0.8079         | 0.6691  | 0.6847                | 0.8005   | 0.6185|
| SegNet    | 0.8332        | 0.8645         | 0.7361  | 0.7601                | 0.8455   | 0.6982|

Therefore, both visual and quantitative evaluation results show that our AFU-Net performs better than FC_U-Net, U-Net, FCN and SegNet on this dataset.

4. Conclusion

In view of the problem that there is excessive background interference in the island remote sensing image semantic segmentation, we propose an image segmentation method based on a fusion network with attention mechanism, called AFU-Net, which is based on the FC_U-Net network structure. This method uses the attention information from the deep network feature maps to guide the extraction of shallow network feature information, pay more attention to target features, and suppress background interference. The visual and quantitative evaluation results of the proposed method on the NWPU-RESSISC45 dataset are greatly improved compared with FC_U-Net, and it is also superior to the other three state-of-the-art methods, U-Net, FCN, SegNet. However, our work still has some limitations. For example, when the image is more complex, the network segmentation accuracy is insufficient. And online training is very time-consuming. Next, we will focus on how to conduct semantic segmentation of remote sensing images more accurately and efficiently.

References

[1] Tianyuan Chen, Hongfei Wang, Hao Liu, Junrong Du, Peng Wu, An Island Remote Sensing Image Segmentation Algorithm Based on FC_U-Net Network[J], 2020 in press.
[2] Wilfried Rickels, Jonas Dovern, Martin Quaas. Beyond fisheries: Common-pool resource problems in oceanic resources and services[J]. Global Environmental Change, 2016, 40:37-49.
[3] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. The 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 2012:1097-1105.
[4] LONG J, SHELHAMER E, and DARRELL T. Fully convolutional networks for semantic segmentation[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3431-3440.
[5] Ronneberger O, Fischer P, Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, 2015: 234-241.
[6] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J], 2015.

[7] Zhang L, Zhang L, Du B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art[J]. IEEE Geoscience and Remote Sensing Magazine, 2016, 4(2):22-40.

[8] Zhang Z X, Liu Q J, Wang Y H. Road extraction by deep residual U-net[J]. IEEE Geoscience and Remote Sensing Letters. 2018, 15(5):749-753.

[9] Yue Xu, Mengru Feng, Jiatian Pi, Yong Chen, Remote sensing image segmentation method based on deep learning model[J], Journal of Computer Applications, 2019, 39(10):2905-2914.

[10] Wang F, Jiang M, Qian C, et al. Residual Attention Network for Image Classification[J]. arXiv preprint arXiv:1704.06904, 2017.

[11] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M.: Attention U-Net: learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018

[12] Huang Z, Wang X, Huang L, et al. CCNet: Criss-Cross Attention for Semantic Segmentation[J], 2018.https://arxiv.org/abs/1811.11721.

[13] Luo W, Li Y, Urtasun R, et al. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks[J], 2017.

[14] Cheng, Gong, Han, et al. Remote Sensing Image Scene Classification: Benchmark and State of the Art[J]. Proceedings of the IEEE, 2017, 10(105):1865-1883.

[15] Garcia-Garcia A, Orts-Escolano S, Oprea S, et al. A Review on Deep Learning Techniques Applied to Semantic Segmentation[J], 2017.

[16] Mehta R, Majumdar A, Sivaswamy J. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures[J]. Journal of Medical Imaging, 2017, 4(2):024003.

[17] Bi L, Feng D, Kim J. Dual-Path Adversarial Learning for Fully Convolutional Network (FCN)-Based Medical Image Segmentation[J]. The Visual Computer, 2018.

[18] Landry Guillaume, Hansen David, Kamp, Florian, Li Minglun, Hoyle Ben, Weller Jochen, Parodi Katia, Belka Claus, Kurz Christopher. Comparing Unet training with three different datasets to correct CBCT images for prostate radiotherapy dose calculations[J]. Physics in medicine and biology, 2019, 64(3).