ADAPTIVE FILTERING REMOTE SENSING IMAGE SEGMENTATION NETWORK BASED ON ATTENTION MECHANISM

Cong zhong Wu[1], Hao Dong[1], Xuan jie Lin[1], Han tong Jiang[1], Li quan Wang[1], Xin zhi Liu[1] and Wei kai Shi[2]

[1]Department of Computer Engineering and Information, Hefei University of Technology, Anhui, P.R.China
[2]Department of Faculty of Information Technology, Macau University of Science and Technology, Macau, P.R.China

ABSTRACT

It is difficult to segment small objects and the edge of the object because of larger-scale variation, larger intra-class variance of background and foreground-background imbalance in the remote sensing imagery. In convolutional neural networks, high frequency signals may degenerate into completely different ones after downsampling. We define this phenomenon as aliasing. Meanwhile, although dilated convolution can expand the receptive field of feature map, a much more complex background can cause serious alarms. To alleviate the above problems, we propose an attention-based mechanism adaptive filtered segmentation network. Experimental results on the Deepglobe Road Extraction dataset and Inria Aerial Image Labeling dataset showed that our method can effectively improve the segmentation accuracy. The F1 value on the two data sets reached 82.67% and 85.71% respectively.

KEYWORDS

Convolutional Neural Network, Remote Sensing Imagery Segmentation, Adaptive Filter, Attention Mechanism, Feature Fusion

1. INTRODUCTION

Remote sensing imagery segmentation is an important part of computer vision tasks. It is widely used in environmental monitoring, urban planning, and rescue of natural disasters such as earthquakes, floods, and mountain fires. Especially in natural disaster rescue, if remote sensing imagery can be segmented faster and more accurately, more rescue time can be obtained and thus damage can be minimized. Roads and buildings are often the objects captured by remote sensing satellites. And backgrounds of these objects are much more complex and diverse. The road image, contains different scenes, such as cities, towns. The building image contains town buildings and sparse country buildings (Figure 1). These factors make it difficult for the network to accurately locate and identify the foreground features of remote sensing imagery.

The research methods for semantic segmentation of remote sensing imagery are mainly two types: traditional methods based on manual feature extraction and deep learning methods based on convolutional neural networks. The traditional segmentation methods include based on region, edge, threshold, etc. These methods can only extract the low-level features of the image. While it cannot fully express the high-level features of the image. With convolutional neural networks

David C. Wyld et al. (Eds): ITCSE, ICDIPV, NC, CBIoT, CAIML, CRYPIS, ICAIT, NLCA - 2021
pp. 23-36, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110903

such as VGGNet [1], GoogleNet [2], ResNet [3], etc. widely used in computer vision tasks, a large number of research works on remote sensing imagery segmentation are based on deep learning methods. At this time, the network can extract the features images faster and more accurately by combining convolution, downsampling, and activation functions. Among these operations, downsampling can reduce the number of parameters and computation of the network. What's more, it can also expand the receptive field of the network. However, it can also arise as aliasing which cause the use information of object to be lost.

In traditional digital signal processing, aliasing refers to the distortion of the sampled signal due to the low sampling frequency, resulting in the inability to recover the original signal. In this case, according to Nyquist's sampling theorem, the sampled signal can recover the original signal completely when the sampling rate must be at least twice the highest frequency of the original signal. In the deep neural network, aliasing also exists due to the downsampling layer. Inspired by the traditional method in which a low-pass filter can recover or reconstruct the original signal, Richard Zhang [4] proposed the concept of filter and applied Gaussian blur layer (filter) before downsampling. We can avoid aliasing by applying a Gaussian filter. However, as the high-frequency noise, such as background, needs to be blurred more compared to the lower frequency edges when using a single Gaussian filter tuned for the noise, the edges are over- blurred leading to significant information loss. To solve this issue, what we need is to apply different Gaussian filters to the foreground and background separately, so that we can avoid aliasing while preserving useful information.

Long proposed a fully convolutional network (FCN)[5] to replace the fully connected layers at the end of the network, making the successful application of deep learning in image segmentation. For example, Shunping Ji [6] successfully applied FCN to building segmentation. Skip-connection proposed by Ronneberger in U-Net [7] was widely used in codec network structure which can help to recover image details and edge information. The U-Net was successfully applied to segment road image by Zhengxin Zhang [8]. Cambridge proposed downsampling index in SegNet [9]. The key component of SegNet is the decoder network which consists of a hierarchy of decoders corresponding to each encoder. Of these, the appropriate decoders use the max-pooling indices received from the corresponding encoder to perform non-linear upsampling of their input feature maps. Chaurasia A proposed to fuse the features between encoder and decoder by pixel summation in LinkNet [10]. While D-LinkNet
[11] improved LinkNet and successfully applied it to road segmentation. However, all these networks ignore the aliasing caused by downsampling. Meanwhile, it also ignores the probable semantic gap between the corresponding levels of Encoder-Decoder as proposed by Nabil [12].

For the large-scale variation in remote sensing imagery, pooling or dilation convolution are two effective ways to deal with. Zhao used different sizes of pooling kernels in PSPNet [13] to increase the receptive field of the network and fuse different scale features. By aggregating information from different regions, the purpose of fully mining the global information is achieved. Dilated convolution[14], firstly proposed by Yu et al, can support the exponential expansion of the receptive field without loss of resolution or coverage. The LFE[15] proposed by Ryuhei Hamaguchi uses dilated convolution to effectively segment building remote sensing imagery. However, as the dilated rate increases, the receptive field of the network increases exponentially and there may be redundant information. Hence, the design of dilated rate can affect the performance of the network. In DeepLabv3 [16] and HDC [17] are improved by dilated rate.

The introduction of an attention mechanism is an effective way to improve remote sensing imagery segmentation. The FarSeg [18] proposed by Zhuo Zheng is based on the correlation between the distribution of remote sensing image data, by capturing the different dimensions of

the feature map to highlight the foreground feature information and suppress irrelevant redundant background information. Unlike previous works that capture contexts by multi-scale feature fusion, Dual Attention Network (DANet)[19] adaptively integrate local features with their global dependencies, which model the semantic interdependencies in spatial and channel dimensions respectively. SCAttNet [20] proposed by Haifeng Li combines spatial attention with

channel attention to segment remote sensing imagery. In summary, the following problems still exist when using deep learning networks to segment remote sensing imagery:

1. Operations such as pooling and downsampling can cause aliasing.
2. When facing larger-scale variance, dilated convolution can expand the receptive field to aggregate multi-scale contextual information but also bring redundant information of background.
3. Semantic gap between the corresponding levels of Encoder-Decoder.

Figure 1. Example road extraction, building detection training images, and corresponding labels

2. RELATED WORK

It is a classification task that makes dense prediction of all pixels in an image. Remote sensing image segmentation methods are mainly divided into two categories. The first kind is based on the traditional artificial feature extraction and segmentation method. The second kind is the image segmentation method based on deep learning. In the traditional segmentation methods, the common image feature extraction includes: based on color feature, based on texture feature, and based on shape feature. Typical feature extraction algorithms include HOG algorithm, SIFT algorithm, and so on. The basic idea of Histogram of Oriented (HOG) is that the detected local object can be described by the distribution of light intensity gradient or edge. It constructs

features by calculating and counting the histogram of gradient direction in the local area of an image. HOG feature combined with SVM classifier is widely used in image recognition. SIFT algorithm obtains features by finding descriptors of feature points in an image and their related dimensions and directions and then carries out image feature point matching. SIFT algorithm is a local feature extraction algorithm, which can maintain invariance in the rotation and scaling of the image, and also can maintain certain stability to noise. The advent of deep learning has revolutionized industries from software to manufacturing. At the same time, it greatly promotes the development of remote sensing image segmentation. The concept of deep Learning originates from the artificial neural networks, which is a research branch in the field of Machine Learning. Through deep learning, low-level features can be combined to form more abstract high-level features to discover the distribution characteristics of data. Among them, Convolutional Neural Network is a typical deep learning model for image segmentation. Based on the classification of training data, deep learning can be divided into supervised learning and unsupervised learning. In recent years, a new semi-supervised learning method has emerged, which combines partially labeled data with partially unlabeled data to train neural networks. In this paper, our approach is based on supervised deep learning.

3. RESEARCH OBJECTIVES

There are a large number of objects in remote sensing images, but the size and shape of the objects are often different. Compared with natural image segmentation, remote sensing image segmentation is more difficult. Especially in the segmentation of small objects and the edge of the object, it is easy to have the situation of wrong segmentation and missing segmentation, which leads to low segmentation accuracy. There are various reasons for low segmentation accuracy, and the common one is the loss of spatial information caused by above and below sampling. However, in the process of subsampling, the sampled signal may be distorted, which makes it impossible to recover spatial details, which is easy to be ignored in remote sensing image segmentation based on deep learning. In addition, this paper proposes a remote sensing image segmentation model based on adaptive filtering.

4. METHOD

4.1. Network Architecture

In this paper, the segmentation of our research is a pixel-level two-class classification. Pixels in the need to be divided into two different parts: foreground object and background. At the same time, each pixel in the foreground is assigned a uniform semantic label. We use pre-trained ResNet-34 as the encoder. The decoder is transposed convolution [21] as for upsampling. The backbone of our network is LinkNet. Firstly, we apply the proposed anti-aliasing module(AFM) before each downsampling operation in the network. Then, we insert the RFM module to eliminate the semantic gap between the corresponding levels of Encoder-Decoder. Thirdly, in the central part of the network, the GAM module is added to aggregate multi-scale contextual information while also avoid redundant information of background. Finally, function of sigmoid is used to classify the output of the network. Figure 2 shows our final model (ARG-Net).

$$sigmoid(x) = \frac{1}{1+e^{-x}} \quad (1)$$

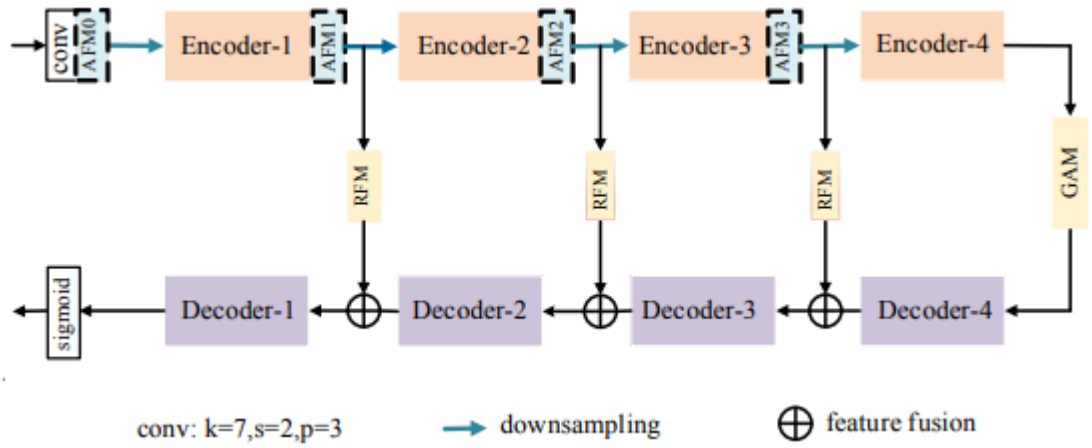

Figure 2. Architecture of ARG-Net. (1)

4.2. Adaptive Filter Model (AFM)

To enable anti-aliasing for ConvNets, we apply the proposed AFM module before each downsampling operation in the network. Inside the module, we first generate filters(a 3x3 conv filter) for different spatial locations and channel groups. Then we apply the predicted filters back onto the input features for anti-aliasing (Figure 3). In remote sensing imagery, low- frequency information tends to be relatively smooth, while high-frequency information tends to have obvious intensity. As frequency components can vary across different spatial locations in an image, the network needs to learn different filters across spatial locations.With the predicted filter, we apply it to input X:

$$Y_{i,j}^{g} = \sum_{p,q\in\Omega} \omega_{i,j,g}^{p,q} \cdot X_{i+p,j+q}^{c} \quad (3)$$

where Yi, j denotes output features at location i, j and Ω points to the set of locations surrounding i, j

In this way, the network can learn to blur higher frequency content more than lower frequency content, to reduce undesirable aliasing effects while preserving important content as much as possible.

Different channels of a feature map can capture different aspects of the input that vary in frequency. Therefore, in addition to predicting different filters for each spatial location, it can also be desirable to predict different filters for each feature channel. Motivated by the observation that some channels will capture similar information, we group the channels into k groups and predict a single low-pass filter for each group. Then, we apply a filter to the input X:

$$Y_{i,j}^{g} = \sum_{p,q\in\Omega} \omega_{i,j,g}^{p,q} \cdot X_{i+p,j+q}^{c} \quad (3)$$

where g is the group index to which channel c belongs.

Figure 3 shows a filtering process on a channel group. Each channel group has the same number of channels. In a channel group, different filters are applied at different spatial locations on a channel (where the filter size is k x k and the feature map size is h x w). For this c/g continuous channel, there is a corresponding consistency in the filter at different positions on each channel. Finally, all channel groups are synthesized into a complete filtered feature map channel through concatenate operation.

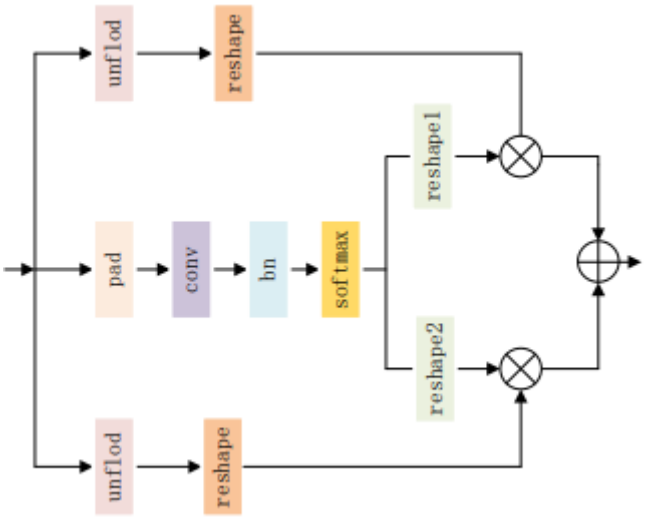

Figure 3. Architecture of AFM.

4.3. Global Attention Model (GAM)

For the large-scale variation in the remote sensing imagery, D-LinkNet used dilated convolution to increase the receptive field of the network,which the output feature map of each dilated convolution contains a larger range of object information. By adopting different dilated rates (where a small dilated rate can extract local information, larger dilated rate extract long-distance information), the network can extract useful feature information of different scales from different receptive fields. Meanwhile, it is helpful for enhancing the learned feature representation ability. However, the local information of the iamge will be lost when the dilated rate becomes larger and larger. What's more, the data sampled from the input becomes sparser, which is not conducive to the convolutional learning of small objects. And there is interference from redundant information in the larger receptive field information. To reduce this influence and further improve the expressive ability of features, an attention guidance module is proposed, as shown in Figure 5. In the original cascaded dilated convolution of D-LinkNet, we removed the dilated convolution block with r=8 and retained the part of r=1,2,4. In the spatial dimension, the global information of the foreground is adaptively captured by operations such as global pooling(GP) and 1x1 conv on the first branch. In the second branch, through 1x1 conv, the foreground information is further learned by the improved dilated convolution, while suppressing redundant and irrelevant information. Finally, features on each branch are fused to extract information of different scales.

$$I * k'(i,j) = \sum_{h}\sum_{w} I(h,w)k'(i-rh, j-rw) \quad (4)$$

where I is the input of the remote sensing imagery, k' is the size of dilated convolution kernel, h, w is the height and width of the image respectively, and r is the dilated rate.

$$k' = (k-1)\cdot(r-1)+k \quad (5)$$

where k is the size of the oridinary convolution kernel.

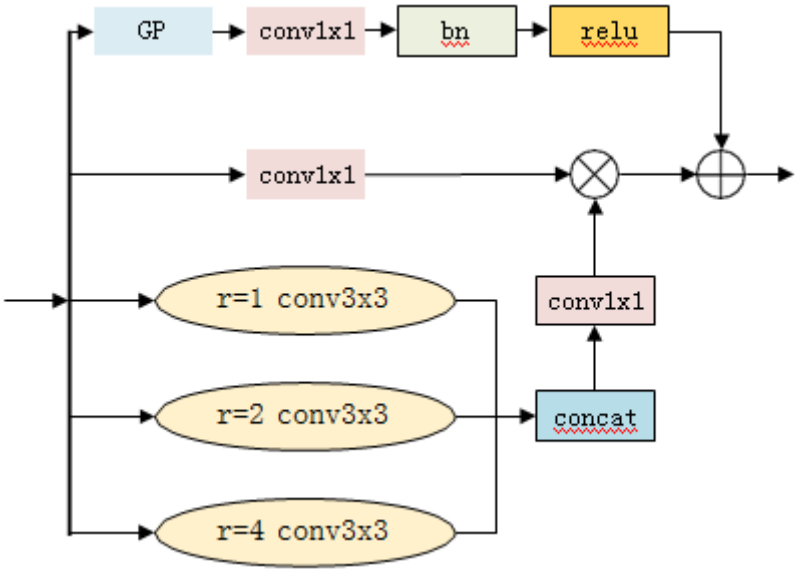

Figure 4. Architecture of GAM.

4.4. Residual Fusion Model (RFM)

D-LinkNet directly used skip-connection to connect the encoder with the decoder, which helps to reduce the loss of spatial information needed to restore the details of the image. The information of the encoder part is of a lower level, while the information of the decoder part is more high-level. There may be a semantic gap between the corresponding levels of Encoder- Decoder. Instead of combining the encoder feature maps with the decoder feature in a straight- forward manner, we pass the encoder features through a sequence of convolutional layers. These additional non-linear operations are expected to reduce the semantic gap between encoder and decoder features. Furthermore, residual connections are also introduced as they make learning easier and are very useful in deep convolutional networks.

The module structure is shown in Figure 5. Firstly, we reduce the number of channels through 1x1 conv to avoid the redundancy of calculations. Then, the low-level feature is learned through two 3x3 convs to reduce the semantic gap between the encoder and decoder.

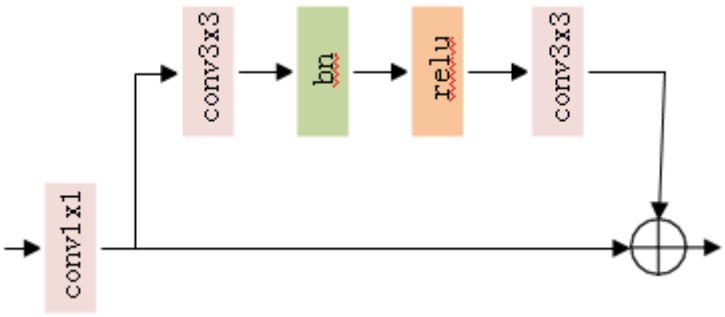

Figure 5. Architecture of RFM.

5. EXPERIMENTS

5.1. Dataset

DeepGlobe Road Extraction (Demir I, etc.) [22] is a road segmentation dataset, including covering cities, towns, suburbs, seashores, tropical rain forests, and other roads in different three countries, such as Thailand, India, Indonesia. The image size is 1024x1024, and the ground resolution is 0.5m. There are 6226 images in the dataset, which are randomly divided into 5226 and 1000 images. Among these images, 5226 images are used for training and the other 1000 images are used for testing. The RGB image of the road is in jpg format, and the corresponding label is in png format. Due to the large size of the original image, we crop all the images to a size of 512x512.

Inria Aerial Image Labeling is a building segmentation dataset, including 187,000 buildings in Christchurch, New Zealand. The image size is 512x512 and the ground resolution is 0.3m. The dataset used in the experiment has a total of 5736 images, of which 4736 are used for training and 1000 are used for testing. The RGB image format of the building is tiff, and the corresponding label format is also tiff. In remote sensing imagery segmentation, there are relatively few public datasets. Even in some public data, the number of images is far from the requirements for training the network. Therefore, we often use data augmentation to optimize training while preventing network overfitting. Image morphological transformation and color transformation are two common ways of data enhancement. In morphological transformation, there are horizontal and vertical flips, rotations of 90 degrees, 180 degrees, 270 degrees, and scale scaling. The color conversion includes the adjustment of saturation, brightness, and contrast.

5.2. Implementation details

All models are trained on the Intel Xeon (R) CPUE5-2640 v4@2.40Hz, which has two graphics cards of GeForceGTX1080Ti with 184.4GB RAM. We use Pytorch as a deep learning framework. We define the initial learning rate of the network as 0.0002. If the loss function does not decrease in 3 training epochs, the learning rate is reduced to 1/5 of the current one. The batch size is set to 8 and the optimizer is Adam[23], where α=0.9, ,β=0.999, eps=1e-8.

5.3. Evaluation metrics

To accurately evaluate the segmentation effect of the module, we calculated four quantitative indicators. These indicators are precision, recall, F1 score, and IoU. Among them, P represents the proportion of the number of correctly predicted objects to all predicted objects. R represents the ratio of the number of correct objects to all positive samples and measures the classifier's ability to recognize positive classes. Since accuracy and recall are not conducive to the comparison of ablation experiments, F1 is often used as the harmonic average of them. IoU is another standard metric for segmentation, which represents the ratio of the intersection of the true value and the predicted value.

$$P = \frac{TP}{TP + FP} \quad (6) \qquad R = \frac{TP}{TP + FN} \quad (7)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (8) \qquad IoU = \frac{TP}{TP + FP + FN} \quad (9)$$

Where TP represents the total number of pixels correctly classified as the foreground object. FP represents the total number of pixels which the background is predicted to be the foreground object. TN represents the total number of pixels which the background is correctly determined as the background. FN represents the total number of pixels where the foreground object is predicted as the background.

5.4. Results and Analysis

We conduct comparative experiments on the DeepGlobe Road Extraction and Inria Aerial Image Labeling datasets.

We use ResNet-18 as our model encoder when we conduct ablation studies on AFM. In this experiment, the size of the filter is set to 3, to match the size of the convolution kernel in the ordinary convolution, which helps to improve the spatial dimension of the filter. As shown in Table 1, through grouping experiments with different channels, it is found that as the number of groups increases, the filtering performance of the network gradually improves. When g=8, it reaches the optimum. If the number of groups still increases, the performance may be degraded due to the over-fitting of the network. Of course, there could be other reasons.

The test result of DeepGlobe Road Extraction is shown in Figure 6. From left to right, it is the original image, label, LinkNet34 segmentation result, adding GAM segmentation result, adding GAM, adding RFM segmentation result, and the final ARG-Net (GAM+RFM+AFM) segmentation result. Among them, white represents the road foreground object and black represents the background. In the figure, the background in the first and second original images occupies a larger proportion. The first and third original images have large background differences. Besides, the roads vary in shape. These characteristics increase the difficulty of road segmentation.

As shown in the segmentation results of the first and second rows in the figure, by adding our module, the occlusion caused by the imbalance between background and foreground can be gradually improved. The interference of the redundant information in the complex background can be reduced. As shown by the segmentation results of the third and fourth rows, the contour of the small object can be segmented step by step by adding different modules. In the end, the ARG-Net improves the overall road segmentation and makes the road more connected. To quantitatively verify the road segmentation performance of the model, recall and F1 are used as evaluation indicators. The results are shown in Table 2. Compared with the original LinkNet-34, our final model increases the recall and F1 by approximately 3.2% and 3.6%, respectively.

Table 1. Ablation comparison experiment of different parameters in AFM on the test of DeepGlobe Road Extraction dataset.

Model	**P**	**F1**
LinkNet18	0.7720	0.7868
LinkNet18(g=2)	0.7769	0.7890
LinkNet18(g=4)	0.7851	0.7905
LinkNet18(g=8)	0.7924	0.7981
LinkNet18(g=16)	0.7855	0.7938

Table 2. Comparative ablation on the test of Deep Globe Road Extraction dataset

Model	**R**	**F1**
LinkNet34	0.7897	0.7906
LinkNet34+GAM	0.8101	0.8083
LinkNet34+GAM+RFM	0.8137	0.8097
ARG-Net(g=8)	0.8204	0.8267

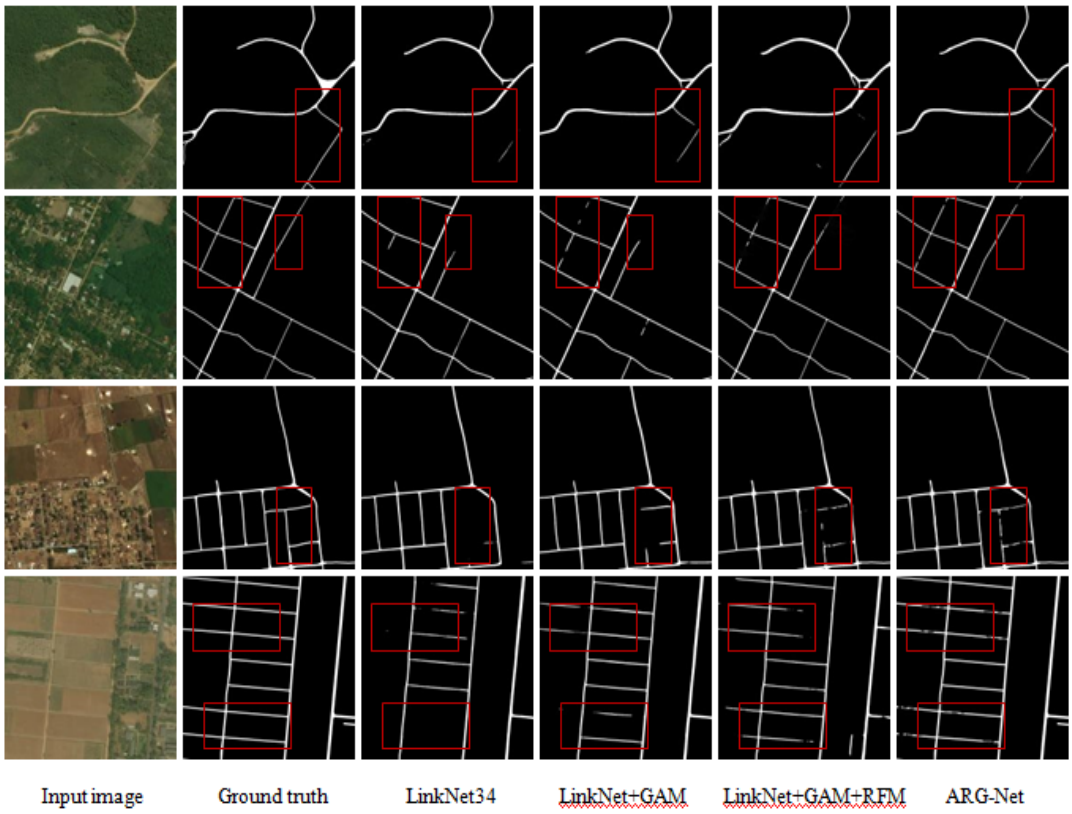

Figure 6. Example results on the test of DeepGlobe Road Extraction dataset.

The test results of Inria Aerial Image Labeling are shown in Figure 7. From left to right, they are the original image, label, LinkNet34 segmentation result, adding GAM test result, adding GAM, RFM test result, and final ARG-Net (GAM+RFM+AFM) test results. Among them, white represents the foreground of the building, and black represents the background.

As shown in the original picture, the size, color, and different backgrounds of the building image increase the difficulty of segmentation. As shown in Figure 7, the first line is an example showing that our module can gradually improve the missing points caused by the low contrast between background and foreground. The second line is an example showing that our method improves the jagged phenomenon that the original network will segment at the edge of the red building (enlarge the image to obtain high-resolution edges). In the last two lines, we can

gradually improve the omission of small target buildings. At the same time, we can also improve the misclassification of some small buildings. Compared with the original LinkNet34, the final ARG-Net model can improve the edge segmentation of small buildings. It can be seen from the last column of Figure 7 that the result segmentation of the building image has more regular,

smooth, and complete. To fully verify the segmentation performance of each module, IoU and F1 are used as the quantitative evaluation index for segmentation. As shown in Table 4, the final network ARG-Net is about 4.3% and 3.2% higher in IoU and F1 than LinkNet34.

Table 3. Results of different models on the test of DeepGlobe Road Extraction dataset

Model	**F1**
FCN-4s	0.7941
SegNet	0.7818
U-Net	0.7730
LinkNet34	0.7906
ARG-Net(g=8)	0.8267

Table 4. Comparative ablation experiment on the test of Inria Aerial Image Labeling dataset

Model	**IoU**	**F1**
LinkNet34	0.7166	0.8251
LinkNet34+GAM	0.7287	0.8436
LinkNet34+GAM+RFM	0.7433	0.8538
ARG-Net(g=8)	0.7579	0.8571

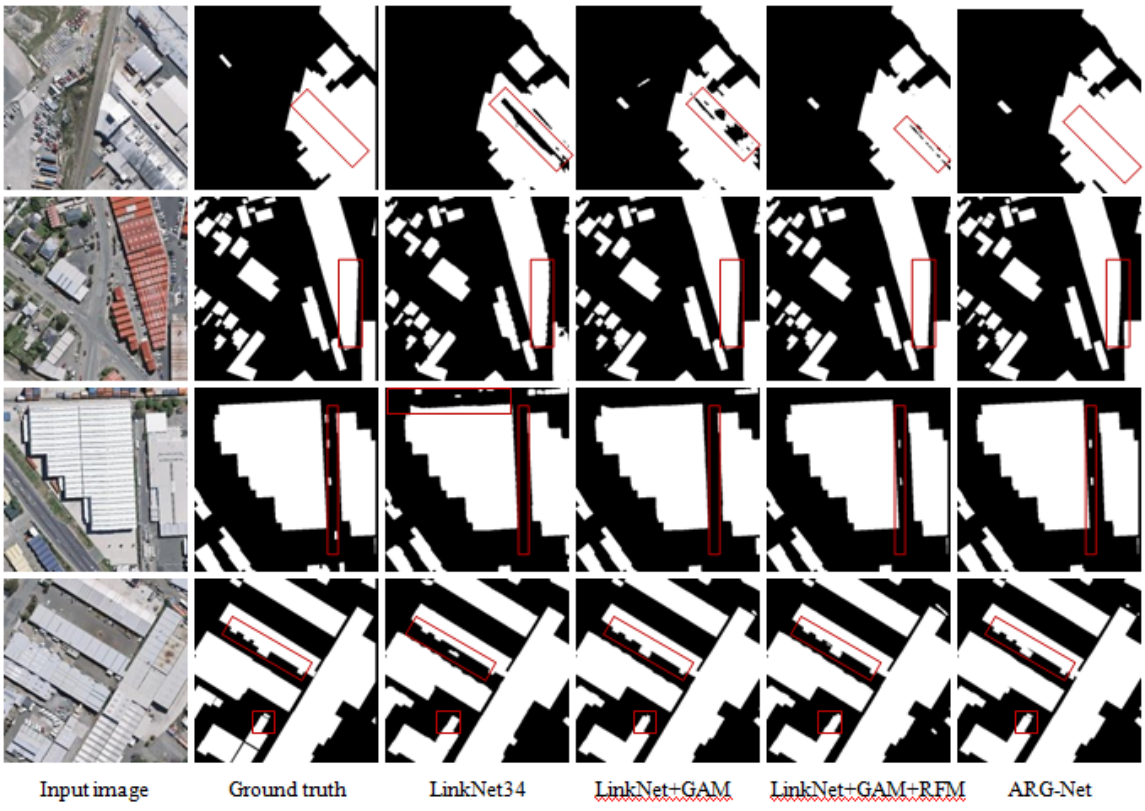

Figure 7. Example results on the test of Inria Aerial Image Labeling dataset.

The limitations or deficiencies observed during testing and evaluation of proposed system are as follows.

It takes a large amount of memory and a long time during training. In the study of the number of channel groups of AFM modules, a large number of network parameters are generated due to the need to predict the filter at each spatial location and each channel group. Therefore, the training will occupy a large amount of memory, at the same time, increase the training time of the network. For the original large remote sensing image, this may not be conducive to network training. The usual solution is to process these images to reduce the number of network training parameters.

Occlusion problem. The ARG-Net proposed by us can reduce the misclassification and missing classification of remote sensing images by the network to a certain extent. However, for some deep occlusion problems (as shown in Figure 6, the road is occluded by trees), the network can not be effectively identified and segmented.

6. CONCLUSION

In this work, we proposed an adaptive filtering layer, which predicts separate filter weights for each spatial location and channel group. This filter can avoid aliasing while preserving useful information. Through a sequence of convolutional during skip-connection between the encoder feature and decoder feature, the semantic gap can reduce. By dilated convolution, the network can ensemble multi-scale features in the center part while avoiding redundant information of background.

Since the production cost of pixel-level image segmentation data sets is relatively high, in the next research, we will focus on adding unlabeled data into the training of the network and using semi-supervised and weakly supervised methods to assist remote sensing image segmentation. In addition, how to balance the relationship between the precision and speed of segmentation network, so as to build an accurate and fast lightweight network model is also the direction of further research in this paper.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China Grant No: 61371156, and Anhui Province Key Scientific and Technological Research Programs Grant No: 201904d07020018. The authors would like to thank the anonymous reviews for their helpful and constructive comments and suggestions regarding this manuscript.

REFERENCES

[1] Muhammad U, Wang W, Chattha S P, et al. Pre-trained VGGNet Architecture for Remote- Sensing Image Scene Classification[C]//2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 1622-1627.

[2] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[4] Zhang R. Making convolutional networks shift-invariant again[J]. arXiv preprint arXiv:1904.11486, 2019.

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[6] Ji S, Wei S, Lu M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 57(1): 574-586.

[7] 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, Cham, 2015: 234-241.
[8] Zhang Z, Liu Q, Wang Y. Road extraction by deep residual u-net[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.
[9] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
[10] Chaurasia A, Culurciello E. Linknet: Exploiting encoder representations for efficient semantic segmentation[C]//2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017: 1-4.
[11] Zhou L, Zhang C, Wu M. D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction[C]//CVPR Workshops. 2018: 182-186.
[12] Ibtehaz N, Rahman M S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural Networks, 2020, 121: 74-87.
[13] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
[14] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.
[15] Hamaguchi R, Fujita A, Nemoto K, et al. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 1442-1450.
[16] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.
[17] Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 1451-1460.
[18] Zheng Z, Zhong Y, Wang J, et al. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 4096-4105.
[19] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.
[20] Li H, Qiu K, Chen L, et al. SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images[J]. IEEE Geoscience and Remote Sensing Letters, 2020.
[21] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1520-1528.
[22] Demir I, Koperski K, Lindenbaum D, et al. Deepglobe 2018: A challenge to parse the earth through satellite images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2018: 172-17209.
[23] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

AUTHORS

Cong zhong Wu, corresponding author, born in 1964, male, professor. His main research direction is computer vision, pattern recognition and image segmentation. E-mail: 2315882652@qq.com

Hao Dong, born in 1994, male, master reading. His main research direction is remote sensing imagery segmentation. E-mail: 2018110990@mail.hfut.edu.cn

Xuan jie Lin, born in 1997, male, master reading. His main research is data mining and machine learning.

Han tong Jiang, born in 1997, male, master reading. His main research direction is data mining and intelligent computation.

Li quan Wang, born in 1994, male, master reading. His main research direction is medical imagery segmentation.

Xin zhi Liu, born in 1994, male, master reading. His main research direction is medical imagery segmentation.

Wei kai Shi, born in 1994, male, master reading. His main research direction is image segmentation.