ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20220421142959
Original scientific paper

MASANet: Multi-Angle Self-Attention Network for Semantic Segmentation of Remote Sensing Images

Fuping ZENG, Bin YANG, Mengci ZHAO*, Ying XING, Yiran MA

Abstract: As an important research direction in the field of pattern recognition, semantic segmentation has become an important method for remote sensing image information extraction. However, due to the loss of global context information, the effect of semantic segmentation is still incomplete or misclassified. In this paper, we propose a multi-angle self-attention network (MASANet) to solve this problem. Specifically, we design a multi-angle self-attention module to enhance global context information, which uses three angles to enhance features and takes the obtained three features as the inputs of self-attention to further extract the global dependencies of features. In addition, atrous spatial pyramid pooling (ASPP) and global average pooling (GAP) further improve the overall performance. Finally, we concatenate the feature maps of different scales obtained in the feature extraction stage with the corresponding feature maps output by ASPP to further extract multi-scale features. The experimental results show that MASANet achieves good segmentation performance on high-resolution remote sensing images. In addition, the comparative experimental results show that MASANet is superior to some state-of-the-art models in terms of some widely used evaluation criteria.

Keywords: global context information; MASANet; multi-scale features; semantic segmentation

1 INTRODUCTION

In recent years, with the continuous development of remote sensing technology, the spatial, temporal, and spectral resolution of remote sensing images has been greatly improved. People can acquire more and more high-resolution remote sensing images [1] (the ground sampling distance is between 5 and 10 cm). In these images, small objects such as cars and buildings can be clearly observed which makes pixel-level semantic segmentation possible. Remote sensing images have great application prospects in the fields of environmental monitoring [2], agriculture [3], forestry [4] and urban planning [5]. Therefore, semantic segmentation based on high-resolution remote sensing images has become a current research hotspot.

Semantic segmentation of remote sensing images is to classify each pixel in the images according to the land cover type. It is an important research direction in the field of pattern recognition. In recent years, with the rapid development of artificial intelligence, CNN-based methods have achieved success in the field of semantic segmentation [6-8]. The fully convolutional networks (FCN) proposed by Long et al. [9] uses standard convolution layer to replace the fully connected layer, realizes regional segmentation and regional object semantic recognition through pooling and convolution, and greatly improves the performance of image segmentation. Many proposed semantic segmentation models, such as SegNet [10] and U-Net [11], are based on FCN.

Other methods to collect context information are realized by using atrous convolution [12] to increase the receptive field. Chen et al. [13] designed an atrous spatial pyramid pooling (ASPP) using parallel atrous convolution with different atrous rates in the proposed DeepLabv3 to obtain more multi-scale context information. The receptive field can be expanded without introducing additional parameters.

The above two solutions have their inherent disadvantages in the segmentation of remote sensing images. On the one hand, most FCN-based methods stack local convolution and pooling operations [14]. Due to the limited receptive field, they cannot deal with various types of complex scenes well. On the other hand, the atrous convolution method leads to the loss of spatial information due to continuous atrous convolution, resulting in the "chessboard effect" [15]. At the same time, the ASPP algorithm is effective for feature extraction of large-scale targets, but small-scale targets will be lost. Common methods to improve long-term dependence modeling ability of CNNs include atrous convolution, global average pooling (GAP) [16], and self-attention [17]. Self-attention adds weighted attention to the original feature graph, and the global dependence of any two positions in the feature graph can be obtained.

In this paper, we propose multi-angle self-attention network (MASANet) to obtain enhanced global context information. We use atrous convolution to increase the receptive field and ASPP to obtain more multi-scale context information. In order to solve the spatial information loss caused by atrous convolution and the small-scale target loss caused by ASPP, we design multi-angle self-attention module (MASAM) to enhance features from different angles, enhance the spatial information lost due to atrous convolution and obtain more global dependencies, and capture the dependencies between long-distance features through self-attention. At the decoding network, we also concatenate and upsample the feature images of different scales obtained in the feature extraction stage and the corresponding feature images output by ASPP, respectively, so as to obtain the final segmentation prediction.

In summary, the contributions of our method are as follows:

1) We propose a multi-angle self-attention network, which uses MASAM and ASPP to obtain richer global dependency and context information.

2) We design a MASAM, which can effectively capture the feature relationship between channels and the dependence between long-distance features.

3) In the decoding network, we integrate the features of different scales from other stages of the encoding network into the current stage, so as to realize the information complementarity between the features of different stages.

4) We conduct quantitative and qualitative comparison experiments using MASANet and three state-of-the-art semantic segmentation methods. Then considering MASAM is adopted for the first time, we carry out some ablation studies to test its effectiveness.

The rest of the paper is structured as follows. Section 2 describes the related work. Our proposed method is illustrated in Section 3. Section 4 introduces our dataset and experimental setup. The experimental results are given in Section 5, and our conclusions are provided in Section 6.

2 RELATED WORK

Although CNNs have achieved good results in semantic segmentation of remote sensing images, there are still problems of limited receptive field and loss of spatial information. Combining attention mechanism with CNNs can overcome this problem. In this section, literature of semantic segmentation and attention mechanism is summarized.

2.1 Semantic Segmentation

Many FCN-based models have been proposed for semantic segmentation. Bhatnagar et al. [18] used a convolutional neural network to map the main vegetation communities of Clara swamp wetland in Ireland in spring. The combination of ResNet50 and SegNet architecture gave the best semantic segmentation results. Wu et al. [19] used U-Net to train the semantic segmentation of remote sensing images and obtained the results of semantic segmentation. Heryadi et al. [20] combined the DeepLabv3 model with two other networks: ResNet and conditional random field network to form a deep network structure, and its semantic segmentation performance is better than other models. However, most of the methods based on FCN cannot deal with various types of complex scenes well because of the limited receptive field. DeepLabv3 based on atrous convolution will lead to the loss of spatial information due to continuous atrous convolution. At the same time, the ASPP algorithm will lead to the loss of small-scale targets.

2.2 Attention Mechanism

In pattern recognition through artificial intelligence, the attention mechanism aims to make the system learn to pay attention to ignore irrelevant information and focus on key information. Common attention modules are CBAM [21] and self-attention. Attention is used in many related works in the field of pattern recognition, such as classification, segmentation and natural language processing. Chen et al. [22] embedded convolution block attention module (CBAM) between convolution blocks of P-Net, constructed CBAM-P-Net and proposed a method to improve the efficiency of P-Net feature extraction. Hou et al. [23] proposed a strip pooling network (SPNet) and introduced a new pooling strategy (called strip pooling) to reconsider the formulation of space pool. This strategy considers a long and narrow kernel, namely, $1 \times N$ or $N \times 1$, to establish new state-of-the-art results. Sheng et al. [14] used the SPNet model to solve the problem of multi-class semantic segmentation of high-resolution remote sensing images. A better semantic segmentation effect is obtained. Wang et al. [24] applied self-attention to the field of computer vision, which improved the computational efficiency.

Inspired by these works, we design a MASAM, which considers the ideas of CBAM, strip pooling, and self-attention to further enhance the features extracted at the encoding network, so as to obtain richer global dependency and context information.

3 METHODOLOGY

The entire MASANet architecture is shown in Fig. 1. It consists of four main components: a backbone network, MASAM, ASPP, and decoder network. This section details the MASANet model for semantic segmentation.

3.1 Backbone Network

ResNet and ResNet-like architectures [13, 25] are powerful visual feature extractors. In order to capture more text information than typical CNN, our encoding network takes the improved ResNet50 as the backbone network for feature extraction. First, the image is convolution of 7×7, and the size of the input images is reduced from 224×224 to 112×112. Then the size of the feature map is continuously reduced through conv2_x, conv3_x and conv4_x to extract effective features. At conv5_x, a 14×14 feature map is obtained by atrous convolution operation to increase the receptive field.

3.2 Multi-Angles Self-Attention Module (MASAM)

As shown in the existing methods, the attention module is effective for semantic segmentation. CBAM proposed by Sanghyun Woo et al. [21] pays attention not only to the feature relationship between channels, but also to the feature relationship between spaces. The pooling kernels used in CBAM are all square pooling kernels. However, the objects in high-resolution remote sensing images vary greatly in size and have different expansibility and orientation, such as long narrow roads and wide grasslands, which pose a great challenge to the traditional square pooling kernel. Sheng et al. [14] introduced the strip pooling to solve the above problems while preventing information from irrelevant areas. As shown in Fig. 2, in our method, we draw lessons from this thought to carry on pool from three angles: side, top, and front. Among them, the side attention module (SAM) is the channel attention module (CAM) in CBAM, while the top attention module (TAM) and front attention module (FAM) borrow the idea of strip pooling. Then we take the three feature maps obtained through SAM, TAM and FAM as the input of subsequent self-attention to further enhance the input features.

SAM is structured as follows, first, the input feature map is passed through the maximum pooling layer and the average pooling layer, respectively. Then the output of the two is passed through the Multilayer Perceptron (MLP) to reduce the parameter overhead, which contains two 1×1 convolutions and a ReLU activation function.

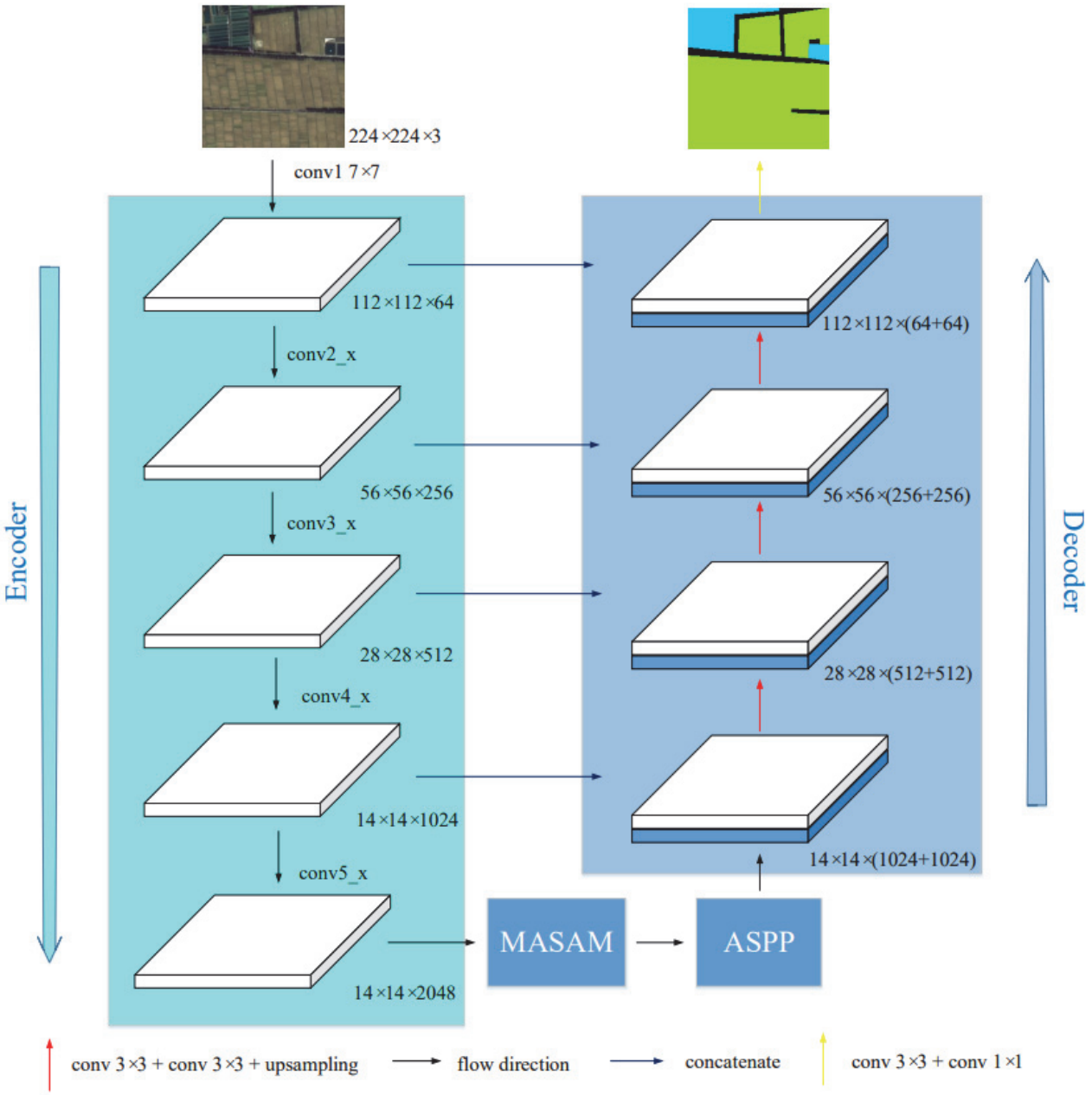

Figure 1 Framework of the MASANet architecture. MASAM denotes the multi-angle self-attention module. ASPP denotes the atrous spatial pyramid pooling

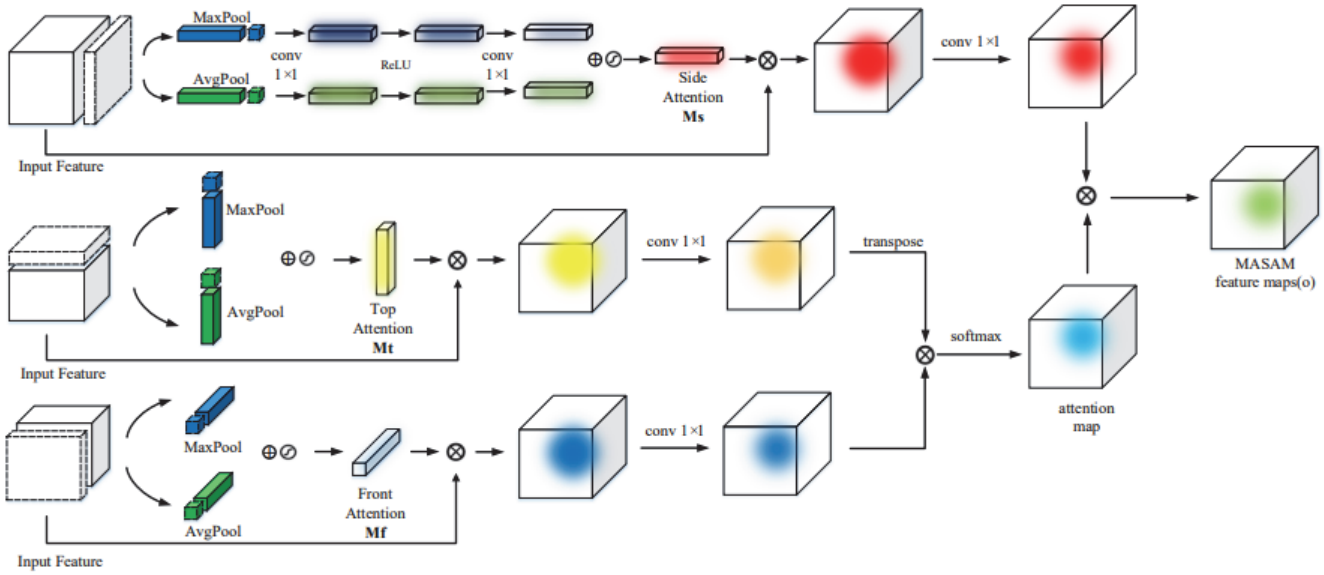

Figure 2 Framework of multi-angle self-attention module

The *MLP* output features are added, and the side attention feature map is output through the sigmoid function. The side attention is computed as:

$$\begin{aligned} M_{\mathrm{s}}(F) &= \sigma\left(MLP\left(AvgPool(F)\right)+MLP\left(MaxPool(F)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{\mathrm{avg}}^{\mathrm{s}}\right)\right)+W_1\left(W_0\left(F_{\mathrm{max}}^{\mathrm{s}}\right)\right)\right) \end{aligned} \tag{1}$$

F is the input feature; σ is a sigmoid operation; $F_{\mathrm{avg}}^{\mathrm{s}}$ and $F_{\mathrm{max}}^{\mathrm{s}}$ denote average-pooled features and max-pooled features, respectively, where W_0 needs to be followed by ReLU activation function; W_0 and W_1 represents the weight matrix of two convolution layers; M_{s} is the side recalibration feature.

TAM is structured as follows: first, the input feature map is displayed in $W \times C$ dimension passes through the maximum pooling layer and the average pooling layer, respectively. Then the outputs of the two are added, and the top attention feature map is output through the sigmoid function. The top attention is computed as:

$$\begin{aligned} M_{\mathrm{t}}(F) &= \sigma\left(AvgPool(F)+MaxPool(F)\right) \\ &= \sigma\left(F_{\mathrm{avg}}^{\mathrm{t}}+F_{\mathrm{max}}^{\mathrm{t}}\right) \end{aligned} \tag{2}$$

F is the input feature; σ is a sigmoid operation; $F_{\mathrm{avg}}^{\mathrm{t}}$ and $F_{\mathrm{max}}^{\mathrm{t}}$ denote average-pooled features and max-pooled features, respectively; M_{t} is the top recalibration feature.

FAM is structured as follows: first, the input feature map is displayed in $H \times C$ dimension passes through the maximum pooling layer and the average pooling layer, respectively. Then the outputs of the two are added, and the front attention feature map is output through the sigmoid function. The front attention is computed as:

$$\begin{aligned} M_{\mathrm{f}}(F) &= \sigma\left(AvgPool(F)+MaxPool(F)\right) \\ &= \sigma\left(F_{\mathrm{avg}}^{\mathrm{f}}+F_{\mathrm{max}}^{\mathrm{f}}\right) \end{aligned} \tag{3}$$

F is the input feature; σ is a sigmoid operation; $F_{\mathrm{avg}}^{\mathrm{f}}$ and $F_{\mathrm{max}}^{\mathrm{f}}$ denote average-pooled features and max-pooled features, respectively; M_{f} is the front recalibration feature.

After three enhanced feature maps are obtained through SAM, TAM, and FAM, the overall feature recalibration process of MASAM is as follows: firstly, the feature maps obtained by TAM and FAM are convolution of 1×1, and the feature maps obtained by TAM and 1×1 convolution are transpose, multiplied by feature maps obtained by FAM and 1×1 convolution, respectively. The feature maps are then passed through softmax to get attention map. On the other hand, multiply SAM over a 1×1 convolution with the attention maps to obtain MASAM feature maps (o). Finally, multiply the MASAM feature maps and the input feature maps to recalibrate the features of the feature maps as a whole.

The specific structure and process of MASAM are shown in the Eq. (4):

$$\begin{aligned} o = {} & f^{1\times1}M_{\mathrm{s}}(F)\times \\ & \times\mathrm{softmax}\left(\mathrm{transpose}\left(f^{1\times1}\times M_{\mathrm{t}}(F)\right)\times f^{1\times1}M_{\mathrm{f}}(F)\right) \end{aligned} \tag{4}$$

F is the input feature; M_{s} is the side recalibration feature; M_{t} is the top recalibration feature; M_{f} is the front recalibration feature; $f^{1\times1}$ represents a convolution operation with the filter size of 1×1 and o is the MASAM recalibration feature maps.

3.3 Atrous Spatial Pyramid Pooling (ASPP)

Inspired by the spatial pyramid pooling [26] and atrous convolution, ASPP applies the resampling method to the features under different atrous rates to capture multi-scale context information efficiently and accurately. We use atrous rates of 6, 12 and 18 in the ASPP module. At the end of the network, the global content information is integrated into the model by using image-level features. The GAP is applied to the final feature mapping of the model.

3.4 Decoder Network

After ResNet50 feature extraction, MASAM feature enhancement and ASPP capturing multi-scale context information, the final feature map with output stride of 16 is finally obtained. It is a challenge to reconstruct the original size segmentation graph from such a small feature graph. Therefore, we designed the decoder network of the U-shaped structure, which does not directly upsample the feature map, but in four steps, as shown in Fig. 1. Firstly, we use bilinear interpolation to upsample the ASPP output feature map by factor 2, then concatenate it with the output feature of conv4_x of ResNet50 with the same resolution, and then execute two 3×3 convolutions; then, the feature map is upsampled by factor 2 again and concatenated with the output feature of conv3_x of ResNet50, and then 3×3 convolution is performed twice; then the obtained feature map is upsampled by factor 2 and concatenated with the output feature of conv2_x of ResNet50, and then twice 3×3 convolution is executed. Finally, the obtained feature map is upsampled by factor 2 and concatenated with the obtained output features of the first 7×7 convolution of ResNet50, then 3×3 convolution and 1×1 convolution are performed to produce the final segmentation output.

4 EXPERIMENTAL SETTINGS

In order to evaluate the proposed method, experiments are carried out on GID [27] dataset. Firstly, the dataset and implementation details are introduced, and then the evaluation criteria we use are introduced. Our method is implemented on PyTorch.

4.1 GID Dataset

In this paper, we test our model and evaluate its performance on an open dataset GID dataset. GID consists of two parts: large-scale classification set and fine land-cover classification set. The fine land-cover classification set used in this experiment is composed of 15 fine classifications. In addition to the 15 classes, we class the

rest as background. In addition, we will discuss paddy field, irrigated land, dry cropland, garden land, arbor forest, shrub land, natural meadow, and artificial meadow as vegetation. The size of the original images is 6800 × 7200 ($H \times W$), the size after cutting is 224 × 224, and the cutting step size is 112. The total number of finally generated images is 37170. 22302/7434/7434 images are randomly selected for training, validation, and testing, respectively. Fig. 3. shows several sample images of training, validation, and testing images in the GID dataset.

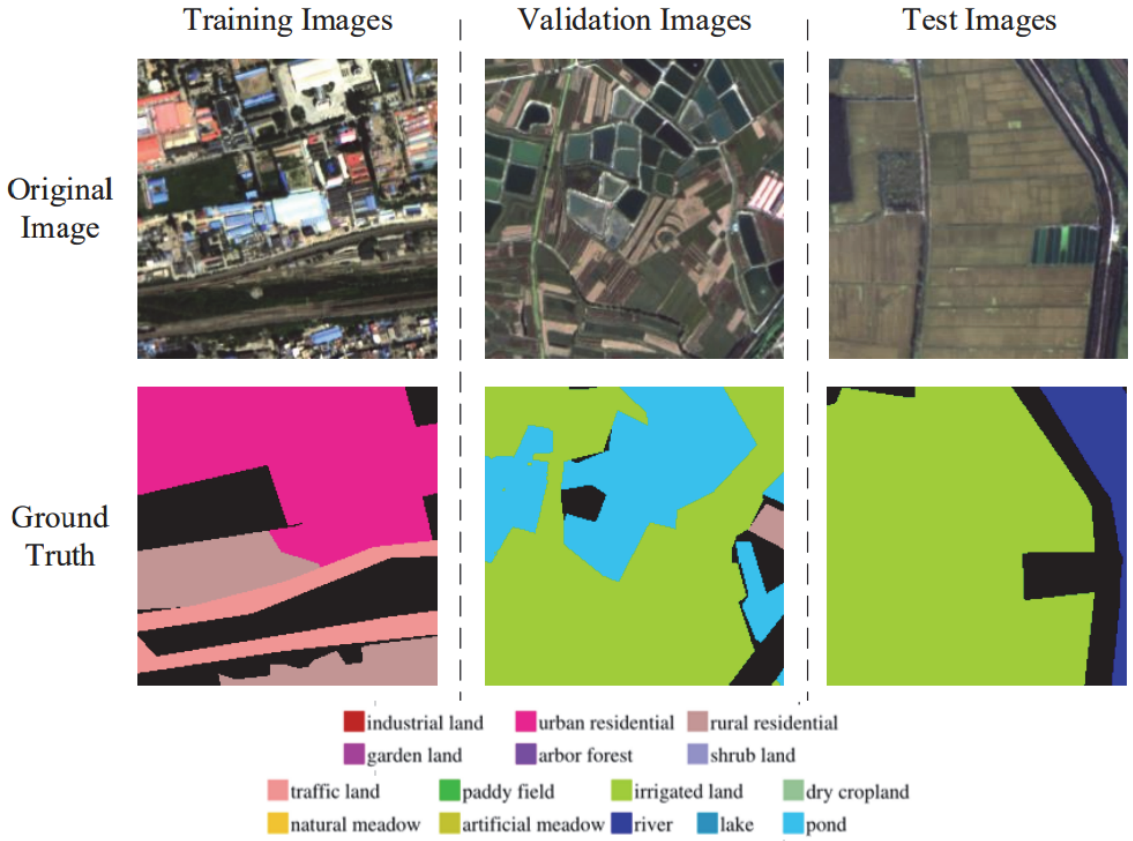

Figure 3 Sample images of training, validation and test images in the GID dataset

4.2 Implementation Details

We train our model on a computer equipped with Intel Core I9-9920X. The GPU type is NVIDIA TITAN V. Because the training model requires a lot of GPU memory, the method in this paper will use 224 × 224 size images as input to the network. Adam optimizer is used to optimize the network and update parameters. In addition, the network proposed in this paper uses NLLLoss as the loss function. When training our model, we set the training epoch to 300 and the learning rate to 0.0001. Our training batch is 8. After training, we integrate the training loss and verification loss in 300 epochs and took the model with the minimum mean value for overall preservation (network structure + parameters).

4.3 Evaluation Metrics

In order to comprehensively evaluate the performance of the model, we use three evaluation criteria widely used to evaluate the performance of semantic segmentation. They are Pixel Accuracy (PA), IoU and mIoU. Where PA is the ratio of the correctly labeled pixels to the total pixels, IoU is the ratio of the intersection and union of the predicted and true values for a class, and IoU is calculated on each class, then averaged, and mIoU is obtained. They are expressed as follows:

(1) PA

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{5}$$

where k is the number of classes minus 1. p_{ii} represents the number of pixels belonging to class i and predicted as class i, p_{ij} represents the number of pixels belonging to class i but predicted as class j.

(2) IoU

$$IoU = \frac{TP}{TP + FN + FP} \tag{6}$$

where TP, FN, FP, and TN denote true positive, false negative, false positive, and true negative, respectively.

(3) mIoU

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} IoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{7}$$

where k is the number of classes minus 1.

5 RESULTS AND DISCUSSION

In this section, we first conduct quantitative and qualitative comparison experiments using MASANet and three state-of-the-art semantic segmentation methods. Then considering MASAM is adopted for the first time, we carry out some ablation studies to test its effectiveness.

5.1 Comparisons and Analysis

In this part, in order to verify the performance of MASANet, we compare MASANet with three representative deep learning network models, namely, U-Net, SegNet, and DeepLabv3 on GID datasets under the same conditions. Tab. 1 lists the IoU of each class on the GID dataset. It can be seen that MASANet obtains higher or similar IoU scores on most classes. In particular, MASANet has made significant improvements in some narrow or wide objects (industrial land more than 1.2%, urban residential more than 0.76%, rural residential more than 2.26%, traffic land more than 3.46%, and pond more than 0.92%). In addition, Tab. 2 lists the overall mIoU and PA on the GID dataset. In general, the PA obtained by MASANet is 0.59%, 2.24% and 0.78% higher than that of U-Net, SegNet, and DeepLabv3, respectively; mIoU is 1.1%, 4.5% and 1.43% higher than U-Net, SegNet, and DeepLabv3, respectively. These results show that our model can obtain more long-range dependencies and more robust multi-scale context information through MASAM, ASPP, and U-shaped, so as to produce better segmentation results.

Table 1 U-Net, SegNet, DeepLabv3, and MASANet get the IoU scores for each class on the GID test set

method	back ground	industrial land	urban residential	rural residential	traffic land	vegetation	river	lake	pond
U-Net	0.8838	0.8689	0.8930	0.8205	0.7886	0.9335	**0.9691**	**0.9526**	0.9045
SegNet	0.8508	0.8272	0.8679	0.7816	0.7500	0.9116	0.9515	0.9115	0.8568
DeepLabv3	0.8793	0.8687	0.8926	0.8197	0.7795	0.9307	0.9633	0.9467	0.9039
MASANet	**0.8978**	**0.8809**	**0.9006**	**0.8431**	**0.8232**	**0.9394**	0.9679	0.9467	**0.9137**

Table 2 PA and mIoU scores of U-Net, SegNet, DeepLabv3, and MASANet on the GID test set

method	PA	mIoU
U-Net	0.9474	0.8905
SegNet	0.9309	0.8565
DeepLabv3	0.9455	0.8872
MASANet	**0.9533**	**0.9015**

Some representative samples of MASANet are shown in Fig. 4. As shown in the figure, the extraction result of MASANet is very complete, almost consistent with the ground truth, and accurately captures the semantic details, which U-Net, SegNet, and DeepLabv3 fail to do. This clearly shows that MASANet provides better performance in capturing multi-scale objects from small details to large-scale objects.

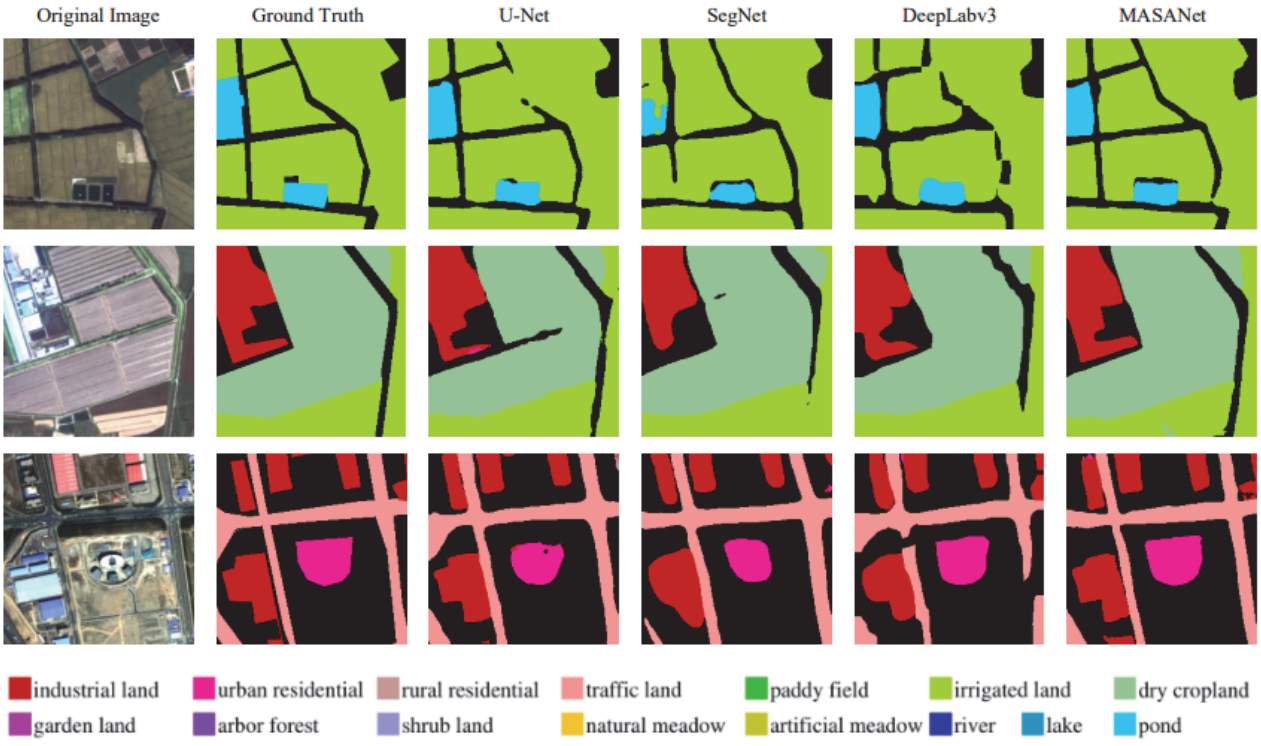

Figure 4 Visualization of U-Net, SegNet, DeepLabv3, and MASANet on the GID dataset

5.2 Effect of the MASAM

In order to verify the importance of MASAM in the segmentation process, under the same training conditions, our MASANet is compared with the original model and the model added with CBAM. In the original model, we deleted MASAM in MASANet. In the comparison model, we put CBAM and MASAM in the same position of the

network. Tab. 3 shows the experimental results of three models on the GID fine land-cover classification dataset. It can be seen that MASANet obtains higher or similar IoU scores on most classes. In particular, MASAM has made significant improvements in some narrow or wide objects (industrial land more than 1.6%, urban residential more than 0.9%, rural residential more than 1.1%, traffic land more than 1.94%, vegetation more than 0.61%). In addition, Tab. 4 lists the overall mIoU and PA on the GID dataset.

In general, the PA obtained by MASANet is 0.65% and 0.63% higher than that original model and the model with CBAM in the same position, respectively; mIoU is 1.14% and 0.96% higher than the other two respectively. These results show that our model obtains more long-range dependencies and more robust multi-scale context information under the guidance of original context information through MASAM, so as to produce a better segmentation effect.

Table 3 Networks with different attention modules get the IoU scores for each class on the GID dataset

method	back ground	industrial land	urban residential	rural residential	traffic land	vegetation	river	lake	pond
original	0.8821	0.8649	0.8915	0.8161	0.7957	0.9333	0.9649	**0.9539**	0.9087
original-CBAM	0.8843	0.8549	0.8916	0.8321	0.8038	0.9301	0.9658	0.9510	0.9136
MASANet	**0.8978**	**0.8809**	**0.9006**	**0.8431**	**0.8232**	**0.9394**	**0.9679**	0.9467	**0.9137**

Table 4 PA and mIoU scores obtained by networks with different attention modules on the GID dataset

method	PA	mIoU
original	0.9468	0.8901
Original-CBAM	0.9470	0.8919
MASANet	**0.9533**	**0.9015**

Some representative samples of MASANet are shown in Fig. 5. As shown in the figure, the extraction result of MASANet is more complete, almost consistent with the ground truth, and accurately captures the semantic details. There will be a small amount of misclassification in the model without attention module and the model with CBAM. This clearly shows that MASANet provides better performance in capturing multi-scale objects from small details to large-scale objects.

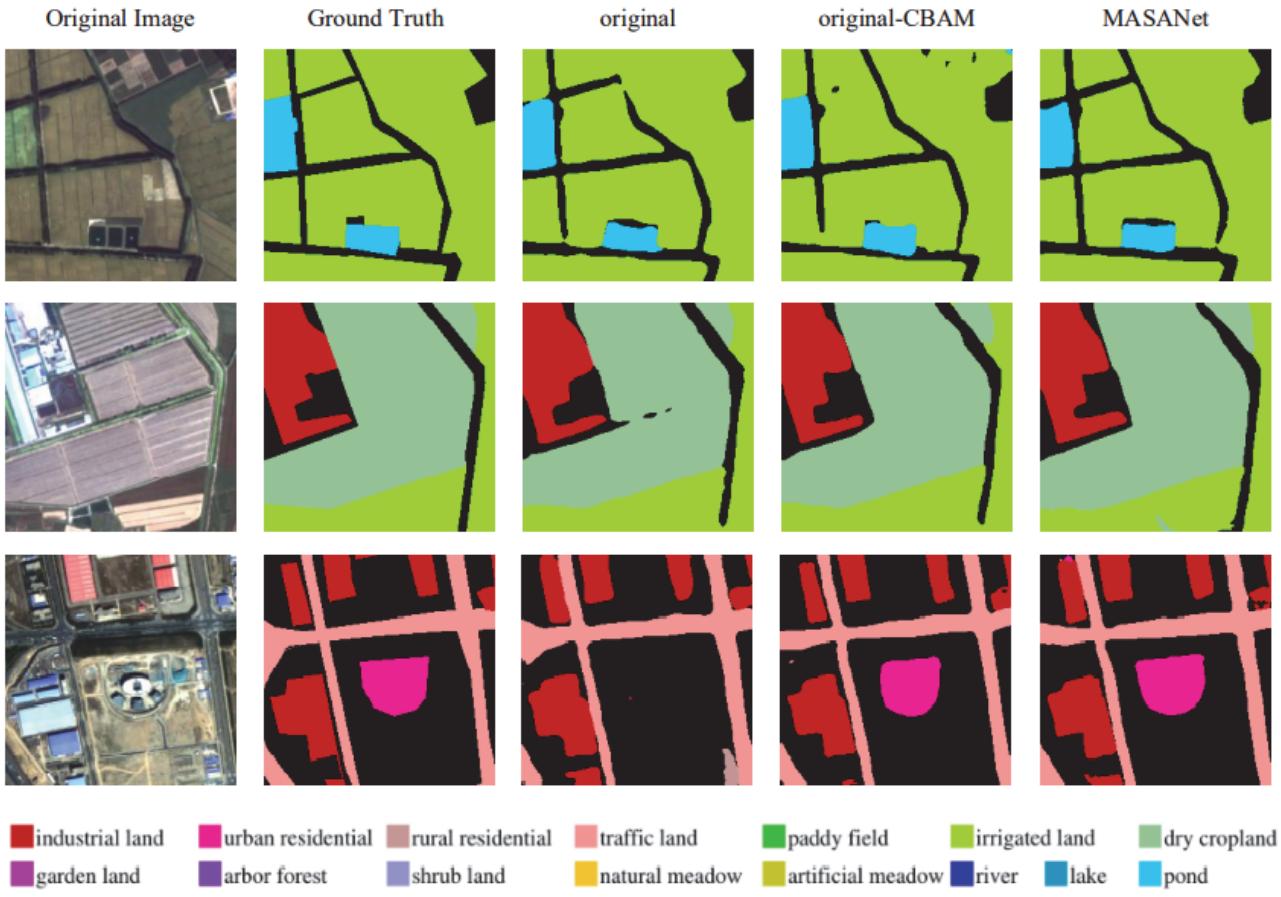

Figure 5 Visualization of original model, original-CBAM, and our MASANet on the GID dataset

6 CONCLUSION

Semantic segmentation of remote sensing images is an important research direction in the field of pattern recognition. However, the effect of semantic segmentation is still incomplete or misclassified due to the loss of global context information. In this paper, we have proposed a multi-angle self-attention network (MASANet) for semantic segmentation. MASANet uses ResNet50 with atrous convolution as the backbone network for feature extraction. We design a multi-angle self-attention module (MASAM) to enhance the extracted features. MASAM

enhances the features from three different angles to further extract the global dependencies of the features. Then, ASPP operation and image-level feature coding global context are carried out to further improve the performance. In addition, we concatenate and upsample the feature maps of different scales obtained in the feature extraction stage and the corresponding feature maps output by ASPP, respectively, so as to obtain the final segmentation prediction. Experimental results show that our method is much better than the competitive method, and our proposed MASAM achieves significant performance gain. Therefore, MASANet is a remote sensing images semantic segmentation model with excellent segmentation performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61702044).

7 REFERENCES

[1] Yuan, W. & Lining, X. (2015). Remote Sensing Satellite Networking Technology and Remote Sensing System: A survey. *2015 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI).* https://doi.org/10.1109/icemi.2015.7494508

[2] Radke, D. & Radke, D. (202). Beyond measurement: Extracting vegetation height from high resolution imagery with deep learning. *Remote Sensing*, *12*, 3797. https://doi.org/10.3390/rs12223797

[3] Yang, M. D., Tseng, H. H., Hsu, Y. C. et al. (2020). Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date uav visible images. *Remote Sensing*, *12*, 633. https://doi.org/10.3390/rs12040633

[4] Guirado, E., Blanco-Sacristán, J., Rodríguez-Caballero, R. et al. (2021). Mask r-cnn and obia fusion improves the segmentation of scattered vegetation in very high-resolution optical sensors. *Sensors*, *21*. https://doi.org/10.3390/s21010320

[5] Ayhan, B., Kwan, C., Budavari, C. et al. (2020). Vegetation detection using deep learning and conventional methods. Remote. *Sensors*, *12*, 2502. https://doi.org/10.3390/rs12152502

[6] Cai, J., Liu, C., Yan, H. et al. (2021). Real-time semantic segmentation of remote sensing images based on bilateral attention refined network. *IEEE Access*, *9*, 28349-28360. https://doi.org/10.1109/ACCESS.2021.3058571

[7] Jiang, Y., Li, M., Zhang, P. et al. (2021). Hierarchical fusion convolutional neural networks for sar image segmentation. *Pattern Recognition Letters*, *147*, 115-123. https://doi.org/10.3390/rs12152502

[8] Xu, J., Lu, K., & Wang, H. (2021). Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, *146*, 179-184. https://doi.org/10.1016/j.patrec.2021.03.015

[9] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[10] Badrinarayanan, V., Kendall, A., & Cipolla, R, (2017). A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[11] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation, *MICCAI*. https://doi.org/10.1007/978-3-319-24574-4_28

[12] Yu, F. & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. https://doi.org/10.48550/arXiv.1511.07122

[13] Chen, L. C., Papandreou, G., Schroff, F. et al. Rethinking atrous convolution for semantic image segmentation. https://doi.org/10.48550/arXiv.1706.05587

[14] Sheng, Y., Yang, J., Lin, Y. et al. (2021). Efficient semantic segmentation method with strip pooling for vhr remote sensing images. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2759-2762. https://doi.org/10.1109/IGARSS47720.2021.9553336

[15] Wu, Q., Luo, F., Wu, P. et al. (2021). Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 3-17. https://doi.org/10.1109/JSTARS.2020.3042816

[16] Zhao, H., Shi, J., Qi, X. et al. (2017). Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230-6239. https://doi.org/10.48550/arXiv.1612.01105

[17] Chen, L. C., Yang, Y., Wang, J. et al. (2016). Attention to scale: Scale-aware semantic image segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3640-3649. https://doi.org/10.1109/CVPR.2016.396

[18] Bhatnagar, S., Gill, L. W., & Ghosh, B, (2020). Drone image segmentation using machine and deep learning for mapping raised bog vegetation communities. *Remote Sensing*, *12*, 2602. https://doi.org/10.3390/rs12162602

[19] Wu, C., Ju, B., Xiong, N. N. et al. U-net super-neural segmentation and similarity calculation to realize vegetation change assessment in satellite imagery. https://doi.org/10.48550/arXiv.1909.04410

[20] Heryadi, Y., Irwansyah, E., Miranda, E. et al. (2020). The effect of resnet model as feature extractor network to performance of deeplabv3 model for semantic satellite image segmentation. *2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*, 74-77. https://doi.org/10.1109/AGERS51788.2020.9452768

[21] Woo, S., Park, J., Lee, J. Y. et al. (2018). Cbam: Convolutional block attention module. *ECCV*. https://doi.org/10.48550/arXiv.1807.06521

[22] Chen, L., Tian, X., Chai, G. et al. (2021). A new cbamp-net model for few-shot forest species classification using airborne hyperspectral images. *Remote Sensing*, *13*. https://doi.org/10.3390/rs13071269

[23] Hou, Q., Zhang, L., Cheng, M. M. et al. (2020). Strip pooling: Rethinking spatial pooling for scene parsing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4002-4011. https://doi.org/10.48550/arXiv.2003.13328

[24] Wang, X., Girshick, R. B., Gupta, A. et al. (2018). Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7794-7803. https://doi.org/10.48550/arXiv.1711.07971

[25] Xie, S., Girshick, R., Dollár, P. et al. (2017). Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987-5995. https://doi.org/10.48550/arXiv.1611.05431

[26] He, K., Zhang, X., Ren, S. et al. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[27] Baheti, B., Innani, S., Gajre, S. et al. (2020). Semantic scene segmentation in unstructured environment with modified deeplabv3+. *Pattern Recognition Letters*, *138*, 223-229. https://doi.org/10.1016/j.patrec.2020.07.029

Contact information:

Fuping ZENG, Lectural
School of Reliability and Systems Engineering,
Beihang University,
Beijing 100191, China
E-mail: zfp@buaa.edu.cn

Bin YANG, PhD
Du Xiaoman (Beijing) Science Technology Co., Ltd.,
Beijing 100094, China
E-mail: researcher_yang@outlook.com

Mengci ZHAO, Master
(Corresponding author)
School of Artificial Intelligence,
Beijing University of Posts and Telecommunications,
Beijing 100876, China
E-mail: zmc2603633412@bupt.edu.cn

Ying XING, Associate Professor
School of Artificial Intelligence,
Beijing University of Posts and Telecommunications,
Beijing 100876, China
E-mail: xingying@bupt.edu.cn

Yiran MA, Master
School of Reliability and Systems Engineering,
Beihang University, Beijing 100191, China
E-mail: 16231236@buaa.edu.cn