Computer Science and Information Systems 22(3):907–926 https://doi.org/10.2298/CSIS240713051T

Image Semantic Segmentation Based on Multi-layer Feature Information Fusion and Dual Convolutional Attention Mechanism

Lin Teng[1], Yulong Qiao[1,⋆], Jinfeng Wang[2], Mirjana Ivanović[3], and Shoulin Yin[1,4]

[1] School of Information and Communication Engineering, Harbin Engineering University
145 Nantong Street, Nangang District, Harbin, 150001, China
qiaoyulong@hrbeu.edu.cn
[2] Weifang Vocational College
Weifang, 261041 China
1057417325@qq.com
[3] Faculty of Sciences, University of Novi Sad, Serbia
mira@dmi.uns.ac.rs
[4] Software College, Shenyang Normal University
Shenyang, 110034 China
yslin@hit.edu.cn

Abstract. Traditional semantic segmentation methods have problems such as poor multi-scale feature extraction ability, weak lightweight backbone network feature extraction ability, lack of effective fusion of context information, resulting in edge segmentation errors and feature discontinuity. In this paper, a novel semantic segmentation model based on multi-layer information fusion and dual convolutional attention mechanism is proposed. In this method, SegFormer network is used as the backbone network, and multi-scale features of encoder output are fused with overlapping features. The feature extraction subnetwork is optimized by constructing the object region enhancement module, and the intermediate feature map is refined adaptively in each convolutional block of the deep network, so as to strengthen the fine extraction of multi-dimensional feature information of complex images. Dual convolutional attention module is used to fusion high-level semantic information to avoid the loss of feature information caused by up-sampling operation and the influence of introducing noise, and refine the effect of target edge segmentation. At the same time, the feature pyramid grid is proposed to process the overlapping features, obtain the context information of different scales, and enhance the semantic expression of features. Finally, the features processed by the feature pyramid grid module are combined to improve the segmentation effect. The experimental results on the public data set show that the proposed method has better performance than the existing methods, and has better segmentation effect on the object edge in the scene.

Keywords: Semantic segmentation, multi-layer information fusion, dual convolutional attention mechanism, feature pyramid grid.

⋆ Corresponding author

1. Introduction

Image semantic segmentation is one of the important research topics in the field of computer vision. Different from object detection and image classification, semantic segmentation processes images at the pixel level, and each pixel is assigned a corresponding label [1,2]. In the field of self-driving, semantic segmentation can segment roads, buildings, pedestrians, obstacles, etc., and give pixel-level annotations for each category. In the medical field, semantic segmentation can segment the location of lesions to assist doctors in diagnosis.

With the development of Convolution Neural Networks (CNN) [3], image semantic segmentation has developed rapidly. Fully Convolutional Networks (FCN) [4] replaced the fully connected layer with the convolutional layer to realize end-to-end semantic segmentation, so that the input image of semantic segmentation did not need a fixed size, which provided a foundation for subsequent methods in the field of semantic segmentation. Reference [5] proposed SegNet (Segmentation Networks), which recorded the positions during the maximum pooling operation during feature extraction and used the position information during up-sampling, but only the maximum position information was saved in this operation, and more information was still lost. The backbone of PSPNet (Pyramid Scene Parsing Networks) proposed in reference [6] adopts ResNet (Residual Networks) and introduced PPM for multi-scale information fusion, which could effectively solve the class confusion problem. DeepLab v3+ was introduced in reference [7] based on DeepLab series models. In the backbone part of this model, deep separable convolution and dilated convolution were adopted to replace traditional convolution and effectively alleviated the problem of parameter number. At the same time, DeepLabV3+ introduced Atrous Spatial Pyramid Pooling (ASPP) [8] to extract multi-scale features, which significantly improved the overall image segmentation accuracy.

The attention-mechanism-based approach is flexible in capturing the connections between global and local information. SENet is a network with pure channel attention mechanism [9], which obtains the weights of different channels through autonomous learning, thus expressing the importance of feature channels through different weights, and modeling the dependencies between channels. CBAM is a simple and effective mixed domain attention module, which processes feature maps in spatial domain and channel domain, and obtains good segmentation results by simple operation. PSANet [10] connects each location in the feature map with other locations through an adaptive learning attention mask to obtain contextual information. At the same time, the bidirectional information propagation path is designed, and the information collected from other locations is used to assist in predicting the current location, which is an efficient method. EMANet effectively aggregates features by introducing a multi-scale information fusion module [11], which uses feature maps at various scales to generate feature fusion results by weighted summation. However, such methods have high computational complexity, easy over-fitting and limited processing of long sequences, so the accuracy of complex images needs to be improved.

Encoder-decoder-based approaches [12,13] typically choose VGGNet [14] as the encoder and then replace its fully connected layer with the decoder structure. FCN solves the problem that convolutional neural networks limit the input image size, but the segmentation results are still rough. SegNet records the spatial position, so that it can accurately recover the image in the up-sampling stage, but the segmentation accuracy of the object

boundary is still not high. UNet [15] introduces fast connections between encoders and decoders, which has achieved good results on medical and remote sensing images, but the application scenarios of this network are limited. DFANer's encoder innovatively uses three sets of Xception networks [16], and the proposed subnet aggregation module optimizes the results. PointRend [17] regards image segmentation as a rendering problem in image processing, refins the rough mask edges generated by the network before, and implements the point-based segmentation module at the adaptive selected position. It can be seen that the accuracy of this kind of network in the segmentation of complex street view images needs to be improved.

The above methods consider the mining of context information and multi-scale information, but do not mine the correlation of semantic information between pixels, resulting in classification errors in complex scenes. Attention mechanism is a kind of mechanism that can establish the dependency relationship between pixels, which is introduced into the semantic segmentation network and plays an important role. SENet (Squeeze and Excitation Networks) was proposed in reference [18], which proved that the attention mechanism could reduce noise while improving classification performance. In reference [19], DANet (Dual Attention Networks) was proposed for semantic segmentation. Self-attention mechanism was also a kind of attention mechanism, self-attention mechanism calculated the similarity between features, and used this to capture the dependency between pixels. However, the computational overhead and memory overhead of self-attention mechanisms were squared complexity. In order to reduce the cost of computation and memory, reference [20] proposed CCNet (Criss-cross Networks), that is, each calculation only considered the row and column where the current pixel was located, and then indirectly connected the global information through the cascade of two cross-attention modules. Reference [21] proposed EMANet (Expectation maximization Attention Networks), which used expectation maximization clustering to optimize the attention mechanism and reduce the computational overhead. Reference [22] proposed EANet(External Attention Networks), which used two linear layers as the K and Q of the attention mechanism to represent the attention mechanism, reducing the computational cost and memory overhead, and the two linear layers could indirectly interact with the global information. Reference [23] proposed the Convolutional Block Attention Module (CBAM), which blended channel attention and spatial attention. Reference [24] rethought the aforementioned attention and proposed coordinate attention, which could achieve better performance improvement with almost no additional computing overhead.

With the application of semantic segmentation in practical projects, researchers have gradually turned their attention to the lightweight models rather than the accuracy of models. EdgeNeXt [25] attempted to combine the advantages of ViTs (Vision Transformers) with traditional convolution by introducing Split Depth-wise Transpose attention (SDT) to efficiently combine the advantages of ViTs and CNN without adding additional parameters and computation. ICNet (Image Cascade Networks) [26] used complex and deep paths to encode small size inputs. The MobileNets series [27,28] used deep separable convolution to replace traditional convolution operations. Xception built an entirely new network architecture using deep separable convolution. The above works focus on the lightweight of the model, which allows the segmentation model to run faster, but the accuracy limits its widespread use in practical applications.

Based on the above analysis, in order to further improve the accuracy of complex image semantic segmentation, a novel semantic segmentation model based on multi-layer information fusion and dual convolutional attention mechanism is proposed, the main contributions are as follows:

1. In this method, SegFormer network is used as the backbone network, and multi-scale features of encoder output are fused with overlapping features.
2. The feature extraction subnetwork is optimized by constructing the object region enhancement module, and the intermediate feature map is refined adaptively in each convolutional block of the deep network, so as to strengthen the fine extraction of multi-dimensional feature information of complex images.
3. The feature extraction subnetwork is optimized by constructing the object region enhancement module, and the intermediate feature map is refined adaptively in each convolutional block of the deep network, so as to strengthen the fine extraction of multi-dimensional feature information of complex images.
4. Dual convolutional attention module is used to fusion high-level semantic information to avoid the loss of feature information caused by up-sampling operation and the influence of introducing noise, and refine the effect of target edge segmentation. At the same time, the feature pyramid grid is proposed to process the overlapping features, obtain the context information of different scales, and enhance the semantic expression of features.
5. Finally, the features processed by the feature pyramid grid module are combined to improve the segmentation effect.

2. Related Works

The semantic segmentation method based on deep learning usually adopts the encoder-decoder structure. Firstly, the image is input to the feature extraction network, and then the extracted semantic features are sent to the decoder to analyze the semantic features and obtain the segmentation graph.

The concept of the attention mechanism stems from the human ability to process external information, that is, when faced with a large amount of information, people will focus their attention on information that is important to them and ignore irrelevant information. In computer science, this mechanism is known as attention, by giving different weights to information, so that the computer can pay more attention to the important parts of the information.

Compared with the traditional attention mechanism, the self-attention mechanism reduces the dependence on external information, is better at capturing the internal correlation of data or features. It can effectively extract global semantic information, thus improving the performance of the model. The self-attention mechanism is calculated as follows:

$$Q = W^Q X; K = W^K X; V = W^V X \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Where, Q (Query) is query vector matrix), K (Key) is key vector matrix) and V (Value) is value vector matrix), they are important matrices in self-attention mechanism. $\sqrt{d_k}$ is the dimension of the key vector. Specifically, it first multiplies the input vector X by the three corresponding initialization weight matrices W^Q, W^K, and W^V respectively to get Q, K, and V. Second, it calculates the dot product of Q and K to obtain the correlation between each of the two input vectors, that is, the similarity. To avoid the result of the dot product being too large or too small, the similarity is divided by $\sqrt{d_k}$. Then, the results are normalized using the softmax function to get the attention weights. Finally, these attention weights are dotted with V to get the output of the self-attention mechanism. In this way, the self-attention mechanism can extract and utilize the global information in the input sequence, and show excellent performance in processing tasks in fields such as natural language processing.

To further improve the performance of the model, Transformer introduces multi-head attention (MHA). This mechanism extends the original attention mechanism to allow the model to focus on different parts of the input sequence to obtain more contextual information. Specifically, the input word vector is transformed by different linear transformations to generate different Q, K and V, and then the self-attention operation is carried out to get the output result. Then, the output of all the "heads" is spliced, and finally through a linear transformation, the final output of the multi-head attention mechanism is obtained. This process takes place independently in multiple "heads," each with its own unique weight matrix. The advantage of this approach is that the model can focus on different features in different representation subspaces, and then integrate these features to get a richer feature representation, and make full use of computing resources, accelerate the training and reasoning process of the model, so as to improve the overall performance of the model.

3. Proposed Image Semantic Model

In this paper, the SegFormer network encoder structure is adopted to optimize the multi-stage feature fusion mode, and the feature pyramid grid module is designed and implemented to improve the feature discontinuity problem and improve the segmentation accuracy of the network by combining the overlapping convolutional dual convolutional attention mechanism to highlight important feature information.

SegFormer network adopts encoder-decoder structure, encoder structure consists of four stages. Each stage is stacked by multiple Transformer blocks. In each Transformer block, overlap patch embeddings module is used for feature extraction of input images. The extracted features are then used to calculate the feature correlation through the efficient multi-head self-attention module. Finally, the calculated features are passed through the mix-feed forward network (MixFFN) module [29]. SegFormer replaces position encoding in MixFFN and uses a depth-separable convolution with a convolution kernel size of 3×3 to provide position information. In the decoder part of the network, the resolution of the four stages of the encoder is 1/4, 1/8, 1/16 and 1/32 of the original image. The number of channels in the feature graph is uniformly adjusted to 256 by a convolution layer with convolution kernel size of 1×1. Then bilinear interpolation is used to up-sample the feature map to the 1/4 size of the original image, and multiple feature maps of the same size are spliced. Two convolution layers with convolution kernel size of 1×1 are used to complete pixelation-level prediction. Finally, bilinear interpolation is used to

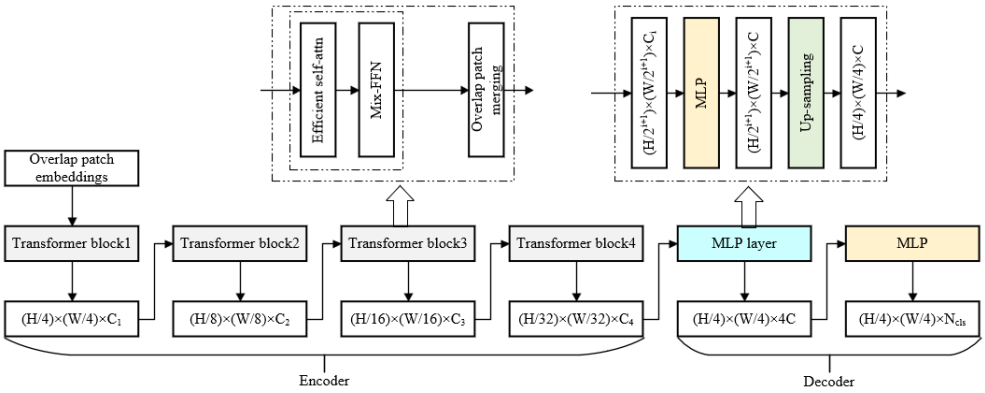

Fig. 1. SegFormer Network structure.

restore the image to the original image size and output the final segmentation result. The segmentation process is shown in Figure 1.

Although the SegFormer network performs well in scene segmentation, it up-samples multiple stage feature maps to the same size at one time in its decoder part, which easily results in insufficient fusion of low-level details and high-level semantic information. At the same time, the high level feature and the low level feature are directly spliced, which inevitably introduces noise and leads to the decrease of segmentation accuracy. In addition, SegFormer network does not introduce multi-scale context processing structure, which is easy to lead to discontinuous features and edge segmentation errors between objects of different scales, resulting in segmentation accuracy problems. Based on the above analysis, this paper proposes innovative improvements in three aspects. The first is to optimize the feature extraction subnetwork by constructing the object region enhancement module (OREM). The second is to use the overlapping feature fusion method to replace the original structure directly connected to the sampling and then merged, and to fuse the low-level features with the high-level features in stages. The dual convolutional attention module is used to calibrate the features of the high-level semantic information before the high-low feature fusion, and to suppress the interference of the low-level redundant information and the noise introduced by the up-sampling process. Third, depthwise separable residual convolution modules are proposed. On this basis, the residual feature pyramid grid (RFPG) is designed and implemented. The semantic information after the fusion of overlapping features is obtained through the feature pyramid grid modules of different scales to obtain the context information of different stages, strengthen the semantic association between features, improve the feature discontinuity problem, and improve the image segmentation accuracy.

3.1. Overlapping Feature Fusion

In this paper, the encoder-decoder structure is followed to build the network model, and the encoder adopts MiT-B2 of SegFormer model as the backbone network. The overall

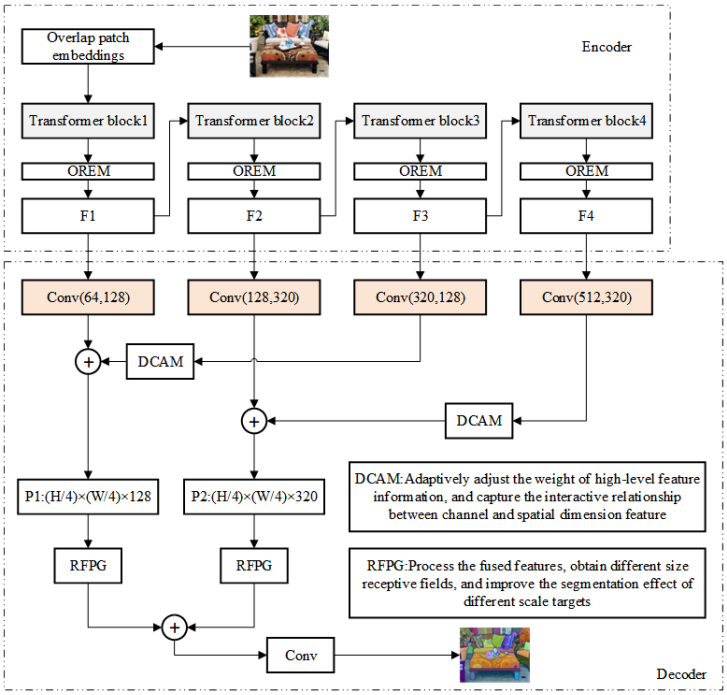

Fig. 2. The proposed model structure in this paper.

architecture of the model is shown in Figure 2. First, an object region enhancement module is established, and the feature maps of each stage are fused with overlapping features to preserve the rich details of low-level features and reduce the interference of noise on high-level semantics. The dual convolutional attention module (DCAM) is used to calibrate the high level feature weights, and then the two parts of the fused results are fed into the RFPG multi-scale feature extraction module with different sizes to extract multi-scale information of different granularity. Then the multi-scale information is merged to improve the segmentation effect of different scales, and the merged multi-scale information is up-sampled by bilinear interpolation to restore the input image size. Finally, the number of channels is adjusted to the number of categories by 1×1 convolutional layer, and the output result of the network is obtained.

As shown in Figure 1, in the decoder part of the original SegFormer network structure, feature graphs of different sizes output by each stage of the encoder are sampled to a unified size, then directly splicing them together, and the final segmentation graph is obtained by adjusting the number of channels through 1×1 convolution. Although this approach has fewer parameters and lower operation cost, it directly fuses low-level detailed features with high-level semantic features, which enricfies the detailed information and inevitably introduces a lot of noise, which affects the segmentation accuracy. For this reason, this paper, based on the MiT-B2 backbone network, carries out overlapping fusion of the extracted feature maps at each stage. As shown in Figure 2, the feature maps of different stages obtained by the backbone network are named F1, F2, F3 and F4, and the overlap-

ping feature fusion is divided into two parts. In P1, the feature map F3 is first adjusted by 1×1 convolution to adjust the number of channels, and the size of the feature map is obtained as $[H/16, W/16, 128]$. Then, the feature map resolution is sampled from 1/16 of the original image to 1/4 of the original image, and the relationship between features is adjusted by the attention module DCAM to enhance the high-level feature channels and spatial information. At the same time, the feature graph F1 is adjusted by 1×1 convolution to get the number of channels, and the size of the feature graph is $[H/4, W/4, 128]$. The two parts of the feature graph are added together to get the output of the first part. Similarly, in the second part of the overlapping fusion, the feature figure F4 is first adjusted to 320 channels by 1×1 convolution, then up-sampled by four times bilinear interpolation, and then added to the feature figure F2 with 320 channels by the attention module DCAM to obtain the output of the second part. Since the feature maps output at different stages of the encoder contain different feature information, the spatial position details of the feature maps F1 and F2 are rich in comparison, which helps to improve the segmentation effect of details such as target edges. However, semantic information is relatively insufficient, while the abstract semantic information contained in feature figures F3 and F4 is rich but lacks spatial details. If multiple feature maps are fused directly, although the detailed information of high-level features is enriched, the low-level feature maps also have a lot of noise in addition to the detailed information, and direct fusion is not conducive to improving the accuracy. The overlapping feature fusion method proposed in this paper fuses the feature maps F1 and F4 with those of the transition stages F3 and F2, and uses dual convolutional attention module DCAM to adaptively adjust the weight of high-level features, which not only realizes the fusion of high-low layer features, but also avoids the interference of lower-level noise on the segmentation results.

3.2. Object Region Enhancement Module (OREM)

The convolutional layer can automatically extract the multi-dimensional feature information from the original data by learning, but the multi-dimensional feature information obtained by most networks is limited. In order to obtain more multi-dimensional feature information and further solve the problem of low precision caused by different object scales in segmentation, this paper constructs an object region enhancement module (OREM), the structure is shown in Figure 3. The OREM module consists of THREE convolution blocks and a DAM, each convolution is connected by a Relu layer and connected by a residual structure, and finally output through a Relu layer [30]. OREM module adaptively refines the intermediate feature graph at each convolution block of feature extraction subnetwork, and dynamically processes the information of each feature graph to help the model learn the features of the data better. At the same time, the features with higher importance can be better expressed, the fine extraction of multi-dimensional features of complex street view images is strengthened, and the sub-network of feature extraction is optimized, which further solves the problem of low segmentation accuracy due to different target scales, occlusion overlap and illumination changes in segmentation, and is suitable for segmentation of complex images.

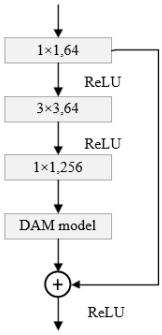

Fig. 3. OREM structure.

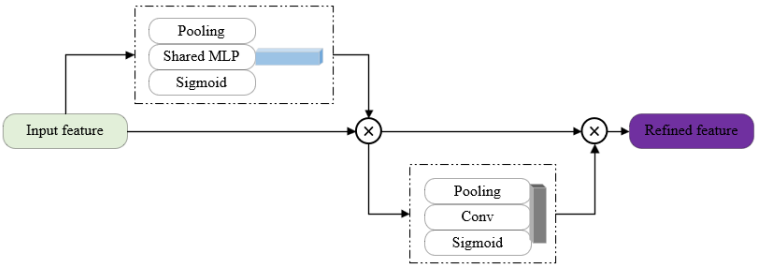

Fig. 4. DCAM structure.

3.3. Dual Convolutional Attention Module (DCAM)

Deep networks can obtain contextual information, but there is no focus on the region and information, resulting in segmentation accuracy has not been a big breakthrough. Therefore, this paper proposes the dual convolutional attention module (DCAM). The DCAM structure is shown in Figure 4. DCAM is a lightweight mixed-domain attention module that improves the accuracy of the network by learning the interrelationships between features. After the feature map is given, DCAM inferences the attention map sequentially along two independent dimensional channel domains and spatial domains, and adaptively weights different features, thereby increasing the weight of useful information and reducing noise and unimportant information.

The role of DCAM in the network is to input the feature map F into the channel attention module, and the input feature map is processed to get the corrected F'. Then the resulting F' is processed with the input F and re-input the spatial attention module to get the corrected feature diagram F”:

$$F' = M_C(F) \otimes F \tag{3}$$

$$F'' = M_S(F') \otimes F' \tag{4}$$

Where C is a channel. M_C is channel attention. M_S is spatial attention. F is the feature graph.

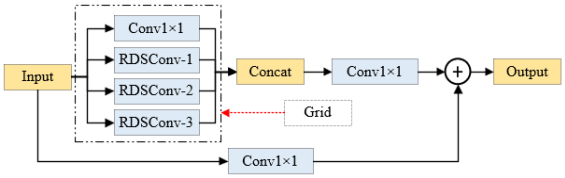

Fig. 5. RFPG module.

3.4. Residual Feature Pyramid Grid (RFPG)

Because the object scale is not uniform in the scene, it is easy to cause the segmentation accuracy is not high, edge segmentation errors and other problems. Therefore, it is very important to obtain multi-scale context information of feature maps to extract features of different scales for improving segmentation accuracy and alleviating feature discontinuity caused by differences between objects of different scales. Therefore, the RFPG module is designed and implemented in this paper. As shown in Figure 5, the RFPG module is mainly built using the depth-separable residual convolution module in this paper. First, the use of depth-separable convolution can reduce the number of parameters in the model, which is conducive to the lightweight of the model, but at the same time, depth-separable convolution separates the spatial dimension operations from the channel dimension operations, which easily leads to the problem of discontinuity between features. Therefore, in this paper, depth-separable residual convolution is proposed to unify the spatial and channel dimensional features, enhance the semantic expression between features without introducing additional parameters, and avoid the phenomenon of gradient disappearance or gradient explosion caused by deepening network hierarchy.

The RFPG module is mainly composed of 1×1 convolution and three RDSConv modules with different expansion rates. The common depth separable convolution operation (DSConv) [31,32] is composed of two parts. First, the input feature graph is separated by the expansion convolution with the convolution kernel size of 3×3, and then the channel dimensions are merged by 1×1 convolution. The specific calculation formula of DSConv is shown in equation (3):

$$output = Conv_{1\times1}(DConv_{3\times3}(input)) \tag{5}$$

$input$ indicates the input feature map of the module. $output$ indicates the output feature map. $Conv1 \times 1$ represents an ordinary convolution with a convolution kernel size of 1×1. $DConv3 \times 3$ represents an expansive convolution with a convolution kernel size of 3×3.

This paper proposes to introduce residual operation into ordinary depth-separable convolution, that is, after performing 3×3 dilated convolution operation on the feature graph, adding the convolution result to the input, and then obtaining the output through 1×1 convolution, which is different from $DSConv$ in that space and channel are operated separately. $RDSConv$ combines spatial dimension operations with channel dimension operations to enhance semantic feature association and alleviate the problems of discontinuous target segmentation and unclear edge segmentation. The specific calculation process of $RDSConv$ is shown in equation (4):

Table 1. RFPG module parameters

Stage	Size of the input feature graph	Expansion rate setting	Input channel number	Output channel number
P1	$(H/4, W/4)$	(1,3,6,9)	128	256
P2	$(H/8, W/8)$	(1,6,12,24)	320	256

$$output = Conv_{1\times1}(input + DConv_{3\times3}(input)) \tag{6}$$

In the RFPG module, the input feature map is concatenated by 1×1 convolution and three $RDSConv$ models with different expansion rates, and the obtained results are concatenated with channel dimensions, and then adjusted to the specified number of channels by 1×1 convolution. At the same time, 1×1 convolution is used to adjust the input feature diagram to the specified number of channels, and the output of the RFPG module is obtained by adding the two together. The specific operation process is shown in Figure 5.

As shown in Figure 2, two parts of output are obtained after overlapping feature fusion of multi-stage encoder feature maps. The first part is obtained by the fusion of feature maps F1 and F3, and the second part is obtained by the fusion of feature maps F2 and F4. In view of the different information contained in these two features, RFPG modules with different expansion rates are used to extract multi-scale features. For the first part, the expansion rate in the RFPG module used is (1,3,6,9); For Part 2, the expansion rate used is (1,6,12,24), and an expansion rate of 1 means that a common convolution of 1×1 is used. The output channel of each convolutional module is set to 256, as shown in Table 1.

For the first part, RFPG module is constructed with a small expansion rate, mainly to avoid the interference of redundant information as far as possible, aiming at the characteristics of rich detailed information in its characteristics. For the second part, which is rich in semantic information, a larger expansion rate is used to increase the receptive field of the network, obtain semantic information at different scales, and improve the segmentation accuracy of the network.

For the two-part feature maps obtained by the fusion of overlapping features, the two-part feature maps with sizes $[H/4, W/4, 256]$ and $[H/8, W/8, 256]$ are obtained after the respective RFPG modules. Then, the feature map of Part 2 is up-sampled with two times bilinear interpolation, and the size of the feature map obtained is $[H/4, W/4, 256]$, which is the same dimension as the feature map of Part 1. Finally, the feature graphs after the fusion of the two parts are added, and the channels of the feature graphs are adjusted by bilinear interpolation, and then the output result of the network is obtained by 1×1 convolution.

3.5. Loss Function

In the field of image semantic segmentation, the loss function has many forms, the most commonly used is the cross-entropy loss function. The function calculates the loss by measuring the similarity between the predicted distribution and the true distribution. The predicted distribution is close to the true distribution, so the smaller the loss function value is smaller. Otherwise it is larger. Its expression is shown in equation (5):

$$Loss_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij}) \tag{7}$$

Where N is the number of samples. C is the number of categories. y_{ij} is the label where sample i belongs to class j, with a value of 0 or 1. p_{ij} is the probability that the model predicts the sample to be of class j, with a value between 0 and 1. The limitation of the cross-entropy loss function is that it does not take into account the unbalanced distribution of labels. When the number of pixels of different categories is very different, the training of the loss function will become more difficult. In addition, the cross-entropy loss function [33] only calculates the loss value of each pixel discretely and then averages it, rather than considering the prediction result of the whole image globally. In order to make up for the deficiency of cross entropy loss function, Dice loss function and its variant Tanimoto loss function are introduced, whose expressions are shown in equation (6) and equation (7) respectively:

$$Loss_{Dice} = 1 - \frac{2\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}p_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}+p_{ij})} \tag{8}$$

$$Loss_{Tanimoto} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}p_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij}+p_{ij}-y_{ij}p_{ij})} \tag{9}$$

The Dice loss function and Tanimoto loss function are numerically equivalent, and both help to solve the problem of difficult training when label distribution is unbalanced. However, when there are more small targets, the loss function is prone to oscillation, and even gradient saturation occurs in extreme cases. Furthermore, as a rule of thumb, a loss function with a quadratic in the denominator is more likely to bring the prediction closer to the true value, regardless of the random initial value of the weight. Therefore, the weighted sum of cross entropy loss function and Tanimoto loss function is selected as the overall loss function, whose expression is shown in equation (8):

$$Loss_{total} = Loss_{CE} + \gamma Loss_{Tanimoto} \tag{10}$$

Where γ is the parameter that balances the effects of the cross-entropy loss function and the Tanimoto loss function, and its value ranges from $(0, +\infty)$.

4. Experimental Results and Analysis

4.1. Data Sets

The performance of the proposed semantic segmentation model is evaluated on two data sets, PASCAL VOC 2012 and Cityscapes. The PASCALvOC 2012 data set is used for training and performance evaluation of the model, and the Cityscapes data set is used for generalization performance testing of the model.

PASCAL VOC 2012 is a widely used public image data set in computer vision. The data set has 21 semantic categories, including 20 object classes and one background class. There are 2913 tagged images. Among them, 2700 images are randomly selected as the

training set, 300 images are selected as the verification set, and the image size of the input semantic segmentation network is set to 512×512.

Cityscapes is one of the well-known data sets of autonomous driving scenarios in urban environments. The data set has 34 semantic categories, of which only 19 are used based on previous work. There are a total of 5000 finely labeled images, each with a resolution of 1024×2048, and in order to be consistent with the VOC2012 data set, 4500 images randomly selected from the data set are used as the training set and 500 images are used as the validation set. The image size of the input semantic segmentation network is set to 512×1024.

4.2. Evaluation Index

In this paper, mean intersection over union (MIoU) and mean accuracy (MAcc) are used to evaluate the experimental results, the expressions are shown in equations (9) and (10):

$$MIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{11}$$

$$MAcc = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{12}$$

Where N represents the total number of categories. TP_i is the number of pixels correctly predicted by category i. FP_i is the number of pixels that predict the other category as category i. FN_i is the number of pixels that predict category i as other categories. $MIoU$ is to calculate the intersection ratio of each semantic class separately, and calculate the average value after summing. $MAcc$ is to calculate the accuracy of each semantic class separately, and calculate the average value after summing.

4.3. Experiment Settings

This experiment is implemented on PyTorch framework, the operating system is Windows11 64-bit operating system, the processor is Intel(R) Xeon(R)Gold 5218R, the graphics card is NVIDIA A10. The memory is 128 GB, and the hard disk is 1TB.

Adam algorithm is used as the optimizer, and the momentum is set to 0.8. Applying the "poly" learning rate strategy, the learning rate gradually decreases with the increase of the number of iterations, and its expression is shown in equation (11).

$$lr - current = lr - initial(1 - \frac{T}{T_{max}})^{mom} \tag{13}$$

Where $lr - current$ is the current learning rate. $lr - initial$ is the initial learning rate, which is defined as 5×10^{-4}. T indicates the current number of iterations and T_{max} indicates the maximum number of iterations. mom is the momentum, it is 0.9. Also, it sets the batch size for each training round to 16, the number of training rounds to 200. In the first 100 rounds of training, the parameters of the backbone network are frozen so that it does not participate in the training, and in the 101-200 rounds of training, the backbone network is thawed.

Table 2. Objective index calculation results

Backbone	MIoU	MAcc
MobileNetv2	77.34/%	86.23/%
ResNet101	80.37/%	89.14/%
Xception	82.74/%	90.50/%
SegFormer	83.78/%	92.61/%

Table 3. Ablation results/%

Method	MIoU	MAcc
Baseline	82.74	90.50
Baseline+OREM	84.49	90.86
Baseline+DCAM	84.73	90.94
Baseline+RFPG	84.58	90.73
Baseline+DCAM+RFPG	85.78	91.45
Baseline+OREM+DCAM+RFPG	85.55	91.42

Table 4. Experimental results of different loss functions/%

Loss function	MIoU	MAcc
$Loss_{CE}$	85.55	91.42
$Loss_{CE} + Loss_{Dice}$	85.73	91.49
$Loss_{CE} + Loss_{Tanimoto}$	85.76	91.54
$Loss_{CE} + 0.5Loss_{Tanimoto}$	85.90	91.64
$Loss_{CE} + 0.3Loss_{Tanimoto}$	86.02	91.73
$Loss_{CE} + 0.25Loss_{Tanimoto}$	85.94	91.57

4.4. Ablation Experiment

Firstly, three deep convolutional neural networks including MobileNetv2, ResNet101 and Xception are selected as the backbone network for the experiment, and the results are shown in Table 2. It can be seen that SegFormer has the highest MIoU and MAcc and the best segmentation effect, so SegFormer is chosen as the backbone network of the model.

Then, ablation experiments are conducted on different modules used in the model, and the results are shown in Table 3. It can be seen that when OREM+DCAM+RFPG module is added on the basis of the baseline model, the $MIoU$ and $MAcc$ of the proposed model are the highest, so this method is adopted as the final network structure.

In the above ablation experiments, the loss function used by the model is the cross entropy loss function. Finally, on the basis of the model in this paper, different loss functions are used for training, and the results are shown in Table 4. It can be seen that when the overall loss function $Loss_{total} = Loss_{CE} + 0.3Loss_{Tanimoto}$, the training effect is the best, indicating that it is easier to use this loss function to converge the model parameters to the optimal value.

4.5. Comparison Experiment results on the PASCAL VOC 2012 data set

This paper is compared with the current popular semantic segmentation models on the PASCAL VOC 2012 data set, and the results are shown in Table 5. It can be seen that

Table 5. Comparison with other methods on the PASCAL VOC 2012 data set/%

Method	MIoU	MAcc
SegNet	60.93	73.53
FCN	62.50	75.31
DeepLab	67.14	76.72
U-Net	73.05	80.69
DeconvNet	74.46	84.77
BiSeNet [34]	79.97	87.65
APCNet	80.67	87.22
PSPNet [35]	82.30	88.68
HRNet [36]	83.91	89.23
DMNet [37]	84.66	90.86
Proposed	**86.02**	**91.73**

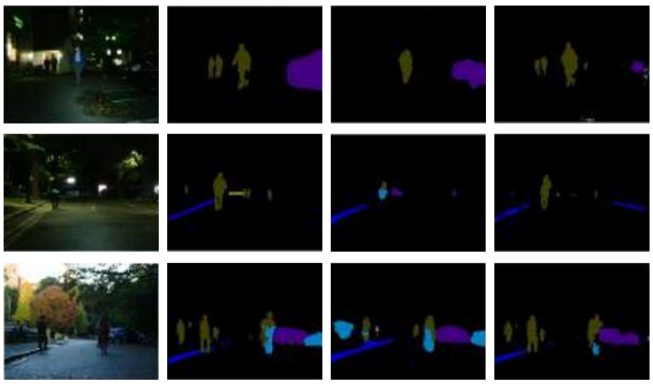

Fig. 6. Comparison of visual results on the PASCAL VOC 2012 data set

the proposed method in this paper achieves the best results on MIoU and MAcc. Compared with DMNet, the MIoU and MAcc of the proposed method increases by 1.32% and 0.87% respectively. Compared with HRNet and PSPNet, MIoU increases by 2.11% and 3.72%, and MAcc increases by 2.50% and 3.05%, respectively. It can be seen that the performance of the proposed method is generally better than that of the current popular semantic segmentation models.

The visualization results of the proposed method and DMNet are shown in Figure 6. The first column is the input image, the second column is the labeled image, and third and fourth column are the segmentation results of DMNet, proposed method, respectively. As can be seen from Figure 6, DMNet makes errors in the segmentation of pedestrians and vehicles, resulting in typical segmentation discontinuities, while none of these errors are found in the proposed method. This is because the proposed method in this paper directly extracts edge features from the original input image and fuses them with the feature image output after sub-sampling by the encoder, so that the feature image has rich shallow spatial information and deep semantic information at the same time. Moreover, the at-

Table 6. Comparison with other methods on the Cityscapes data set/%

Method	MIoU	MAcc
FCN	62.61	70.46
DeepLab	63.18	72.28
U-Net	63.69	71.41
BiSeNet	69.38	76.92
DeepLabv3+	73.87	79.59
PSPNet	74.72	80.77
Proposed	**76.03**	**81.90**

tention mechanism is used to enhance the meaningful information, while other semantic segmentation models do not supplement the spatial details into the feature images after down-sampling. Therefore, the proposed model has significant advantages in object edge segmentation, and the overall segmentation performance is better.

Thereby, by constructing object region enhancement module and dual attention module, the proposed semantic segmentation model in this paper can recover the spatial details of feature images after encoder down-sampling to a certain extent, enhance the accuracy of object edge segmentation, and pay more attention to meaningful information. In addition, by improving the loss function, the parameters of the proposed model in this paper converge more easily to the optimal value, and finally the overall effect of semantic segmentation is improved to some extent. The experimental results show that the proposed model has made remarkable progress in semantic segmentation, especially in object edge segmentation.

4.6. Comparison experiment results on the Cityscapes data set

In order to verify the generalization ability of the model in this paper, Cityscapes data set is used for generalization experiment, and the results are shown in Table 6. The proposed model achieves the best results on MIoU and MAcc, which are 2.16% and 2.31% higher than DeepLabv3+, respectively.

The results of the Proposed visualization with DMNet are shown in Figure 7. It can be seen that the Proposed method can segment the edge part of the object more accurately, the segmentation results are complete and clear, and the overall performance is better.

5. Conclusion

In this paper, a novel semantic segmentation model based on multi-layer information fusion and dual convolutional attention mechanism is proposed to solve the problem of object edge missegmentation and feature discontinuity in scene segmentation. Firstly, the ability to extract multi-dimensional feature information is improved by constructing the object region enhancement module. Combined with the dual convolutional attention mechanism, the high and low level features are fused and the interference of redundant information is suppressed. A pyramid feature grid module combined with residuals is designed and implemented to enhance the semantic expression between features and alleviate the problem of feature discontinuity. The experimental results show that the proposed

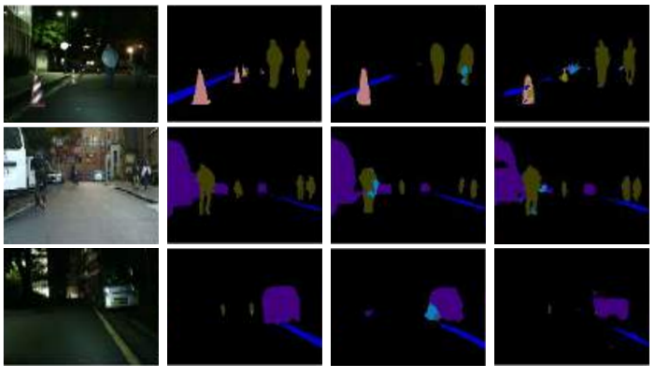

Fig. 7. Comparison of visual results on the Cityscapes data set.

method in this paper can effectively solve the problems of segmentation void and unclear target edge, improve the segmentation accuracy, and the segmentation effect is good. The next step will optimize the lightweight structure and improve the generalization ability of the model in different scenarios.

Acknowledgments. Authors are grateful for the anonymous review by the review experts.

References

1. Luo Z, Yang W, Yuan Y, et al. Semantic segmentation of agricultural images: A survey[J]. Information Processing in Agriculture, 2023.
2. Mo Y, Wu Y, Yang X, et al. Review the state-of-the-art technologies of semantic segmentation based on deep learning[J]. Neurocomputing, 2022, 493: 626-646.
3. Yin S, Wang L, Teng L. Threshold segmentation based on information fusion for object shadow detection in remote sensing images[J]. Computer Science and Information Systems, 2024. doi: 10.2298/CSIS231230023Y.
4. 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.
5. 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.
6. 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.
7. Yuan H, Zhu J, Wang Q, et al. An improved DeepLab v3+ deep learning network applied to the segmentation of grape leaf black rot spots[J]. Frontiers in plant science, 2022, 13: 795410.
8. Lian X, Pang Y, Han J, et al. Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation[J]. Pattern Recognition, 2021, 110: 107622.
9. Li X, Li M, Yan P, et al. Deep learning attention mechanism in medical image analysis: Basics and beyonds[J]. International Journal of Network Dynamics and Intelligence, 2023: 93-116.

10. Zhao H, Zhang Y, Liu S, et al. Psanet: Point-wise spatial attention network for scene parsing[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 267-283.
11. Li X, Zhong Z, Wu J, et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9167-9176.
12. Asadi A, Safabakhsh R. The encoder-decoder framework and its applications[J]. Deep learning: Concepts and architectures, 2020: 133-167.
13. Qamar S, Ahmad P, Shen L. Dense encoder-decoder Cbased architecture for skin lesion segmentation[J]. Cognitive Computation, 2021, 13(2): 583-594.
14. S. Yin, H. Li, Y. Sun, M. Ibrar, and L. Teng. Data Visualization Analysis Based on Explainable Artificial Intelligence: A Survey[J]. IJLAI Transactionss on Science and Engineering, vol. 2, no. 2, pp. 13-20, 2024.
15. Li X, Chen H, Qi X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes[J]. IEEE transactions on medical imaging, 2018, 37(12): 2663-2674.
16. Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
17. Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
18. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
19. Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 299-307.
20. Huang Z, Wang X, Huang L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 603-612.
21. Li X, Zhong Z, Wu J, et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9167-9176.
22. Katafuchi R, Tokunaga T. LEA-Net: layer-wise external attention network for efficient color anomaly detection[J]. arxiv preprint arxiv:2109.05493, 2021.
23. Agac S, Durmaz Incel O. On the use of a convolutional block attention module in deep learning-based human activity recognition with motion sensors[J]. Diagnostics, 2023, 13(11): 1861.
24. Yang Y, Jiao L, Liu X, et al. Dual wavelet attention networks for image classification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(4): 1899-1910.
25. Maaz M, Shaker A, Cholakkal H, et al. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 3-20.
26. Zhao S, Dong Y, Chang E I, et al. Recursive cascaded networks for unsupervised medical image registration[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 10600-10610.
27. Sinha D, El-Sharkawy M. Thin mobilenet: An enhanced mobilenet architecture[C]//2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON). IEEE, 2019: 0280-0285.
28. Qin Z, Zhang Z, Chen X, et al. Fd-mobilenet: Improved mobilenet with a fast downsampling strategy[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 1363-1367.
29. Qin Z, Zhang Z, Chen X, et al. Fd-mobilenet: Improved mobilenet with a fast downsampling strategy[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 1363-1367.

30. Yin S, Li H, Laghari A A, et al. An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G internet-of-everything[J]. IEEE Internet of Things Journal, 2024.
31. Zhang K, Cheng K, Li J, et al. A channel pruning algorithm based on depth-wise separable convolution unit[J]. IEEE Access, 2019, 7: 173294-173309.
32. Dang L, Pang P, Lee J. Depth-wise separable convolution neural network with residual connection for hyperspectral image classification[J]. Remote Sensing, 2020, 12(20): 3408.
33. Li X, Yu L, Chang D, et al. Dual cross-entropy loss for small-sample fine-grained vehicle classification[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4204-4212.
34. Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.
35. Zhou J, Hao M, Zhang D, et al. Fusion PSPnet image segmentation based method for multi-focus image fusion[J]. IEEE Photonics Journal, 2019, 11(6): 1-12.
36. Seong S, Choi J. Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates[J]. Remote Sensing, 2021, 13(16): 3087.
37. Fang K, Li W J. DMNet: difference minimization network for semi-supervised segmentation in medical images[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2020: 532-541.

Lin Teng received the M.A. degree from Shenyang Normal University, Shenyang, China, in 2020. She is currently pursuing the Ph.D. degree with the College of Information and Communication Engineering, Harbin Engineering University, Harbin. Her research interests are image processing and semantic segmentation.

Yulong Qiao was born in 1978. He received the Doctoral degree in engineering from the Harbin Institute of Technology in 2006. In recent years, he has published more than 50 papers in domestic and foreign journals and important conferences, among which more than 20 papers were indexed by SCI, and more than 30 were indexed by EI. His research interests include signal representation and analysis, statistical image processing, image/video processing and applications, and texture analysis and applications. He has won two science and technology awards of Heilongjiang Province. He is a Senior Member of the Chinese Institute of Electronics and IEEE Signal Processing Branch.

Jinfeng Wang received the M.A. degree from Shenyang Normal University, Shenyang, China, in 2016. She currently works at Weifang Vocational College. She has published several academic papers related to her subject. She received several project awards. Her research interests are education management, information processing and big data.

Mirjana Ivanovic (Member, IEEE) has been a Full Professor with the Faculty of Sciences, University of Novi Sad, Serbia, since 2002. She has also been a member of the University Council for informatics for more than 10 years. She has authored or coauthored 13 textbooks, 13 edited proceedings, 3 monographs, and of more than 440 research articles on multi-agent systems, e-learning and web-based learning, applications of intelligent techniques (CBR, data and web mining), software engineering education, and most of which are published in international journals and proceedings of high-quality international conferences. She is/was a member of program committees of more than 200

international conferences and general chair and program committee chair of numerous international conferences. Also, she has been an invited speaker at several international conferences and a visiting lecturer in Australia, Thailand, and China. As a leader and researcher, she has participated in numerous international projects. She is currently an Editor-in-Chief of Computer Science and Information Systems Journal.

Shoulin Yin received the M.A. degree from Shenyang Normal University, Shenyang, China, in 2015. He is currently pursuing the Ph.D. degree with the College of Information and Communication Engineering, Harbin Engineering University, Harbin. His research interests are remote sensing image processing and object detection.

Received: July 13, 2024; Accepted: January 17, 2025.