IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 February 2024, accepted 28 April 2024, date of publication 30 April 2024, date of current version 8 May 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3395541*

RESEARCH ARTICLE

# Research on Image Semantic Segmentation Based on Hybrid Cascade Feature Fusion and Detailed Attention Mechanism

**ZUOQIANG DU**[1] **AND YUAN LIANG**[2]
[1]School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China
[2]Jinan Inspur Data Technology Company Ltd., Jinan 250000, China

Corresponding author: Yuan Liang (yuanliang_hrb@163.com)

This work was supported in part by the Natural Science Foundation of China, under Grant 606750192; in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province, under Grant UNPYSCT-2020212; and in part by the Science Foundation of Harbin Commerce University, under Grant XL0095.

**ABSTRACT** In view of the low segmentation accuracy for small-scale object and insufficient segmentation of local boundary for semantic segmentation methods based on Deep Learning, this paper proposes an image semantic segmentation approach based on attention mechanism and feature fusion. On the basis of ensuring the overall accuracy, the segmentation accuracy of small-scale object and local boundary is improved, and it meets the requirement of accurately segmenting the object in the complex background. Firstly, an image semantic segmentation model based on hybrid cascade and feature fusion is proposed, and the hybrid concatenation and multi-cores pooling methods are used to extract deeper semantic information. Then, a cross-stages fusion approach is designed to divide the backbone network of the encoder stage in the network and the improved Atrous Spatial Pyramid Pooling module into three stages to fully utilize the different semantic information of the shallow and deep layers. Thirdly, the attention mechanism is introduced into the hybrid cascade and feature fusion image semantic segmentation network model, and image semantic segmentation model based on cross-stages and attention mechanisms is explained. Self attention is added to channel attention enhances the connection between feature maps, and one-dimensional convolution in the spatial attention mechanism is used to increase the spatial receptive field. The final results on the public dataset PASCAL VOC2012 and SUIM show that MIoUs have reached 86.68% and 61.55% respectively, and it proved that the overall accuracy of the approach proposed in this paper is higher than other ones.

**INDEX TERMS** Semantic segmentation, hybrid cascade feature fusion, detailed attention mechanism, deep learning.

## I. INTRODUCTION

Image semantic segmentation has always been an important research direction in image processing. Although traditional methods are more mature than Deep Learning (DL), there is still a big gap in accuracy and speed. Meanwhile, semantic segmentation based on Deep Convolutional Neural Networks (CNNs) has attracted more and more attention [1], [2], [3], [4]. Deep CNNs do not only improve the recognition accuracy of image classification, it also plays a great role in promoting the structured output of local tasks, including the generation of target boxes in object recognition and the detection ofobject key points [5], [6], [7], [8]. The application of Deep CNNs to image semantic segmentation technology has great significance for improving the application of this network in scene understanding and classification [1], [9], [10], [11], [12], [13]. Although the current research progress has made a lot of achievements, the approaches on image semantic segmentation still have the problem that some details are lost, which leads to poor segmentation effect of small-scale object and unclear segmentation of object boundary under complex background conditions [12], [14], [15], [16].

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

In 2012, Krizhevsky et al. proposed an AlexNet architecture where ReLU was proposed as the activation function and dropout is introduced to reduce the overfitting phenomenon. At the same time, the local response is standardized to enhance the generalization performance of the network model. The use of large amounts of training data and GPU parallel computation is the AlexNet's innovation which greatly improves the accuracy of image classification [17]. In 2014, Long et al. put forward the Full Convolutional Networks (FCN) [18]. Different from the traditional CNN model, FCN abandoned the full connection layer, and it superimposed the convolutional layer and the pooling layer together to form a complete CNN. In the training process of the network, FCN will adjust the step of each convolutional layer in the network and the size of the pooling layer according to the input image to ensure that the network can process any size of the input image. In the inference phase, FCN gets a pixel-level output of the same size as the original image by propagating the entire image forward. In 2015, Ronneberger et al. proposed an U-Net model, which is often used in medical segmentation [19]. U-Net network adopted the method of copy clipping to connect encoder and decoder, which can effectively transmit low-level information and thus improve the accuracy of segmentation results. However, the disadvantage of U-Net is that the encoder can only perform local operations and cannot integrate the global information of the image.

Lin et al. established the RefineNet network which successfully solved the problem of information loss in 2016 [20]. The RefineNet network model utilized multi-resolution feature fusion techniques to utilize features from different scales preferably. Specifically, RefineNet extracted features at different levels and fused these features together based on a series of submodules to generate semantic segmentation results. However, the disadvantage of this network is local missing images. Chen et al. [21] proposed DeepLabv1, which is the first-generation version of DeepLab series model. DeepLabv1 used the full convolutional network and void convolution to improve the accuracy of segmentation results. This version of the model was trained and tested on the PASCALVOC2012 dataset, on which it achieved state-of-the-art results at the time. In 2017, Zhao et al. established a PspNet network, which used pyramid pooling module to process feature maps in a cascading manner and integrate information among different scales and subregions [22]. The prior information and the original feature map are added and input into the final convolution module to complete the prediction. Chen et al. [23] improved the DeepLabv1 network model and proposed DeepLabv2 in 2017. It adds Atrous Spatial Pyramid Pooling (ASPP) and Conditional Random Field (CRF) to DeepLabv1 model to improve the quality of segmentation results. ASPP can capture different features of object at multiple dimensions, while CRF can add contextual information to the segmentation results.

In December of the same year, Liu et al. proposed DeepLabv3 based on DeepLabv1 and DeepLabv2 [24]. DeepLabv3 further improved the core in DeepLabv2 and introduced new techniques to improve the accuracy and efficiency of the segmentation results. It used Deformable ConvNets to replace empty convolution, and Batch Normalization and Depthwise Separable Convolution were used to improve the efficiency of the model. In 2018, Chen and his team proposed the DeepLabv3+ network model. The network model used the structure of DeepLabv3 as the encoder to extract rich context information of images, and it explored the cavity convolution of different expansion rates in ASPP module to obtain the feature maps of different sizes of sensitive fields. In 2019, Takikawa et al. [25] proposed an approach called ''two-stream'' convolutional networks. The fusion of two kinds of information can improve the accuracy and robustness of image semantic segmentation. The advantage of this approach is able to handle two different types of features simultaneously, and reduces conflict and duplication between features.

Graph-fcn network was proposed by Lu et al. in 2020 [26], which used the middle feature layer of semantic segmentation network to build a Graph network model for solving effectively the problem that local location information may be lost in the process of feature extraction. Recently, Ding et al. [27] proposed SCARF semantic segmentation model in 2021, which adopted multi-level feature extraction, including local feature, global feature and contextual feature. In 2022, Song et al. [28] proposed the FLANet network, which used a single similarity map combined with spatial and location coding to solve the feature loss problem in non-local attention modules.

In recent years, there had been many researches on semantic segmentation based on deep learning, but there are still some problems such as large amounts of computation, poor segmentation of small scale object and discontinuity of local boundary segmentation in complex background. Aim to improve the segmentation precision of small scale object and local boundaries, we make the contributions as follow:

(1) Inspired by DeepLabv3+ network, this paper improves the ASPP module as encoder to network model for obtaining deeper and richer semantic information.

(2) The method of cross-stage feature fusion is used to fuse the shallow and deep features extracted from each level to improve the accuracy of semantic segmentation.

(3) A detailed attention mechanism module is designed to integrate the feature information of different scales to suppress the meaningless features and enhance the meaningful ones.

(4) The network model designed in this paper improves the segmentation accuracy of both small-scale object and local boundary by comparing on the open data set of semantic segmentation.

## II. IMAGE SEMANTIC SEGMENTATION BASED ON CONVOLUTIONAL NEURAL NETWORK

CNNs adopt the design idea of convolutional kernel parameter sharing, which greatly reduces the number of network

parameters compared with other neural networks, thus decreasing the overfitting problem that often occurs in the training stage. DeepLabv3+ is a typical semantic segmentation model based on CNN.

Compared with other DeepLab series semantic segmentation models, DeepLabv3+ has a great improvement in segmentation accuracy and speed. The encoder part of DeepLabv3+ extracts the image feature information from the Deep Convolutional Neural Network (DCNN), and uses the atrous convolution of different expansion rates in the Atrous Spatial Pyramid Pooling (ASPP) module to obtain the feature map of different sizes of sensory fields.

In the decoder part, the low-resolution features extracted from DCNN are fused with the high-resolution features upsampled to aggregate the context information of different regions, and the spatial information is recovered with $3 \times 3$ convolution and 4 times bilinear interpolation to obtain the prediction image with high precision. The network structure is shown as Figure 1.

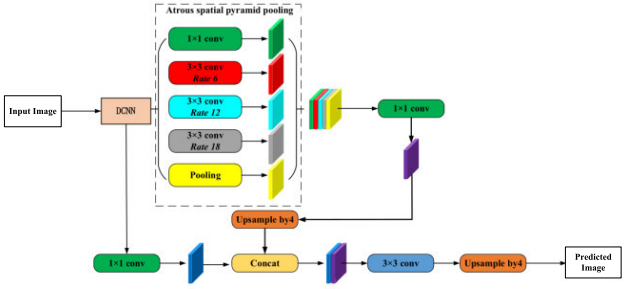



**FIGURE 1.** DeepLabv3+ model structure.

## III. IMAGE SEMANTIC SEGMENTATION MODEL BASED ON HYBRID CASCADE AND FEATURE FUSION

In order to solve the problem that semantic segmentation model is not effective in considering image segmentation under complex background, we propose an image semantic segmentation model based on hybrid cascade and feature fusion. The network is based on DeepLabv3+ [29]. The backbone network adopts a lightweight network Xception developed by MSRA team, which is combined with the improved hollow ASPP module to improve the feature extraction ability of the model. Feature extraction module and image resolution recovery module form the network model structure of the resolver to realize semantic segmentation of image objects. The improved ASPP introduces multi-core pooling module, and extracts deep semantic information by hybrid cascade, so as to improve the model ability to acquire target information under complex background. Figure 2 shows the network structure based on hybrid cascade and feature fusion.

### A. BACKBONE NETWORK

The function of the backbone network is to extract the features of the input image initially, and then carry out convolution processing to take out the rich semantic features of the input image. The current semantic segmentation network models pursue the balance between segmentation accuracy and real-time, so the lightweight network with strong generalization is usually adopted. In this paper, the highest precision lightweight network Xception is used to extract the preliminary features. In order to reduce the number of parameters in the network and the problem of excessive calculation, DSC is used to replace the conventional convolution in the FCN structure. In DSC operation, each input channel will carry out the convolution operation with the corresponding convolution kernel, and get an intermediate feature layer equal to the number of convolution cores. Then in the point by point convolution stage, these intermediate feature layers are added point by point, and finally the output feature map is obtained. Compared with the traditional convolution operation, this convolution operation can reduce the computation amount, which is shown as (1).

$$\frac{D_K \times D_K \times M + M \times N}{D_K \times D_K \times M \times N} = \frac{1}{N} + \frac{1}{D_K^2} \tag{1}$$

where the size of the convolution kernel is represented by $D_K$, $M$ and $N$ respectively represent the number of channels and the convolution kernel in the input feature graph. In (1), the denominator and numerator represent the computations for ordinary and DSC respectively, and the number of parameters for a DSC is at most $1/D_K^2$ of that for the ordinary convolution. In addition, the structure of Xception adds batch normalization processing and activation functions after each $3 \times 3$ depth separation of convolutional layers to speed up network training. Figure 3 shows the structure of Xception.

### B. IMPROVED ASPP MODULE

ASPP module uses three DSCs with different convolution kernel sizes to extract features of different scales, as shown in Figure 4. However, due to the sparse distribution of sampling points in the feature layer of ASPP, a lot of detail information will be lost. Moreover, only extract shallow semantic information can be extracted by ASPP module, it affects the segmentation effect of the network in the complex background.

The improved ASPP module introduced in this paper uses the hybrid cascade approach to realize the feature information sharing among the branches, so as to obtain more rich deep semantic information, and improves the ability of the network model to acquire the object in the complex background. After the first pooling operation, the multi-scale feature map is obtained by average pooling of different scales to further get the relevant information of local features and edge details. The structure of improved ASPP module is shown in Figure 5.

The improved ASPP module takes the feature map extracted by the deep CNNs as the input, and extracts the feature by using the atrous convolution with three different expansion coefficients (6,12,18). In the atrous convolution, one or more pixels are separated between the pixels inside the kernel, which can increase the Effective Receptive

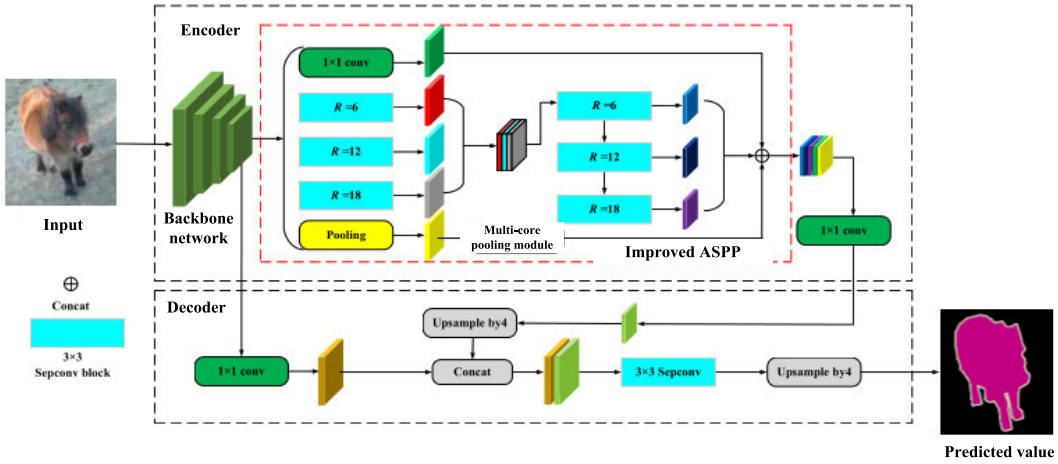



**FIGURE 2.** Image semantic segmentation model based on hybrid cascade and feature fusion.

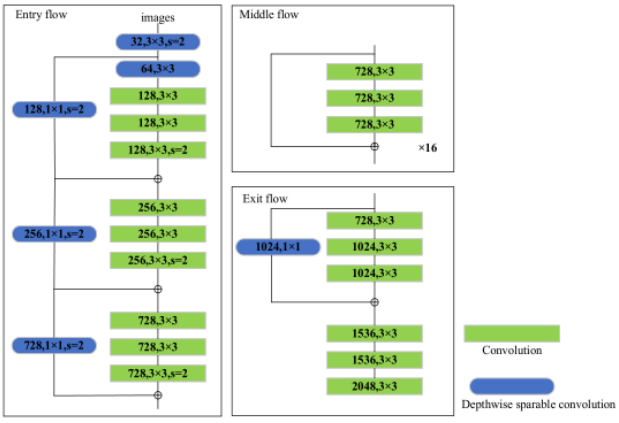



**FIGURE 3.** Network structure of Xception.

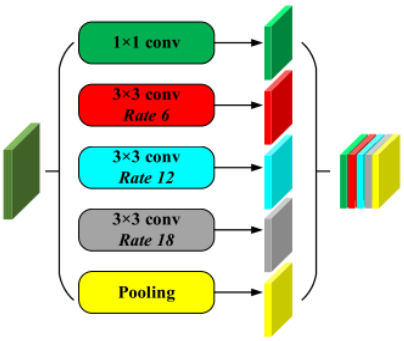



**FIGURE 4.** ASPP module.

Field (ERF) of the convolutional kernel. The calculation formula of atrous convolution is shown in (2).

$$g_{i,j}(x_\ell) = \sum_{C=0}^{C_\ell} \theta_{k,r}^{i,j} * x_\ell^C \tag{2}$$

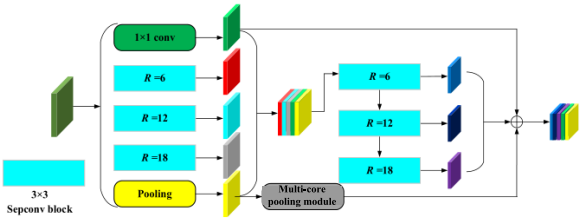



**FIGURE 5.** ASPP module.

where $g_{i,j}$ is the convolution operation of output characteristic figure, $x_\ell$ is the characteristic map belong to channel $c$ of $i$th row and $j$th column, and $\theta_{k,r}$ is the atrous convolution with convolution kernel size $k$ and expansion rate $r$.

The extraction of semantic information of different scales by atrous convolution can effectively solve the problem of information loss. Subsequently, the obtained feature maps are mixed cascaded using the cavity convolution of three expansion coefficients (6,12,18), which aims to increase the deeper features of model, expand the ERF of the network model, and strengthen the model ability to acquire object in complex background. At the same time, the improved ASPP uses multi-core pooling module to further extract features after pooling operation and strengthen the relation of each subinterval. Finally, the feature map obtained by the multi-core pooling module is fused with the feature map obtained by the hybrid cascade mode.

The improved ASPP module extracts the semantic information of depth features after the original feature map obtained by three times atrous convolutions, which can improve the recognition ability of the overall object features in the complex background. The structure diagram of different extraction methods is shown in Figure 6.

Although the parallel type can achieve multi-scale feature extraction, the pixel utilization rate is low, however, the series structure can not obtain multi-scale semantic information.

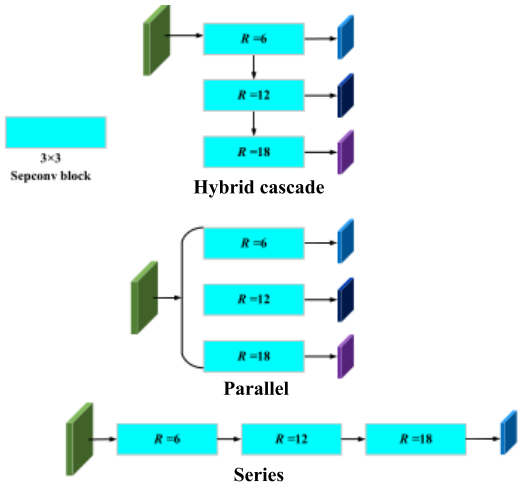



**FIGURE 6.** Cascade operations in different ways.

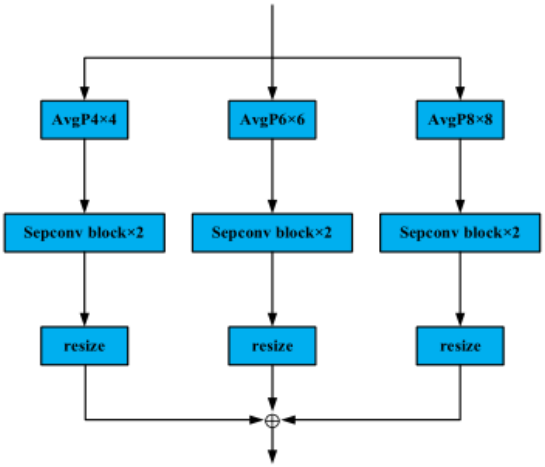



**FIGURE 7.** Multi-core pooling module.

Therefore, we adopt a hybrid cascade approach to extract multi-scale features from images and expand the receptive field of the network. The extracted output result of each layer with a smaller cavity rate will serve as the input for the next layer with a larger cavity rate, thereby improving the utilization rate of image feature information. The sensitivity calculation formulas of parallel and hybrid cascades are shown as (4) and (5) respectively.

$$P = K + (K-1) \times (d-1) \tag{3}$$

$$P_{\text{parallel}} = \max[P_1, P_2, P_3] \tag{4}$$

$$P_{\text{hybrid}} = P_1 + P_{12} + P_{123} \tag{5}$$

where $K$ is the size of the convolution kernel, $P$ is the size of $K$ after the atrous convolution, and $d$ is the expansion coefficient. Meanwhile, $P_1, P_2$, and $P_3$ represent the receptive field sizes at $R = 6$, $R = 12$, and $R = 18$, respectively. $P_{12}$ represents the size of the receptive field after $R = 6$ and $R = 12$. $P_{123}$ represents the size of the receptive field after $R = 6$, $R = 12$, and $R = 18$. Obviously, it can be obtained that $P_{\text{hybrid}} > P_{\text{parallel}}$.

In the improved ASPP, the multi-core pooling module is introduced to process the feature image. The input feature map is average pooled according to different scales, and then the pooled results are spliced together to obtain a multi-scale feature representation. The specific operation mode is shown in Figure 7. The function of the multi-core pooling module is to enhance the sensing ability of NN for targets of different scales and realize the gradual refinement of target features. More detailed local information can be obtained by the feature block averaging pooling, which makes the model understand the image more comprehensively.

### C. DECODER AND LOSS FUNCTION

The task of decoder is to recover and supply the information by the resolution of the feature map. In order to improve the feature information of the image, it is necessary to combine the high and low layer semantic feature maps to make up the information that may be lost by the encoder. Bilinear interpolation is a commonly used upsampling method. A continuous function on a discrete grid is firstly decomposed into two one-dimensional continuous functions in the horizontal and vertical directions. Then, for a given discrete coordinate point, the four points adjacent to it among the four nearest discrete grid points around it is found, and the function values of these points should be interpolated. Two values of the adjacent points are obtained by horizontal linear interpolation, and then the final result is obtained by linear interpolation of these two values in the vertical direction, as shown in Figure 8.

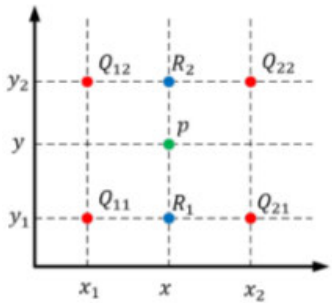



**FIGURE 8.** Schematic diagram of bilinear interpolation.

Suppose that $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$, $Q_{22} = (x_2, y_2)$, the linear interpolations of the $x$ direction are shown as (6) and (7).

$$f(R_1) = \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}) \tag{6}$$

$$f(R_2) = \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \tag{7}$$

The linear interpolation of the y direction is shown as (8).

$$f(P) = \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2) \tag{8}$$

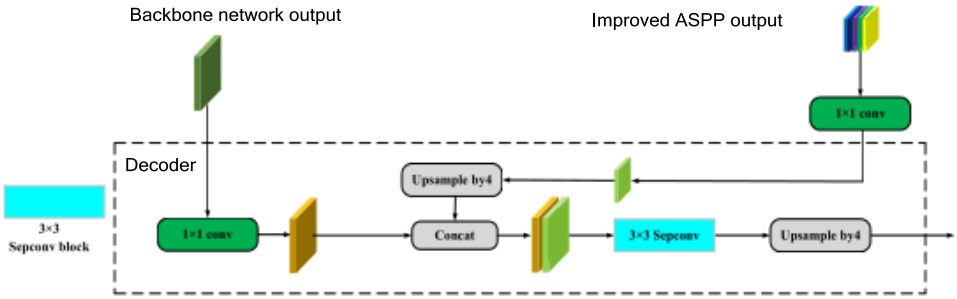



**FIGURE 9.** Decoder module.

The decoder structure is shown in Figure 9. Convolution is used to reduce the dimensions of the feature map. The feature map obtained from the backbone network is combined with output based on the improved ASPP. Softmax activation function is used to predict the final result.

The cross-entropy loss function can naturally handle class imbalance problems during training because of assigning higher loss to misclassified samples. The function checks all pixels one by one and compares the predictions and label vectors.

$$P_L = -\sum_C y_t \log(y_p) \tag{9}$$

The loss of entire image is the average of the loss per pixel, where $y_p$ is the predicted result and $y_t$ is the correct result. When the predicted probability distribution of the network model designed in this paper is exactly the same as that of the actual label, the cross entropy loss is 0. Otherwise, the greater difference between the probability distribution predicted by the model and the probability distribution of the actual label, the more the cross-entropy loss. Therefore, during the training process, the model optimizes the prediction effect by minimizing the cross-entropy loss of all pixel positions.

## IV. SEMANTIC SEGMENTATION MODEL BASED ON CROSS-STAGE AND DETAILE ATTENTION MECHANISM

In order to solve the problems of small-scale object and unclear boundary segmentation in semantic segmentation, we introduce the Cross-Stage Feature Fusion and Detailed Attention Mechanism (CFF-DAM) module is introduced on the basis of the hybrid cascade and feature fusion image semantic segmentation model. In CNNs, shallow feature maps have higher resolution and contain more small-scale object information and detail.

We divide the network model into three stages which fuse the image semantic information of different stages, and use two skip connections to splice the feature maps. It can not only solve the problem of gradient explosion effectively, but also strengthen the correlation between low and high level features.

The feature map output in the encoder stage contains detailed feature information such as target positions and edges, although there are also a lot of meaningless feature information which greatly affects the effect of semantic segmentation. The attention mechanism module can be adjusted the weight of each pixel according to different context information, which helps the network to focus more effectively on meaningful features.

Based on convolutional attention mechanism, the fully connected layer is replaced by the self-attention mechanism which can strengthen the feature dependence of image channel dimension. The alternative approach can extract useful feature more efficiently and enhance the representation of the model. Then, a $7 \times 7$ convolution of the spatial attention mechanism is fused with the feature graph obtained by one-dimensional convolution to increase the receptive field of the network model, thus improving the segmentation ability of small-scale object. The overall structure of model is shown as Figure 10.

### A. DESIGN OF CROSS-STAGE FEATURE FUSION

The ASPP improved by encoder is divided into three stages. With two jump connections, the feature map is spliced into channels to compensate for the information lost due to insignificance, which realizes CFF.

As shown in Figure 11, Stage1, Stage2 and Stage3 represent shallow semantic information feature extraction stage, middle semantic information feature extraction stage, and deep semantic information feature extraction stage. CFF enables the network model to fully acquire the semantic information of different stages, and enhances the ability of the network model to obtain the feature information of small-scale object. The calculation formula of cross-stage is shown as follow.

$$H(x) = Concat(H(x_1); H(x_2); H(x_3)) \tag{10}$$

where $H(x)$ is the output feature map, $H(x_1)$, $H(x_2)$ and $H(x_3)$ are the output of feature maps of Stage1, Stage2 and Stage3, and $Concat(\ )$ is the channel splicing. It can retain the semantic information of the first two stages through cross-stage fusion.

### B. IMPROVED CONVOLUTION ATTENTION MECHANISM

Each channel of convolutional attention mechanism is a response to a certain feature, and there is some relationships

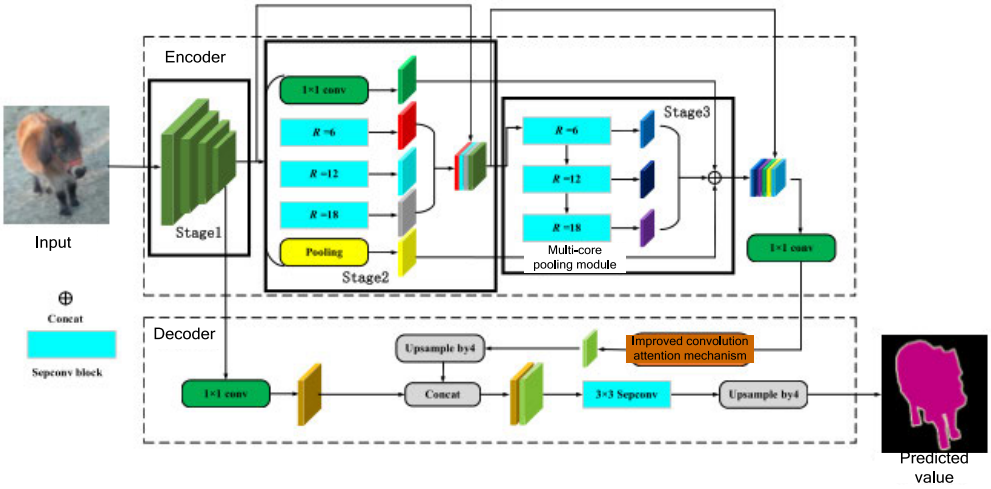



**FIGURE 10.** Image semantic segmentation model based on cross-stages and detailed attention mechanism.

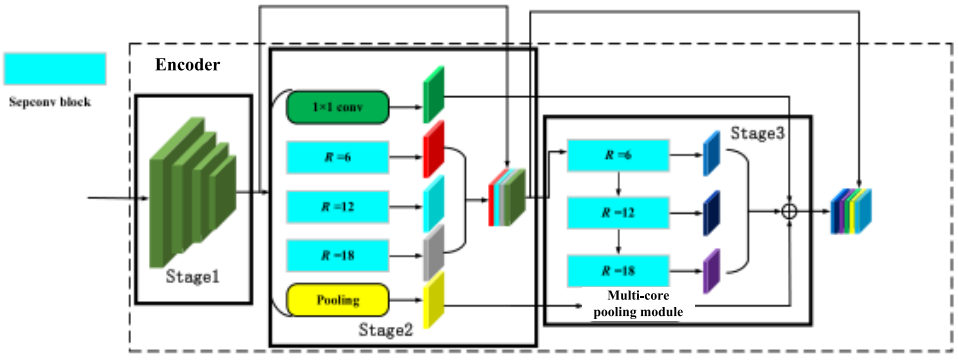



**FIGURE 11.** Cross-stage feature fusion.

between these features. In the channel attention mechanism, the fully connected layer is used to obtain the channel weight. However, it is difficult to take into account the correlation between features, and the non-local method of self-attention mechanism can directly capture a larger range of dependencies from the relative positions between semantic pixels.

In this paper, the middle fully connected layer of channel attention module is replaced with a self-attention mechanism to improve the representation ability of the model, so that the model can learn richer and more representative features from the input data. The function of spatial attention mechanism in convolutional attention mechanism is to focus on the location of key information, and the size of receptive field directly affects the segmentation effect of small-scale object. The improved convolutional attention mechanism module is shown in Figure 12.

Suppose that the input feature map be $F \in R^{C*H*W}$, through the improved channel attention, the feature graph obtains the weights and multiplies them by weighted input tensors. Secondly, the feature graph $F_1$ obtained in the previous stage is input into the improved spatial attention module, and the spatial attention weights are obtained and weighted to obtain the final feature graph $F_2$.

$$F_1 = M_C(F) \otimes F \tag{11}$$

$$F_2 = M_S(F_1) \otimes F_1 \tag{12}$$

The improved channel attention module first converts the input feature graph into the tensor form of $C \times 1 \times 1$, and then goes through the operation of the mean-pooling and max-pooling to extract the important information of different feature maps. The pooled feature graph is mapped to three feature vectors Q(Query), K(Key) and V(Value) through three different $1 \times 1$ convolutions. A weight matrix is obtained by matrix multiplication of the Query and Key branches. Then, the relationship feature graphs between each channel are obtained by Softmax function. By multiplying the feature map with Value, invalid features are suppressed and important features are enhanced, and the global context information in semantic segmentation task is get. The formula for calculating the self-attention score matrix **A** is shown

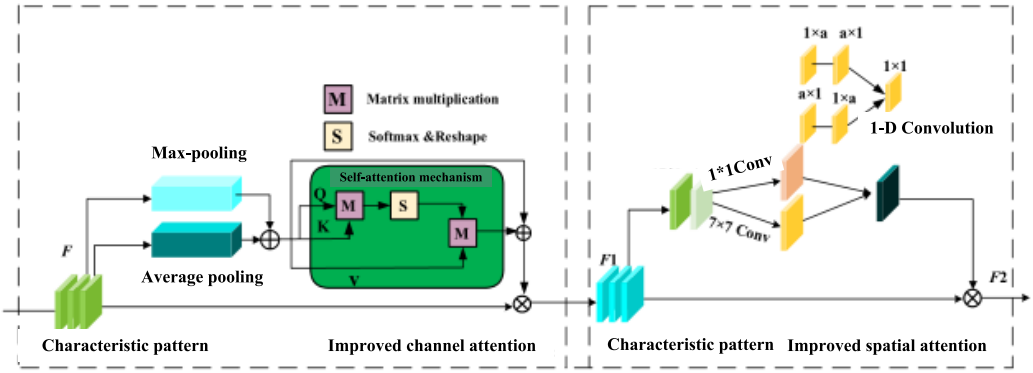



**FIGURE 12.** Improved convolutional attention mechanism module.

as follow.

$$A = soft\ \max(\frac{QK^{\mathrm{T}}}{\sqrt{d}}) \tag{13}$$

$$M_C(F) = \sigma(A \otimes V) + M_{in} \tag{14}$$

where $M_C(F)$ be the feature map of output channel attention mechanism module, $M_{in}$ be the feature maps with mean-pooling and max-pooling, and d be the dimension of the feature map.

The spatial attention module firstly performs global max-pooling and mean-pooling operations respectively in channel dimensions, and then fuses feature graphs from different pooling modes through convolution layers with convolution kerns of $7 \times 7$. The weight $M_C$ of spatial attention is acquired by calculating the Sigmoid nonlinear activation function. However, the approach is still insufficient for dividing the receptor field of small-scale object. It is divided into two sub-branches. While retaining the $7 \times 7$ convolution, the convolution kernel of $1 \times a + a \times 1$ and $a \times 1 + 1 \times a$ are respectively used in the second branch for convolution operation, and they are fused to form a dense connection in the $a \times a$ region. This operation is more suitable for receiving global information, enlarging the receptive field of the model, and improving the ability of network to acquire small-scale object. The expression of the calculation process is shown in (15).

$$M_S(F) = \sigma(f^{7\times7}(F^S_{\max}; F^S_{avg}) \otimes f^1(F^S_{\max}; F^S_{avg})) \tag{15}$$

where $f^{7\times7}$ be a convolution operation with a size of $7 \times 7$ convolution kernel, be $f^1$ the Sigmoid nonlinear activation function. The output feature graph $F_2$ can be expressed by the input feature graph $F$ through the channel space attention network.

$$F_2 = M_C(F) \otimes M_S(F) \tag{16}$$

where $M_S(F)$ and $M_C(F)$ are the improved spatial attention mechanism and improved channel attention mechanism respectively.

## V. EXPERIMENTAL RESULTS AND ANALYSIS

### A. DATASET OF PASCAL VOC2012 AND SUIM

PASCALVOC2012 dataset focuses on the object in the actual scene, and the data provided is available for supervised learning. The pictures and corresponding labels are provided to complete image classification and detection, image segmentation, human action classification and human body parts detection.

The dataset can be divided into 4 categories and 20 subcategories. Among the 2913 images, 1464 ones are used as the training set of the network. Each subcategory of PASCALVOC2012 dataset uses a different color value, for example, black represents the background, red refers to the plane, green represents the train, purple represents the dog, and white refers to the unknown area in the label. The image visualization is shown in Figure 13.

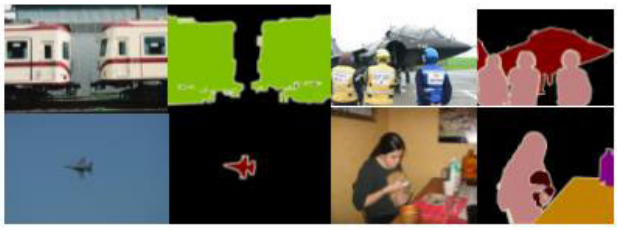


**FIGURE 13.** Visualization of PASCAL VOC2012.

SUIM is a dataset for semantic segmentation of underwater image. In underwater environment, semantic segmentation faces special challenges, such as light attenuation, scattering, noise, and marine organisms, which make it more difficult to correctly extract semantic information from underwater image. SUIM dataset contains 1525 training and verification pictures and 110 test set pictures. The visualization of SUIM dataset is shown in Figure 14.

Since the two datasets above provide less training data, data enhancement techniques can be used to extend the sets. Data enhancement can not only enhance the generalization and robustness of network, but also reduce the risk of overfitting.

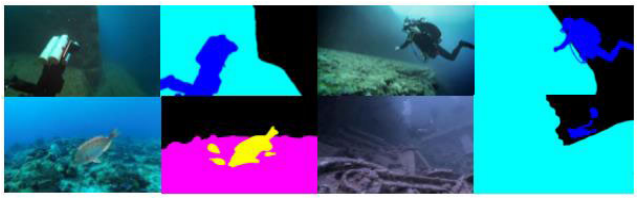


**FIGURE 14.** Visualization of SUIM.

### B. EXPERIMENTS ENVIRONMENT

The detailed configurations of experiments are shown in Table 1. The image resolution is 512 × 512, and the setting of hyper-parameters is as follows: batch_size = 4, epoch = 300, and the initial value of the learning rate is 0.001.

With the increasing of the number of training epoch, the learning rate will be reduced to 1/2 of the original one if the loss value of the validation set does not change after 3 consecutive epoch. In addition, the network model in this paper uses cross-entropy as a loss function and sets the parameter to finish training early. When the validation set has no change in loss value after 20 epoch, the training will stop, and this operation can effectively reduce the repetition of training.

**TABLE 1.** Experiments environment configuration.

| Parameter | Configuration |
|---|---|
| CPU | Intel Core i7-10700KF |
| Memory | 16G |
| GPU | NVIDIA GeForce RTX 2080 Ti |
| Operation System | Windows 10 |
| Tensorflow | 2.4.0 |
| CUDA | 11.0 |
| CuDNN | 8.0 |
| Python | 3.7.8 |

### C. EVALUATION INDEX

In this paper, the Mean Intersection over Union (MIoU)is used to evaluate the experimental results. MIoU operates sum and average of each type of IoU [30], [31].

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \tag{17}$$

$$\mathrm{MIoU} = \frac{\sum_{k=1}^{classes} \mathrm{IoU}_i}{classes}) \tag{18}$$

IoU represents the intersection between the results values predicted by the network and the true value. *TP* is the true positive, *FP* represents the false positive, and *FN* represents the false negative. The sets relation is shown in Figure15, and *label* is the set of true labels, and *pred* is the set of predicted result.

$$\mathrm{MIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{19}$$

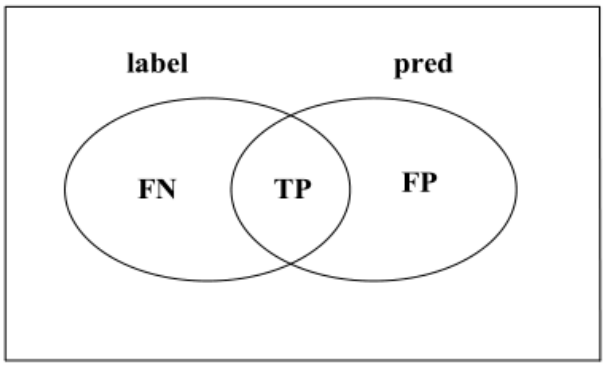



**FIGURE 15.** Representation of sets relation.

### D. EXPERIMENTAL RESULTS ANALYSIS OF IMAGE SEMANTIC SEGMENTATION MODEL BASED ON HYBRID CASCADE AND FEATURE FUSION

From traditional ML approaches, we choose Term Frequency–Inverse Document Frequency (TF-IDF), Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) as baseline methods, and three other models of TextCNN, TextRCNN and Transformer from Deep Neural Networks.

In this paper, VGG16, ResNet50, MobileNetV2 and Xception are selected to carry out experiments as backbone network of encoder, and Xception with the highest accuracy is selected as the backbone network of the design model in this paper. The comparison results are shown in Table 2, which lists the values of MIoU and frames per second respectively. It can be seen that the MIoU of Xception reaches 82.45%.

**TABLE 2.** Comparison of different backbone networks.

| Backbone Network | MIoU(%) | fps |
|---|---|---|
| VGG16 | 75.36 | 5.52 |
| ResNet50 | 79.93 | 22.73 |
| MobileNetV2 | 78.53 | 37.39 |
| Xception | **82.45** | **23.55** |

In order to verify the impact of the improved ASPP module on the model performance, the ablation and comparison experiments of the multi-core pooling module and different parallel cascade modes are carried out. The experimental results are shown in Table 3. Compared with the parallel and series approaches, the MIoU of hybrid cascade one is 0.22% and 0.73% higher respectively. The network model

**TABLE 3.** Ablation experiments of image semantic segmentation model based on hybrid cascade and feature fusion.

| ASPP | Series | Parallel | Hybrid Cascade | Multi-core Pooling | MIoU(%) |
|---|---|---|---|---|---|
| √ | | | | | 82.45 |
| √ | | | | | 83.54 |
| √ | √ | | | | 84.39 |
| √ | | √ | | | 84.90 |
| √ | | | √ | | 85.12 |
| √ | | | √ | √ | **85.48** |

**TABLE 4.** Evaluation results based on F1-score index.

| Class | TF-IDF | NB | SVM | RF | TextCNN | TextRCNN | Transformer | Ours1 |
|---|---|---|---|---|---|---|---|---|
| Aerop | 90.0 | 91.8 | 76.8 | 78.4 | 87.0 | 73.6 | 93.7 | **94.2** |
| Bicycle | 40.8 | **71.9** | 34.2 | 33.1 | 44.9 | 37.6 | 68.6 | 68.8 |
| Bird | 84.2 | **94.7** | 68.9 | 78.2 | 87.5 | 62.0 | 88.5 | 91.3 |
| Boat | 67.3 | 71.2 | 49.4 | 55.6 | 72.9 | 46.8 | **82.3** | 82.1 |
| Bootle | 70.7 | 75.8 | 60.3 | 65.3 | 75.3 | 58.6 | 77.6 | **83.2** |
| Bus | 90.9 | **95.2** | 75.3 | 81.3 | 91.9 | 79.1 | 94.7 | 95.0 |
| Car | 84.8 | 89.9 | 74.7 | 75.5 | 85.7 | 70.1 | 89.7 | **90.0** |
| Cat | 87.4 | **95.9** | 77.6 | 78.6 | 90.5 | 65.4 | 93.8 | 94.7 |
| Chair | 34.8 | 39.3 | 21.4 | 25.3 | 52.2 | 23.6 | 61.5 | **64.7** |
| Cow | 83.0 | 90.7 | 62.5 | 69.2 | 87.7 | 60.4 | 92.7 | **93.9** |
| Table | 58.7 | **71.7** | 46.8 | 52.7 | 69.6 | 45.6 | 70.6 | 69.1 |
| Dog | 82.3 | 90.5 | 71.8 | 75.2 | 87.8 | 61.8 | 91.6 | **93.5** |
| Horse | 87.1 | **94.5** | 63.9 | 69.0 | 85.7 | 63.5 | 83.5 | 93.4 |
| Motorbike | 86.9 | **88.8** | 76.5 | 79.1 | 77.7 | 75.3 | 86.8 | 87.0 |
| Person | 82.4 | 89.6 | 73.9 | 77.6 | 83.4 | 74.9 | 89.6 | **90.5** |
| Plant | 64.5 | **72.8** | 45.2 | 54.7 | 52.5 | 42.6 | 52.8 | 64.2 |
| Sheep | 84.6 | 89.6 | 72.4 | 78.3 | 85.5 | 63.7 | 90.5 | **92.2** |
| Sofa | 54.9 | 64.0 | 37.4 | 45.1 | 75.8 | 42.5 | 76.2 | **80.7** |
| Train | 77.5 | 85.1 | 70.9 | 73.3 | 87.7 | 67.8 | 91.5 | **92.1** |
| Tv/monit | 64.1 | 76.3 | 55.1 | 56.2 | 72.6 | 52.7 | **82.1** | 81.4 |
| MIoU(%) | 74.8 | 82.6 | 62.2 | 66.4 | 77.7 | 59.9 | 83.5 | **85.48** |

with multi-core pooling module has improved the evaluation index by 0.36% compared with the model without the module. The experiments prove that the improved ASPP module designed in this paper is 1.94% higher on MIoU.

In order to verify the validity of the hybrid cascade and feature fusion model (Ours1) proposed in this paper, IoU and MIoU are compared with seven classical semantic segmentation network models such as DeconNet, PspNet, FCN and DeepLab. The comparison results are listed in Table 4.

As shown from Table 4, the IoUs of the model proposed in this paper are superior to other models in most categories such as Aerop, Bootle, Chair, Car, Cow, Dog, Person, Sheep, Sofa and Train. Compared with DeconNet, PspNet, FCN, DeepLab, MsefNet, SegNet and DeepLabv3+, the overall accuracy of the model we designed is 10.68%, 2.88%, 23.28%, 19.08%, 7.78%, 25.58% and 1.98% higher respectively. The MIoU of the network model proposed in this paper reaches 85.48%, which is higher than the seven classical semantic segmentation networks, and it proves that our model can effectively improve the overall accuracy.

The visualization comparison of the image semantic segmentation based on hybrid cascade and feature fusion(Ours1) with different network models such as FCN, PspNet, DeconNet is shown in Figure 16. The first images in each row are the original input images, the second ones are the labels, and the third to last images are the segmentation result maps based on different approaches respectively.

Compared with other models, the improved image segmentation method (Ours1) has higher overall accuracy, better visualization effect of semantic segmentation, which reduces the incidence of false segmentation in images.

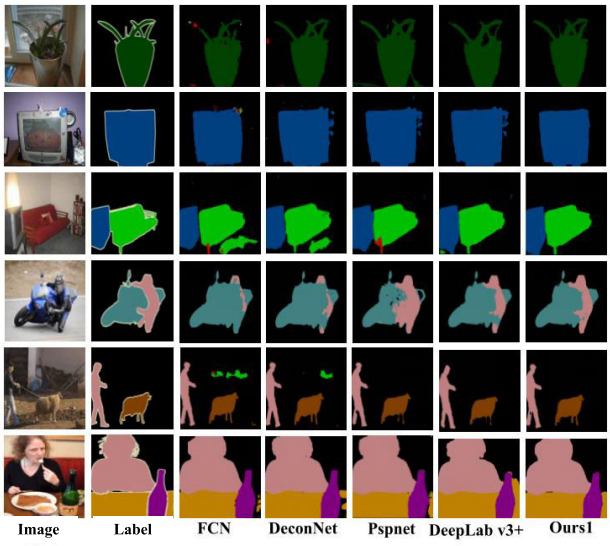



**FIGURE 16.** Visualization comparisons of different network models.

### *E. EXPERIMENTAL RESULTS ANALYSIS OF IMAGE SEMANTIC SEGMENTATION MODEL BASED ON CROSS-STAGE AND DETAILDE ATTENTION MECHANISMS*

In order to verify the effectiveness of the cross-stage feature fusion, we conduct ablation experiments on two jump connections divided into three stages, where x represents the cross-stage fusion from Stage1 to Stage2, and y represents the cross-stage fusion from Stage2 to Stage3. The final experimental data are shown in Table 5.

**TABLE 5.** Comparison of cross-stage fusion.

| Model | x | y | MIoU(%) |
|---|---|---|---|
| Ours1 | √ | | 85.59 |
| Ours1 | | √ | 85.65 |
| Ours2 | √ | √ | **85.72** |

From Table 5, the MIoU of the image semantic segmentation model based on hybrid cascade and feature fusion reaches 85.72% through cross-stage fusion. The experiment proves that cross-stage feature fusion can effectively fuse

semantic information of different stages and improve the accuracy of network.

In order to verify the influence of the improved convolutional attention mechanism module, the ablation experiments are conducted on the basis of the image semantic segmentation approach based on hybrid cascade and feature fusion after cross-stage fusion(Ours2), and other attention mechanisms were selected for comparison experiments. For the improved spatial attention mechanism (Ours), the effect of the one-dimensional convolution value $a = 9$ is the best one. The final experimental data are shown in Table 6.

**TABLE 6. Comparisons between improved attention mechanisms and other ones.**

| Model | DataSet | MIoU(%) |
|---|---|---|
| Ours2 | VOC2012 | 85.72 |
| Ours2+SE | VOC2012 | 85.88 |
| Ours2+ Coordinate attention | VOC2012 | 86.06 |
| Ours2+CBAM | VOC2012 | 86.16 |
| Ours2+Ours(a=5) | VOC2012 | 86.15 |
| Ours2+Ours(a=7) | VOC2012 | 86.53 |
| Ours2+Ours(a=9) | VOC2012 | **86.68** |
| Ours2+Ours(a=11) | VOC2012 | 86.58 |
| Ours2 | SUIM | 55.10 |
| Ours2+SE | SUIM | 57.38 |
| Ours2+ Coordinate attention | SUIM | 59.85 |
| Ours2+CBAM | SUIM | 60.12 |
| Ours2+Ours(a=9) | SUIM | **61.55** |

It can be seen that the attention mechanism can effectively improve the overall accuracy of the network model. The accuracy of the attention mechanism module designed in this paper added to the SUIM dataset reaches 61.55%. Compared with SE, coordinate attention and CBAM, the improved attention mechanism model are improved by 4.17%, 1.7% and 1.43% respectively on SUIM dataset. In the PASCALVOC2012 dataset, the accuracy reached 86.68%.

The visual comparison results of the final network (Ours3) with other network models are shown in Figure 17. The visual comparisons of the first three rows are the PASCALVOC2012 dataset and the other rows are the SUIM dataset. It can be seen that the segmentation results of the proposed model for the human head part and the horn part in the images are more precise than other models on the PASCALVOC2012 dataset.

SUIM dataset is a dataset image taken underwater with complex background. Compared with other models, our model designed has improved the effect of false segmentation. The reason is that the network model in this paper uses long-distance dependency information and focuses on important features. It improves the characterization ability of networks and enhances the ability to distinguish categories.

In the segmentation results of small-scale object, the performance of the network models are shown in Figure 18. Among them, the first column represents the original images of the small-scale object, the second column are the labels, and the third one are the segmentation results of Ours3 model. It is clear that Ours3 can effectively segment small-scale object.

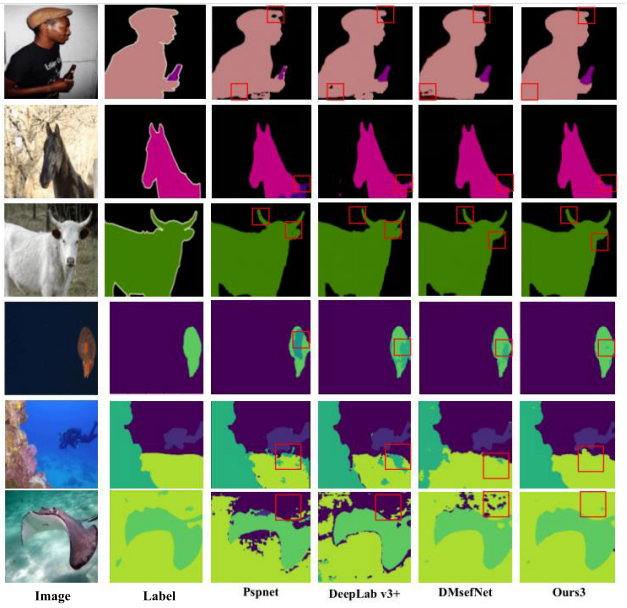



**FIGURE 17. Visualization comparisons of different models.**

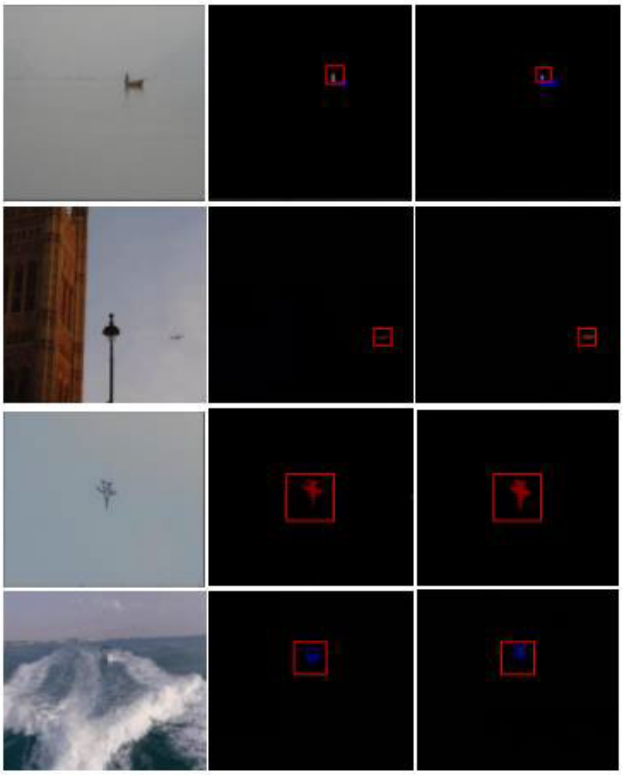


**FIGURE 18. Visualization of small scale object segmentation.**

Figure 19 is the visualization segmentation effects of Ours3 on the PASCALVOC2012 dataset and SUIM underwater dataset. The first and fourth columns are the original images of PASCALVOC2012 and SUIM dataset, the second and fifth columns are the labels, and the third and sixth columns are the operation result graphs.

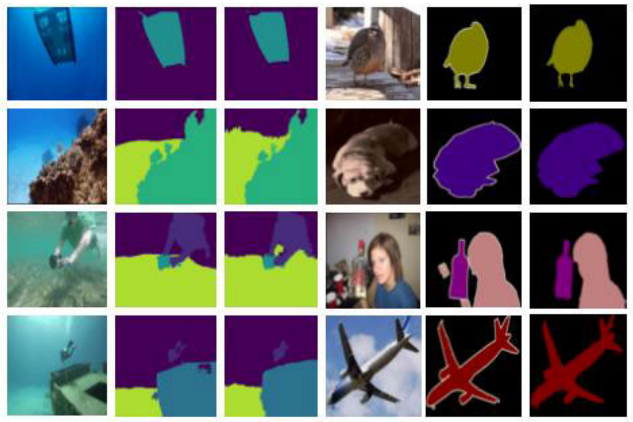


**FIGURE 19.** Segmentation visualization of Ours3.

**TABLE 7.** Precision comparisons of semantic segmentation models in PASCAL VOC2012 dataset.

| Approach | MIoU(%) |
|---|---|
| DeconNet | 74.82 |
| DenseASPP | 74.61 |
| GCRF | 73.27 |
| NISNet | 78.13 |
| DFN | 80.60 |
| DANet | 80.94 |
| APCNet | 80.71 |
| ResNet-GCN | 81.00 |
| PspNet | 82.60 |
| RefineNet | 83.43 |
| SDN | 83.50 |
| DeepLabv3+ | 83.54 |
| SOTA | 85.23 |
| ExFuse | 85.44 |
| **Ours3** | **86.68** |

**TABLE 8.** Precision comparisons of semantic segmentation models in SUIM dataset.

| Approach | MIoU(%) |
|---|---|
| DeepLabv3 | 41.72 |
| PspNet | 44.41 |
| DeepLabv3+ | 52.43 |
| GCANet | 57.34 |
| DmsefNet | 58.23 |
| PANet | 58.42 |
| SOTA | 58.97 |
| **Ours3** | **61.55** |

In order to verify the effectiveness of Ours3, the MIoU evaluation index is compared with other approaches, and the experimental results are shown in Table 7. It can be seen that the accuracy of Ours3 is the highest among all models. Compared with semantic segmentation methods such as PspNet, DeconNet, DeepLabv3+ and GCRF, the results using attention mechanism and feature fusion are 4.08%, 11.86%, 3.14% and 13.41% higher, respectively. It is 1.45% higher than the current SOTA method [15].

The training and test results under SUIM dataset are shown in Table 8. It can be seen that the accuracy of Ours3 is superior to other networks. Compared with DeepLabv3, PspNet, DeepLabv3+ and GCANet, the results have respectively increased by 19.83%, 17.14%, 9.12% and 4.21%. Moreover, experimental results show that the proposed method has reached the current level of SOTA.

## VI. CONCLUSION

In this paper, the image semantic segmentation model based on hybrid concatenation and feature fusion is proposed. The overall architecture of network adopts an encoder-decoder structure, where the ASPP module is improved in the encoder. The hybrid concatenation and multi-core pooling approach is used to extract deeper semantic information, better integrate global multi-scale context information, and improve the overall segmentation accuracy of object. Cross-Stage Feature Fusion is designed to divide the backbone network of the encoder and the improved ASPP into three stages. By using the jump connection, the feature map of each stage is fused. While it enables the network model to fully utilize the different semantic information of the shallow and deep layers, and enhances the dependency between features. Thereby the segmentation accuracy of small scale object is improving.

The image semantic segmentation model based on cross stage and detailed attention mechanisms is proposed. Attention mechanism is introduced into the hybrid cascade and feature fusion image semantic segmentation network, and the improvements are made on the CBAM. Adding self attention to channel attention enhances the connection between feature maps, and using one-dimensional convolution in the spatial attention mechanism to increase the spatial receptive field, which enriches the information of feature map and enhances the representation ability of network. MIoU of the model we presented has reached 86.68% on PASCAL VOC2012 and 61.55% on SUIM dataset. Experiment results have shown that the overall accuracy of our model is higher than existing methods, which proves the effectiveness of ours.

There are still some follow-up works for improvement in this research. The image semantic segmentation approach designed in this paper adopted decoding structure, and introduced a variety of feature fusion techniques to obtain features at different stages. However, there has been no taking into account the importance difference of the features at each stage. In the future, the loss functions of multi-stages can be added to training for learning. The network model we proposed had relatively strong segmentation ability for small-scale target and boundary, and the model will be lightweight design for pruning and KD to reduce the redundant information in the network.

## REFERENCES

[1] Q. Wang, W. Li, and Z. Jin, "Review of text classification in deep learning," *Open Access Library J.*, vol. 8, no. 3, pp. 1–8, 2021.

[2] G. Diraco, A. Leone, A. Caroppo, and P. Siciliano, "Deep learning and machine learning techniques for change detection in behavior monitoring," in *Proc. AI* AAL@ AI* IA*, vol. 2019, 2019, pp. 38–50.

[3] C. P. Bara, M. Papakostas, and R. Mihalcea, "A deep learning approach towards multimodal stress detection," in *Proc. AffCon@ AAAI*, 2020, pp. 67–81.

[4] A. F. M. S. Saif and Z. R. Mahayuddin, "Vision based 3D object detection using deep learning: Methods with challenges and applications towards future directions," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 11, pp. 203–214, 2022.

[5] A. Saif and Z. R. Mahayuddin, "Crowd density estimation from autonomous drones using deep learning: Challenges and applications," *J. Eng. Sci. Res.*, vol. 5, no. 6, pp. 1–6, 2021.

[6] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, pp. 447–453, May 2020.

[7] S. Mirjalili, H. Faris, and I. Aljarah, "Introduction to evolutionary machine learning techniques," in *Evolutionary Machine Learning Techniques*. Singapore: Springer, 2020, pp. 1–7, doi: 10.1007/978-981-32-9990-0.

[8] J. Atwan, M. Wedyan, Q. Bsoul, A. Hamadeen, R. Alturki, and M. Ikram, "The effect of using light stemming for Arabic text classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, pp. 768–773, 2021.

[9] H. Amazal and M. Kissi, "A new big data feature selection approach for text classification," *Sci. Program.*, vol. 2, pp. 1–10, Apr. 2021.

[10] X. Luo, "Efficient English text classification using selected machine learning techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021.

[11] D. Liciotti, M. Bernardini, L. Romeo, and E. Frontoni, "A sequential deep learning application for recognising human activities in smart homes," *Neurocomputing*, vol. 396, pp. 501–513, Jul. 2020.

[12] A. Saif and Z. R. Mahayuddin, "Moving object detection using semantic convolutional features," *J. Inf. Syst. Technol. Manag.*, vol. 7, no. 29, pp. 24–41, 2022.

[13] A. F. M. S. Saif and Z. R. Mahayuddin, "An efficient method for hand gesture recognition using robust features vector," *J. Inf. Syst. Technol. Manage.*, vol. 6, no. 22, pp. 25–35, Sep. 2021.

[14] T. Jain, V. K. Verma, A. K. Sharma, B. Saini, N. Purohit, B. H. Mahdin, M. Ahmad, R. Darman, S.-C. Haw, S. M. Shaharudin, and M. S. Arshad, "Sentiment analysis on COVID-19 vaccine tweets using machine learning and deep learning algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 32–41, 2023.

[15] A. F. M. S. Saif, E. D. Wollega, and S. A. Kalevela, "Spatio-temporal features based human action recognition using convolutional long short-term deep neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 1–15, 2023.

[16] K. Wang, Z. Meng, and Z. Wu, "Deep learning-based ground target detection and tracking for aerial photography from UAVs," *Appl. Sci.*, vol. 11, no. 18, p. 8434, Sep. 2021.

[17] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2015, pp. 234–241.

[20] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[24] C. Liu, L. C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 82–92.

[25] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5229–5238.

[26] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-FCN for image semantic segmentation," in *Proc. Int. Symp. Neural Netw.*, 2020, pp. 97–105.

[27] X. Ding, C. Shen, Z. Che, T. Zeng, and Y. Peng, "SCARF: A semantic constrained attention refinement network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3002–3011.

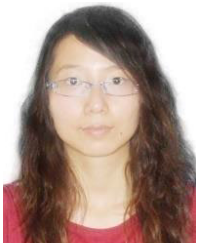
[28] Q. Song, J. Li, C. Li, H. Guo, and R. Huang, "Fully attentional network for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2280–2288.

[29] Z. Wang, J. Wang, K. Yang, L. Wang, F. Su, and X. Chen, "Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+," *Comput. Geosci.*, vol. 158, Jan. 2022, Art. no. 104969.

[30] F. Chen, H. Liu, Z. Zeng, X. Zhou, and X. Tan, "BES-Net: Boundary enhancing semantic context network for high-resolution image semantic segmentation," *Remote Sens.*, vol. 14, no. 7, pp. 1638–1648, Mar. 2022.

[31] Y. Alghamdi, A. Munir, and H. M. La, "Architecture, classification, and applications of contemporary unmanned aerial vehicles," *IEEE Consum. Electron. Mag.*, vol. 10, no. 6, pp. 9–20, Nov. 2021.

**ZUOQIANG DU** received the M.D. degree from the College of Computer Science and Technology, Harbin Engineering University, China, in 2005. He is currently a Professor with the School of Computer and Information Engineering, Harbin University of Commerce. He has published with the identity of first author and corresponding author more than ten articles indexed by SCI/EI in journals. His main research interests include the application of quantum computing and information processing, artificial intelligence, and machine learning.



**YUAN LIANG** received the M.D. degree from the College of Automation, Harbin Engineering University, China, in 2011. She is currently a Researcher with Jinan Inspur Data Technology Company Ltd., and responsible for the research and development of multiple host security products. She has published with the identity of first author and corresponding author more than ten articles indexed by SCI/EI in journals. Her main research interests include EDR technology, eBPF technology, artificial intelligence, and machine learning.