



Coastal Aquaculture Area Extraction Based on Self-Attention Mechanism and Auxiliary Loss

Bo Ai , Heng Xiao, Hanwen Xu, Feng Yuan , and Mengyun Ling

Abstract—With the development of deep learning in satellite remote sensing image segmentation, convolutional neural networks have achieved better results than traditional methods. In some full convolutional networks, the number of network layers usually increases to obtain deep features, but the gradient disappearance problem occurs when the number of network layers deepens. Many scholars have obtained multiscale features by using different convolutional calculations. We want to obtain multiscale features in the network structure while obtaining contextual information by other means. This article employs the self-attention mechanism and auxiliary loss network (SAMALNet) structure to solve the above problems. We adopt the self-attention strategy in the atrous spatial pyramid pooling module to extract multiscale features while considering the contextual information. We add auxiliary loss to overcome the gradient disappearance problem. The experimental results of extracting aquaculture areas in the Jiaozhou Bay area of Qingdao from high-resolution GF-2 satellite images show that, in general, SAMALNet achieves better experimental results compared with UPS-Net, SegNet, DeepLabv3, UNet, DeepLabv3+, and PSPNet network structures, including recall 96.34%, precision 95.91%, F1 score 96.12%, and MIoU 92.60%. SAMALNet achieved better results extracting aquaculture area boundaries than the other network structures listed above. The high accuracy of the aquaculture area can provide data support for the rational planning and environmental protection of the coastal aquaculture area and promote more rational usage of the coastal aquaculture area.

Index Terms—Aquaculture area extraction, auxiliary loss, self-attention mechanism, deep learning.

I. INTRODUCTION

AS A rapidly expanding industry, aquaculture produces high-protein fish, seafood, and other aquatic products that feed hundreds of millions of people [1], [2]. Aquaculture production is increasing rapidly around the world. According

Manuscript received 8 October 2022; revised 24 November 2022 and 4 December 2022; accepted 11 December 2022. Date of publication 21 December 2022; date of current version 2 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071279, and in part by the SDUST Research Fund under Grant 2019TDJH103. (Corresponding author: Feng Yuan.)

Bo Ai, Heng Xiao, and Hanwen Xu are with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: aibo@sdust.edu.cn; xiaoheng1002@163.com; xu5hanwen@163.com).

Feng Yuan is with the Guangdong Ocean Development Planning Research Center, Guangzhou 510220, China (e-mail: yuanfeng2323@163.com).

Mengyun Ling is with the School of Geomatics, East China University of Technology, Nanchang 330013, China (e-mail: 15107977985@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3230081

to the Food and Agriculture Organization of the United Nations (FAO), global production in 2019 exceeded 85 million tons, worth \$275 billion [1]. Aquaculture provides tremendous economic benefits in the food and financial sectors but also causes many new ecological and environmental issues. The rapid growth of aquaculture production in India has led to the gradual replacement of mangrove wetlands, directly undermining the role of mangroves in stabilizing shorelines and maintaining biodiversity [3], [4]. The use of feed, antibiotics, pesticides, and nutrient-rich effluents in aquaculture have ultimately led to eutrophication and ecosystem degradation in coastal areas and estuaries [1], [5]. It is critical to utilize aquaculture areas, coastal environmental protection efficiently, and food safety to accurately obtain the distribution of aquaculture areas.

Remote sensing technology is widely used in resource monitoring due to its low cost, high efficiency, and comprehensive coverage [6]. In recent years, with the advancement of remote sensing technology, the amount of remote sensing data has grown exponentially, and the data resolution has become higher and higher [7]. Many researchers use high-resolution optical image data to extract aquaculture areas, such as Sentinel-1, Sentinel-2 [8], and GF-2 data.

Currently, the most common methods used to extract aquaculture areas from remotely sensing data include visual interpretation method, methods based on a combination of image features and decision trees or random forests, and deep learning methods. The visual interpretation method is time-consuming and dependent on the experience of the interpreter, [9], and is usually used to produce labeled maps in other segmentation methods. Google Earth Engine (GEE) cloud computing platform, can quickly extract spectral features, texture features, and geometric features of satellite images [10]. The extracted features are combined with random forest or decision tree to extract aquaculture areas. Xu et al. [11] extracted the distribution of aquaculture areas using texture features, normalized vegetation index, normalized water body index, and normalized stacking index features of gray scale cooccurrence matrix (GLCM) as input to the random forest algorithm. Sun et al. [12] used geometric indexes obtained from water index, texture, and radar backscatter to extract the distribution of aquaculture ponds. Yu et al. [13] used pixel selection techniques and image segmentation methods to automatically identify water bodies that are flooded throughout the year; based on this, water bodies were classified into fish ponds and nonfish ponds based on geometric features, such as the area and perimeter of the objects. The extraction of aquaculture

areas based on artificially selected image features is certainly a straightforward method. But it directly depends on which image features we select artificially, and people are subjective about feature selection. The deep learning process directly avoids the selection of image features and obtains them through the computation of the network, while allowing direct end-to-end pixel-level feature extraction.

As a result, many researchers are attempting to use deep learning methods to extract image features and realize image classification. The ResU-Net [14] network can employ satellite images from Sentinel-2, ALOS-DEM, and NOAA-DEM as inputs. The outputs specified nine wetland types; the ResU-Net model was used to map northeast Vietnam's coastal wetland region correctly. Dang et al. [14] created a UPS-Net network structure based on the UNet network structure that can fuse boundary and contextual information to reduce "adhesion" in raft culture areas. Liu et al. [15] used the RCF network structure to extract the raft aquaculture area's boundary line, then vectorized the boundary line in ArcGIS, converted it to surface data, and finally obtained the aquaculture area's scope. They have excellent experimental results, whether using deep learning to extract aquaculture areas directly or extracting the boundary lines of the aquaculture areas first and then extracting the aquaculture areas. To alleviate the difficulty of pixel labeling in hyperspectral images, Hang et al. [16] proposes an unsupervised feature learning model that exploits the relationship between hyperspectral and LiDAR data to extract features and avoid the use of label information. A dual fine-tuning strategy is designed on top of this, which can obtain both semantic and intrinsic structural information of the samples and achieve very good classification performance. Hang et al. [17] proposed a multiscale progressive segmentation network MPSEgNet to solve the problem of simultaneously segmenting large-scale changing objects in high-resolution remote sensing images. MPSEgNet employs three subnetworks to segment large, small, and other scale objects; a position-sensitive module is used to combine the three subnetworks to consider the contribution of each self-network, which achieves a great improvement compared to other segmentation models. This is a significant improvement over other segmentation models.

We try to construct our own network structure to accomplish the extraction of aquaculture areas. The backbone network structure is very essential for the extraction of image features. In view of the excellent performance of ResNet in the image field, we chose ResNet 50, which has a smaller number of parameters and better extraction results, as the backbone network. In the model feature extraction of convolutional neural network (CNN), convolutional computation can be used to obtain multiscale feature maps and contextual information. The self-attentive mechanism can also accomplish the acquisition of contextual information. For this purpose, we construct an improved atrous spatial pyramid pooling (ASPP) module based on ASPP combined with the self-attentive mechanism. Gradient disappearance is a very common problem, and we alleviate this problem by means of auxiliary loss. The combination of the main branch loss and the auxiliary loss is used as the final loss. The above then forms the core of self-attention mechanism and auxiliary loss network (SAMALNet).

II. RELATED WORK

A. Auxiliary Loss in Neural Network

Auxiliary loss is essential in classifying remote sensing images, and many researchers have used auxiliary loss structure in neural networks. In the work on remote sensing image scene classification, Bazi et al. [18] discovered that the CNN network has a gradient disappearance problem, which they solved by using the auxiliary loss to inject the gradient earlier in the network. The global branch of the HR-GLNet structure uses HRNet as the backbone network, while the local branch uses FPN and residual network. When calculating the loss, the primary loss and auxiliary loss are used to aggregate global and local features, resulting in more accurate lunar crater detection than networks like UNet, HRNet, and others [19]. Wang et al. [20] proposed the center loss function as an auxiliary objective function to test its effectiveness as an auxiliary loss function that can improve hyperspectral image classification results. The problem of gradient disappearance in the network structure can be solved by auxiliary loss. We compare the impact of feature maps from different layers, as auxiliary losses on the final experimental results in SAMALNet structure and then discover the perfect network structure.

B. Attention Modules

The attention mechanism is used in the visual tasks of remote sensing image processing to simulate human attention. John and Zhang [21] proposed Attention UNet, a fully CNN structure that adds an attention mechanism to the UNet structure. The network's attention mechanism allows it to retain high-level spatial data. Remote sensing deforestation detection has produced better results than other network structures. Kim et al. [22] presented a multiscale, multilevel attention network (MSMLANet). MSMLANet's attention (MLA) module combines local and global information from multilevel spatial and spectral enhancement for context aggregation. MSMLANet improves the classification results of local climate zones with a more flexible implementation. For building footprint extraction, Liu et al. [23] proposed a multiscale geoscience network (MS GeoNet) that added a CBAM module of channel attention module and spatial attention module to improve the network's generalization ability. Hang et al. [24] proposed a spectral attention module and a spatial attention module in their work on the classification of hyperspectral images. They integrated these two modules into the original CNN, which focused on spectral information and spatial information, respectively, and finally fused the two types of information to achieve better performance than the original CNN. For classification problems in multispectral images, Le Sun et al. [25] proposed the KELM (MSAF-KELM) structure to achieve accurate fusion of multiple classifiers. The KELM structure to achieve fast classification and the WSAFS structure to achieve efficient fusion results. The experimental results show the excellent performance on ultrasmall sample rate. In order to enable the neural network to acquire both low-level features and high-level features, Le Sun et al. [26] proposed the SPANet network structure.

The SPAM structure in SPANet is used to extract deeper multiscale features and salient features, and the FFM structure is used to complete the fusion of low-order features and high-order features. In the field of deep learning for remote sensing change monitoring, Song et al. [27] proposed ACABFNet. ACABFNet extracts the local information of the image by CNN branch and the global information of the image by transformer branch, and then fuses the two by using bidirectional fusion. The global feature information from image height and width is fused using the axial cross-attention module. Excellent experimental results are obtained on three datasets. Yuan et al. [28] proposed OCNNet structures. The difference between SAMALNet and OCNNet is that OCNNet contains interlaced sparse self-attention module. However, SAMALNet adds an auxiliary loss module on top of the self-attention structure, where the auxiliary loss module is added taking into account the pixel categories.

We use a CNN to extract image features to achieve a high-precision extraction of coastal aquaculture areas. The feature map of the backbone feature extraction network is used as an auxiliary loss, so the gradient disappearance problem is alleviated when the network is backpropagated. In constructing the neural network, SAMALNet employs both the auxiliary loss module to mitigate the issue of gradient messages and the self-attention module to obtain richer contextual information, achieving better experimental results than other network architectures.

The main contents of this article can be summarized as follows.

- 1) A neural network structure called SAMALNet is constructed during the segmentation of GF-2 satellite images in the Jiaozhou Bay area of Qingdao, and the experimental results are superior to other neural network structures.
- 2) Using the self-attention module to substitute different branches of the ASPP structure, building modified ASPP structures, and evaluating the impact of several new ASPP structures on the network extraction results.
- 3) Calculating the auxiliary loss between the feature maps and label maps of four different phases in the ResNet50 backbone feature network structure. Comparing the impact of different stages' auxiliary losses on extraction outcomes.

III. MATERIALS AND METHOD

A. Study Area

Jiaozhou Bay is located in the southern part of the Shandong Peninsula, China. It is a natural semienclosed bay [29], [30]. The Dagu River, Mohe River, Baisha River, Licun River, and other rivers flow into the Jiaozhou Bay annually [30]. Jiaozhou Bay is an important coastal aquaculture base in the Shandong Province, which is abundant in fisheries resources and is a key location for the reproduction of a variety of economically important fish, shrimp, and crabs.

The coastal aquaculture area of Jiaozhou Bay is essential in Qingdao. According to the Shandong Province Statistical Yearbook, the aquaculture output of Qingdao exceeds 1.1 million tons, and the aquaculture area exceeds 300 000 hectares from 2015 to 2020. In Fig. 1, the red rectangle indicates the location of

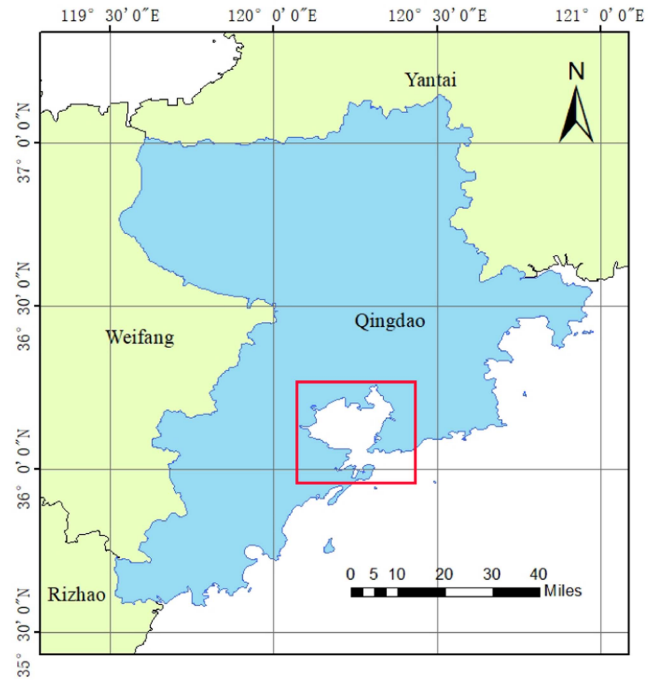


Fig. 1. Overview map of the study area. The red rectangle indicates the sea area of Jiaozhou Bay.

the Jiaozhou Bay area in Qingdao, and the GF-2 satellite image data used in the experiment are from the coastal aquaculture areas along this sea.

B. Dataset and Preprocessing

The experiment used the GF-2 satellite to capture the satellite image of the Jiaozhou Bay area in Qingdao City, Shandong Province, China, from 2015 to 2020. The data observation width is 45 km, and there are five bands in total, with the panchromatic band having a resolution of 1 m and the remaining four bands of blue, green, red, and near-infrared having a resolution of 4 m. The distribution of aquaculture areas only occupies a small fraction of the space due to the huge image width. We trimmed and saved the area containing aquaculture sites in order to obtain 11 satellite images containing aquaculture areas. The data are subjected to geometric correction and atmospheric correction successively. The combined image's resolution was 1 m after merging the blue, green, red, and panchromatic bands. In each satellite image, ArcGIS software was used to delineate the aquaculture areas and preserve them as single-channel image files as semantic segmentation labels, where the background area is denoted by 0 and the aquaculture area is denoted by 1. The fused images and labels are cropped to 256×256 pixels, respectively, and the cropped results are randomly divided into 1622 training sets, 541 validation sets, and 537 test sets. These data are in TIF format, and each cropped image corresponds to its label.

As can be seen from Fig. 2, the aquaculture pond in Fig. 2(a) and the river in the image are more similar, the area of the similar aquaculture pond in Fig. 2(b) shows a gradient of color, not uniform, the color difference of the adjacent aquaculture

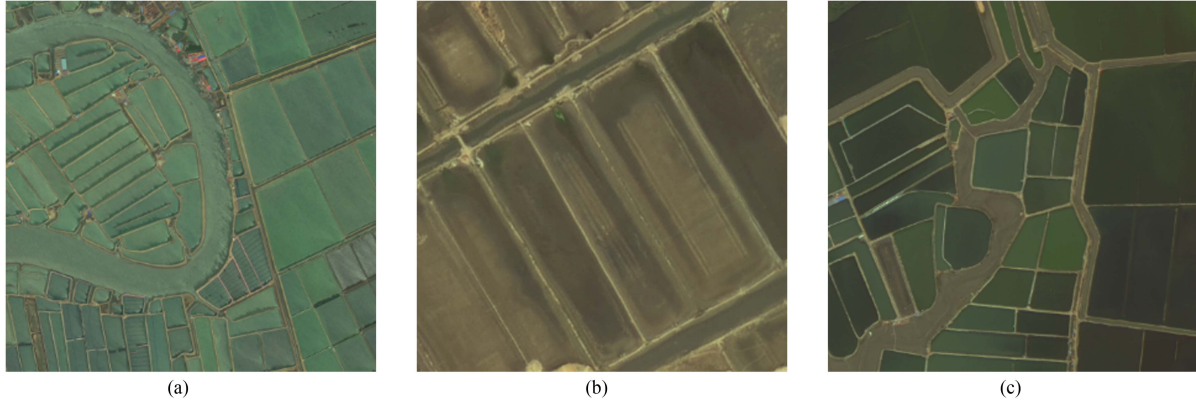


Fig. 2. Detail images of some coastal zone aquaculture areas.

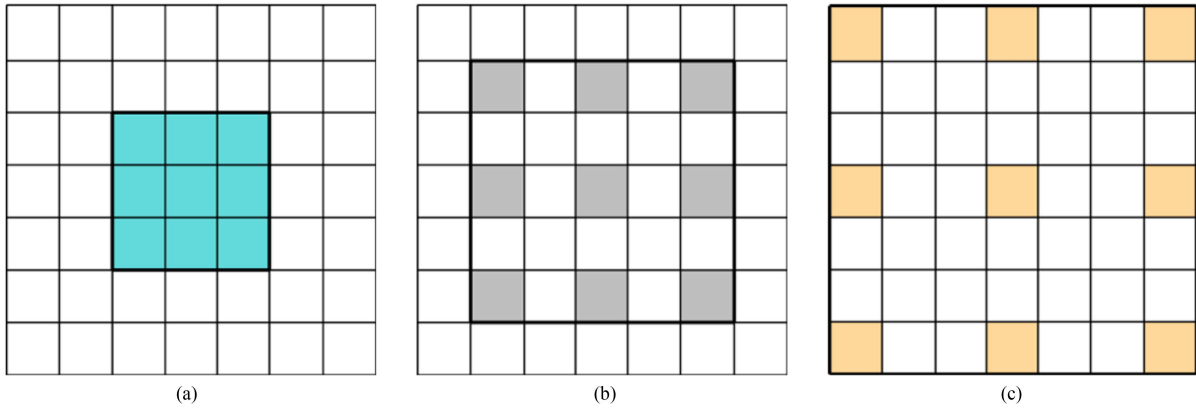


Fig. 3. Comparison between ordinary convolution and dilated convolution receptive field. The black boundary line represents the outer boundary of the receptive field; d stands for expansion coefficient. (a) Ordinary convolution, $d=1$. (b) Dilated convolution, $d=2$. (c) Dilated convolution, $d=3$.

pond in Fig. 2(c) is quite different, in addition, the shape of the aquaculture pond in the image is irregular. It is not appropriate to distinguish between aquaculture areas by color and shape characteristics. Since the aquaculture areas in the image are distributed along the edges of the seawater, the classification should take into account not only the characteristics of the aquaculture zones but also the context information of the aquaculture areas.

The deep neural network consists of multiple layers, including convolutional layers, pooling layers, output layers, etc. The neural network uses a stack of multilayer structures to achieve feature extraction. Low-level rough features are extracted from the shallow network, and fine high-level features are extracted from the deep network. When extracting features, SAMALNet structure is not limited to convolution computation but can obtain more contextual information and employs an additional loss to prevent gradient vanishing.

C. Dilated Convolution and ASPP

In the deep CNN image segmentation task, convolution computation is usually used to extract image features. The size of the convolution kernel directly determines the receptive field of ordinary convolution calculation. To have a larger receptive

field during the convolution calculation, we use the dilated convolution method for feature extraction, which has a larger receptive field than ordinary convolution.

Ordinary convolution is shown in Fig. 3(a). We can also assume that the expansion coefficient of convolution is one and that the 1×1 feature point after convolution corresponds to the 3×3 receptive field in the previous layer's feature map. Fig. 3(b) shows that the expansion coefficient of convolution is 2, and the 1×1 feature point after convolution corresponds to the 5×5 receptive field in the previous layer's feature map. In Fig. 3(c), the expansion coefficient of convolution is 3, and only a 1×1 feature point of convolution corresponds to the previous layer feature map's 7×7 receptive field. The dilated convolution of the calculation formula is as follows:

$$F = (d - 1) \times (k - 1) + k. \quad (1)$$

d stands for the expansion coefficient, k for the convolution kernel size, usually 3×3 or 5×5 , and F for the receptive field size in the formula above. The feature map of the upper layer corresponds to the receptive field, as shown in the bold black box in Fig. 3. Although a dilated convolution can have larger receptive fields and more context information, convolution with a fixed size can only obtain partial features. Multiple

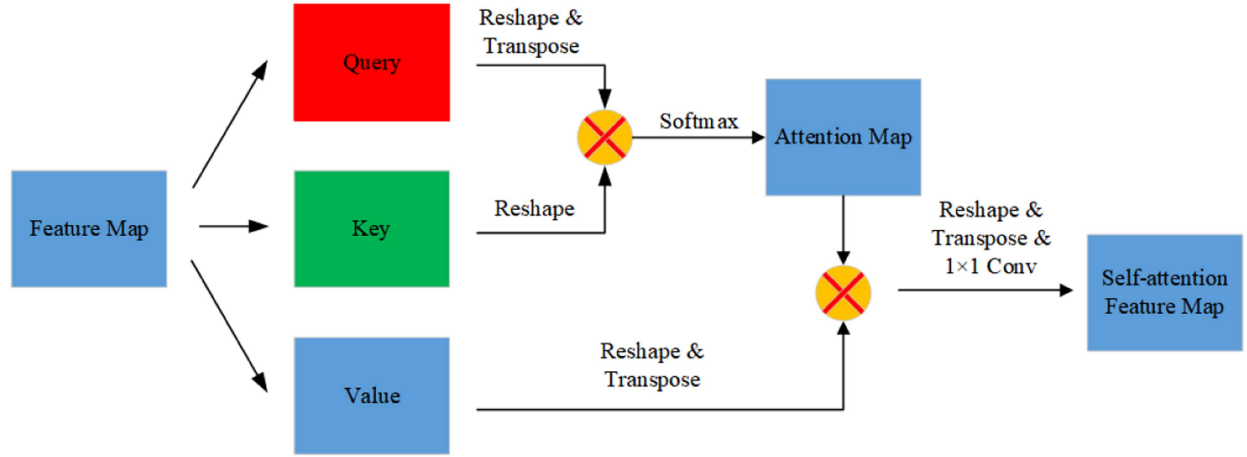


Fig. 4. Calculation flow chart of self-attention. Query, Key and Value correspond to Q, K, and V, respectively.

convolution kernels are used to extract image features with different expansion coefficients from the same feature map, thus producing multiscale features of the same feature map. In the ASPP structure, the feature extraction is usually performed with different expansion coefficients of convolution, which are generally 6, 12, and 18 [31]. Padding is usually specified in the convolution calculation to keep the size of the feature map constant before, and after the calculation [20]. The multiscale feature extraction in the SAMALNet structure adopts 1×1 global average pooling, 1×1 convolution, 3×3 convolution with an expansion coefficient of 6, and 3×3 convolution with an expansion coefficient of 18.

D. Object Context Self-Attention

Convolution computation is used to extract features in both the PSPNet and DeepLab series algorithms. Traditional computations frequently employ convolution kernels of size 3×3 , 5×5 , or even 7×7 . The convolution kernel's size directly affects the receptive field's size. Based on convolution, expansion coefficients of different sizes are added to form dilated convolution to obtain multiscale features and context information to improve the final network output effect. The self-attention mechanism evaluates the relationship between two feature points, receives more contextual information, and generates a new feature map. The following is the formula for calculating the self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{Q \cdot K^T}{\sqrt{d_k}} \right) V. \quad (2)$$

In the above equation, Q (Query), K (Key), and V (Value) are obtained by multiplying the feature map with the matrices W_q , W_k , and W_v , respectively, where W_q , W_k , and W_v are obtained by network learning and d_k is the dimension of the matrix. In Fig. 4, the feature map X is obtained from the calculation of the previous step. We assume that the shape of the feature map X is $B \times C \times H \times W$. Then the shapes of Q and K obtained are $B \times C/2 \times W \times H$, and the shape of V is $B \times C \times H \times W$. For Q , the reshaped shape becomes $B \times N \times C/2$ ($N = H$

$\times W$), and for each channel, each row in the matrix represents the value of each single feature point on each channel. For K , the reshaped shape becomes $B \times C/2 \times N$. On each channel, each column in the matrix represents its value on each channel. Multiplying the processing result of Q with that of K , we can obtain a $B \times N \times N$ feature map with $N \times N$ size on each channel of the feature map, at which time the dependence of any two feature points on the feature map can be obtained. In order to prevent the results in the matrix from being too large and affecting the results of Softmax, let the product of the two be divided by $\sqrt{d_k}$, and after Softmax calculation, the correlation coefficients of Q and K are obtained, which is the Attention map. For V , the shape obtained after reshaping and transposing is $B \times N \times C$. At this time, the Attention Map is multiplied with the result of V , and this result correlation coefficient will be added to V . Then the feature map of size $B \times C \times H \times W$ is obtained by reshaping processing and convolution calculation, and at this time the feature map is equipped with contextual information. The SAMALNet structure employs self-attention to replace the convolution of 3×3 with an expansion factor of 12.

E. Network Structure

We adopt ResNet50 [32] as the feature extraction network, the core skeleton of ResNet50 is the inverse residual structure. ResNet50 uses a reverse residual structure as a robust feature extraction network to overcome the problem of parameter gradient diminishing as the network deepens. It performs excellently in image classification, semantic segmentation, object detection, and other domains.

As shown in Table I, in Stage0, both convolution and max pooling calculations are included. In the convolution part, we first perform two $64 \ 3 \times 3$ convolutions with stride two on the image and $128 \ 3 \times 3$ convolutions with stride two once. Szegedy et al. [33] argued that smaller convolution kernels reduce the computation and number of parameters, so we use multiple 3×3 convolutional computations in Stage0 instead of one 7×7 convolutional computation. In the pooling part, we perform max-pooling of 3×3 feature maps with stride 2. To reduce the

TABLE I
MAIN COMPUTATIONAL PROCEDURE OF RESNET50

Stages of the ResNet50	Stage0	Stage1	Stage2	Stage3	Stage4
The calculation process of each stage	3×3, 64, stride=2 3×3, 64, stride=2 3×3, 128, stride=2 3×3 max pool, stride=2	1×1, 64 3×3, 64 1×1, 256	1×1, 128 3×3, 128 1×1, 512	1×1, 256 3×3, 256 1×1, 1024	1×1, 512 3×3, 512 1×1, 2048
Repeat times	1	3	4	6	3

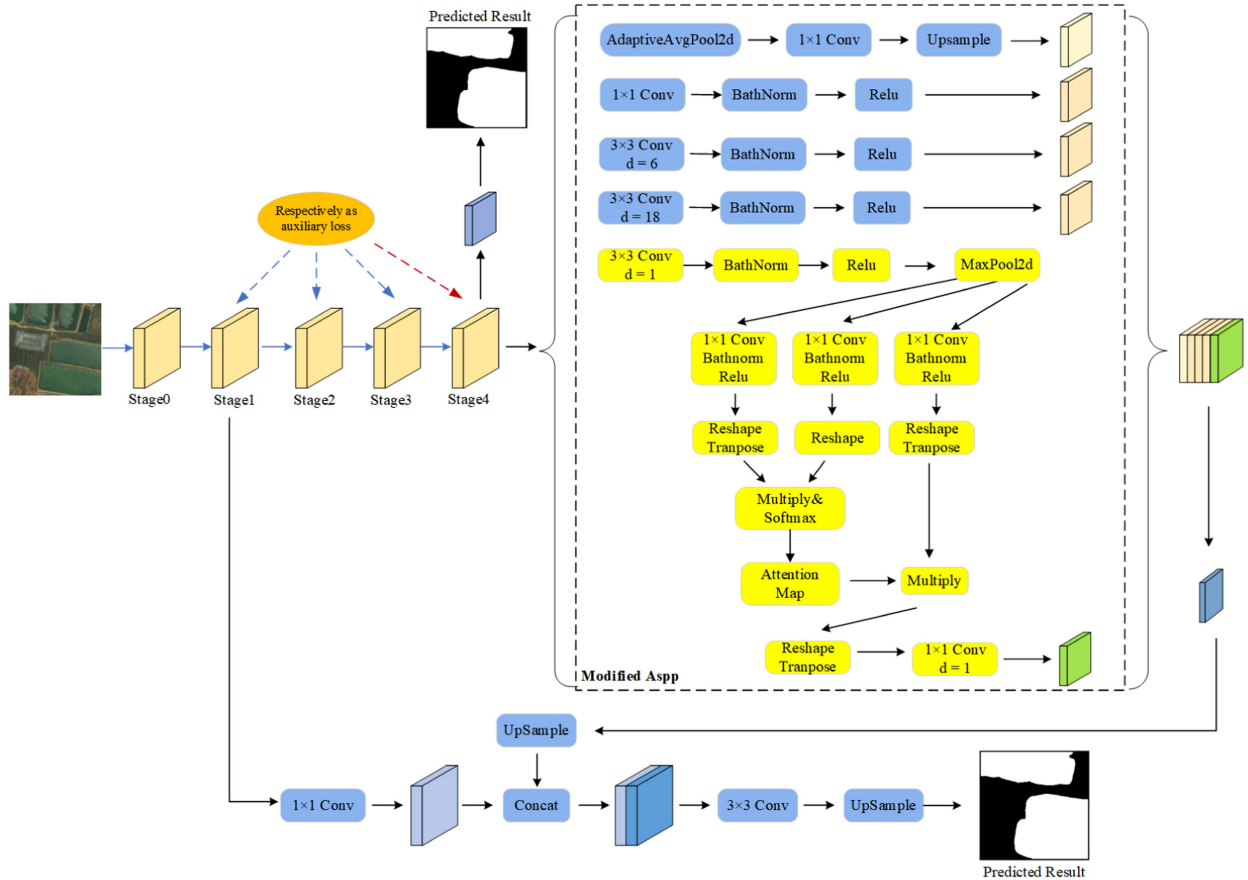


Fig. 5. Structure of SAMALNet. Conv denotes the convolution calculation, d denotes the expansion coefficient, Upsample denotes upsampling calculation, Concat denotes stacking of feature maps, and Image Pool means global average pooling. The red dotted line on the far right of the orange ellipse points to Stage4 as the auxiliary loss.

computational effort, we substitute the convolutional computation with 7×7 convolutional kernels with the convolutional computation with three smaller kernels.

In Stage1, we take the results of the previous computation and perform mainly the convolutional computation with 64 convolutional kernels of 1×1 size and 64 convolutional computations with 3×3 , and 256 convolutional computations with 1×1 convolutional kernels. This Stage1 convolution calculation has been repeated three times, and the result of the previous calculation is used as input for the following calculation. The computation of Stage 2, 3, 4 is similar to that of Stage1.

The whole Stage1 to Stage4 can be regarded as the extraction process in Fig. 6. Supposing the feature map is x , and the residual calculation process is $F(x)$. The output of the residual network

is $H(x)$, then the each process can be expressed by the following formula:

$$H(x) = F(x) + x. \quad (3)$$

We apply 3×3 convolution to the output of Stage4 and adjust the channel number of the feature graph to 512. At this point, the height and width of the feature map are h and w . And then put the resulting map into the module of the modified ASPP. The feature map is entered into the computation of five different branches, respectively. In the dashed box of Fig. 5, the branch in the first row includes 1×1 global average pooling, 1×1 convolution calculation, and upsampling of the feature map to (h, w) size. The second line branch contains 1×1 convolution, batch, and relu

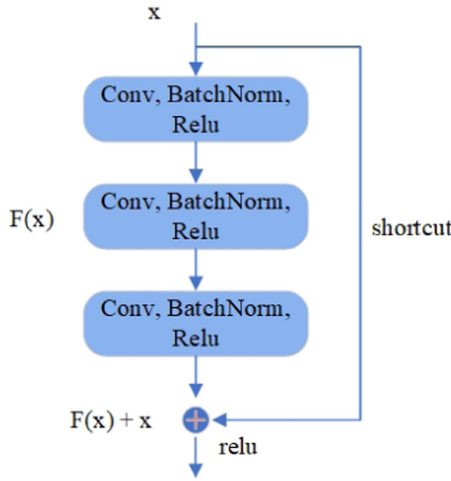


Fig. 6. Residual structure.

function computation. The branch of the third and fourth lines contains 3×3 convolution, batch processing, and relu function calculation, where the $d = 6$ of the convolution part of the third line and the $d = 18$ of the convolution part of the fourth line. The branch of the fifth line mainly contains 3×3 convolution and the self-attention module mentioned in Fig. 4 above. After multiscale image feature extraction, we use the features after Stage1 to keep the shallow features in the network. The results of the above five branches are stacked, and then the stacked results are convolved to adjust the number of channels of the feature map and superimposed with the 1×1 convolution results of Stage1, and then 3×3 convolution and 1×1 convolution are performed in turn to obtain the classification results. The feature map of Stage4 is adjusted and upsampled to obtain another extraction result.

F. Loss Function and Auxiliary Loss

The loss function describes the difference between the network's segmentation result and the actual label. The loss function for evaluating the merits of the segmentation result is usually cross-entropy. The binary cross-entropy loss function is defined as follows:

$$\text{loss}_{\text{main}} = -\frac{1}{n} \sum_{i=0}^n y \cdot \log(y') + (1 - y) \cdot \log(1 - y'). \quad (4)$$

For each image, in the above formula, $y \in \{0, 1\}$ is the true value of the i th pixel, $y' \in \{0, 1\}$ represents the predicted value of the i th pixel, and n represents the number of pixels on the image. In various deep neural network tasks, the auxiliary loss is utilized to prevent gradient vanishing. In PSPNet [34] backbone feature extraction network, the feature map is upsampled to the original image size, and the loss is calculated between the upsampled result and the labeled map. The obtained loss is combined with the loss of the main branch in a specific ratio as the final loss. In the test task, only the extraction results of the main branch are used as the final output. In our experimental data, there are more than 83 million pixels in the background

of the training set, accounting for 78.22%, and more than 23 million pixels in the aquaculture area, accounting for 21.78%, due to the imbalance of data categories. The following formula is used to calculate the weight:

$$\text{weight} = \log_e(N). \quad (5)$$

When calculating the weights of different categories, N denotes the number of pixels in the aquaculture area and the background, respectively.

$$\begin{aligned} \text{loss}_{\text{auxiliary}} = & -\frac{1}{n} \sum_{i=0}^n [w_1 \cdot y \cdot \log(y') \\ & + w_2 \cdot (1 - y) \cdot \log(1 - y')]. \end{aligned} \quad (6)$$

In the above equation, w_1 and w_2 represent the weights of background and aquaculture area, respectively. The ordinary loss function is used on the main branch of the network, while the weighted cross entropy loss function is used on the auxiliary loss branch. The auxiliary loss optimizes the learning process during the training of the network. Only the prediction results of the main branch are used in the testing phase, and this strategy is widely used in fully CNNs based on the structure of ResNet [34]. For the coefficients of the auxiliary loss function, we use 0.4 as the coefficient of the auxiliary loss function, and the choice of coefficients is described in more detail in Section V.

$$\text{loss}_{\text{total}} = \text{loss}_{\text{main}} + 0.4 \times \text{loss}_{\text{auxiliary}}. \quad (7)$$

G. Evaluation Metrics

In a supervised classification task, there are a variety of evaluation indexes for neural network structure algorithms, most of which are calculated using the confusion matrix. For evaluation, we use confusion matrix-based precision, recall, F1 score, and MIoU. In Table III, true positive (TP) indicates that the pixels in the prediction result and the actual label is the number of pixels in the aquaculture area. False negative (FN) indicates that the prediction result is the background, and the true label is the number of pixels in the aquaculture area. False positive (FP) indicates that the predicted result is the aquaculture area and the true label is the number of pixels in the background. True negative (TN) indicates the number of pixels where the predicted result and the actual label are aquaculture areas.

Recall stands for recall rate.

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (8)$$

Precision stands for precision.

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (9)$$

F1 Score is the harmonic mean based on precision and recalls.

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (10)$$

TABLE II
EXTRACTION EVALUATION INDEX OF JIAOZHOU BAY DATA BY SAMALNET
AND OTHER NEURAL NETWORKS (%)

Network Structure	Backbone	Recall	Precision	F1	MIoU
SegNet	VGG16	94.77	94.96	94.87	90.35
DeepLabv3	ResNet101	95.42	94.87	95.14	90.84
UNet	VGG16	95.25	95.49	95.37	91.24
DeepLabv3+	ResNet101	95.56	95.20	95.38	91.26
PSPNet	ResNet101	95.34	95.94	95.63	91.72
UPS-Net	—	95.34	95.28	95.31	91.15
SAMALNet	ResNet50	96.34	95.91	96.12	92.60

SAMALNet results and best results are in bold.

TABLE III
CONFUSION MATRIX

Confusion Matrix		Predicted Class	
		Aquaculture area	Background
Label Class	Aquaculture area	True Positive	False Negative
	Background	False Positive	True Negative

MIoU represents the mean of the crossover ratio.

$$\text{MIoU} = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + FP + TP}. \quad (11)$$

The number of categories in the foreground is denoted by k in the above equation. The total number of classifications is $k+1$ because the background occupies one category, and coastal aquaculture areas are assigned to another category.

IV. EXPERIMENT

A. Experiment Details

We use 0.5–2.0 times random scaling, random horizontal flipping, transformation into tensors, and regularization in the training stage. The 256×256 random center clipping is used to transform into tensors and regularize in the verification and testing stages. The initial learning rate of 0.1, batch size of 16, 20 000 iterations, optimizer selected SGD, the momentum of 0.9, weight attenuation of $1E-4$, validation set interval of 100.

The operating system is CentOS Linux Release 7.6.1810 (Core), the disk storage is 20 GB, the CPU is Intel(R) Xeon(R) Platinum 8163 CPU@2.50 GHz, the GPU is NVIDIA Tesla T4 16 GB, the deep learning framework is Pytorch version 1.2.0, and the environment for writing and running code in Python 3.6.8.

B. Experiment Result

Under the different network structures and the exact experimental details, our and other neural network structures are used to train and verify the Jiaozhou Bay wetland data. Table II summarizes the final test set results. Average values of recall, precision, F1 score, and MIoU in background and aquaculture areas are used as comparison indexes to simplify data comparison. The results of the experiments demonstrate that SAMALNet design based on self-attention mechanism and auxiliary loss structure outperforms SegNet [35], DeepLabv3 [31], UNet [36], DeepLabv3+ [37], UPS-Net [38], and PSPNet in four indicators.

In Table II, the experimental results show that SAMALNet achieves the most in recall, F1, and MIoU metrics compared to other networks in the suburban area, and still only a 0.03%

difference with PSPNet in precision metrics. DeepLabv3+ has excellent ASPP structure and decoding network structure; it surpasses other CNN network structures in the recall metric, reaching 95.56%, and still has a 0.78% difference with SAMALNet. PSPNet has an excellent pyramidal pooling module and outperforms other CNN network structures in precision, F1, and MIoU metrics. The precision exceeds SAMALNet by 0.03%, while F1 and MIoU are 0.49% and 0.88% different from SAMALNet. SAMALNet has both a module for contextual information and a module for mitigating auxiliary loss, which is an obvious advantage over the full CNN network structure.

Fig. 7 shows the results of using various neural networks to segment GF-2 satellite images in the Jiaozhou Bay area of Qingdao. The extraction results of SAMALNet and PSPNet outperform other neural networks when the extraction results in the first column are compared with the regions inside the pink rectangles in the label map. The extraction results of SAMALNet, DeepLabv3, and DeepLabv3+ in the green rectangular box are great, while the extraction results of other regions are relatively poor. Meanwhile, UNet's green rectangular block diagram contains some aquaculture areas that are incorrectly divided. In the extraction result graph in the second column, SAMALNet and DeepLabv3 in the purple rectangle area are better, whereas PSPNet, UNet, DeepLabv3+, and SegNet are significantly different from the label graph. DeepLabv3+, UPS-Net, and UNet have divided the aquaculture area portion of the purple ellipse in the figure into the background. Other network extraction results, particularly PSPNet extraction results, are relatively poor, whereas our results in the third column extraction result graph are closest to the label graph in the red circle area. SAMALNet, DeepLabv3+, and DeepLabv3 have good extraction results in the blue rectangular region of the fourth column's extraction result graph, whereas other methods have significant differences with the label graph. Only the SAMALNet method is closest to the label graph in the red right triangle region, and other extraction result graphs differ significantly from the label graph.

Combining the quantitative comparison of the four evaluation indicators in Table II and the above extraction results, it is not difficult to see that the SAMALNet structure has a significant advantage over other networks in terms of the values of the evaluation indicators, and SAMALNet achieves better experimental results than other networks in the edge part of the aquaculture area extraction results.

V. DISCUSSION

In this section, we discuss the experiments in four main aspects as follows.

- Evaluate the impact of different backbone feature networks on SAMALNet on the experimental results.
- In constructing the global loss, we assign different coefficients to the auxiliary loss to evaluate the impact of coefficient differences.
- Evaluate the impact of different ASPP structures on the experimental results.
- Evaluate the impact of different stages of feature maps as auxiliary losses on the experimental results.

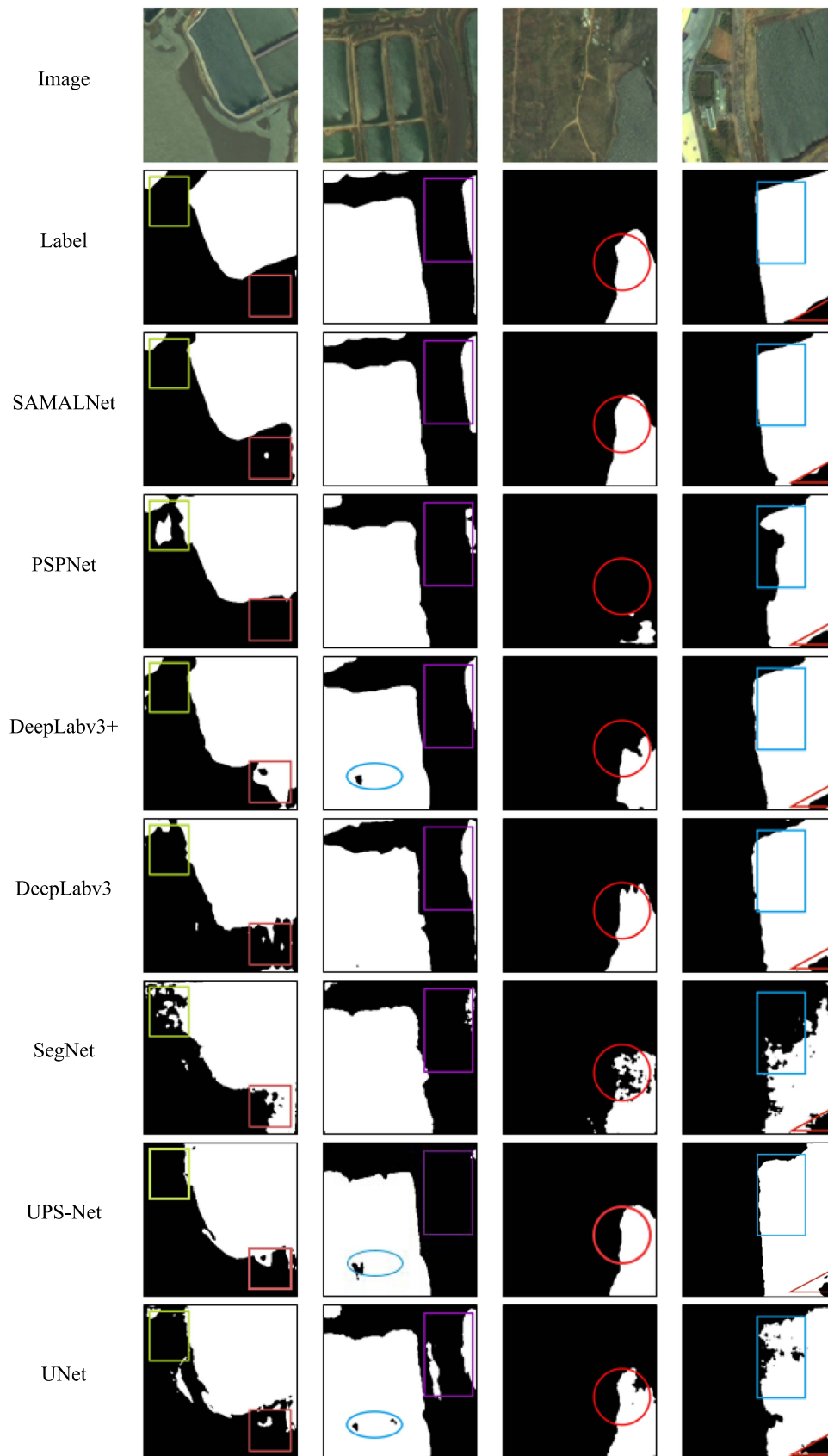


Fig. 7. Comparison of segmentation results of Jiaozhou Bay data by several neural networks. In the images, the black part represents the background, and the white part represents the aquaculture area. We add black edges to each segmentation result to distinguish the boundary of the image.

TABLE IV
EXTRACTION EVALUATION INDEX OF JIAOZHOU BAY DATA BY SAMALNET
AND OTHER NEURAL NETWORKS (%)

Network Structure	Backbone	Recall	Precision	F1	MIoU
SAMALNet	EfficientNet	93.47	91.31	92.31	85.95
SAMALNet	ShuffleNet	96.00	95.93	95.96	92.31
SAMALNet	MobileNetV2	95.16	95.48	95.32	91.15
SAMALNet	ResNet18	95.55	96.05	95.80	92.02
SAMALNet	ResNet34	95.90	95.45	95.68	91.79
SAMALNet	ResNet101	95.92	95.22	95.56	91.60
SAMALNet	ResNet50	96.34	95.91	96.12	92.60

Best result and best backbone in bold.

TABLE V
EFFECT OF THE COEFFICIENTS OF THE AUXILIARY LOSS FUNCTION ON THE
EXPERIMENTAL RESULTS (%)

Network Structure	Coefficient	Recall	Precision	F1	MIoU
SAMALNet	0.2	95.70	95.30	95.50	91.47
SAMALNet	0.3	95.29	95.09	95.19	90.93
SAMALNet	0.4	96.34	95.91	96.12	92.60
SAMALNet	0.5	95.99	95.62	95.80	92.02

Best results in bold.

A. Ablation Study of Different Backbone

In the construction of SAMALNet, we selected ResNet50 as the backbone network structure due to the extraction results. We also conducted experiments to evaluate the impact of other backbone networks on the experiments, and the main networks selected were EfficientNet [39], ShuffleNet [39], MobileNetV2 [40], ResNet18 [32], ResNet34 [32], and ResNet101 [32]. The results of this part of the experiments are shown in Table IV.

From the table, we can see that when ResNet50 is used as the backbone network The extraction of aquaculture areas achieves the best experimental results in recall, F1, and MIoU, but there is a difference of 0.14% with ResNet18 in terms of precision metrics. We chose ResNet50 as the backbone network structure by considering all the metrics.

B. Ablation Study of Different Auxiliary Loss Function Coefficients

In constructing the final loss function as the experimental results, we used different auxiliary loss function coefficients to evaluate the experimental results. 0.2, 0.3, 0.4, and 0.5 were used as coefficients to construct the network structure, and the final results are shown in Table V.

The results show that when 0.4 was chosen as the coefficient, better results were obtained for all indicators.

C. Ablation Study of Different ASPP Structures

Image global average pooling, 1×1 convolution, 3×3 convolution with an expansion coefficient of 6, 3×3 convolution with an expansion coefficient of 12, and 3×3 convolution with an expansion coefficient of 18 are the five modules that make up the ASPP module. The self-attention module replaces the convolution module with a 3×3 expansion factor of 12 to create a modified ASPP module, as shown in the neural network structure diagram in Fig. 5. In order to obtain better experimental results, other experimental attempts are required. On the premise that the following other network modules remain unchanged,

we adjusted the ASPP structure and built five different network structures. For each row in Table VI, the leftmost part represents the mode's name, and the other parts on the right represent the modules contained in the combined method. Compared with other modes, the ASPP constructed by mode 4 is the best.

According to the network structure composed of different ASPP combinations in Table VI, the same experimental environment and network parameters were used to train, verify, and test the wetland data in the Jiaozhou Bay area of Qingdao. Table VII shows that in terms of recall, mode two and mode three achieve 95.65% and 95.65%, respectively, while mode four outperforms them by 0.22% and 0.69%, respectively. Regarding accuracy, mode four fails to achieve the best performance and is 0.27% off the best indicator. For F1, mode one and mode two achieve 96.08% and 95.89%, respectively, while mode four is 0.04% and 0.23% higher than them. For the MIoU indicator, mode two and mode one achieve 92.19% and 92.52%, while mode four is 0.41% and 0.08% higher, respectively. Among these four evaluated metrics, mode IV has the best experimental results in the three metrics. It became an improved version of ASPP, part of the SAMALNet structure.

D. Ablation Experiments With Different Stages as Auxiliary Loss

In the above, the final selected modified ASPP includes image pool, a 1×1 convolution, a convolution with an expansion coefficient of 3×3 , self-attention, and a 3×3 convolution with an expansion coefficient of 18. In Fig. 5, the four feature maps pointed by the dotted line in the orange ellipse are used as auxiliary losses to construct different network structures under the premise that other network structures remain unchanged, and the Stage4 pointed by the red dotted line on the far right of the current orange ellipse is used as an auxiliary loss. In ResNet50, the channels of the output feature maps of Stage1 and Stage2 are 256 and 512, respectively, and we directly use 1×1 convolution for classification; the channels of the output feature maps of Stage3 and Stage4 are 1024 and 2048, respectively, and we first use 1×1 convolution to adjust the channels of the feature maps and then perform classification. In the figure, the four characteristic diagrams pointed by the orange oval dotted line are, respectively, used as auxiliary losses under the premise that other network structures remain unchanged to build different network structures. The characteristic diagram of Stage4 is currently displayed as auxiliary losses. Comparative experiments with different stages as auxiliary losses are also carried out on the premise that the improved ASPP remains unchanged and other network structures and parameters remain unchanged.

For the above four different kinds of feature maps for classification as auxiliary loss, we combined the structure of mode four with the Stage1, Stage2, Stage3, and Stage4 outputs to construct four different network structures, respectively. We used the data from the aquaculture area in the Jiaozhou Bay area of Qingdao and trained, validated, and tested them separately, and the test result is shown in Table VIII. We can see that the feature map of Stage4 in ResNet50 as the auxiliary loss has achieved the best experimental results in recall, F1 Score, and MIoU, while

TABLE VI
ASPP STRUCTURE FORMED BY FIVE COMBINATION MODES

Combination	Module name				
mode one	self-attention	1×1 Conv	3×3 Conv, d=6	3×3 Conv, d=12	3×3 Conv, d=18
mode two	Image Pool	self-attention	3×3 Conv, d=6	3×3 Conv, d=12	3×3 Conv, d=18
mode three	Image Pool	1×1 Conv	self-attention	3×3 Conv, d=12	3×3 Conv, d=18
mode four	Image Pool	1×1 Conv	3×3 Conv, d=6	self-attention	3×3 Conv, d=18
mode five	Image Pool	1×1 Conv	3×3 Conv, d=6	3×3 Conv, d=12	self-attention

Best mode is bold.

TABLE VII
EXPERIMENTAL RESULTS OF FIVE MODES (%)

Combination	Recall	Precision	F1	MIoU
mode one	95.97	96.18	96.08	92.52
mode two	95.65	96.14	95.89	92.19
mode three	96.12	95.80	95.96	92.31
mode four	96.34	95.91	96.12	92.60
mode five	96.03	95.80	95.91	92.22

Best mode and best result in bold.

TABLE VIII
ASPP OF MODE FOUR AND DIFFERENT LAYER FEATURE MAPS ARE SELECTED AS AUXILIARY LOSS EXPERIMENTAL RESULT (%)

Auxiliary loss layer	Recall	Precision	F1	MIoU
Stage1	95.05	96.12	95.57	91.60
Stage2	95.47	95.85	95.66	91.76
Stage3	96.04	95.99	96.02	92.41
Stage4	96.34	95.91	96.12	92.60

Best results in bold.

it is still different from the best results in the precision index. In the feature extraction network part of this experiment, we compare the results of different layers of network feature maps as auxiliary losses and finally choose the output of Stage4 in ResNet50 as the auxiliary loss.

VI. CONCLUSION

ResNet50 is used as the feature extraction network, and a new neural network structure SAMALNet is constructed. By combining the self-attention mechanism module and an auxiliary loss module, we employ SAMALNet, UNet, SegNet, DeepLabv3, DeepLabv3+, UPS-Net, and PSPNet to extract coastal aquaculture areas from satellite remote sensing images of the Jiaozhou Bay Qingdao. According to the experimental results, SAMALNet, based on auxiliary loss and self-attention, is generally better than the above network structure in evaluation indicators. The extraction results in the edge part of the aquaculture area are better than the above network structure.

Meanwhile, we compared the different network structures obtained by replacing each module in the ASPP module using the self-attention mechanism separately. After the experimental comparison, we chose the convolutional module with a convolutional kernel of 3×3 size and an expansion coefficient of 12, replaced by the self-attention module as the final network structure for the improved ASPP. With the above modified ASPP modules fixed unchanged, we extracted the feature maps at different levels in the network modules as the auxiliary loss modules. The experimental ablation results show that the optimal results can be obtained using the feature map of Stage4 in ResNet50 as an auxiliary loss. In this article, we construct a neural network using a self-attention mechanism and auxiliary loss to achieve

better experimental results than other networks in aquaculture area extraction work. We use the relatively simple ResNet50 as a feature extraction network, but a network structure like Swin Transformer has robust feature extraction capabilities and works very well in areas, such as image segmentation and object detection. In our future work, we will attempt to use a structure similar to Transformer for image processing work.

REFERENCES

- [1] M. Ottinger, F. Bachofer, J. Huth, and C. Kuenzer, "Mapping aquaculture ponds for the coastal zone of Asia with Sentinel-1 and Sentinel-2 time series," *Remote Sens.*, vol. 14, no. 1, 2022, Art. no. 153.
- [2] M. Ottinger, K. Clauss, and C. Kuenzer, "Large-scale assessment of coastal aquaculture ponds with Sentinel-1 time series data," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 440.
- [3] K. A. Prasad, M. Ottinger, C. Wei, and P. Leinenkugel, "Assessment of coastal aquaculture for India from Sentinel-1 SAR time series," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 357.
- [4] D. Wang et al., "Estimating aboveground biomass of the mangrove forests on northeast Hainan Island in China using an upscaling method from field plots, UAV-LiDAR data and Sentinel-2 imagery," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 85, 2020, Art. no. 101986.
- [5] D. Stiller, M. Ottinger, and P. Leinenkugel, "Spatio-temporal patterns of coastal aquaculture derived from Sentinel-1 time series data and the full Landsat archive," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1707.
- [6] A. T. N. Dang, L. Kumar, M. Reid, and H. Nguyen, "Remote sensing approach for monitoring coastal wetland in the mekong delta, Vietnam: Change trends and their driving forces," *Remote Sens.*, vol. 13, no. 17, 2021, Art. no. 3359.
- [7] Q. Zhao et al., "An overview of the applications of Earth observation satellite data: Impacts and future trends," *Remote Sens.*, vol. 14, no. 8, 2022, Art. no. 1863.
- [8] W. Yuan and W. Xu, "MSST-Net: A multi-scale adaptive network for building extraction from remote sensing images based on Swin transformer," *Remote Sens.*, vol. 13, no. 23, 2021, Art. no. 4743.
- [9] X. Zhang, S. Ma, C. Su, Y. Shang, T. Wang, and J. Yin, "Coastal oyster aquaculture area extraction and nutrient loading estimation using a GF-2 satellite image," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4934–4946, 2020.
- [10] Y. Duan et al., "Tracking changes in aquaculture ponds on the China coast using 30 years of Landsat images," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 102, 2021, Art. no. 102383.
- [11] Y. Xu, Z. Hu, Y. Zhang, J. Wang, Y. Yin, and G. Wu, "Mapping aquaculture areas with multi-source spectral and texture features: A case study in the pearl river basin (Guangdong), China," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4320.
- [12] Z. Sun et al., "Nation-scale mapping of coastal aquaculture ponds with Sentinel-1 SAR data using Google earth engine," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 3086.
- [13] Z. Yu, L. Di, M. S. Rahman, and J. Tang, "Fishpond mapping by spectral and spatial-based filtering on Google earth engine: A case study in singra upazila of Bangladesh," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2692.
- [14] K. B. Dang et al., "Coastal wetland classification with deep U-Net convolutional networks and Sentinel-2 imagery: A case study at the Tien Yen estuary of Vietnam," *Remote Sens.*, vol. 12, no. 19, 2020, Art. no. 3270.
- [15] Y. Liu, X. Yang, Z. Wang, C. Lu, L. Zhi, and Y. Fengshuo, "Aquaculture area extraction and vulnerability assessment in Sanduao based on richer convolutional features network model," *J. Oceanology Limnology*, vol. 37, pp. 1941–1954, 2019.

- [16] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532812.
- [17] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012.
- [18] Y. Bazi, M. M. Al Rahhal, H. Alhichri, and N. Alajlan, "Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2908.
- [19] Y. Jia, L. Liu, S. Peng, M. Feng, and G. Wan, "An efficient high-resolution global-local network to detect lunar features for space energy discovery," *Remote Sens.*, vol. 14, no. 6, 2022, Art. no. 1391.
- [20] W. Wang, S. Dou, and S. Wang, "Alternately updated spectral-spatial convolution network for the classification of hyperspectral images," *Remote Sens.*, vol. 11, no. 15, 2019, Art. no. 1794.
- [21] D. John and C. Zhang, "An attention-based U-Net for detecting deforestation within satellite sensor imagery," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 107, 2022, Art. no. 102685.
- [22] M. Kim, D. Jeong, and Y. Kim, "Local climate zone classification using a multi-scale, multi-level attention network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 345–366, 2021.
- [23] T. Liu et al., "Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 109, 2022, Art. no. 102768.
- [24] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.
- [25] L. Sun, Y. Fang, Y. Chen, W. Huang, Z. Wu, and B. Jeon, "Multi-structure KELM with attention fusion strategy for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539217.
- [26] L. Sun, S. Cheng, Y. Zheng, Z. Wu, and J. Zhang, "SPANet: Successive pooling attention network for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4045–4057, 2022.
- [27] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, and B. Chen, "Axial cross attention meets CNN: Bibranch fusion network for change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 32–43, 2023.
- [28] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context for semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 2375–2398, Aug. 2021, doi: [10.1007/s11263-021-01465-9](https://doi.org/10.1007/s11263-021-01465-9).
- [29] X. Ji, B. Yang, and Q. Tang, "Acoustic seabed classification based on multibeam echosounder backscatter data using the PSO-BP-adaboost algorithm: A case study from Jiaozhou bay, China," *IEEE J. Ocean. Eng.*, vol. 46, no. 2, pp. 509–519, Apr. 2021.
- [30] Z. Wang, H. Wang, S. Fan, M. Xin, and X. Sun, "Community structure and diversity of macrobenthos in Jiaozhou bay," *Mar. Pollut. Bull.*, vol. 171, 2021, Art. no. 112781.
- [31] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: <http://arxiv.org/abs/1706.05587>
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
- [34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239, doi: [10.1109/CVPR.2017.660](https://doi.org/10.1109/CVPR.2017.660).
- [35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017, doi: [10.1109/TPAMI.2016.2644615](https://doi.org/10.1109/TPAMI.2016.2644615).
- [36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Med. Image Comput. Comput.-Assist. Interv. MICCAI*, pp. 234–241, 2015.
- [37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Comput. Vis.*, pp. 833–851, 2018.
- [38] B. Cui, D. Fei, G. Shao, Y. Lu, and J. Chu, "Extracting raft aquaculture areas from remote sensing images via an improved U-Net with a PSE structure," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 2053.
- [39] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, vol. 97, pp. 6105–6114.
- [40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520, doi: [10.1109/CVPR.2018.00474](https://doi.org/10.1109/CVPR.2018.00474).



modeling, maritime search, and rescue decision analysis.



Bo Ai received the B.S. degree in printing engineering and the M.S. degree in cartography and geographic information systems from Wuhan University, Wuhan, China, in 2001 and 2005, respectively, and the Ph.D. degree in cartography and geographic information engineering from the Shandong University of Science and Technology (SDUST), Qingdao, China, in 2011.

He is currently a Professor with the SDUST. He is the Director of the Geographic Information Department, SDUST. His research interests include coastal natural resources monitoring, ocean spatial-temporal

Heng Xiao received the bachelor's degree in surveying and mapping engineering from Nanyang Normal University, Henan, China, in 2020. He is currently working toward the master's degree with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao, China.

His research interests include remote sensing image processing.



Hanwen Xu received the B.S. degree in geographic information science from Shandong University of Science and Technology, Shandong, China, in 2019, and the M.S. degree in geography from Shandong University of Science and Technology, Qingdao, in 2022.

His research interest is remote sensing image processing.



marine industry.

Feng Yuan received the B.S. degree in physical geography from Wuhan University, Wuhan, China, in 2003, and the M.S. degree in physical geography from Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Beijing, China, in 2006.

He is a Senior Engineer and is currently the Director of Marine Information Office of Guangdong Ocean Development Planning Research Center. His research interests include marine information technology, marine data analysis, marine economy, and



Mengyun Ling received the bachelor's degree in surveying and mapping engineering from Nanyang Normal University, Henan, China, in 2020. She is currently working toward the master's degree in surveying and mapping engineering with the East China University of Technology, Jiangxi, China.

Her research interests include UAV image processing, deep learning, and artificial intelligence.