MDOAU-net: A Light and Robust Deep Learning Model for Aquaculture SAR Image Classification

Jichao Wang, Jianchao Fan, *Member, IEEE*, Jun Wang, *Fellow, IEEE*

***Abstract*—Offshore aquaculture area information extraction from synthetic aperture radar (SAR) images is important for large-scale marine resource exploitation and protection. In this letter, a deep learning model called Multi-scaled Attention U-net with Dilated convolution and Offset convolution (MDOAU-net) is proposed for aquaculture SAR image classification. The U-net backbone and attention gate of the Attention U-net are used in the MDOAU-net model. In addition, the MDOAU-net model consists of three distinctive parts. First, a multi-scale feature-fusion block is adopted in the input of MDOAU-net to extract features from raw images. Moreover, adapted from the Attention U-net for SAR image classification, fewer channels are used in each convolution layer of the MDOAU-net to match latent features in SAR images. Furthermore, nine dilated convolution blocks are adopted in the encoder-decoder structure to extract semantic features in the presence of speckle noises. In addition, offset convolution blocks are developed to convert spatial information into channel information for the precise classification of blurry boundaries. Four skip connections of the U-net backbone are replaced by four offset convolution blocks. Experiment results are elaborated to demonstrate that the MDOAU-net model with much fewer parameters significantly outperforms six existing methods in terms of classification accuracy.**

***Index Terms*—SAR, marine aquaculture, classification, offset convolution, dilated convolution**

I. Introduction

AQUACULTURE is a major part of the global economy. It provides us with more than 82 million tons of fish and 32 million tons of seaweeds yearly [1]. For managing the marine resources and protecting the environment, we need to monitor the scale of aquaculture areas closely.

Aquaculture area monitoring is commonly carried out via image classification using synthetic aperture radars (SAR) to obtain remote-sensing images. Existing image classification methods at the pixel-level can be classified as conventional algorithms and machine learning algorithms. Among the conventional algorithms, the watershed algorithm [2] is a region growing algorithm originated in mathematical morphology. The simple linear iterative clustering algorithm (SLIC) [3] is an algorithm for clustering pixels sharing common characteristics into the same super-pixels. Although these algorithms are good at interpretability, they are not effective in classifying images with complex backgrounds and noises, because they can only extract few special features.

Many machine learning algorithms, such as support vector machine [4], are also applied for image classification. Although these algorithms are good at classification accuracy, they can handle images with simple contents only. The fully convolutional network [5] is the first deep learning model for pixel-level image classification. U-net [6] is a well-known deep learning model for image classification. Its encoder-decoder structure is robust as demonstrated in a series of applications [7] [8]. Seg-net [9] is developed with less memory by using pool indices to replace skip connections. The self-attention module is developed [10] to highlight useful features and suppress others. Attention U-net [11] is a combination of self-attention modules and U-net for cancer image classification. In the aforementioned studies, image classification performance is improved by adding convolution layers in skip connections for the fusion of shallow features and abstract features. Furthermore, GoogLeNet [12], VGG [13], and Residual Net [14] are very deep nets for image classification with small kernels, and they provide pretrained modules whose kernel weights are common features for classification. The atrous convolution (i.e., dilated convolution) is proposed in Deep-lab-v3 to obtain a large receptive field [15]. The aforementioned deep learning models are for general image classification without addressing the characteristics of SAR images. Autoencoder-based DCSCN is developed for raft information extraction and achieves almost 10% accuracy improvement over some machine learning methods [16]. Newly proposed U-net-like models for object detection in SAR images imply that a light version of U-net is also capable for SAR image classification [17].

Such related gernel image classification works inspired us to use U-net-based deep learning methods for aquaculture SAR image classification. Two main contributions of this letter are summarized as follows:

- A light and robust U-net-like model, called MDOAU-net, is proposed to classify SAR images for offshore aquaculture.
- Offset convolution blocks are developed in MDOAU-net to extract information from noisy SAR images of dense aquaculture areas with blurred boundaries.

The work described in the paper was supported in part by the National Natural Science Foundation of China under Grant 42076184, 41876109,41806207, 41706195, in part by the National Key Research and Development Program of China under Grant 2017YFC1404902 and Grant 2016YFC1401007, in part by the National High Resolution Special Research under Grant 41-Y30F07-9001-20/22, in part by the Research Grants Council of the Hong Kong Special Administrative Region of China under General Research Fund Grant 11202318. (*Corresponding authors: Jianchao Fan and Jun Wang.*)

Jichao Wang is with the School of Data Science, City University of Hong Kong, Hong Kong, China (e-mail: jichawang2-c@my.cityu.edu.hk).

Jianchao Fan is with the Key Laboratory of Sea-Area Management Technology, Department of Ocean Remote Sensing, National Marine Environmental Monitoring Center, Dalian, China. (e-mail: jcfan@nmemc.org.cn).

Jun Wang is with the Department of Computer Science and the School of Data Science, City University of Hong Kong, Hong Kong, China (e-mail: jwang.cs@cityu.edu.hk).

The remainder of this letter is organized as follows. Section II describes the proposed MDOAU-net. Section III delineates experimental results of marine aquaculture SAR image classification using six methods. Finally, Section IV concludes the letter.

II. Proposed Method

A. Overview of MDOAU-net

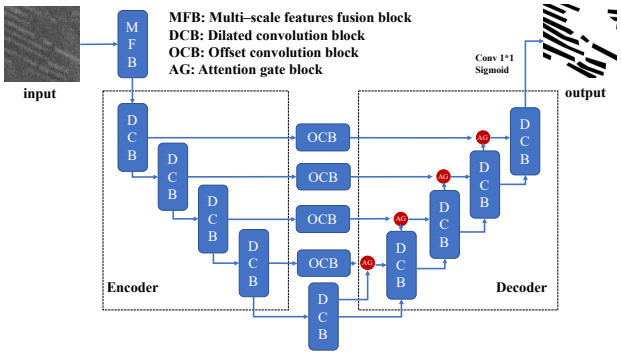


Fig. 1. Architecture of the MDOAU-net model

The architecture of MDOAU-net is shown in Fig. 1. The MDOAU-net model consists of one multi-scale feature-fusion block, nine dilated convolution blocks, four offset convolution blocks, and four attention gates. In view of the widespread deployment of attention gates in deep learning, the attention gate module in [11] is also used in MDOAU-net.

B. The number of channels

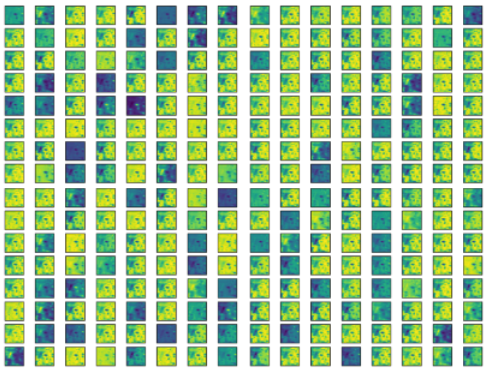


Fig. 2. Visualization of the output feature maps in the 16th convolution layer of the Attention U-net, captured from a forward propagation pass for classifying an SAR image

Fig. 2 shows a visualization of features belonging to a trained Attention U-net model for SAR image classification. Some feature maps displayed in Fig. 2 are similar. As similar feature maps represent similar features, it is not necessary to set as many as 256 channels in the layer. This situation also happens on other convolution layers. To avoid learning redundant features, the number of channels in each layer of MDOAU-net is only one-eighth of that in the Attention U-net.

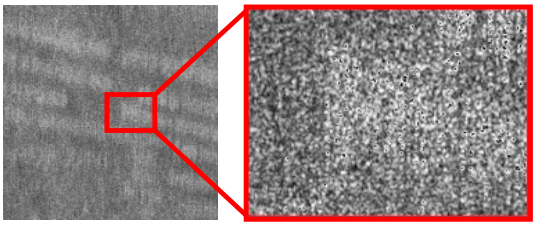


Fig. 3. An SAR source image and a regional enlarged view

C. Dilated convolution blocks

Harsh noises in an SAR image, as shown in Fig. 3, bring three challenges. The first one is that strong noises randonly make some pixels in the receptive field of a convolution kernel to become outliers. Generally, the kernel size of a convolution kernel is normally three or five. So, the receptive field contains no more than 25 pixels. Several outliers account for a considerable proportion of a receptive field, resulting a fatal threat to feature extraction. The second challenge is that harsh noises in SAR images make the shapes of aquaculture areas more irregular. The last one is that noises make texture features unclear and difficult to be extracted from SAR images in grayscale. A good way to overcome speckle noises is to expand the receptive field of convolution kernels.

The formula to compute the receptive field side is

$$Side = \gamma(\kappa - 1) + 1, \quad \gamma, \kappa \in \mathbb{N}_+ \tag{1}$$

where γ is the dilated rate, and κ is the kernel size.

An straightforward approach is to increase the size of convolution kernels to expand the receptive field. Despite a large convolution kernel achieves the purpose, it reduces feature extraction efficiency with increased computational effort, because valuable information in pixels is corrupted by heavy speckle noises. As shown in (1), the other way to expand the receptive field is by increasing γ. It is a spaced sampling policy of convolution operations and called dilated convolution. The dilated convolution operator is defined as

$$(F *_{dc} k)(p) = \sum_{s+\gamma t=p} F(s)k(t)$$

where $*_{dc}$ is the operative symbol of dilated convolution, F is the picture matrix, k is a convolution kernel, γ is the dilated rate, κ is the convolution kernel matrix and p is the superimposed limit of position coordinates s and t.

Since γ and κ are independent variables, γ can be increased without changing κ. So, the dilated convolution can expand the sampling range without adding parameters. Table I lists the sizes of receptive fields of blocks in the encoder.

Because SAR images hardly contain continuous linear textures, breaking the default continuous sampling policy would lose little information. Given the above explantion, we adopt dilation convolution kernels in basic convolution blocks and name them dilated convolution blocks. The formula described dilated convolution block is written as

$$DCB(X, \kappa, \gamma) = ReLU(BN(Conv(X, \kappa, \gamma)))$$

TABLE I
SIZES OF RECEPTIVE FIELDS OF BLOCKS IN THE ENCODER OF MDOAU-NET AND ATTENTION U-NET ($pixel$)

Measured position	MDOAU-net	Attention U-net
Encoder block1	21×21	5×5
Encoder block2	41×41	9×9
Encoder block3	61×61	13×13
Encoder block4	81×81	17×17

where DCB and X are input and output, respectively; $ReLU$, BN, and $Conv$ denote the ReLU active function, two-dimensional batch norm function, and two-dimensional convolution function, respectively.

D. A multi-scaled feature-fusion block

The shape of the offshore aquaculture area is approximately rectangle, whereas the size of rectangles varies. For this reason, a multi-scaled feature fusion block is added in the front of the encoder. A broad receptive-field convolution can classify big objects and a small-scaled receptive-field one can classify small ones. In addition, in pixel-level classification, each pixel is supposed to be associated with a label. Semantic information is usually displayed by many pixels and it is too abstract to be described by using a formula. Without using semantic information, image classification performance is often compromised, as images contain both semantic and geometric information and different types of objects may have the same geometric characteristics. So the combination of abstract semantic information and shallow geometric information is helpful. A multi-scaled feature-fusion block consists of five convolution kernels with different combinations of kernel sizes and dilation rates. Because the biggest receptive field in the multi-scaled feature-fusion block contains 43 pixels, the reflection padding operation is used to avoid the occurrence of artifacts appearing at the image edges.

For a give input matrix X with m rows and n columns, the formula for a reflection padding operation is written as

$$RP(X,l,r,u,d) = \begin{bmatrix} X_{1+u,1+l} & \cdots & X_{1+u,1} & \cdots & X_{1+u,n} & \cdots & X_{1+u,n-r} \\ \vdots & \ddots & \vdots & & \vdots & \iddots & \vdots \\ X_{1,1+l} & \cdots & X_{1,1} & \cdots & X_{1,n} & \cdots & X_{1,n-r} \\ \vdots & & \vdots & & \vdots & & \vdots \\ X_{m,1+l} & \cdots & X_{m,1} & \cdots & X_{m,n} & \cdots & X_{m,n-r} \\ \vdots & \iddots & \vdots & & \vdots & \ddots & \vdots \\ X_{m-u,1+l} & \cdots & X_{m-u,1} & \cdots & X_{m-u,n} & \cdots & X_{m-u,n-r} \end{bmatrix}$$
$$l,r,u,d \in \mathbb{N}_+$$

where l, r, u, and d are the padding lengths in the left, right, top, and bottom direction of X, respectively.

Next, the multi-scaled feature-fusion block is written as

$$\begin{aligned} DCB_n &= DCB(RP(X,\lambda,\lambda,\lambda,\lambda),\kappa_n,\gamma_n), \\ MFB(X) &= [DCB_1, DCB_2, DCB_3, DCB_4, DCB_5], \end{aligned}$$

where X is the input matrix of the block, λ is the padding size, n is the index of scales, κ_n and γ_n are the parameters of dilated convolution kernels.

E. Offset convolution blocks

The boundaries of aquaculture areas in SAR images are usually unclear due to noises. As convolution kernels scan over the image of a border area, the classification results transit from a high possibility of labeling aquaculture areas to a high possibility of labeling other areas. For the convolution operation with its receptive field covering pixels in two sides of an aquaculture area boundary, the activation results are error-prone. Shrinking the error-prone border areas could improve classification performance. A convolution kernel convolving over only one side of boundary yields undoubted activation results. By comprehensively analyzing local features from boundary areas, clear features characterizing boundaries can be extracted.

Because the aforementioned convolution extracts the features offsetting from the center of the convolution kernel, we define it as offset convolution. In implementing the offset convolution, a reflection padding is added to the original feature map in one direction. Then a dilated convolution is applied to the padded feature map. The offset convolution is written as

$$OC(X,\kappa,\gamma,l,r,u,d) = DCB((RP(X,l,r,u,d)),\kappa,\gamma)$$

where X, l, r, u, d are the same as RP; l, r, u, and d are offset control parameters. If any three of the control parameters equal 0, the kernel extract features of the center pixel in one edge. If any two of them are 0, the feature is extracted by the kernel for the pixel of a corner in a reception field.

In SAR images, aquaculture areas are usually inclined. So, four offset convolutions with different directions, such as northeast, northwest, southwest, and southeast, are used to obtain surrounding information. Then four feature maps are concatenated together. In the output of the block, another convolutional kernel is used to extract the information contained in all the channels. The offset convolution block is described as

$$\begin{aligned} OC_1 &= OC(X,\kappa,\gamma,Side(\kappa,\gamma),0,Side(\kappa,\gamma),0), \\ OC_2 &= OC(X,\kappa,\gamma,0,Side(\kappa,\gamma),Side(\kappa,\gamma),0), \\ OC_3 &= OC(X,\kappa,\gamma,Side(\kappa,\gamma),0,0,Side(\kappa,\gamma)), \\ OC_4 &= OC(X,\kappa,\gamma,0,Side(\kappa,\gamma),0,Side(\kappa,\gamma)), \end{aligned}$$

$$OCB(X,\kappa,\gamma) = DCB([OC_1,OC_2,OC_3,OC_4],\kappa,\gamma).$$

Fig. 4 illustrates the offset convolution block. A specific pixel in the input feature map, such as pixel A in Fig. 4, deviates from the corresponding element in the four new feature maps. The offsets allow convolution kernels to learn the features of the specific pixel in a direction, transferring blurred boundary features into latent channel features.

III. EXPERIMENTS

A. Experiment settings

The experiments are based on 109 raw SAR images captured by the GF-3 satellite in the coastal area near Changhai,

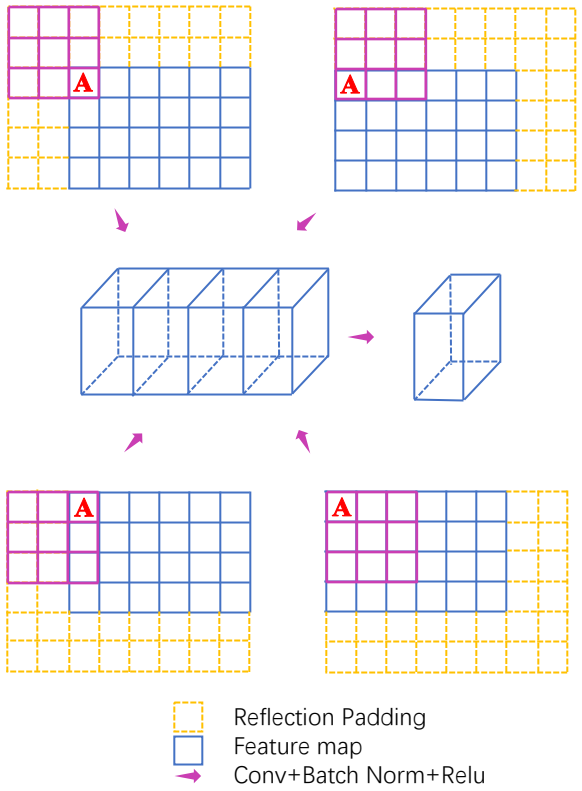


Fig. 4. Offset convolution block schematic diagram

Dalian, China. An offshore aquaculture dataset is constructed from these images via manual labeling. In the experiments, the dataset is partitioned to a training set with 89 images and a test set with 28 images. In addition, six baselines (i.e., Seg-net, U-net, Attention U-net, U-net++, ResNet50, and ResNet101) are coded in Pytorch. All the models are optimized by using the ADAM optimizer with the learning rate 10^{-3}. The source codes and pre-trained models are available on https://github.com/Jichao-Wang/MDOAU-net.

Overall accuracy (OA) is used as the evaluation metric, as in [18] and [19]. OA is defined as

$$OA = \frac{N_{pre}}{N_{total}}$$

where N_{pre} represents the number of pixels with right classification and N_{total} represents the total number of input pixels.

B. Ablation results

The results of three experiments are discussed below to compare the classification performance of the offset convolution and multi-scaled feature-fusion block proposed in Section II. Experiments 1 and 2 are based on pre-trained VGG blocks [13] and the proposed offset convolution blocks, respectively, by replacing skip connections. Experiment 3 is based on MDOAU-net (i.e., the model in Experiment 2 with the multi-scaled feature-fusion block added to its input). Compared the result of Experiment 1 with that of Experiment 2 in Table II, we can see from that offset convolution blocks outperform general convolutional blocks replacing skip connections. Compared the result of Experiment 2 with that of Experiment 3 in Table II, we can see that the multi-scaled feature-fusion block improves the classification accuracy.

TABLE II
ABLATION EXPERIMENT RESULTS FOR THE OFFSET CONVOLUTION BLOCK AND THE MULTI-SCALE FEATURE-FUSION BLOCK

Ablation experiment number	#1	#2	#3
Pre-trained VGG block replace skip connection	✓		
Offset convolution block replace skip connection		✓	✓
Fused multi-scaled block			✓
OA(%)	90.43	91.52	92.02

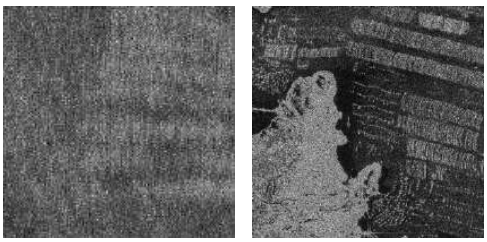
C. Performance comparisons

From Table III, we can see that the overall accuracy of MDOAU-net is about 3%-4% higher than those of Seg-net, U-net, and Attention U-net. In addition, although the performance by MDOAU-net is is improved 0.4% only over those by the state-of-the-art model, U-net++, in the U-net family, the number of MDOAU-net parameters is less than a half of U-net++ parameters.

TABLE III
NUMBER OF PARAMETERS AND OVERALL ACCURACY OF THE MODELS

Model	# of parameters (million)	OA(%)
Seg-net [9]	30	89.19
U-net [6]	31	88.08
Attention U-net [11]	34	88.14
U-net++ [7]	9	91.58
ResNet50 [14]	38	77.15
ResNet101 [14]	71	63.98
MDOAU-net	**4**	**92.02**

D. Case studies



(a) (b)

Fig. 5. Two SAR image samples: (a) sample A, (b) sample B

In the following, the differences in classification performance among the competing models are explained based on two SAR image samples shown in Fig. 5. Classification probability maps are generated by the models before binarization to become classification result maps. Both the classification probability maps and classification result maps are used to explain differences among model performances. Since the overall accuracy of ResNet50 and ResNet101 are significantly

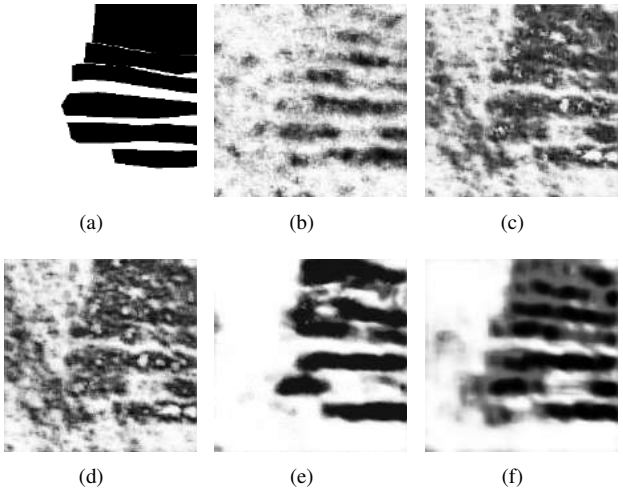


Fig. 6. Classification probability maps for sample A: (a) Ground truth of sample A, (b) Seg-net, (c) U-net, (d) Attention U-net, (e) U-net++, (f) MDOAU-net.

lower than those of other models as shown in Table III, their results are not shown herein.

The ground truth image in Fig. 6(a) shows that the aquaculture area with floating rafts is dense and its shape is slightly deformed from a regular rectangle. The boundaries of the aquaculture area are blurred due to the low image resolution and quality. Figs. 6(b), (c), and (d) show that Seg-net, U-net and Attention U-net result in high classification errors. Figs. 6(e), and (f) show that MDOAU-net outperforms U-net++ in terms of classification accuracy in the presence of blurry boundaries in the dense aquaculture area.

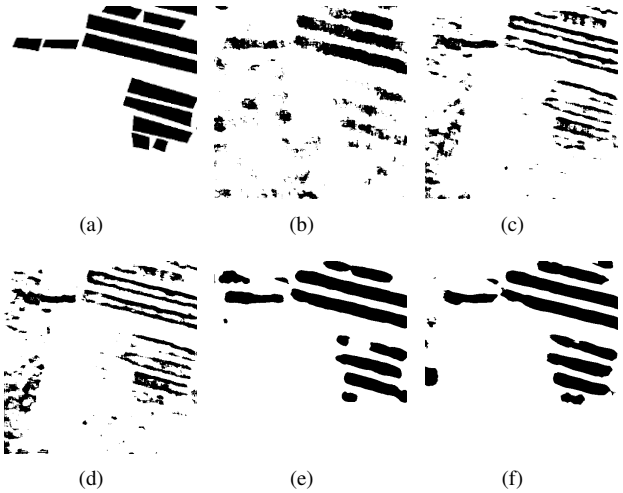


Fig. 7. Classification results for sample B: (a) Ground truth of sample B, (b) Seg-net, (c) U-net, (d) Attention U-net, (e) U-net++, (f) MDOAU-net.

As shown in Fig. 5(b), the grayscale values of island and aquaculture areas in sample image B are similar. As shown Fig. 7(f), MDOAU-net almost accurately distinguishes the mainland, operational floating rafts, and inoperative rafts. This example fully demonstrates the high robustness and accuracy of MDOAU-net for SAR image classification in complex backgrounds.

IV. Conclusions

In this letter, a deep neural network model called MDOAU-net is proposed for SAR image classification. Compared with six popular deep learning models, MDOAU-net has the fewest parameters and perform the best in terms of overall accuracy. In addition, MDOAU-net is shown to be more robust than other competing models for noisy image classification in dense aquaculture areas.

References

[1] FAO, "Fisheries and Aquaculture Software. FishStatJ: Software for Fishery and Aquaculture Statistical Time Series." [Online]. Available: http://www.fao.org/fishery/statistics/software/fishstatj/en

[2] F. Meyer, "Color image segmentation," in *Proc. of International Conference on Image Processing and Its Applications*, 1992, pp. 303–306.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[4] H. Xu, G. Zhu, J. Tian, X. Zhang, and F. Peng, "Image segmentation based on support vector machine," *J. Electron. Sci. Technol.*, vol. 3, no. 3, pp. 226–230, 2005.

[5] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2016.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[7] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[8] F. H. Wagner, R. Dalagnol, Y. Tarabalka, T. Y. Segantine, R. Thomé, and M. Hirye, "U-Net-Id, an instance segmentation model for building extraction from satellite images—case study in the Joanópolis city, Brazil," *Remote Sensing*, vol. 12, no. 10, p. 1544, 2020.

[9] 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.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, and B. Kainz, "Attention u-net," in *Proc. of 1st Conference on Medical Imaging with Deep Learning*, 2018.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[15] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[16] J. Geng, J. Fan, J. Chu, and H. Wang, "Research on marine floating raft aquaculture SAR image target recognition based on deep collaborative sparse coding network," *Acta Automatica Sinica*, vol. 42, no. 4, pp. 593–604, 2016.

[17] J. Li, C. Guo, S. Gou, Y. Chen, M. Wang, and J. W. Chen, "Ship segmentation on high-resolution SAR image by a 3d dilated multiscale u-net," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020.

[18] Y. Hu, J. Fan, and J. Wang, "Classification of PolSAR images based on adaptive nonlocal stacked sparse autoencoder," *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 7, pp. 1050–1054, July 2018.

[19] Z. Guo, H. Liu, Z. Zheng, X. Chen, and Y. Liang, "Accurate extraction of mountain grassland from remote sensing image using a capsule network," *IEEE Geosci. Remote. Sens. Lett.*, vol. 18, no. 6, pp. 964–968, 2020.