An Attention-Fused Deep Learning Model for Accurately Monitoring Cage and Raft Aquaculture at Large-Scale Using Sentinel-2 Data

Yunci Xu and Lizhen Lu

***Abstract*—Cage and raft aquaculture (CRA) is vital for the coastal economy and provides high-quality aquatic products. Accurately monitoring large-scale CRA lays the foundation for predicting CRA product yield and mitigating environmental impacts. This study, focusing on the challenges of detecting large-scale. CRA from freely downloaded, multispectral remote sensing imagery due to the complexity of both CRA and marine environment, proposed an attention-fused deep learning model for accurately retrieving large-scale CRA in China's offshore sea using open-source Sentinel-2 (S2) satellite data. We first downloaded the cloud-free preprocessed S2 images in selected study areas. Manual labeling of cage, raft, and background areas was performed using high-resolution remote sensing images, with labeled images clipped into 32×32 patches. To enhance the perception ability of the feature, the convolutional block attention module was integrated into the well-performing UNet++ by incorporating both channel and spatial attention in each convolutional block of the encoder as well as the Level 1 convolutional blocks of the decoder. Using the sample dataset in 2021, the proposed AF-UNet++ was trained, compared to four mainstream convolutional neural networks, and then adopted to map CRA in both 2021 and 2018 in the study areas, as well as four additional sites. Experimental results demonstrate: 1) our model has the highest OA, F1, and m intersection over union (IoU), with IoU for cage 4.15% higher than other models. 2) Visual comparison illustrates that AF-UNet++ best excels in extracting CRA. 3) Extraction results both in 2021 and 2018 confirm the proposed model can effectively monitor large-scale CRA and has the spatio-temporal stability.**

***Index Terms*—Attention-fused deep learning, cage and raft aquaculture (CRA), China's offshore sea, convolutional block attention module (CBAM), Sentinel-2 (S2) imagery, UNet++.**

I. Introduction

GLOBAL aquaculture production increased by 5.78% from 2018 to 2020, retaining its growth trend amid the worldwide spread of the COVID-19 pandemic [1]. China has produced more farmed aquatic animals and algae than any other country in the world since 1991 and has witnessed an increase in the entire volume of aquaculture, reaching 66.90 million tons in 2021 [2]. Offshore aquaculture, which is commonly equipped with cages and floating rafts, is one of the most important seafood sources for the growing population of our nation and world, while also serving as a significant catalyst for economic development. However, the development of cage and raft aquaculture (CRA) has resulted in a series of eco-environmental problems, such as seawater pollution caused by excessive breeding baits and spreading fertilizers [3], eutrophication caused by cultured crop metabolites [4], quality decline of marine products caused by antibiotic abuse [5], and the impact of the uncontrolled expansion of breeding areas on coastal ecosystems and ship traffic [6].

Accurately and timely monitoring CRA is the first and also essential step for alleviating the aforementioned problems. Remote sensing technology, due to its benefits of large-scale coverage and real-time availability, proves to be the most effective approach for dynamically monitoring land cover, environment, and also offshore aquaculture. Thus, a considerable number of researches on accurately identifying CRA from satellite imagery have been conducted since the start of this century. In terms of the classification methods, these researches can be categorized into the following types [7]: visual interpretation [8], [9], information enhancement [10], [11], feature learning [12], [13], and object-oriented classification [14], [15], [16]. In recent years, deep learning has been preferred to precisely detect CRA from satellite imagery owing to its ability to automatically capture abstract features [17]. This has allowed it to achieve more accurate classification results than traditional machine learning methods such as random forest and support vector machine. Among the considerable number of deep learning models, especially among the convolutional neural networks (CNNs), U-Net along with its modified versions has attracted popular attention in semantic segmentation for its elegant architecture and relatively effective while efficient performance, etc. [18], [19], [20]. For instance, Lu et al. [21] improved the U-Net model and applied it for extracting offshore aquaculture areas with medium-resolution remote sensing imagery; Yu et al. [22] proposed a framework that integrates UNet++ with marker-controlled watershed segmentation for the accurate and refined delineation of aquaculture ponds, etc.

Meanwhile, the attention mechanism, which focuses on important features and suppresses unnecessary ones, has been widely integrated into CNNs, especially in U-Net-like or other variants of the encoder–decoder architecture, for improving the representation of interests and the segmentation results. For example, Cui et al. [23] created a reverse attention module that

Manuscript received 9 October 2023; revised 6 January 2024 and 6 February 2024; accepted 7 April 2024. Date of publication 18 April 2024; date of current version 1 May 2024. *(Corresponding author: Lizhen Lu.)*

The authors are with the School of Earth Science, Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: 12238037@zju.edu.cn; llz_gis@zju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3390762

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

suppresses seawater features, enabling the learning characteristics for both apparent and inapparent aquaculture sites. Qin et al. [24] embedded the convolutional block attention module (CBAM) [25] into the decoder of the network they proposed to gain accurate feature maps for offshore farm extraction, etc.

Recently, the existing studies on identifying CRA are mainly using high spatial resolution (HSR) remote sensing images. For example, Deng et al. [26] designed a multiscale-fusion superpixel segmentation optimization module and combined it with DeepLabV3+ network for extracting CRA from fused 0.8-m-resolution Gaofen-2 images; Wang et al. [27] developed an incremental double unsupervised deep learning model consisting of a feature extraction network and a fully convolutional semantic segmentation network for characterizing unlabeled CRA from 3-m-resolution Gaofen-3 images; Ma et al. [28] adopted a combination method of piecewise linear stretching and R^3Det network to discriminate CRA from the fused 2-m- resolution Gaofen-6 images; Chen et al. [29] applied DeepLabV3+ and U-Net models for recognition of marine ranching from multitemporal 1-m-resolution Gaofen-1 data, and the results show that U-Net performs more stably. Relevant research works on pixel-level classification of CRA with CNNs also include using HSR Gaofen-2 [30], [31], [32], Google Earth [33], [34] and Gaofen-1 [35] satellite images, as well as using unmanned aerial vehicle images [36]. Though the aforementioned studies on CRA detection by CNNs can obtain high classification accuracies, their proposed methods are still unsuitable for monitoring CRA in large-scale areas for the following reasons.

1) HSR satellites commonly need to take a couple of months for revisiting the same large-scale area while CRA information needs to be quickly updated.
2) HSR satellites often have an insufficient number of spectral bands, resulting in a significant loss of spectral information.
3) Massive data and computer power are needed to extract large-scale CRA using HSR imagery.
4) HSR images are almost not free to use which results in a considerable expenditure for updating large-scale CRA.

Therefore, a few studies applied CNNs for mapping or even monitoring CRA from Landsat [37] and 16-m resolution Gaofen-1 [38] images. Nevertheless, most of the aforementioned research works gained good performance only within their specific research regions, while some studies [39], [40], focusing on extracting CRA over large-scale, encountered limitations in terms of capturing fine-grained details and discerning individual objects. Accurately monitoring CRA at a large scale still faces challenges because of the diversity of cage and raft materials, the differences in CRA shapes, colors, and sizes, and the confusion of cage and raft target identification caused by the complex background interference of land and sea [26].

To address the abovementioned challenges, this study proposes a new attention-fused deep learning model for accurately monitoring CRA at large-scale (i.e., the study area covers more than 1500 km^2) by utilizing open-source 10-m-resolution Sentinel-2 (S2) satellite remote sensing data. The main objectives of this study are as follows.

1) We construct a representative sample dataset of CRA for deep learning, which includes a wide range of positive samples encompassing CRA instances of various sizes, shapes, spectral characteristics, and spatial distributions, as well as negative samples consisting of different seawater backgrounds that do not contain CRA.
2) We propose an attention-fused deep learning model named AF-UNet++ by flexibly integrating the CBAM module into the well-performed CNN skeleton, UNet++. The proposed model, employing dense connections to facilitate comprehensive information learning across all layers and incorporating the attention modules to enhance the efficient learning of valuable features related to CRA, harnesses the strengths of both the CNN and attention module, and thus enables the accurate detection of large-scale CRA.

II. Study Area and Sample Dataset

A. Study Areas

China's extensive coastline and abundant marine areas confer innate geographical advantages for the thriving development of mariculture. We selected the following three large aquaculture zones as the study areas (red boxes in left map of Fig. 1).

1) Sansha Bay in the northeastern region of Fujian Province, which covers 1770 km^2 and is surrounded by mountains and islands, an interspersed cages, and floating rafts.
2) Haizhou Bay with 1860 km^2 and diverse seawater characteristics in Jiangsu Province, which is known as the "Capital of China's Seaweed" and is mainly arranged by square-shaped rafts.
3) The eastern coastal area encompassing Huangjuzi Bay and Ludao Island in Liaoning Province, which covers an area of 2024 km^2 and is mainly for sea cucumber farming with floating rafts, its seawater depths varying from east to west.

The variety of materials, colors, and shape sizes of cage and raft, and the diversity of geographical and seawater conditions in the study areas, lead to the difficulty in accurately monitoring CRA.

B. Data and Preprocessing

The freely available S2 multispectral instrument from the European Space Agency provides researchers great opportunities for exploring the use of those satellite images to monitor land cover and targets of interest like CRA in coastal areas. S2 multispectral images contain 13 bands with different resolutions of 10, 20, and 60 m. The 10-m-resolution bands are Blue, Green, Red, and Near InfraRed. The 20-m-resolution bands are two Short-Wave InfraRed bands and four Vegetation Red Edge bands. The 60-m-resolution bands are the Water Vapor band, the Coastal Aerosol band, and the Cirrus band. Taking into account the spatial resolution, we used four 10-m and six 20-m bands of S2 level-2A images for further analysis and resampled them to 10-m-resolution using the bilinear interpolation method.

Considering the growth cycle of marine plants, and the fact that the coastal zones are often covered by thick clouds in the

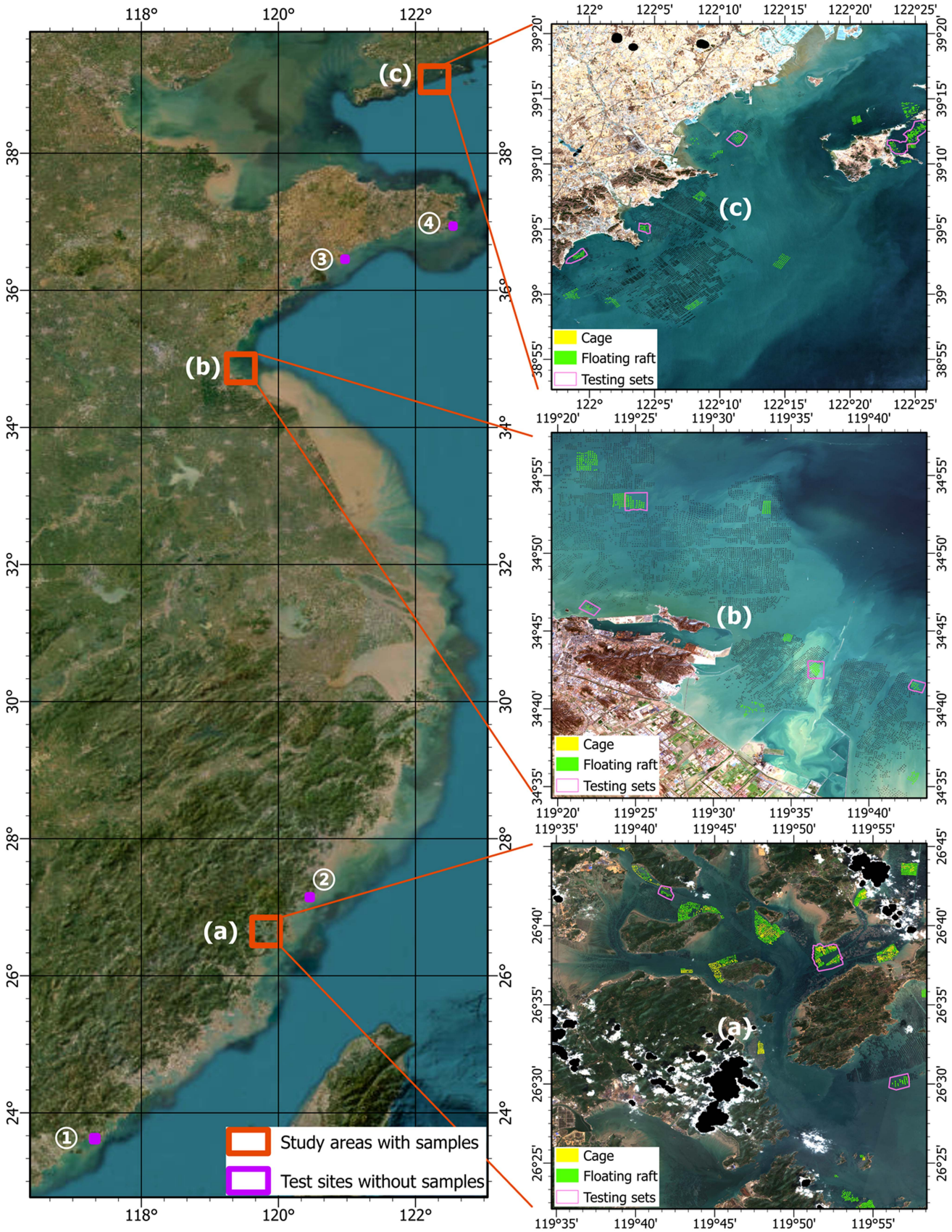

Fig. 1. Location of our study areas and test sites (left), the distribution of samples in the study areas of (a) Fujian, (b) Jiangsu, and (c) Liaoning (the samples in pink boxes are for testing).

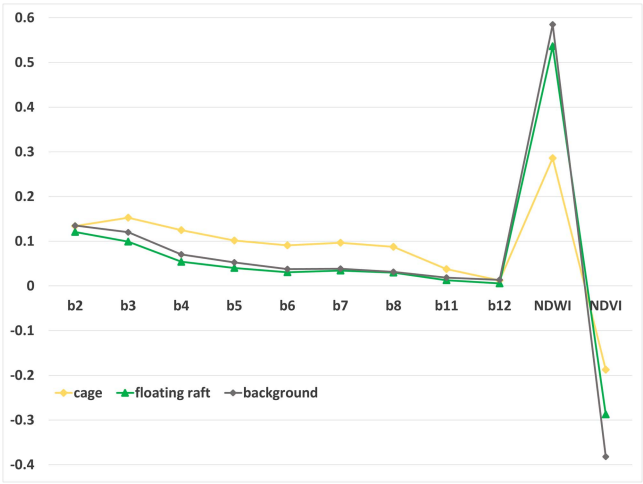

Fig. 2. Spectral curves of cage, floating raft, and background.

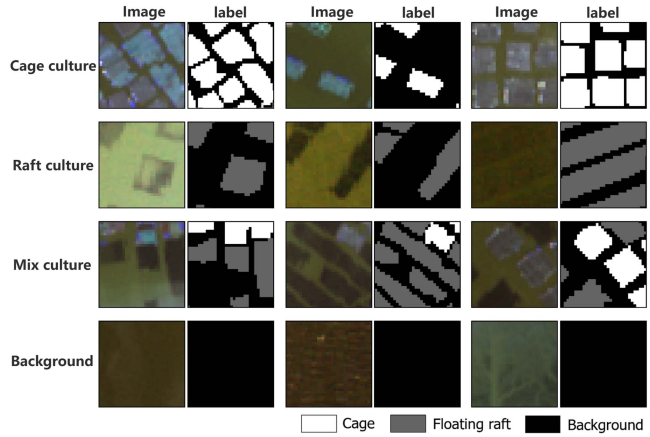

Fig. 3. Examples of true-color image patches and the corresponding white-gray-black labels of cage, floating raft, and background.

summer, we narrowed the image dates within February to April. With the filtering condition of "cloud coverage $\leq 10\%$", the S2 images on March 24, 2021 for the Fujian study area, February 20, 2021 for the Jiangsu study area, and February 7, 2021 for the Liaoning study area were preprocessed and downloaded using Google Earth Engine platform.

To determine which bands or indexes are needed to input CNN for mapping CRA, we delineated spectral/index feature curves (see Fig. 2) of three main categories, cage, floating raft, and background within the study areas, using the mean values of the samples. The considered indexes are the normalized difference water index (NDWI) and the normalized difference vegetation index (NDVI), since the former is capable of identifying water body and the latter is useful for discriminating certain rafts cultivating aquatic plants. They can be calculated by (1) and (2), respectively

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR}}{\mathrm{Green} + \mathrm{NIR}} \tag{1}$$

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}}. \tag{2}$$

Combining the visual interpretation of Fig. 2 and the measuring results of feature importance in [16], we applied bands 2, 3, 4, 8, 11, and 12, as well as NDVI and NDWI for extracting CRA. Due to the unique structure and spatial arrangement of cages and floating rafts, the texture feature of local binary patterns (LBP) [41] was obtained based on the spectral bands 2, 3, and 4. We designed five combination schemas of the above-mentioned bands and indexes for tests as in Table III, and plan to choose the best performance schema for CRA mapping.

C. Sample Dataset

With the aid of HSR remote sensing images in Google Earth, we manually labeled cage and raft samples from the background in the ArcGIS Pro platform. The samples were considerably created in terms of the spatial distribution, the representativeness of shape, size, color, edge clarity of targets, and the diversity of background (see Fig. 3). The locations of labeled samples are illustrated in Fig. 1. Considering the complex marine environment due to waves, seawater color, sediment concentration, etc., a certain number of negative samples excluding CRA were also created. We also took into consideration the diversity and geographical variability of the background when selecting negative samples. The inclusion of negative samples helps the model avoid misidentifying seawater bodies with similar spectral features as CRA and accurately delineates CRA boundaries by understanding background characteristics. Samples are transformed into maps where the values of two, one, and zero represent cage, raft, and background, respectively. These representative samples are prepared for training and validating the proposed model.

By comparing the model accuracies among four patch sizes of 32 pixels $\times$ 32 pixels, 64 pixels $\times$ 64 pixels, 128 pixels $\times$ 128 pixels, and 256 pixels $\times$ 256 pixels, we determined to crop the images, as well as the labels into 32 pixels $\times$ 32 pixels patches. The details of our sample dataset are show in Table I. The training and validating sets contain a total of 635 patches, including 100 background patches. These patches were randomly divided into training and validation sets in the ratio of 7:3. In addition, there are 304 labeled patch images (in pink boxes of Fig. 1) purposes for testing.

III. Methodology

A. Attention-Fused Deep Learning Model

The main idea of this study is to combine the advantages of CNN and attention module to construct a fine-performance deep learning model for accurate detection of large-scale CRA. Therefore, we considerably design the architecture of AF-UNet++ model by flexibly embedding CBAM modules into UNet++ baseline. The proposed model is detailed in this section.

1) Baseline: U-Net is a simple U-shaped CNN with a clear principle [42] and yields outstanding results on image semantic segmentation [35], [37], [43]. Its architecture is composed of

TABLE I
NUMBERS OF LABEL OBJECTS AND IMAGE PATCHES IN SAMPLE DATASET

Province	*Training & Validation Sets*				*Testing Sets*			
	Number of Label Objects		*Number of Image Patches*		*Number of Label Objects*		*Number of Image Patches*	
	raft	*cage*	*cage & raft*	*background*	*raft*	*cage*	*cage & raft*	*background*
Liaoning	510	8	159	30	570	62	142	0
Jiangsu	404	0	123	15	287	0	82	0
Fujian	1157	781	253	55	291	135	80	0
TOTAL	**2071**	**789**	**635**		**1148**	**197**	**304**	

an encoder subnetwork and a decoder subnetwork. The encoder applies four blocks of convolution and down-sampling (pooling) layers to obtain multiscale feature maps, while the decoder uses four blocks of convolution and up-sampling layers to acquire refined feature maps. In order to retain more detailed information about the targets, the feature maps in every block from the encoder are copied and concatenated to the corresponding up-sampling layers in the decoder. Since its proposal, U-Net has been successfully adopted in numerous studies.

The emergence of U-Net's modified versions boosts its applications in the segmentation of medical, natural, and remote sensing images. UNet++ is one of the successfully modified U-Net versions [44]. By replacing the plain, same-level concatenations with nested dense skip connections which brings smooth gradient descent for inner convolution blocks, UNet++ mitigates the semantic gap between the feature maps of the encoder and decoder prior to fusion. This multiscale feature fusion enhances the representation capability of the network and improves segmentation accuracy. Another major advantage of UNet++ is its deep supervision. Adopting the idea of adding loss functions (LFs) at each hidden layer, UNet++ enables feature learning at early levels, thus making it possible for users to check the performance of the network at the intermediate layers. Meanwhile, a few popular CNNs that use dilated convolutions, residual modules, and residual connections—like ResNet-18 [45] and DeepLabV3+ [46]—have also achieved exceptional semantic segmentation performance.

Since the fact that UNet++, comparing with some commonly used CNNs like U-Net and DeepLabV3+, performs more effectively on end-to-end image classification [29], [47], [48], we chose it as the baseline CNN.

2) Convolutional Block Attention Module: With the nested dense connection structure, UNet++ alleviates, to some extent, the fusion gap of features from the decoder and the encoder. However, its experimental results of detecting CRA demonstrate that it could not effectively detect the targets of blurred edge, mixed cage, and raft, or "weak" raft. This phenomenon may arise due to the incorporation of excessive redundant information by UNet++, making it challenging to capture the essential features of CRA. In order to highlight the distinctive CRA features for correctly detecting these types of targets, we introduced the widely used, lightweight CBAM into UNet++.

The CBAM consists of a channel attention submodule (CAM) followed by a spatial attention submodule (SAM), and sequentially infers attention maps in the channel dimension and spatial dimension (see Fig. 4). CAM concentrates on exploiting the interchannel relationship of the input image. CAM first incorporates two operations, namely global maximum pooling and global average pooling, to generate two adjusted feature vectors. These vectors are then fed into a shared network consisting of a multilayer perceptron with one hidden layer using shared weights. Subsequently, elementwise addition operations, followed by sigmoid activation, are performed to obtain the channel attention feature map. SAM focuses on acquiring the interspatial relationship of the input image. In SAM, max-pooling and average-pooling operations along the channel dimension are applied to obtain two spatial attentions. A spatial attention map is produced by concatenating and convolving two spatial attention using a conventional convolutional layer. Then a sigmoid function is applied to obtain the final spatial attention. By multiplying the attention maps outputted from CBAM to the input feature map of CNN, the model can reach adaptive feature refinement.

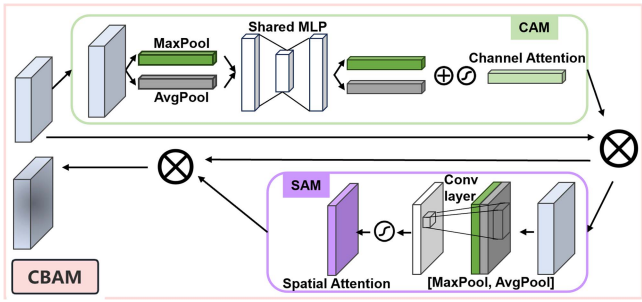

Fig. 4. Structure of CBAM [24].

3) Model Architecture: We integrated CBAM into UNet++ and named the improved model as AF-UNet++. The overall architecture of the AF-UNet++ is presented like web-link, inversed pyramid (see Fig. 5). Briefly, CBAM is induced into UNet++ in two ways: a) A CBAM is added into each of the four blocks in the encoder, located after two consecutive convolutional layers (Consecutive_Conv in Fig. 5). In other words, each block of the encoder in the proposed model, which is named Conv_CBAM, consists of two convolutions and one CBAM. By locating in this way, CBAM facilitates the refinement of previously extracted features, thereby enhancing the spatial representation of features and the discriminative capability of the block. b) A CBAM is added to each of Level 1 blocks in the decoder, located before two consecutive convolutional

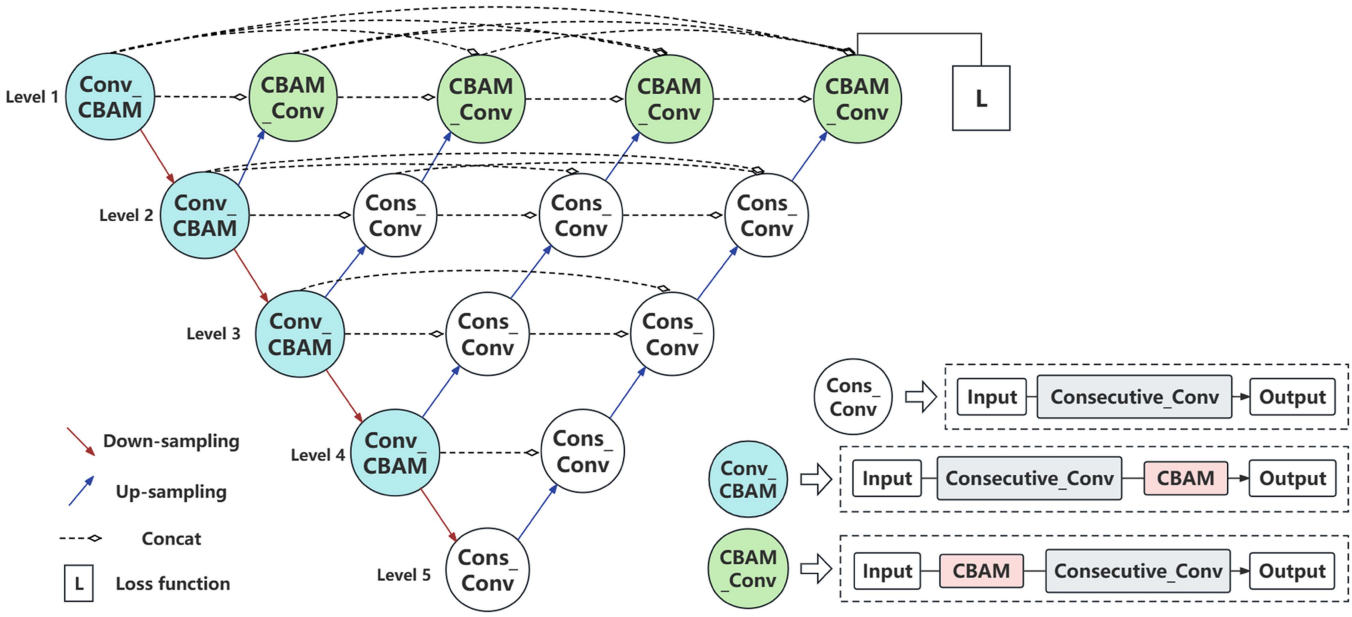

Fig. 5. Architecture of AF-UNet++ with CBAM.

layers. This kind of block is now called CBAM_Conv, which is composed of one CBAM and two convolutions. By embedding each block in this manner, CBAM acts as a feature recalibrator and helps to improve the quality of the features extracted by the subsequent convolutional layer.

The CBAM in each block of encoder improves the model's ability to assign important weights in both spatial and channel dimensions to the targets of interest at all levels. Meanwhile, in the decoder feature fusion process, we have features with multiple levels and channels. For ease of reference in subsequent descriptions, we designate the layers of the model from top to bottom as Level 1 to 5 (see Fig. 5). The CBAM in each Level 1 block of decoder further helps the model better allocate weights to these heterogeneous features, enabling more efficient subsequent feature learning. There are several motivations behind only combining CBAM into the encoder as well as the Level 1 blocks of the decoder. First, this structure is clear and simple to describe. Second, it facilitates model transfer learning by just unfreezing the blocks of Level 1 layer containing CBAM. Third, it keeps the balance between accuracy level and resource load.

4) Loss Function: LF is applied to measure the fitting degree between the predicted results from CNNs and the ground trues. It is also the key for optimizing CNNs. One of the most commonly used LFs is cross-entropy LF (L_{CE}), it can be computed by as

$$L_{\mathrm{CE}} = -\sum_{c=1}^{M} y_c \log\left(p_c\right) \tag{3}$$

where M, y_c, and p_c represent the number of classes, the class values of labels, and the probability that the samples are assigned to class c, respectively.

However, L_{CE} is not competent for optimizing CNNs when the pixel ratio of targets is significantly less than that of the background. Thus, dice loss (L_D) [49] is developed to handle this issue. It can be calculated as

$$L_D = \frac{2 \times \mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TP} + \mathrm{FP}} \tag{4}$$

where TP, TN, FP, and FN indicate the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Meanwhile, since detecting edges of CRA is important, we also applied boundary loss (L_B) [50] in (5) to assess and improve our proposed model

$$L_B = 1 - \frac{2P_cR_c}{P_c + R_c} \tag{5}$$

where P_c and R_c represents precision and recall for class c, respectively.

Therefore, to aim at obtaining CRA more precisely, we used the combination of L_{CE}, L_D, and L_B as final LF in (6) for AF-UNet++

$$L = L_{\mathrm{CE}} + L_D + L_B. \tag{6}$$

B. Evaluation Metrics

The evaluation metrics of overall accuracy (OA) in (7), F1 score in (10), and intersection over union (IoU) in (11) are used to validate the performance of our proposed model, as well as to compare the accuracies of AF-UNet++ with the baselines

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{10}$$

TABLE II
NETWORK PARAMETERS OF AF-UNET++ AND BASELINES

Network Parameters	*Optimal Parameters*
Optimizer	SGD
Momentum	0.9
Weight decay	0.0001
Activation	ReLU
Initial learning rate	0.001
Epoch	200
Batch Size	4

TABLE III
COMPARISON OF AF-UNET++ ACCURACIES WITH DIFFERENT COMBINATION SCHEMAS OF FEATURES

Schema		*OA (%)*	*IoU (%)*		
			cage	*raft*	*background*
1	Bands 2,3,4	88.04	55.43	56.22	84.51
2	Bands 2,3,4,8	89.25	69.57	59.01	85.72
3	Bands 2,3,4,8, 11,12	90.88	70.42	65.47	87.42
4	Bands 2, 3,4,8,11,12, NDWI,NDVI	91.72	73.04	72.92	88.67
5	Bands 2, 3,4,8,11,12, **NDWI, NDVI, LBP texture**	**91.32**	**80.01**	**74.12**	**88.37**

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}. \tag{11}$$

In addition, the commonly used frames per second (FPS) is adopted to compare the computational speeds of different fusion structures of UNet++ with CBAM. A higher FPS value indicates a higher efficiency.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The experiments of AF-UNet++ and the mainstream CNN models were conducted on the server with an NVIDIA GeForce RTX 2050 GPU. The parameters of the proposed model and other CNNs for training are shown in Table II.

B. Selecting Features for Deep Learning

We design five schemas of the preselected six bands and two indexes, along with the inclusion of LBP texture for bands 2, 3, and 4, as part of our training and validation process for the proposed model. The experimental results (see Table III) show that schema 5 achieves the highest IoUs, indicating that the integration of indexes and texture features with spectral information effectively improves the segmentation IoUs of our model. This improvement can be attributed to the fact that these indexes, along with LBP texture, capture the distinguishing characteristics between CRA and the background significantly. As a result, we have utilized this schema for further analysis in our study.

C. Evaluation of the Fusion Ways of CBAM and UNet++

In this section, we design an experiment to assess the effect of the fusion ways for CBAM and UNet++ and, thus, validate that our proposed model can balance between efficiency and performance. Therefore, we focus on whether the number and location of CBAM modules fused into the baseline influence the models' performance.

We design five different fusing architectures (see Fig. 6) as follows:

a) CBAMs added to the decoder convolutional blocks at Level 1;
b) CBAMs added to the convolutional blocks in the encoder;
c) CBAMs added to the convolutional blocks at Level 1 and the first column;
d) CBAMs added to the convolutional blocks at Levels 1 and 2, and in the encoder;
e) CBAMs added to all convolutional blocks.

It is worth mentioning that (c) represents our final proposed model, while the architecture (e) is the resemblance to that in [51].

The experimental results are presented in Table IV. The table illustrates that as the number of CBAM modules increases, both the number of parameters and the model's architecture complexity increase. This, in turn, leads to a decrease in the computational speed of the model. Based on the results of (a) and (b), it is evident that incorporating CBAMs with decoder blocks of UNet++ at Level 1 enhances the model's ability for extracting cages, while adding CBAMs in the encoder improves the recognition of rafts. This could be attributed to the likelihood of rafts being confusable with the background, demanding the assignment of greater significance for their accurate identification. Consequently, the early integration of CBAM in the encoder modules plays a pivotal role in enhancing the recognition of rafts. Comparing the OAs of (c), (d), and (e), it is evident that simply increasing the number of CBAM modules does not lead to significant improvements in the accuracies of the models but adversely affects the practicality of the models. Compared with the other four architectures, (c) demonstrates the best OA, with the highest IoUs for raft and background. It effectively balances the recognition of raft and cage, which aligns well with the practical requirements of the three-class classification problem. This observation indicates that the significant improvement in our model's performance is not solely due to the increase in parameters, but rather the result of architectural enhancements. Therefore, we have determined the adoption of the structure (c) as the final proposed model for addressing the CRA extraction task.

D. Comparisons of CRA Extraction Results

To quantitatively evaluate the performance of our proposed model, we used the same dataset and network parameters to train it along with four state-of-the-art CNNs namely U-Net, UNet++, ResNet-18, and DeepLabV3+. The five trained models were evaluated with testing sets, and the quantitative comparisons of their results are listed in Table V. It can be seen that AF-UNet++ has the highest OA, F1, and mIoU compared

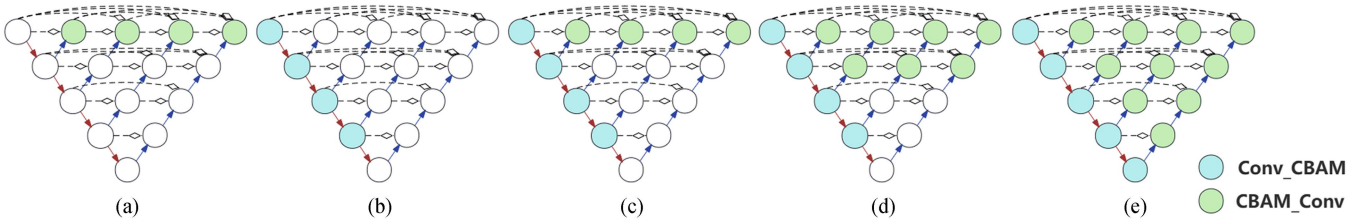

Fig. 6. Five architectures of fusing CBAM with UNet++. (a) Adding CBAMs in decoder blocks at Level 1. (b) Adding CBAMs in encoder blocks. (c) Architecture of our proposed AF-UNet++. (d) Adding CBAMs in encoder and decoder blocks at Level 1&2. (e) Resemblance to the architecture in [51].

TABLE IV
QUANTITATIVE COMPARISONS OF THE ACCURACIES, NUMBER OF PARAMETERS, AND COMPUTATIONAL SPEEDS BETWEEN OUR PROPOSED MODEL(C) AND THE OTHER FUSION ARCHITECTURES OF CBAM WITH UNET++

Architecture	*Number of parameters(M)*	*FPS*	*OA (%)*	*F1 Score*	*IoU (%)*		
					cage	raft	background
a	51.43	96 149.41	89.80	86.24	74.91	63.29	87.61
b	51.44	98 214.84	90.18	86.00	71.94	65.74	87.84
c(our model)	**51.47**	**91 560.34**	**90.83**	**86.67**	**71.88**	**69.38**	**88.49**
d	51.52	81 214.18	89.90	85.36	70.27	64.58	87.78
e	51.86	79 189.10	90.46	86.84	75.2	67.02	88.15

TABLE V
QUANTITATIVE COMPARISONS OF THE ACCURACIES BETWEEN OUR PROPOSED MODEL AND OTHER MAINSTREAM MODELS

Model	*OA (%)*	*F1 Score*	*IoU (%)*		
			cage	raft	background
ResNet-18	83.68	0.76	44.79	60.97	79.84
DeepLabV3+	89.62	0.85	65.26	74.71	86.33
U-Net	89.75	0.87	75.86	69.33	86.52
UNet++	90.78	0.87	72.09	72.92	88.38
AF-UNet++	**91.32**	**0.89**	**80.01**	**74.12**	**88.37**

with the other state-of-the-art models. Its OA reaches 91.32% with 7.64%, 1.7%, 1.57%, and 0.54% over those of other CNNs, and its F1 attains 0.89, which is 0.2 higher than those of others. Moreover, our model demonstrates a remarkable advantage in terms of IoU of the cage type, where our model achieved an improvement of 4.15% compared to other mainstream models. These findings show that the proposed model surpasses the commonly used models from the perspective of quantitative comparison.

Fig. 7 presents the visual comparison of the CRA extraction results among the proposed model and two effective baseline models from the UNet family. In the first row, U-Net produces many false identifications (red boxes) and UNet++ decreases the number of false classification pixels, while the proposed model almost perfectly detects raft objects. In the second row, both U-Net and UNet++ incorrectly interpret considerable island pixels (red boxes), while our model can interpret these pixels correctly. In the third row, U-Net detects objects with incomplete shapes and UNet++ with erroneous adhesion of

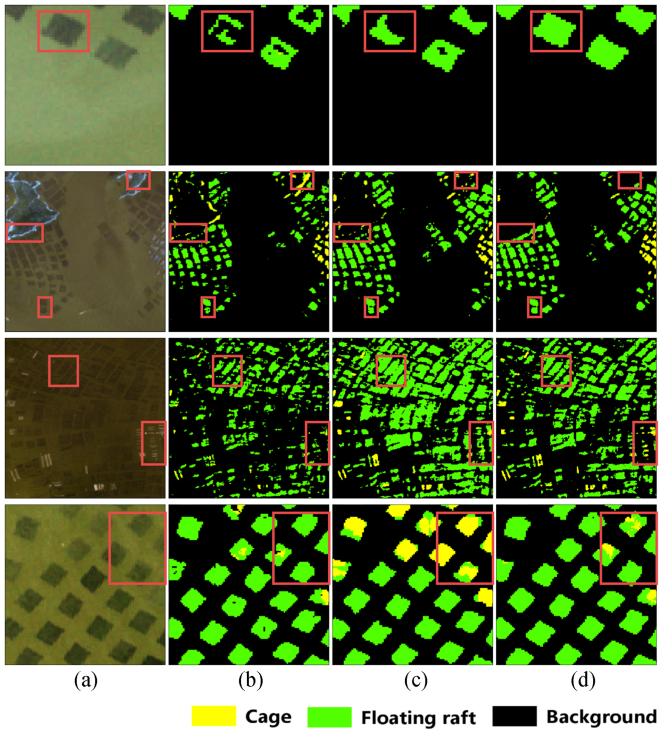

Fig. 7. Visual comparison of CRA extraction results. (a) True color S2 images. (b), (c), and (d) Results of U-Net, UNet++, and AF-UNet++, respectively.

nearby CRA areas (red boxes), while AF-UNet++ best characterizes CRA shape and structure. Besides, U-Net fails to recognize slender cages and U-Net++ incorrectly classifies gaps as rafts (red boxes), while only AF-UNet++ partially captures cage shapes and gaps. In the fourth row, though the three models

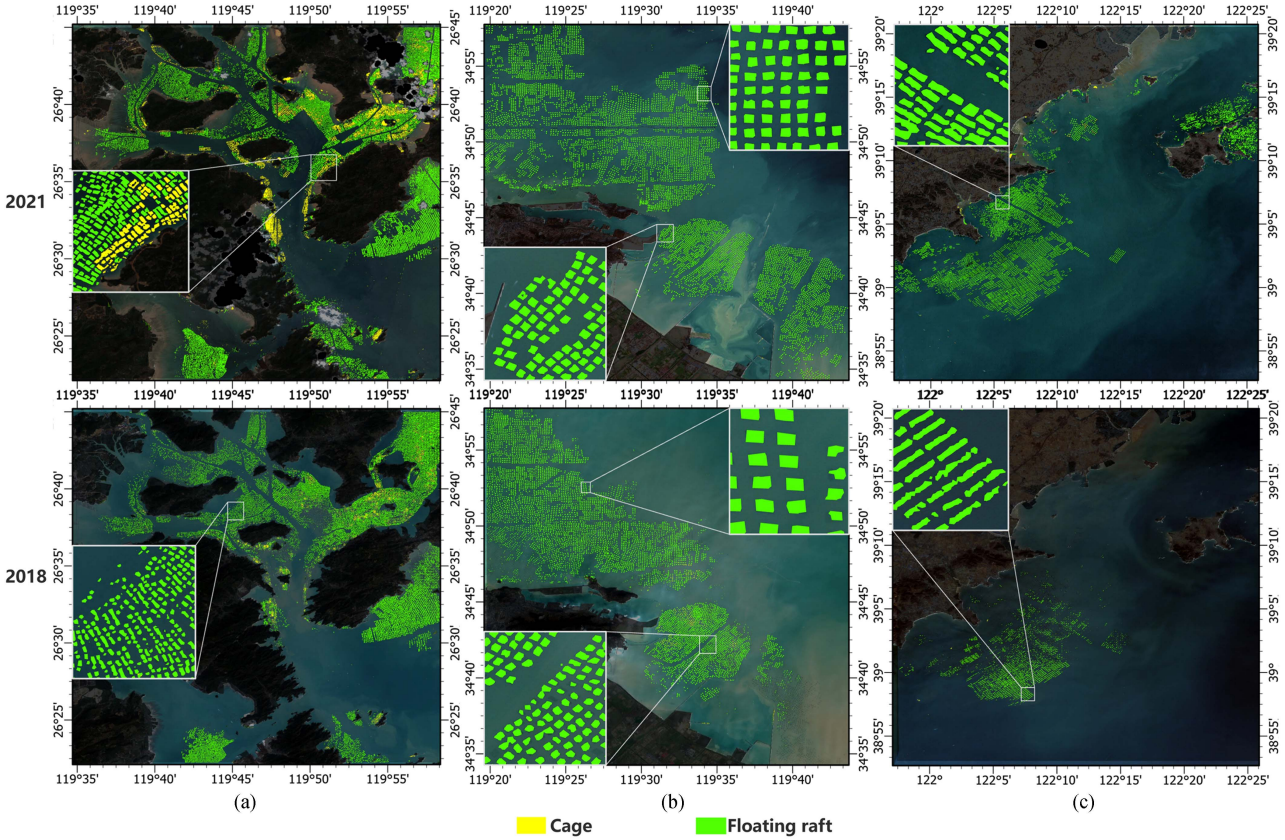

Fig. 8. Using AF-UNet++ for large-scale mapping in the study areas in 2021 and 2018. (a) Fujian. (b) Jiangsu. (c) Liaoning.

exhibit the difficulty of recognizing "weak" rafts (red boxes), our model outstands the boundary of objects best.

In summary, both the quantitative and visual comparisons demonstrate that, comparing some state-of-the-art CNN models, AF-UNet++ best excels in extracting CRA.

E. Temporal-Spatial CRA Monitoring

We adopted the trained AF-UNet++ model with the best performance to automatically detect CRA in the three study areas from S2 data in 2021, the results are exhibited in Fig. 8. Visually interpreting the results illustrates that cages and rafts are precisely delineated, and shorelines and open water are correctly excluded. It proves that our proposed model has the capacity for mapping large-scale CRA.

We directly applied the best-performing model trained on the 2021 dataset to extract CRA in the three study areas from S2 images taken in 2018. The detection results are shown in Fig. 8. Carefully visual interpretation of the 2018 results tells that, though it makes some omissions or errors for identifying ambiguous and small rafts, as well as objects in cage and raft mixed zones, the proposed model can identify most cage and raft objects and guarantees to effectively monitoring CRA using remote sensing images from different time periods.

In order to justify the spatial stability of our model for the identification of CRA, we extended our validation efforts to four distinct sites (in purple boxes, Fig. 1) beyond the scope of the three study areas. These additional sites, located in Shandong Province, Zhejiang Province, and Fujian Province, were carefully selected to encompass diverse seawater background for further evaluating the generalizability of our model. The trained model was directly employed to extract CRA in these four sites. The experimental results (see Fig. 9) show that, though there are some misclassification and omission CRAs with "weak" or "blur" edges, the model still can successfully identify most CRAs. The conclusion that our proposal can transform to other regions can be made.

F. Discussions

Though the proposed model can map CRA in the different years and regions well, some improvement works still need to be done. First, the model's performances of extracting cages in the three study areas are different, the cages in the Fujian study area seem to be easily misclassified as rafts than those in the other two study areas [see Fig. 10(a)]. It implies that more cage samples should be constructed in the Fujian study area. Second, the identification results in 2018 show there are some floating rafts that were misrecognized as cages due to the significant difference in these raft features between 2021 and 2018. The model, to some content, failed to capture the intrinsic features of the objects that remain consistent over time. It indicates that sample dataset in different years need to be added to ensure more precisely monitoring CRA [see Fig. 10(b)]. Nevertheless, raft objects in those sites with no samples tend to be relatively easier misrecognized than those with samples, especially some weak

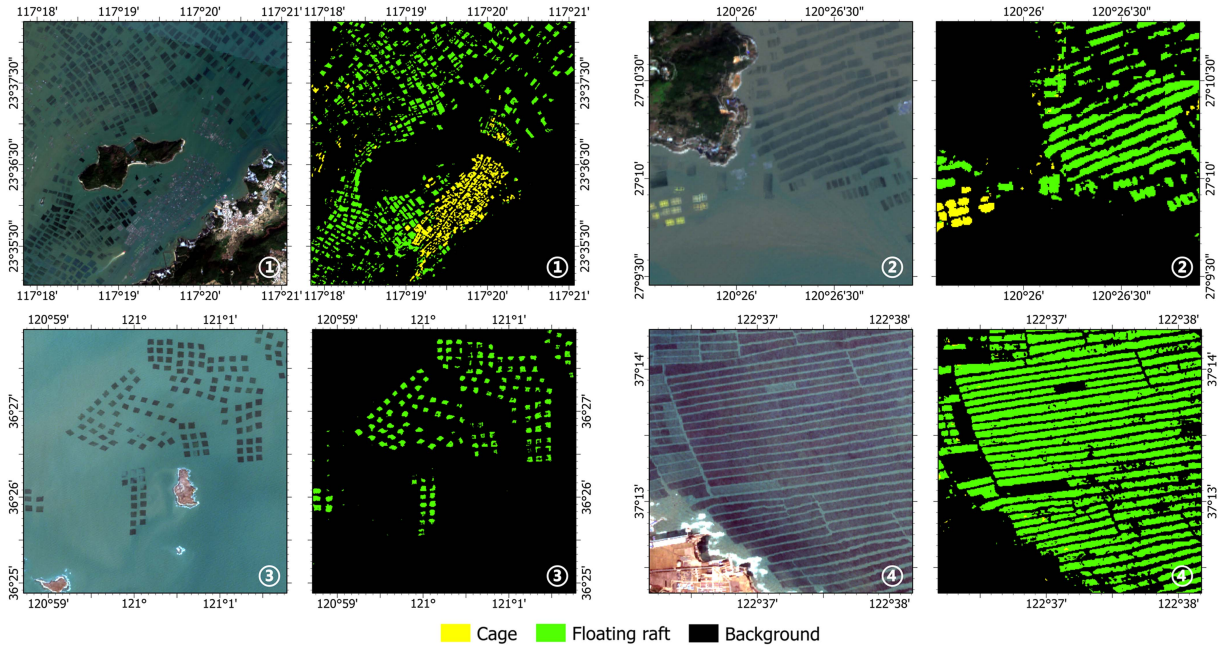

Fig. 9. Extraction results from AF-UNet++ in the four selected testing sites (purple boxes in Fig. 1).

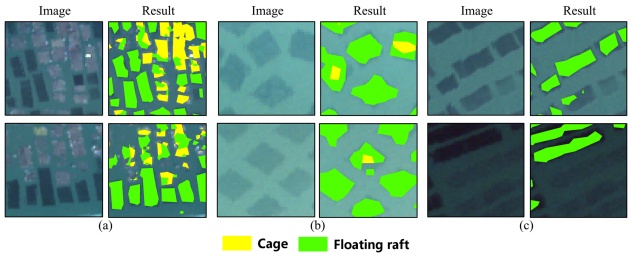

Fig. 10. Details of CRA extraction results in 2018 from AF-UNet++.

objects [see Fig. 10(c)]. This suggests that a sample dataset with a more reasonable spatial distribution is essential to be updated for more accurate monitoring of CRA in the future.

Other future works include following.

1) Construct more reasonable sample dataset in both spatial and temporal dimensions which covers the entire offshore area in our country and multiyears.
2) Given the rapid advancements in semantic segmentation methods for remote sensing imagery, we aim to explore more advanced models, such as the transformer and its variants [52], [53], [54], which have gained popularity in recent times. Then apply it to map spatial-temporal distribution of large-scale CRA in the whole coastal seawater region, China.
3) Discover a model that is suitable for cross-data-source applications, such as validating whether AF-UNet++ can monitor CRA using Landsat data.

V. Conclusion

The complexity of both CRA and its marine environment makes monitoring large-scale CRA still face challenges. Focusing on the challenges, we proposed the attention-fused deep learning model named AF-UNet++ by deliberately inducing a channel-spatial attention module called CBAM into the well-performed baseline UNet++. The proposed model was trained and validated by our self-constructed sample dataset in 2021, and adopted to map the large-scale CRA in the three selected areas in both 2021 and 2018, as well as the four selected sites out of the scope of aforementioned areas. Moreover, comparison experiments on AF-UNet++ and other mainstream models were conducted. The results demonstrate that our model detects CRA more accurately than the others and can accurately monitor large-scale CRA. Yet further works, including building sufficient and more reasonable spatial-temporal-distributed sample dataset, monitoring CRA in the entire offshore sea of China in recent decades with the aid of Landsat imagery, etc., still need to be carried out.

References

[1] Food and Agriculture Organization of the United Nations (FAO), *The State of World Fisheries and Aquaculture 2022: Towards blue Transformation*. Rome, Italy: FAO, 2022.
[2] China Statistical Bureau, *China Statistical Yearbook*. Beijing, China: China Statistical Bureau, 2022.
[3] A. Rico and P. J. Van den Brink, "Probabilistic risk assessment of veterinary medicines applied to four major aquaculture species produced in Asia," *Sci. Total Environ.*, vol. 468, pp. 630–641, 2014.
[4] Y. Zheng, R. Jin, X. Zhang, and J. Chen, "The considerable environmental benefits of seaweed aquaculture in China," *Stochastic Environ. Res. Risk Assessment*, vol. 33, no. 5, pp. 1203–1221, 2019.
[5] Q. Gao, Y. Li, Z. Qi, F. Du, and H. Zhao, "Diverse and abundant antibiotic resistance genes from mariculture sites of China's coastline," *Sci. Total Environ.*, vol. 630, pp. 117–125, 2018.
[6] M. Ottinger, K. Clauss, and C. Kuenzer, "Aquaculture: Relevance, distribution, impacts and spatial assessments - A review," *Ocean Coastal Manage.*, vol. 119, pp. 244–266, 2016.
[7] C. Liu, T. Jiang, Z. Zhang, J. Chen, and J. Yang, "Extraction method of offshore mariculture area under weak signal based on multisource feature fusion," *J. Mar. Sci. Eng.*, vol. 8, no. 2, 2020, Art. no. 99.

[8] J. Fan, H. Huang, and H. Fan, "Extracting aquaculture area with RADASAT-1," *Mar. Sci.-Qingdao-Chin. Ed.*, vol. 29, no. 10, 2005, Art. no. 40.

[9] M. Wang, G. Li, Y. Liu, and J. Wang, "Dynamic changes of mariculture areas in eastern Shandong Peninsula in recent 20 years," *J. Appl. Oceanogr.*, vol. 36, no. 2, pp. 319–326, 2017.

[10] Y. Wu, F. Chen, Y. Ma, and R. Li, "Research on automatic extraction method for coastal aquaculture area using Landsat8 data," *Remote Sens. Land Resour.*, vol. 30, no. 3, pp. 96–105, 2018.

[11] J. Geng, J. Fan, and H. Wang, "Weighted fusion-based representation classifiers for marine floating raft detection of SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 444–448, Mar. 2017.

[12] P. Liu and Y. Du, "A CBR approach for extracting coastal aquaculture area," *Remote Sens. Technol. Appl.*, vol. 27, no. 6, pp. 857–864, 2013.

[13] J. Chu, D. Zhao, and F. Zhang, "Wakame raft interpretation method of remote sensing based on association rules," *Remote Sens. Technol. Appl.*, vol. 27, no. 6, pp. 941–946, 2013.

[14] M. Wang, Q. Cui, J. Wang, and F. Ma, "Raft cultivation area extraction from high resolution remote sensing imagery by fusing multi-scale region-line primitive association features," *ISPRS J. Photogram. Remote Sens.*, vol. 123, pp. 104–113, 2017.

[15] J. Kang, L. Sui, X. Yang, and Z. Li, "Sea surface-visible aquaculture spatial-temporal distribution remote sensing: A case study in Liaoning province, China from 2000 to 2018," *Sustainability*, vol. 11, no. 24, 2019, Art. no. 7186.

[16] Y. Xu and L. Lu, "Spatiotemporal distribution of cage and raft aquaculture in China's offshore waters using object-oriented random forest classifier," in *Proc. 10th Int. Conf. Agro-Geoinformat.*, 2022, pp. 1–6.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[19] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high-resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, 2019, Art. no. 1382.

[20] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 190, pp. 196–214, 2022.

[21] Y. Lu, W. Shao, and J. Sun, "Extraction of Offshore Aquaculture areas from medium-resolution remote sensing images based on deep learning," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3854.

[22] J. Yu et al., "Coastal aquaculture extraction using GF-3 fully polarimetric SAR imagery: A framework integrating UNet++ with marker-controlled watershed segmentation," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2246.

[23] B. Cui, Y. Zhao, M. Yang, L. Huang, and Y. Lu, "Reverse attention dual-stream network for extracting laver aquaculture areas from GF-1 remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5271–5283, 2023.

[24] G. Qin, S. Wang, F. Wang, Y. Zhou, Z. Wang, and W. Zou, "U_EFF_NET: High-precision segmentation of offshore farms from high-resolution SAR remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8519–8528, 2022.

[25] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[26] J. Deng, Y. Bai, Z. Chen, and Y. Wang, "A convolutional neural network for coastal aquaculture extraction from high-resolution remote sensing imagery," *Sustainability*, vol. 15, no. 6, 2023, Art. no. 5332.

[27] X. Wang, J. Zhou, and J. Fan, "Idudl: Incremental double unsupervised deep learning model for marine aquaculture SAR images segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4209412.

[28] Y. Ma, X. Qu, C. Yu, and M. Gong, "Automatic extraction of marine aquaculture zones from optical satellite images by R3Det with piecewise linear stretching," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4430.

[29] Y. Chen, G. He, R. Yin, and C. Yang, "Comparative study of marine ranching recognition in multi-temporal high-resolution remote sensing images based on deeplab-v3+ and U-Net," *Remote Sens.*, vol. 14, no. 22, 2022, Art. no. 5654.

[30] C. Liu, T. Jiang, Z. Zhang, J. Chen, and J. Yang, "Extraction method of offshore mariculture area under weak signal based on multisource feature fusion," *J. Mar. Sci. Eng.*, vol. 8, no. 2, 2020, Art. no. 99.

[31] Z. Jiang and Y. Ma, "Accurate extraction of offshore raft aquaculture areas based on a 3D-CNN model," *Int. J. Remote Sens.*, vol. 41, no. 14, pp. 5457–5481, 2020.

[32] Y. Liu, X. Yang, Z. Wang, and Z. Chen, "Aquaculture area extraction and vulnerability assessment in Sanduao based on richer convolutional features network model," *J. Oceanol. Limnol.*, vol. 37, no. 6, pp. 1941–1954, 2019.

[33] C. Yu, Z. Hu, R. Li, and M. Gong, "Segmentation and density statistics of mariculture cages from remote sensing images using Mask R-CNN," *Inf. Process. Agriculture*, vol. 9, no. 3, pp. 417–430, 2022.

[34] C. Yu, Y. Liu, X. Xia, and M. Gong, "Precise segmentation of remote sensing cage images based on SegNet and voting mechanism," *Appl. Eng. Agriculture*, vol. 38, no. 3, pp. 573–581, 2022.

[35] B. Cui, D. Fei, G. Shao, and W. Ke, "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.

[36] W. Han, J. Li, S. Wang, and L. Zhang, "A context-scale-aware detector and a new benchmark for remote sensing small weak object detection in unmanned aerial vehicle images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102966.

[37] H. Su, S. Wei, J. Qiu, and H. Huang, "RaftNet: A new deep neural network for coastal raft aquaculture extraction from Landsat 8 OLI data," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4587.

[38] T. Shi, Q. Xu, Z. Zou, and J. Chen, "Automatic raft labeling for remote sensing images via dual-scale homogeneous convolutional neural network," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1130.

[39] Y. Fu et al., "A new satellite-derived dataset for marine aquaculture areas in China's coastal region," *Earth Syst. Sci. Data*, vol. 13, no. 5, pp. 1829–1842, 2021.

[40] X. Liu et al., "Mapping China's offshore mariculture based on dense time-series optical and radar data," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1326–1349, 2022.

[41] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, 1994, pp. 582–585.

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

[43] N. Zang, Y. Cao, Y. Wang, B. Huang, L. Zhang, and J. Chen, "Land-use mapping for high-spatial resolution remote sensing image via deep learning: A review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8159–8178, 2021.

[44] 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. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2019.

[45] 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.

[46] L. C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[47] A. Aizatin and I. G. B. B. Nugraha, "Comparison of semantic segmentation deep learning methods for building extraction," in *Proc. Int. Conf. Comput. Eng., Netw., Intell. Multimedia*, 2022, pp. 1–5.

[48] F. Barrientos-Espillco et al., "Semantic segmentation based on deep learning for the detection of cyanobacterial harmful algal blooms (CyanoHABs) using synthetic images," *Appl. Soft Comput.*, vol. 31, 2023, Art. no. 110315.

[49] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[50] A. Bokhovkin and E. Burnaev, "Boundary loss for remote sensing imagery semantic segmentation," in *Proc. Adv. Neural Netw.*, 2019, pp. 388–401.

[51] Z. Zhao, K. Chen, and S. Yamane, "CBAM-Unet++: Easier to find the target with the attention module ' CBAM'," in *Proc. IEEE 10th Glob. Conf. Consum. Electron.*, pp. 655–657, 2021.

[52] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, Art. no. 30.

[53] Z. Xu et al., "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3585.

[54] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.