Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2054877, 17 pages
https://doi.org/10.1155/2022/2054877

Research Article

Extraction and Classification of the Supervised Coastal Objects Based on HSRIs and a Novel Lightweight Fully Connected Spatial Dropout Network

Yan Chen,[1] Jiahua Wan,[2] Yantao Xi,[3] Wenxiang Jiang,[1] Mengyuan Wang,[1] and Menglei Kang[1]

[1] *School of Artificial Intelligence and big Data, Hefei University, Hefei 230601, China*
[2] *School of big Data and Artificial Intelligence, Anhui Xinhua University, Hefei 230088, China*
[3] *School of Resources and Geosciences, China University of Mining and Technology, Xuzhou 221006, China*

Correspondence should be addressed to Yan Chen; chenyan090501@126.com and Jiahua Wan; jiahwan@163.com

Received 2 July 2022; Accepted 20 July 2022; Published 12 August 2022

Academic Editor: Chia-Huei Wu

For the protection and management of coastal ecosystems, it is crucial to monitor typical coastal objects and examine their characteristics of spatial and temporal variation. There are limitations to the conventional object-oriented and spectrum-based approaches to HSRIs interpretation. The majority of recently conducted studies on semantic segmentation based on DCNNs concentrate on improving the accuracy of single objects at local scales. The completeness, generalization, and edge accuracy of the extraction and classification of multiple objects with the complex background at regional scales still need to be improved. We created a benchmark dataset CSRSD for coastal supervision using HSRIs and GIS in this study to address the aforementioned problems. In the meantime, by combining the traditional U-Net and DeepLabV3+ feature fusion strategies, we propose a novel fully connected fusion pattern by switching to deepwise separable convolution from conventional convolution and introducing spatial dropout to create a brand new CBS module. The LFCSDN, a new lightweight fully connected spatial dropout network, has been suggested. The findings demonstrate that our constructed semantic segmentation dataset, which has produced reliable results on U-Net and DeepLabV3+, can be used as a benchmark for applications based on DCNNs for coastal scenes. While maintaining high accuracy, LFCSDN can significantly reduce the number of parameters. Our suggested CBS module can increase the model's generalization by reducing overfitting. In order to analyze the spatiotemporal characteristics of target changes in the study area, tests on expansive remote sensing imagery were also conducted. The findings can be applied to ecological restoration, coastal area mapping, and integrated management. Additionally, it serves as a resource for studies on multiscale semantic segmentation in computer vision.

1. Introduction

The coastal area is a transition zone between land and sea, consisting of mudflats, swamps, and mangroves, which has the functions of wind and wave prevention, shore protection, water conservation, climate regulation, and prevention of seawater invasion and maintenance of biodiversity. Due to urbanization, mariculture, and extreme weather, ecological environment issues have existed in certain developing countries' coastal areas. Monitoring the central ecology and mariculture objects and gaining information on their spatial and temporal changes are crucial for coastal resource protection and utilization. They also provide decision-making support for coastal urban management and planning. However, due to the wide coverage of the coastal area and the complexity of landscape and feature categories, the traditional means based on an on-site survey is time-consuming. A variety of automatic technologies used for remote sensing interpretation can rapidly capture the characteristics of spatiotemporal distribution of the objects on a large scale, which has been a principal means of ecological environment monitoring in coastal areas. Optical and radar images with medium and coarse spatial resolution have been widely used. With the rapid development of remote sensing technologies, the high spatial resolution images (HSRIs) containing richer features gradually increase, and the costs have been reduced. Researchers have implemented numerous approaches to extract and classify the monitored objects from HSRIs in the coastal areas, such as spectrum-based classification [1, 2] and object-oriented image segmentation [3–5]. However, the elaborate spatial details require adequate spectral responses, but the spectrum distinguishability of HSRIs is relatively insufficient, which would lead to inferior classification in the spectral domain. For example, the classified pixels might be occupied by an amount of salt-pepper noise. An object-oriented trick first performs superpixel segmentation that can fully use features such as shape, texture, and positional relation of the targets to classify the HSRIs more accurately. However, due to the lack of unified criteria for the superpixel segmentation, deviations might be exported into the classification stage. In addition, the two-stage process would restrict the overall efficiency.

In recent years, AlextNet [6], VGGNet [7], ResNet [8], etc., those deep learning-based deep convolutional neural networks (DCNNs) have achieved great success in the field of computer vision. To compare with the classic convolutional neural networks (CNNs) such as LeNet-4 and LeNet-5 [9], DCNNs introduce several innovative tricks such as ReLU, dropout, batch normalization, multiple max-pooling layers, GPUs-based training, etc., which support further extension on a network depth and improvement on the model training efficiency. Regarding remote sensing applications, DCNNs have been extensively studied and achieved significant results in HSRIs oriented scene classification [10–12] and object detection [13–16]. Likewise, the researchers have further explored the capability of DCNNs for achieving a pixel-wise prediction/semantic segmentation on HSRIs classification. Fully convolutional networks (FCNs) [17], a set of novel end-to-end semantic segmentation models that use convolutional layers to replace the fully connected layers of DCNNs-based image recognition are the most widely used for pixel-wise prediction in the remote sensing domain for HSRIs classification recently. Several improved networks based on FCNs, such as U-Net [18] and DeepLab series [19–21], demonstrate superior results in HSRIs. Although the DCNNs-based methodology is commonly employed for classifying and extracting targets in HSRIs, it mainly focuses on urban scene applications [22], such as building and road extraction, vehicle and aircraft detection, and land use/cover classification. Few pieces of research have been launched on the coastal scene. We made a brief overview to summarize the related work involving certain representational costal objects.

Sun et al. [23] apply the VHR optical images and LiDAR data to construct a three-stage mapping framework based on deep learning for tree species diversity assessment in tropical wetlands. They enhance three DCNNs (AlexNet, VGG-16, and ResNet-50) to better utilize spatial contextual information for tree species classification. The results show the effectiveness of the DCNNs-based solutions for mapping species diversity. Li et al. [24] propose a multilayer mangrove mapping method considering the upper and lower vegetation detection and species classification using the combined dataset of multispectral WorldView-3 data, airborne hyperspectral imagery, and LiDAR data. Random forest (RF) and support vector machine (SVM) are compared in the company with DCNNs. The results show that LiDAR-based RF and SVM can obtain a greater accuracy on the species mapping, while spectral features are more sensitive to DCNNs. Guo et al. [25] design an innovative DCNN, Capsule-U-Net, which couples the capsule networks [26] with U-Net to achieve high accuracy extraction of mangroves by learning the spatial locations of pixels between objects in images. Guo et al. [27] develop a multiscale context embedding module to extract multiscale contextual information and propose a deep learning-inspired pixel classification model, which obtains an improved mIoU on Sentinel-2A data. Wan et al. [28] adopt the VHR images, and fuse multiscale DCNNs for sapling detection. Diniz et al. [29] test the application of a U-Net classifier over the coastal extension in the BCZ to evaluate its robustness in multitemporal identification of the coastal artificial aquaculture ponds. Cui et al. [30] use multiple combined convolutional layers, pooling layers, and nonlinear ReLU activation functions to build a deep network to extract nonlinear and invariant high-level features of aquaculture raft areas. Liu et al. [31] propose a deep learning RCF (richer convolutional features) network to extract the aquaculture rafts in the bay from the high-resolution GF-2 images. The results show that the proposed method does need not to separate the areas from land and sea in advance, and it still can maintain good extraction in the area with more sediment in the water and more giant waves. The accuracy reaches more than 93%, which is appropriate for large-scale mariculture applications. Based on the GF-2 images, Zheng et al. [32] propose an improved double-branch network method for marine cage extraction. The model consists of densely connected blocks on the spatial encoding path and can quickly obtain global context information of objects using the global average pooling. Feature fusion is used to recover the spatial details to improve the extraction accuracy of the marine cage.

Currently, DCNNs-based research for HSRIs classification in a coastal scene generally focuses on the local scale and accuracy improvement of a certain single object. Studies on the multiple object extraction and spatiotemporal distribution analysis on a regional scale are comparatively scarce. In one respect, sparse multi-object datasets are inadequate for many applications. On the other hand, HSRIs with more spatial details might lead to additional misclassification caused by the spectrum-class inconsistency. A variety of nontarget classes in the background makes it difficult to obtain a generalized accuracy for the entire target objects, especially for certain objects possessing diverse sizes, shapes, and structures. Meanwhile, sharp edges and accurate completeness of the simple and multiscale objects deserve more concern. Considering the storage characteristics of HSRIs, a lightweight network could provide support to extend the practical deployment. Therefore, we take the mangrove, aquaculture raft, and pond that present a multiscale characteristic and concerned attention in a coastal region of China as target classes to create a benchmarked multi-object semantic segmentation dataset for the coastal monitored research and applications. To further refine the object edge and promote the accuracy improvement of multiscale objects, we reference the strategies of U-Net feature fusions that connect the low-level and high-level features and of DeepLab that connects multiple receptive fields to propose a parallel fully connected feature encoder fusing the various resolutions of feature maps. To minimize the model parameters, we construct several depthwise separable convolutions [33] instead of the standard convolutions. Consequently, a novel lightweight fully connected spatial dropout network (LFCSDN) with an innovative CBS module stacking the cascaded layers of feature maps, batch normalization layers, and spatial dropout [34] layers for moderating the overfitting and promoting the generalization is proposed. Additionally, we make a classification on a large scale extent and evaluate the spatiotemporal distribution characteristics of the monitored coastal targets in the study area based on the proposed network and the ground truth images with diverse time stamps. The dataset would be open-sourced later as needed, and our proposed methodology and research conclusions could be applied to the coastal environment's eco-environmental management. They might provide a broadening thought for the research of multiscale semantic segmentation in the domain of computer vision.

2. Study Area and Dataset

The study area, about 217 km^2, is located in the western bay of Fangchenggang City of Guangxi Zhuang Autonomous Region of China, consisting of Dong Bay, Xi Bay, and Gangkou District of the city shown in Figure 1. Fangchenggang is a coastal prefecture-level city. It is one of the 25 major coastal ports in China and the unique city of China that connects the ASEAN (Association of Southeast Asian Nations) by land and sea.

In recent years, due to illegal construction and sewage discharge by the factories, the mangroves (as shown in Figure 1) in the coastal areas of Guangxi are degraded to a certain degree, and the regional ecosystem has been disruptive to some extent [35]. In addition, a few fishermen and farmers have been fascinated by the benefits of illegally occupying the offshore areas to breed the fish or shellfish using aquaculture ponds or rafts (illustrated in Figure 1), which endangers voyage safety and the sustainable regular fishing ecosystem. Therefore, carrying out the research based on HSRIs and DCNNs-based intelligent extraction of the coastal target objects in the current area would provide significant support for the managers to supervise the illegal activities, which is meaningful to the city's sustainable development.

DCNN-based intelligent extraction or classification methods rely on several image samples. Besides paying a high price to purchase, collect, and process the raw images with several bands, it is an augmented and practical way to gain a batch of RGB images from the online map service such as Google Earth, Gaode Maps, and Baidu Maps. We collected four aerial HSRIs with a spatial resolution of 0.58 m (UTM projection) from the Google Earth service. The sizes of those images are all 25,856 × 25,344 by pixel, and their timestamps are 2003, 2007, 2015, and 2018, respectively, with a proximity month. In addition to the target objects of mangroves, aquaculture rafts, and aquaculture ponds, non-target backgrounds that consist of built-up areas, seawater, common plants, and bare areas have been included as well. These images have been labelled semi-automatically or manually based on ArcGIS to create a batch of ground truth samples.

Due to covering a large surface, a remote sensing image generally presents a large size that cannot be input in a DCNNs-based model directly with standard memory. Therefore, after comprehensively evaluating the target objects' scale characteristics, we choose a 512 × 512 pixel window to clip the large image into several small patches sequentially. The sliding window is set with a certain overlapping to ensure more accurate edges while recombining those patches. Ultimately, we created 12540 patches with the size of 512 × 512 by pixel. According to the machine learning criteria, a dataset is generally divided into a training set, a test set, and a validation set by the ratio of 6 : 2 : 2. We build the Coastal Supervision Remote Sensing Dataset (CSRSD), including 7524 patches as a training set and 2508 patches as a test set and a validation set, respectively. Currently, it does not seem to have an open-source multi-object dataset oriented to the monitored coastal scene, so we would open the CSRSD as needed shortly.

3. Methodology and Experiments

3.1. Lightweight Fully Connected Spatial Dropout Network. DCNNs-based semantic segmentation for remote sensing classification or extraction exists in two predominant patterns: patch-wise and pixel-wise approaches [36]. A patch-wise pattern first crops several smaller patches, e.g., 8 ×8 or 16 ×16, from the original larger remote sensing image randomly or in a certain order. These patches are then fed to a DCNNs-based image recognition model for training and testing. The centric pixel's category of a patch is labelled as ground truth. A sliding window of the same size as the patches is used to traverse the whole image area to predict each centric pixel's category. The drawback of this method is its low efficiency. In contrast, a pixel-wise semantic segmentation does not require a set of pre-trained patches and can perform an end-to-end pixel-level classification. FCNs, U-Net, and DeepLabV3+ are the representational pixel-wise models or networks. FCNs are the first proposed pixel-wise end-to-end semantic segmentation approaches, which have been widely used as baselines of the modified methods as the extracted objects' edges by FCNs are commonly blurred. U-Net is applied for medical image segmentation in the early days. It introduces a skip connection to optimize an object's edges by fusing low-level and high-level features. However, U-Net employs max pooling to expand the receptive field, resulting in losses in a certain feature's position. A dilated/atrous convolution has been proposed instead of max pooling for downsampling to obtain high-level features in DeepLab series models. In addition,

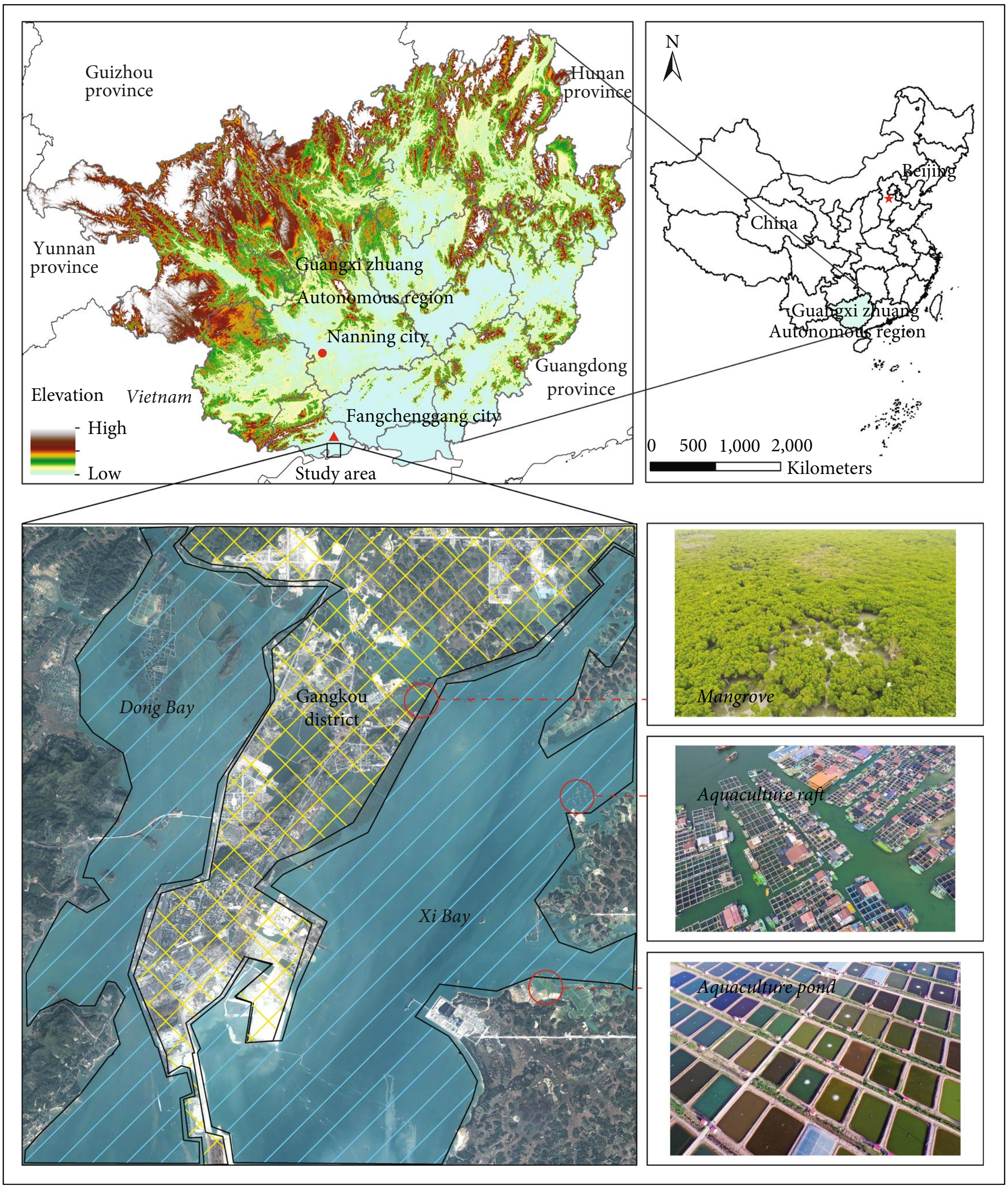


Figure 1: Study area and the target objects.

an Atrous Spatial Pyramid Pooling (ASPP) [19] in DeepLapV2 takes effect for multiscale object segmentation. To further optimize the segmentation accuracy, we propose a lightweight fully connected spatial dropout network (LFCSDN) by integrating the U-Net and DeepLab series feature fusion strategies. The architecture of our proposed network is shown in Figure 2.

3.1.1. Lightweight Fully Connected Encoder. As shown in Figure 2, the input image size and the number of channels are assumed to be W × H × C. Notation *s* denotes a convolutional stride in the figure, and *k* denotes a kernel or filter size. The colorful lines with arrows and block layers represent the meanings shown in the upper right corner of the figure. The regular convolution operation is denoted by the term “Conv2D” in Figure 2’s legend. The figure labels the stride and the size of the convolution kernel that was used for convolution. “ReLU” and “BN” stand for batch normalization and nonlinear activation, respectively. The upsampling process is identified as “UpSample2D.” The deepwise separable convolution “SepConv2D” is identified by the graph’s markings for the stride and convolution kernel size. “Softmax” refers to the convolutional maps’ multiclassification output, which is akin to the activation function. The new module we propose, which will be discussed below, is referred to as “CBS.” Firstly, low-level features are extracted by several

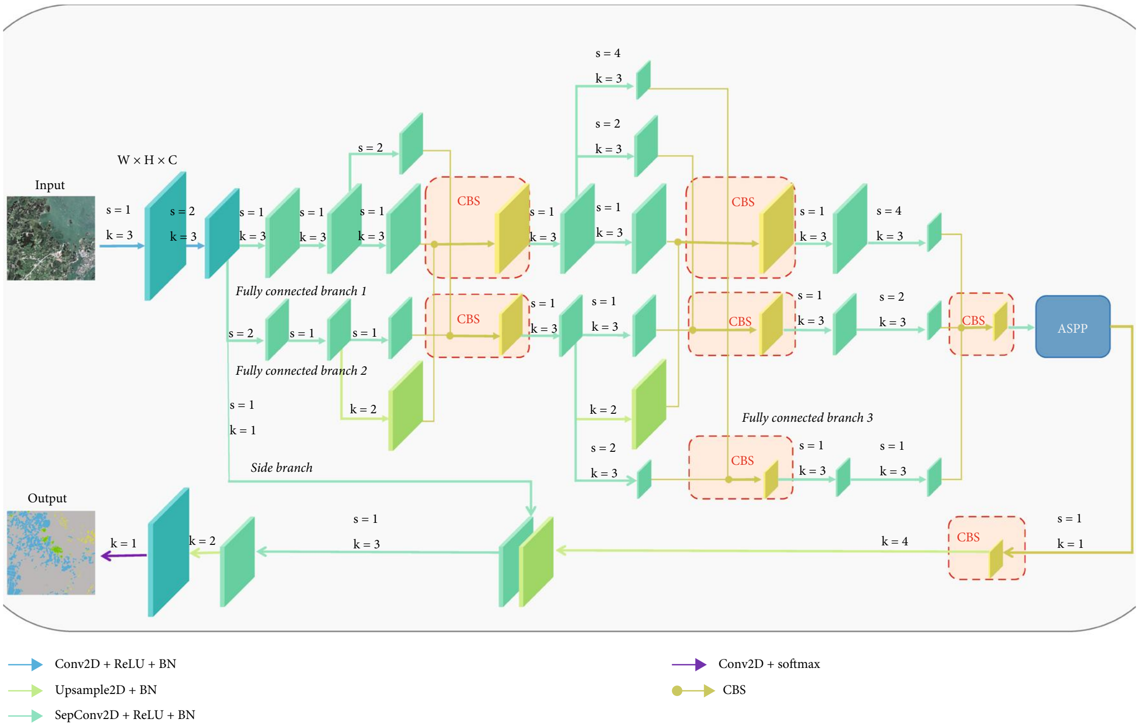


Figure 2: The architecture of LFCSDN.

standard convolutions. Convolution with a stride greater than one instead of a max pooling has been adopted for downsampling, which is notated as "downsampling convolution." It reduces the loss of positional information while obtaining a larger receptive field. To further decrease parameters and achieve a lightweight model, we construct three branches of deepwise separable convolutions [33] based on the first downsampled feature maps. The branches contain two main branches and a side branch for later feature fusion, where the main branch is denoted as a "fully connected branch." The initial convolutional kernel size in the fully connected branch 1 is 3 × 3, and the convolutional stride is 1. The convolutional stride in the fully connected branch 2 is set to be 2, i.e., a feature map is subjected to two-fold downsampling. The kernel size in the side branch is 1 × 1. Namely, it is compressed and downsampled with the same size for a subsequent feature fusion. After downsampling, output feature maps of the fully connected branch 1 and fully connected branch 2 can be used as a new input for the encoding stage, and the image resolution will be maintained in the respective branches, which is denoted as the "constant resolution feature map." A convolution that remains the feature maps unchanged in image resolution is denoted as "constant resolution convolution."

After several constant resolution convolutions, the output feature maps in the fully connected branch 1 and fully connected branch 2 are downsampled and upsampled, respectively, to two times in size for constructing a new branch feature layer. Meanwhile, a constant resolution convolution is also performed on the original fully connected branch. The depthwise separable convolution with a stride greater than 1 is still used for downsampling, and the bilinear interpolation is employed to upsample the feature map. Downsampled feature maps from the fully connected branch 1 and constant feature maps from the fully connected branch 2 are connected by a CBS module to generate a new fully connected branch. Suppose a feature map generated by the aforementioned operations is treated as a neuron. In that case, the crossing fusion between several feature maps with diverse scales builds a structure similar to a fully connected neural network, which retains more features than the standard VGGNet and ResNet using step-by-step downsampling. The proposed model further downsamples the output feature maps of the fully connected branch 2 to generate a new fully connected branch, given the complexity of the non-target background. The feature maps of branch 3 are reduced by 4 times. Feature fusing and CBS are performed on the fully connected branches 1, 2, and 3 to output new feature maps for each branch. To generate advanced semantic features, the feature maps in fully connected branches 1 and 2 are downsampled to be four times and two times as small, respectively, and then cascaded to the feature maps of the fully connected branch 3, which couples the CBS module in the end to obtain the final output feature maps.

To obtain multiscale and elaborate features from certain hierarchical feature maps, a skip connection and ASPP have also been coupled in our network. Atrous convolutions with diverse dilated rates are used in the encoded feature maps. Its output is fused with the previous side branch, followed by another CBS module and upsampling operation. The ultimate predicted output of the network is generated following the depthwise separable convolutions, upsampling and softmax operation in turn.

3.1.2. CBS Module. There are two kinds of feature fusion methods in general, an "add" way of channel-aligned fusion and a "concatenate" way of channel-cascaded fusion. The former requires the same number of channels to be fused, which can rich certain feature information of the original channel and lead to few parameters for the subsequent calculation. The latter can extend the feature dimension, resulting in a larger number of parameters but provides many more features. Therefore, as the computation condition allows, "concatenate" is more appropriate for a multi-object application. Although "concatenate" can gain a greater variety of features, it might also introduce certain redundant details leading to overfitting or poor generalization. Accordingly, we innovatively couple the spatial dropout [34] layer into concatenated feature maps to diminish the detailed noise by randomly dropping the entire feature channel of feature maps.

A spatial dropout abbreviated as "SPD" in this paper is a dropout [37] method proposed by Tompson et al. for the image processing application. A standard dropout randomly zeroes certain neurons' values, but a spatial dropout zeroes all pixels' values of certain feature maps, as shown in Figure 3. This approach has proven to be effective in practice in image recognition [38]. The standard dropout method randomly zeroes values of the independent neurons in a certain feature map, which cannot normalize the output if a strong correlation exists between neighboring neurons. At the same time, SPD can help improve the contrast's independence. To speed up the model's convergence and avoid the gradient explosion and gradient vanishing, a CBS module is ultimately constructed by adding a batch normalization layer between concatenation and spatial dropout. In subsequent experiments, several ablation studies, including the combinations of CBS modules and U-Net and DeepLabV3+, have been set up for further discussion.

3.2. Evaluation Metrics. In addition to using the evaluation metrics commonly used in remote sensing image classification, such as the overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), kappa, and F1-score (F1), it should be noted that OA is primarily the probability that every category is correctly predicted in the predicted outcome, whereas PA is calculated in accordance with the recall used in machine learning or deep learning, i.e., the probability that a category is correctly predicted in the real outcome, or the proportion of correctly classified pixels to the real total pixels. The probability that a category is correctly predicted in the predicted outcome, or the percentage of correctly classified pixels out of the total predicted pixels, is how "UA," which is used in machine learning or deep learning, is calculated. At the same time, F1 reconciles and balances UA and PA. Kappa is a useful remedy for the skewed evaluation caused by sample imbalance. We also choose the mIoU, which is an abbreviation of "mean intersection over union" used to measure the accuracy of a DCNNs-based semantic segmentation network. It calculates an average ratio of the intersection and union of the predicted pixels and labelled pixels (i.e., ground truth). It takes into account the number of pixels and pixel positions. Equation (1) shows the mathematical expression:

$$mIoU = \frac{1}{n}\sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n}\left(p_{ji} - p_{ii}\right)}, \qquad (1)$$

where $\boldsymbol{n}$ is the number of categories, $\boldsymbol{i}$ is a labelled category, $\boldsymbol{j}$ is a predicted category, $\boldsymbol{p_{ij}}$ denotes the probability of a predicted pixel $\boldsymbol{i}$ as $\boldsymbol{j}$, $\boldsymbol{p_{ii}}$ denotes the probability of a predicted $\boldsymbol{i}$ like $\boldsymbol{i}$, and $\boldsymbol{p_{ji}}$ denotes the probability of a predicted $\boldsymbol{j}$ as $\boldsymbol{i}$.

In addition, to measure the lightweightness of diverse ablation models, the parameters' amount has been used as an additional evaluation metric. Innovatively, we calculate the mean square error of each epoch between the training and validation set to measure the degree of overfitting or generalization. It is identified as a "loss coefficient" in current research.

3.3. Experiments. Three experiments have been implemented one after another. Firstly, we use the representative DCNNs-based semantic segmentation networks to conduct a benchmarking experiment on our constructed CSRSD dataset. Then, we analyze and discuss the results based on the evaluation mentioned above metrics. The results could be used as a baseline. Secondly, to verify the suppression of overfitting and the improvement of generalization of the CBS module, we design several ablation experiments by adding and removing BN and SPD following the concatenation fusion. It should be noted that the BN always follows a block of convolutions and ReLU functions in each designed experiment containing BN. To ensure the integrity of the feature information, all CBS modules are placed only at the fully connected branch junctions. Thirdly, to describe the model's generalization and efficiency in classifying a larger-scale image, we select another image with the same date and sensor but near the study area for the experiment. We analyze and discuss the characteristics and potential causes of spatial and temporal changes of three coastal target objects in the study area based on the labelled images from a regional scale. A Conda virtual environment, Pycharm IDE, several third-party libraries of Python, Tensorflow framework, and NVIDIA RTX3090 GPU have been adopted to support the experiments. All hyperparameters, cost, and optimization functions remain the same.

4. Results and Discussion

4.1. Effectiveness Evaluation of the Proposed Networks. Table 1 shows the major ablation study's overall accuracy (OA), kappa, mIoU, and parameter amount of storage occupied on a disk. The loss coefficient of each network is also presented in Table 1. It can be seen that the DCNNs-based classification for HSRIs can achieve high accuracy. An optimal OA, kappa, and mIoU appear in the DeepLabV3+ (CBS) method, while our LFCSDN obtains the best parameters and loss coefficient. The results show that the U-Net and DeepLabV3+ with the addition of a CBS module as well as our

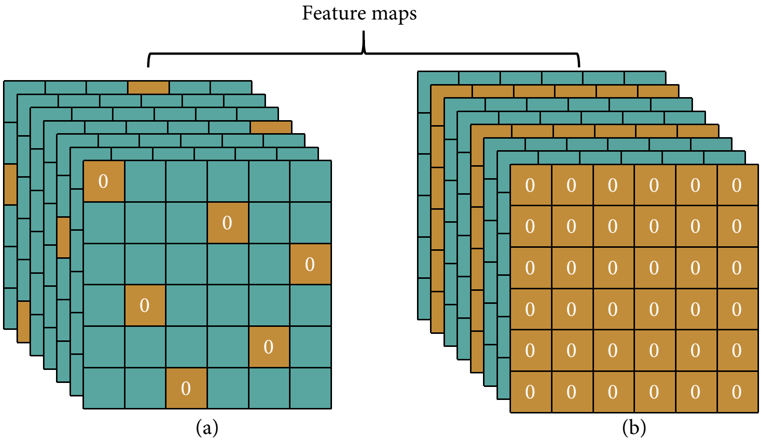


Figure 3: Feature maps generated by the dropouts. (a) Standard dropout. (b) Spatial dropout.

Table 1: Results of the overall effectiveness evaluation.

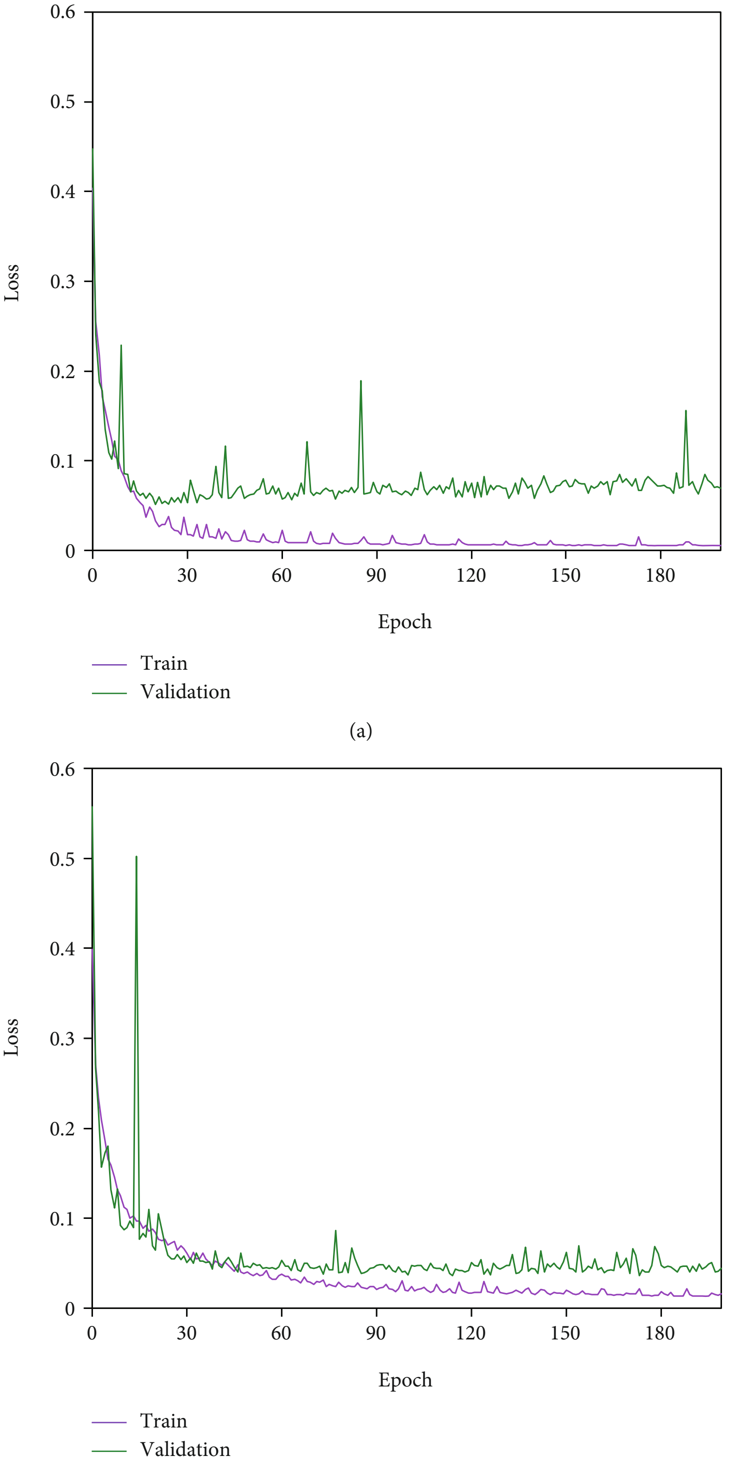
Methods	OA (%)	Kappa	mIoU (%)	Parameters (MB)	Loss coefficient
U-Net	96.34	0.73	76.96	252	0.0275
U-Net (CBS)	98.12	0.87	87.34	252	0.0051
DeepLabV3+	98.33	0.88	86.77	65	0.0043
DeepLabV3+ (CBS)	**98.50**	**0.90**	**88.47**	65	0.0028
LFCSDN (ours, BN)	98.24	0.87	87.72	**16**	0.0040
LFCSDN (ours, CBS)	98.22	0.88	86.72	**16**	**0.0012**

proposed LFCSDN have enhanced the accuracy significantly. All loss coefficients are decreased, demonstrating an improved generalization. DeepLabV3 + (CBS) performs best in accuracy, while our proposed models are lighter overall due to the deepwise separable convolutions. Meanwhile, it possesses the best generalization, followed by DeepLabV3+ (CBS).

More intuitively, Figure 4 shows the trend of loss changes in the training and validation sets of DeepLabV3+ and LFCSDN before and after the addition of the CBS module, respectively. It can be seen that the overall variability of the training loss and validation loss per epoch is reduced with a less volatile. During the training, CBS can promote the consistency of the losses of training and validation sets that further validates the improved generalization, which has a certain mitigation effect on overfitting.

Table 2 shows the PA, UA, IoU, and F1 of three target objects of the mangrove, aquaculture raft, and aquaculture pond, demonstrating the single category accuracy. To elaborate on the PA concerning the mangrove, all models achieve good classification except for the standard U-Net, which does not reach above 90%. The best model is our proposed LFCSDN with 94.88%, which has an encoder that uses fully connected feature fusion and BN. For the aquaculture raft, the results of all models are also high. The best one is the improved U-Net combined with a CBS module. DeepLabV3+ with just over 90% is the best aquaculture pond classification. Using the UA as an evaluation metric, the best three models for the categories are DeepLabV3 + (CBS) (94.45%), LFCSDN (94.32%) and DeepLabV3 + (CBS) (87.92%), respectively. The best F1 for the mangrove is 94.24%, produced by DeepLabV3 + (CBS). The best F1 of 95.49% of the aquaculture raft is obtained by U-Net (CBS). 88.40% of the aquaculture pond classification is still obtained by DeepLabV3+ (CBS). While the evaluation metrics described above the results in terms of the number of pixels, IoU takes into account the spatial position characteristics of the pixels and therefore requires a higher level of pixel-wise classification. The vast majority of IoU results for each method are below 90%, with the best IoU classification for the mangrove and aquaculture pond both produced by the method of DeepLabV3+ (CBS) at 89.11% and 79.2%, respectively, while the best IoU for aquaculture raft is produced by U-Net (CBS) at 91.38%.

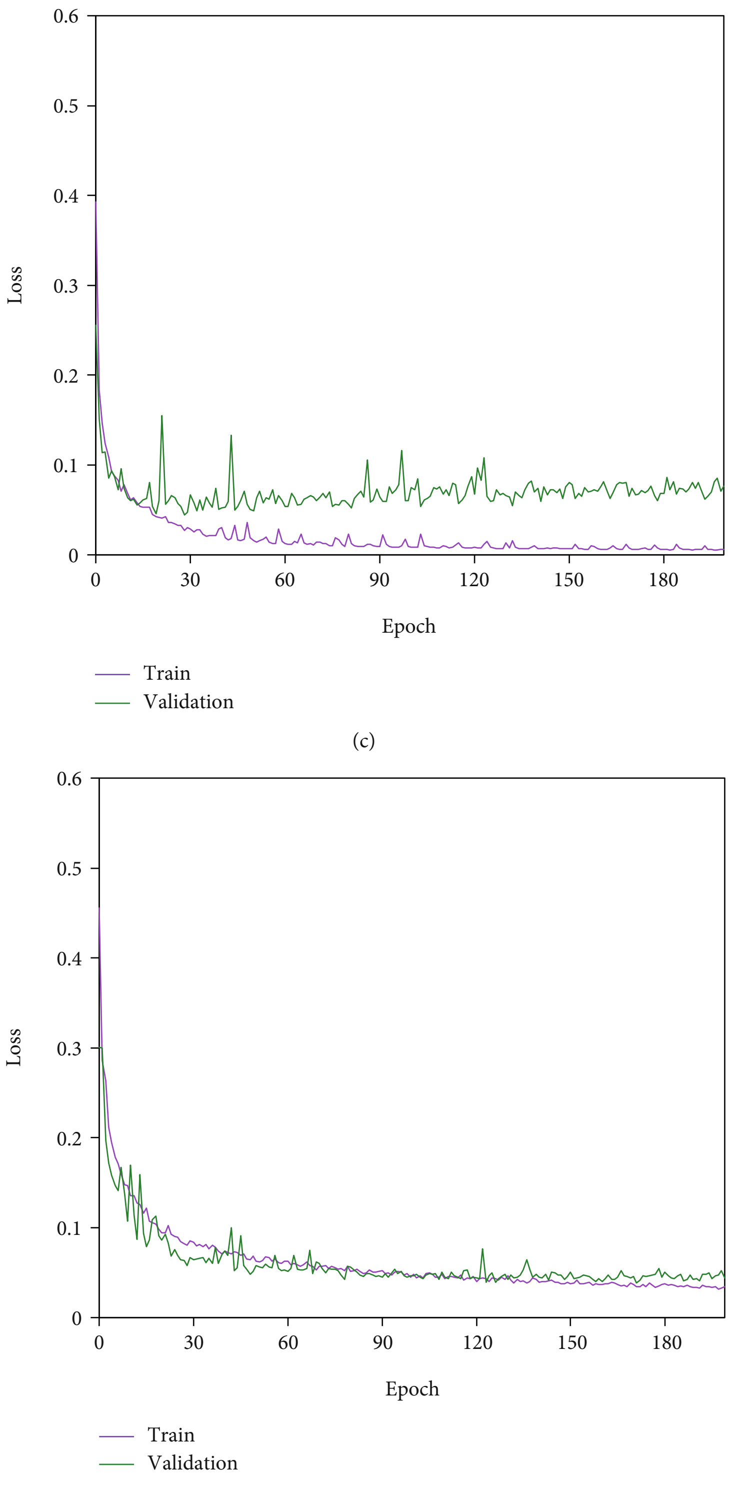
Regardless of the evaluation metric, three categories are classified in descending order of accuracy: aquaculture raft, mangrove, and aquaculture pond. When a CBS module is added, the values of most of the indicators have been increased. The best accuracy in terms of the target categories is probably due to the aquaculture raft's special texture and color structure, which has fewer features with a similar structure to the other two categories and the complex background features. In terms of shape features, buildings in the background are similar, but the textural and dimensional features differ more between them. Therefore, the architectural features in the background have less influence on the classification of aquaculture rafts. In the case of the mangrove, it has the same color characteristics as the regular vegetation in the background but has a certain textural difference. In addition, the mangroves are mostly found in



(a)

(b)

Figure 4: Continued.



(c)

(d)

Figure 4: Comparisons of the generalization performance. (a) DeepLabV3+. (b) DeepLabV3 + (CBS). (c) LFCSDN(BN). (d) LFCSDN(CBS).

mudflats with saltwater intrusion, which is a robust spatial feature. The worst classification accuracy of the aquaculture pond relative to the others might be due to the presence of large areas of seawater in the background category that have the same texture and color characteristics as aquaculture ponds, resulting in obvious misclassification.

Table 3 and Figure 5 show the mangrove misclassification rate, i.e., the number of mangrove pixels that have been misclassified into the aquaculture raft, pond, and background categories as a proportion of all true mangroves' pixels. It can be seen that regardless of the modelling approach, the mangrove is misclassified into other categories

Table 2: Results of the individual target's effectiveness evaluation.

Metrics	Methods	Mangrove	Aquaculture raft	Aquaculture pond
PA (%)	U-Net	86.64	94.33	78.77
	U-Net (CBS)	94.88	**96.80**	89.77
	DeepLabV3+	94.50	94.95	**90.71**
	DeepLabV3+ (CBS)	94.04	94.40	88.90
	LFCSDN (ours, BN)	**94.88**	93.76	89.85
	LFCSDN (ours, CBS)	93.50	90.99	88.49
UA (%)	U-Net	88.98	89.10	56.26
	U-Net (CBS)	91.65	94.23	79.23
	DeepLabV3+	92.45	88.58	83.05
	DeepLabV3+ (CBS)	**94.45**	91.91	**87.92**
	LFCSDN (ours, BN)	91.06	**94.32**	82.19
	LFCSDN (ours, CBS)	92.74	94.25	83.63
F1 (%)	U-Net	87.80	91.64	65.64
	U-Net (CBS)	93.24	**95.49**	84.17
	DeepLabV3+	93.47	91.66	86.71
	DeepLabV3+ (CBS)	**94.24**	93.14	**88.40**
	LFCSDN (ours, BN)	92.93	94.04	85.85
	LFCSDN (ours, CBS)	93.12	92.59	85.99
IoU (%)	U-Net	78.25	84.58	48.86
	U-Net (CBS)	87.33	**91.38**	72.67
	DeepLabV3+	87.74	84.60	76.54
	DeepLabV3+ (CBS)	**89.11**	87.16	**79.22**
	LFCSDN (ours, BN)	86.79	88.74	75.20
	LFCSDN (ours, CBS)	87.13	86.21	75.43

Table 3: Misclassification rates of the mangroves.

Methods	Aquaculture raft	Aquaculture pond	Background
U-Net	0.006395	0.262101	10.747875
U-Net (CBS)	0.000132	**0.008735**	8.345952
DeepLabV3+	0.001192	0.051232	7.494523
DeepLabV3+ (CBS)	**0.000038**	0.029939	**5.522487**
LFCSDN (ours, BN)	0.013667	0.048097	8.883216
LFCSDN (ours, CBS)	0.000366	0.039204	7.215670

in the order of background, aquaculture pond, and aquaculture raft, with the highest number being misclassified into the background. In the study area, the background consists of buildings, bare areas, farmlands, and regular vegetation, a complex composition. Due to the spectral similarity between the farmlands and vegetation, a mangrove could likely be misclassified into the background, resulting in the highest percentage of misclassification. On the other hand, the mangrove and aquaculture pond, with distinctly different characteristics, are near each other in terms of spatial position and are likely to be misclassified due to a very small number of incorrectly labelled samples. For example, the true pixel belongs to the mangrove, but a small number of pixels distributed on edges are labelled as the aquaculture pond. The model identifies those pixels as mangroves based on the correct labelled samples' statistics, resulting in misclassification. The mangrove is uncorrelated with the aquaculture raft with the least misclassification. Remarkably, the misclassification rate has been improved in all three networks with an additional CBS module.

Table 4 and Figure 6 show the proportion of misclassified aquaculture rafts in mangroves, aquaculture ponds, and the background. The background is still the most misclassified due to its complexity. For example, some bare areas in the background resemble aquaculture rafts. In addition, aquaculture rafts are spatially located in seawater, and their labelled edges are inevitably subject to certain errors. The least misclassification of the six models is

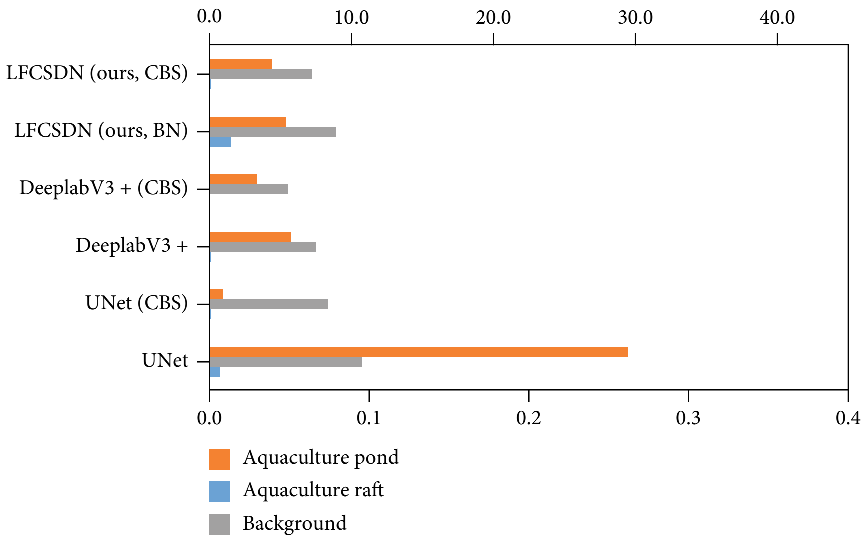


FIGURE 5: Misclassification rates of the mangroves.

TABLE 4: Misclassification rates of the aquaculture rafts.

Methods	Mangrove	Aquaculture pond	Background
U-Net	0.093508	0.022565	10.782253
U-Net (CBS)	0.099160	**0.000000**	5.674210
DeepLabV3+	0.037697	0.027594	11.352514
DeepLabV3+ (CBS)	**0.004317**	0.000979	8.084409
LFCSDN (ours, BN)	0.019538	0.001602	5.662638
LFCSDN (ours, CBS)	0.101653	0.025324	**5.619378**

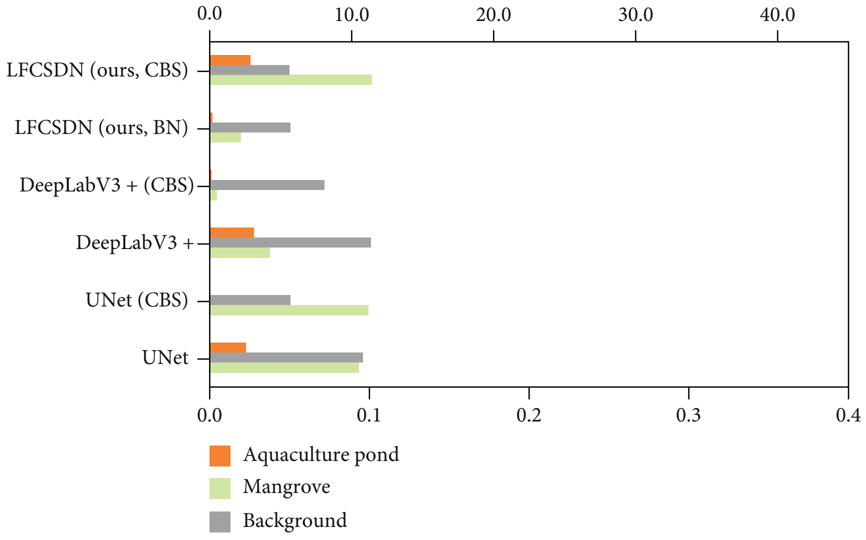


FIGURE 6: Misclassification rates of the aquaculture rafts.

DeepLabV3 + (CBS) for the aquaculture raft and U-Net (CBS) for the aquaculture pond and LFCSDN (ours, CBS) for the background.

Table 5 and Figure 7 show the misclassification rate for the aquaculture pond. Overall, there is a significant increase in the misclassification of the aquaculture pond into the background relative to the mangrove and aquaculture raft, reaching over 10%. This is mainly because an aquaculture pond consists of seawater. In contrast, the seawater and regular ponds in the background coincide with the category and are difficult to distinguish from each other in terms of a coarse spectrum. The shape feature of the aquaculture pond

Table 5: Misclassification rates of the aquaculture ponds.

Methods	Mangrove	Aquaculture raft	Background
U-Net	0.134942	0.002740	43.599243
U-Net (CBS)	0.012220	0.001137	20.759122
DeepLabV3+	**0.003515**	**0.000000**	16.947220
DeepLabV3+ (CBS)	0.008277	0.000012	**12.074753**
LFCSDN (ours, BN)	0.008212	0.000160	17.802990
LFCSDN (ours, CBS)	0.010225	0.000027	16.359246

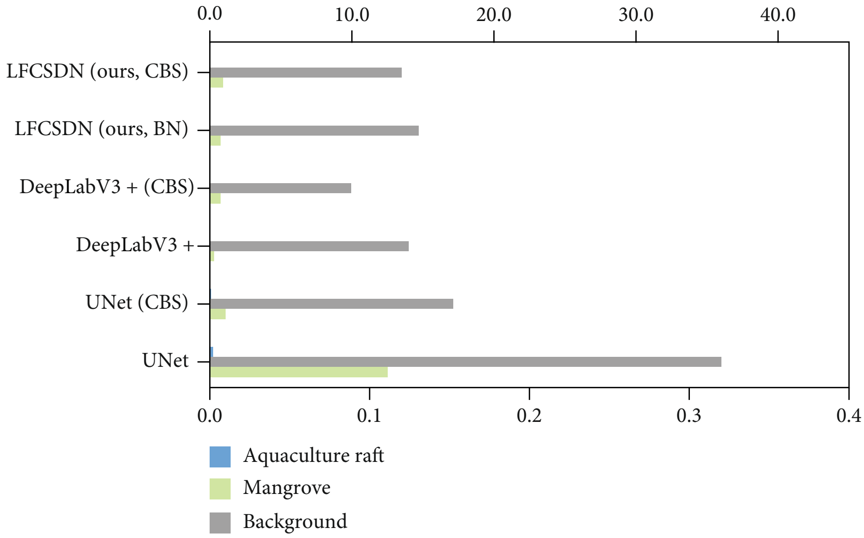


Figure 7: Misclassification rates of the aquaculture ponds.

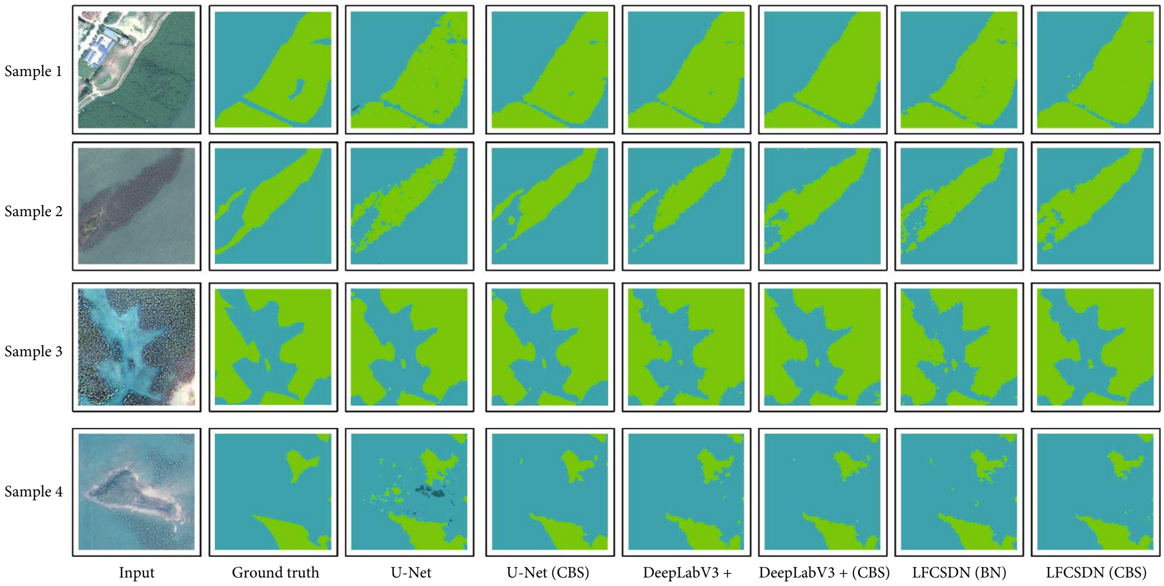


Figure 8: Comparisons of the mangroves extracted by the models.

can be used to distinguish it from the seawater, whereas it is relatively difficult to distinguish with a regular pond in the background. In addition, as the labelled samples are collected from the dynamic temporal scale, some of the ponds have been abandoned, whose bare bottoms are similar to the bare areas of the background, so this might contribute to a large number of misclassification of the aquaculture pond.

Several representative samples of the mangrove, aquaculture pond, and raft containing diverse orientations, shapes, colors and textures have been selected from the test

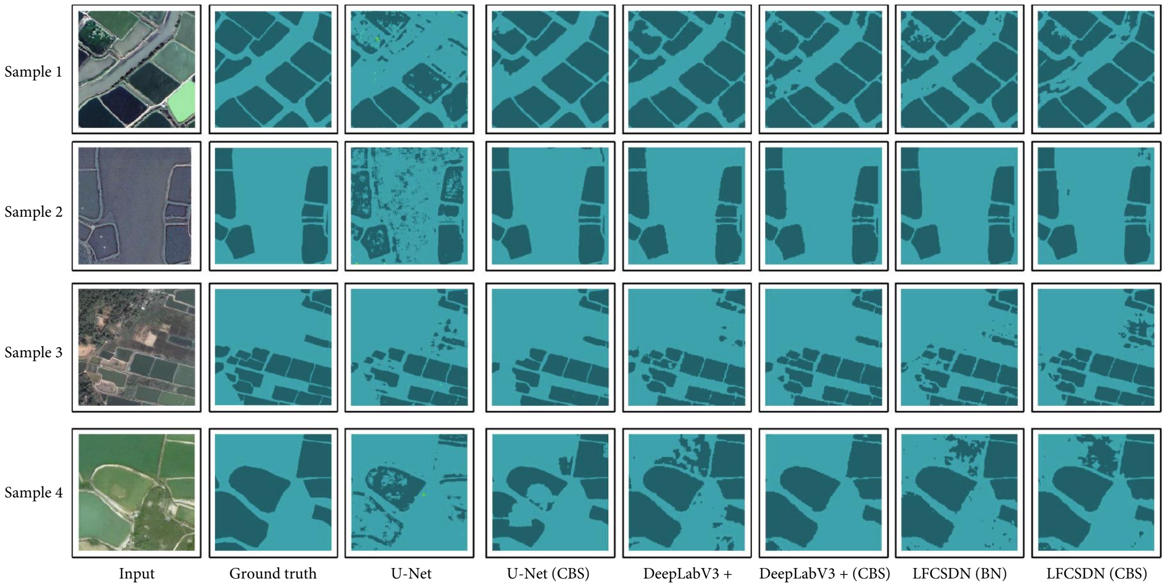


FIGURE 9: Comparisons of the aquaculture ponds extracted by the models.

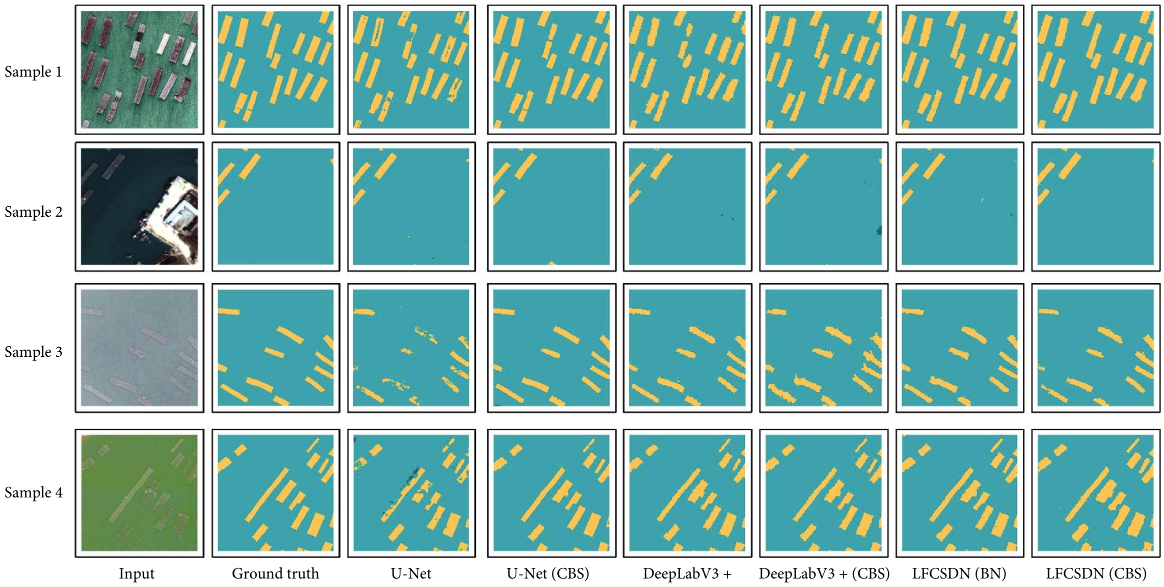
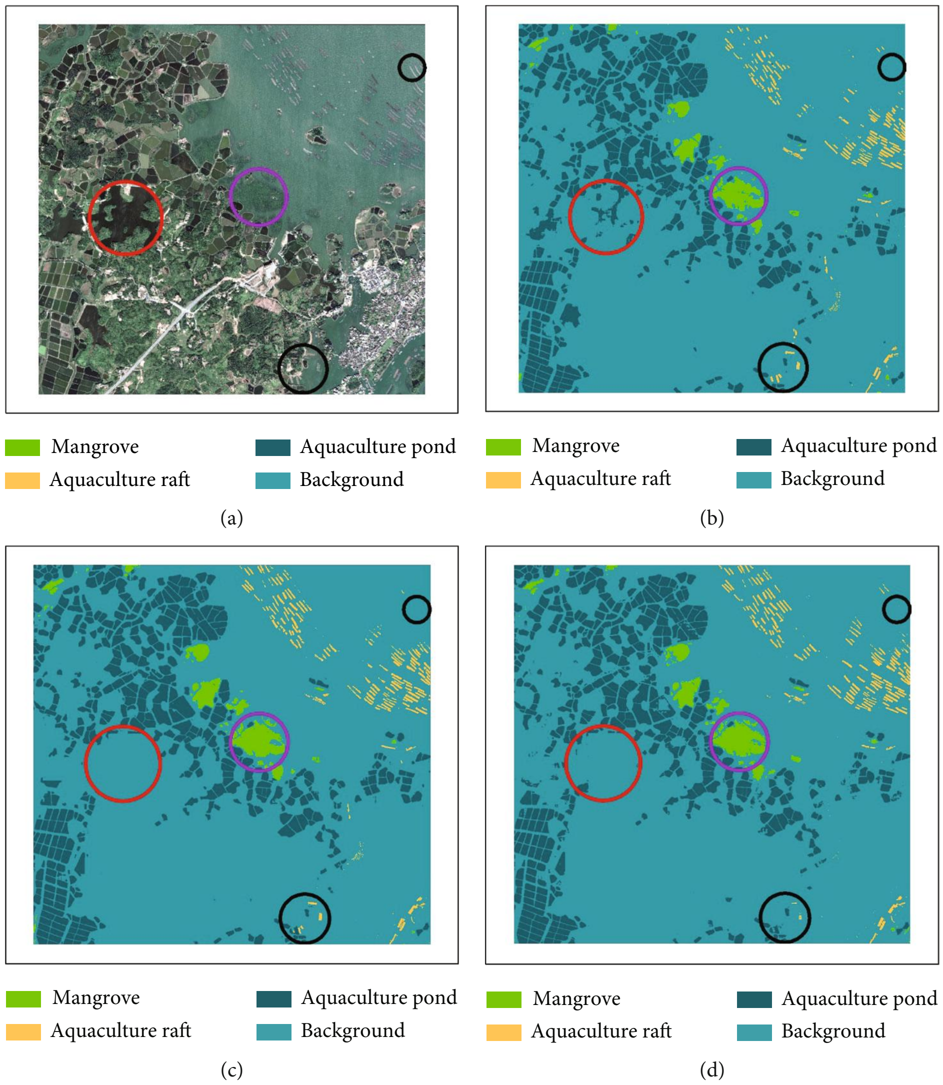


FIGURE 10: Comparisons of the aquaculture rafts extracted by the models.

set and visualized in the individual category, as shown in Figures 8, 9, and 10, respectively. The mangrove samples include samples that are easily confused with the typical vegetation in addition to the essential characteristics. The aquaculture pond samples also include the samples that are easily confused with the seawater and regular ponds. The results are consistent with the data in the previous tables, with the misclassification mainly concentrated on the background. In addition, the results show that U-Net is more prone to classification incompleteness than DeepLabV3+ and our proposed method, mainly because it does not consider multi-scale feature fusion. But the edge obtained by U-Net is more refined due to its skip connection which unites all low-level features. DeepLabV3+ uses only one feature fusion in the upsampling stage, resulting in a coarser pixel resolution on the edge. Our proposed method falls in the middle of U-Net and DeepLabV3+.

To test the generalization of the model and the classification efficiency on a larger scale image, we select an area outside the study area. Its image has been collected from the same sensor and possesses the same spatial resolution and date. The image covers an area of approximately 12.5 km^2 and has an image resolution of 6180×5901 by pixel. This area contains mangroves, aquaculture rafts, aquaculture ponds, and the background. As the area is used as a prediction, it has not been labelled. We describe the classification quality through a visualization result map. We only use U-Net and Deeplabv3+, which contain the CBS module and our proposed model. Under the same hardware and software environment, three models spent 146, 148, and 158 seconds



(a)

(b)

(c)

(d)

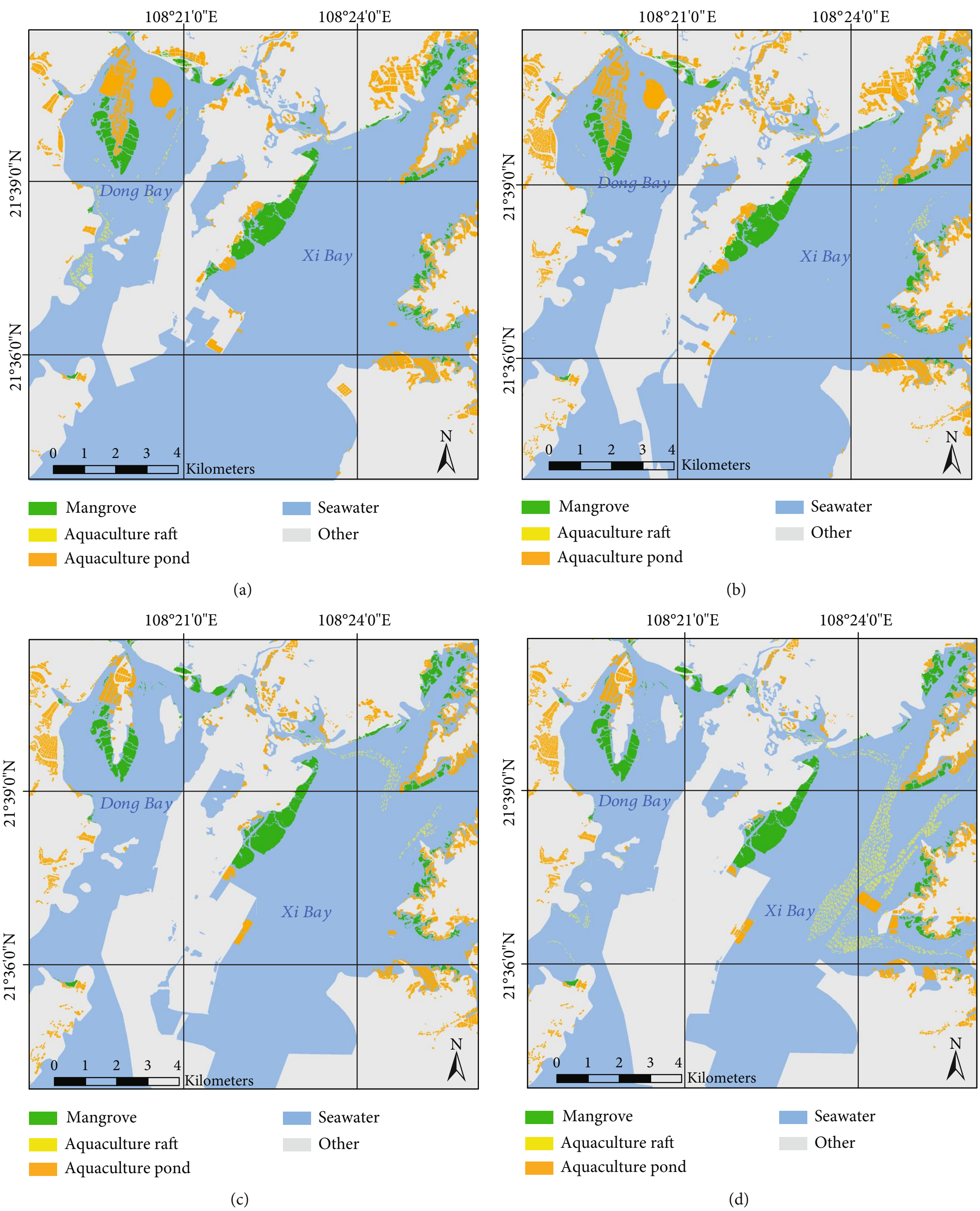
Figure 11: Extraction and classification in a large scale area. (a) Input image. (b) Output by U-Net (CBS). (c) Output by DeepLabV3+ (CBS). (d) Output by LFCSDN (CBS).

to extract and classify the area, respectively. Our proposed model takes the most time overall, mainly since the model requires multiple times upsampling and downsampling operations when predicting.

The results of the visual extraction and classification are shown in Figure 11. The results show that all three methods perform well in terms of the overall accuracy of the interpretation and classification of the large-scale image from the same sensor on the same date and possessing the same spatial resolution. In the red circles of the figure, the U-Net (CBS) method misclassified several regular ponds into the background. At the same time, both DeepLabV3+ (CBS) and LFCSDN (CBS) can correctly identify and classify them. In the purple circles, LFCSDN (CBS) can better preserve the integrity of a larger mangrove. For the decoding and classification of aquaculture rafts, all three methods show a significant loss in the misclassification of rafts into the background, as shown in the black circles in the upper right corner of the figure. In the black circles in the lower right corner, U-Net (CBS) performs best in classifying the aquaculture rafts that are non-concentrated ones. In addition, we have also experimented with other large-scale images collected from diverse sensors with various dates and resolutions. The results show that many coastal datasets are still needed for real applications.

4.2. Spatiotemporal Distribution Characteristics. The spatial and temporal distribution of the three target objects in the study area in 2003, 2007, 2015, and 2018 is shown in Figure 12 and Table 6, which gives the total area.

On the temporal scale, the total area of the mangroves in the study area remains stable without a large change in fifteen years. A slight decrease after 2015 might be related to the damage to mangroves due to illegal construction and sewage discharge from the factories mentioned in the introduction section. The area of the aquaculture ponds used for breeding fish and shrimp



(a)

(b)

(c)

(d)

Figure 12: The spatiotemporal distribution of the target objects (a) in 2003, (b) in 2007, (c) in 2015, and (d) in 2018.

has decreased by 20% between 2003 and 2018, probably due to the rise of oysters and scallops markets as well as the impact of the widespread use of new aquaculture technologies such as aquaculture rafts and net cages. On the other hand, strengthening governance for the illegal activities by fishers and farmers has been done well. The number of the aquaculture raft, a critical oyster breeding method, increased significantly between 2003 and 2018, mainly probably due to the increased demand for oysters in inland areas, prompting the fishers to increase their investment in aquaculture rafts. On the spatial scale, the mangroves, aquaculture ponds, and

Table 6: Temporal distribution of the three target objects.

Class and date	Area (km^2)			
	2003	2007	2015	2018
Mangrove	4.92	5.12	4.85	4.75
Aquaculture raft	0.29	0.20	0.24	1.65
Aquaculture pond	10.30	12.67	8.83	8.07

aquaculture rafts distribute in both Xi bay and Dong bay of the Gangkou District of Fangchenggang. Mangroves are more in the east than in the west. Aquaculture rafts showed explosive growth in the east bay in 2018. As seen in the figure, since 2015, some of the aquaculture ponds have been replaced by other lands.

5. Conclusions

Coastal typical object monitoring based on HSRIs is one of the important means of coastal ecological environment supervision. However, the complexity of high-resolution remote sensing images makes the traditional methods not very efficient. DCNNs-based semantic segmentation provides a new pattern to improve the extraction and classification accuracy. However, due to the lack of coastal scene datasets and the large-scale characteristics of remote sensing images, as well as the current semantic segmentation models still need to be improved in terms of accuracy and generalization, the research on the classification or extraction of coastal scene HSRIs based on DCNNs generally focuses on the local scale and the accuracy improvement of a single object. Few studies focus on multi-object extraction and spatiotemporal distribution analysis at the regional scale. Objects with different sizes, shapes, and structures have difficulty obtaining universal accuracy for the whole target object due to various nontarget categories in the background. Meanwhile, sharpened edges and accurate completeness of simple and multiscale objects deserve more attention.

This study constructs a multi-object coastal dataset CSRSD using HSRIs and GIS. Meanwhile, we modify the feature fusion strategies of U-Net and DeepLabV3+, replace the standard convolution with deepwise separable convolution, and introduce the spatial dropout to build a new CBS module. As a result, a new lightweight fully connected, connected spatial dropout network, LFCSDN, has been proposed based on the abovementioned modification. The experimental results demonstrate that the semantic segmentation dataset we constructed, with credible results on baseline-based U-Net and DeepLabV3+, can be used as a benchmark dataset for a DCNNs-based application on coastal scenes. LFCSDN can greatly reduce the number of parameters while ensuring good accuracy. The proposed CBS module can alleviate overfitting and improve the model's generalization. In addition, testing on large-scale remote sensing images and analyzing the spatial and temporal characteristics of the target changes in the study area could be used for ecological restoration, mapping, and comprehensive management in coastal areas. It also provides a reference for the research of multiscale semantic segmentation in computer vision.

Data Availability

All data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Key Scientific Research Foundation of the Education Department of Province Anhui (Grant No. KJ2020A0658), the Hefei University Talent Research Funding (Grant No. 20RC13), and the Hefei University Scientific Research Development Funding (Grant No. 20ZR03ZDA).

References

[1] Q. Xia, C. Z. Qin, H. Li, C. Huang, and F. Z. Su, "Mapping mangrove forests based on multi-tidal high-resolution satellite imagery," *Remote Sensing*, vol. 10, no. 9, p. 1343, 2018.

[2] Y. Hirata, R. Tabuchi, P. Patanaponpaiboon, S. Poungparn, R. Yoneda, and Y. Fujioka, "Estimation of aboveground biomass in mangrove forests using high-resolution satellite data," *Journal of Forest Research*, vol. 19, no. 1, pp. 34–41, 2014.

[3] J. Chen, Z. Mao, B. Philpot, J. Li, and D. Pan, "Detecting changes in high-resolution satellite coastal imagery using an image object detection approach," *International Journal of Remote Sensing*, vol. 34, no. 7, pp. 2454–2469, 2013.

[4] D. Ventura, A. Bonifazi, M. F. Gravina, A. Belluscio, and G. Ardizzone, "Mapping and classification of ecologically sensitive marine habitats using unmanned aerial vehicle (UAV) imagery and object-based image analysis (OBIA)," *Remote Sensing*, vol. 10, no. 9, p. 1331, 2018.

[5] A. Zaki, I. Buchori, A. W. Sejati, and Y. Liu, "An object-based image analysis in QGIS for image classification and assessment of coastal spatial planning," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 25, no. 2, pp. 349–359, 2022.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[9] Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] Q. Liu, R. Hang, H. Song, and Z. Li, “Learning multi-scale deep features for high-resolution satellite image scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 117–126, 2018.

[11] W. Wu, H. Li, L. Zhang, X. Li, and H. Guo, “High-resolution PolSAR scene classification with pretrained deep convnets and manifold polarimetric Parameters,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 6159–6168, 2018.

[12] R. Dong, D. Xu, L. Jiao, J. Zhao, and J. An, “A fast deep perception network for remote sensing scene classification,” *Remote Sensing*, vol. 12, no. 4, p. 729, 2020.

[13] R. A. Marcum, C. H. Davis, G. J. Scott, and T. W. Nivin, “Rapid broad area search and detection of Chinese surface-to-air missile sites using deep convolutional neural networks,” *Journal of Applied Remote Sensing*, vol. 11, no. 4, article 042614, 2017.

[14] W. Liu, D. Cheng, P. Yin et al., “Small manhole cover detection in remote sensing imagery with deep convolutional neural networks,” *ISPRS International Journal of Geo-Information*, vol. 8, no. 1, p. 49, 2019.

[15] Y. You, Z. Li, B. Ran, J. Cao, S. Lv, and F. Liu, “Broad area target search system for ship detection via deep convolutional neural network,” *Remote Sensing*, vol. 11, no. 17, p. 1965, 2019.

[16] D. Li, Y. Li, Q. Xie, Y. Wu, Z. Yu, and J. Wang, “Tiny defect detection in high-resolution aero-engine blade images via a coarse-to-fine framework,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[17] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, 2015.

[18] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Cham, 2015.

[19] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[20] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, https://arxiv.org/asb/1706.05587.

[21] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, Munich, Bavaria, Germany, 2018.

[22] T. Hoeser, F. Bachofer, and C. Kuenzer, “Object detection and image segmentation with deep learning on earth observation data: a review—part II: applications,” *Remote Sensing*, vol. 12, no. 18, p. 3053, 2020.

[23] Y. Sun, J. Huang, Z. Ao, D. Lao, and Q. Xin, “Deep learning approaches for the mapping of tree species diversity in a tropical wetland using airborne LiDAR and high-spatial-resolution remote sensing images,” *Forests*, vol. 10, no. 11, p. 1047, 2019.

[24] Q. Li, F. K. K. Wong, and T. Fung, “Mapping multi-layered mangroves from multispectral, hyperspectral, and LiDAR data,” *Remote Sensing of Environment*, vol. 258, p. 112403, 2021.

[25] Y. Guo, J. Liao, and G. Shen, “Mapping large-scale mangroves along the maritime silk road from 1990 to 2015 using a novel deep learning model and landsat data,” *Remote Sensing*, vol. 13, no. 2, p. 245, 2021.

[26] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] M. Guo, Z. Yu, Y. Xu, Y. Huang, and C. Li, “ME-net: a deep convolutional neural network for extracting mangrove using sentinel-2A data,” *Remote Sensing*, vol. 13, no. 7, p. 1292, 2021.

[28] L. Wan, H. Zhang, M. Liu, Y. Lin, and H. Lin, “Early monitoring of exotic mangrove sonneratia in Hong Kong using deep convolutional network at half-meter resolution,” *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 2, pp. 203–207, 2021.

[29] C. Diniz, L. Cortinhas, M. L. Pinheiro et al., “A large-scale deep-learning approach for multi-temporal aqua and salt-culture mapping,” *Remote Sensing*, vol. 13, no. 8, p. 1415, 2021.

[30] B. G. Cui, Y. Zhong, D. Fei et al., “Floating raft aquaculture area automatic extraction based on fully convolutional network,” *Journal of Coastal Research*, vol. 90, no. sp1, pp. 86–94, 2019.

[31] Y. Liu, X. Yang, Z. Wang, C. Lu, Z. Li, and F. Yang, “Aquaculture area extraction and vulnerability assessment in Sanduao based on richer convolutional features network model,” *Journal of Oceanology and Limnology*, vol. 37, no. 6, pp. 1941–1954, 2019.

[32] Z. T. Zheng, H. Fan, J. Wang, Y. Wu, B. Wang, and T. Huang, “An improved double-branch network method for intelligently extracting marine cage culture area,” *Remote Sensing for Land and Resources*, vol. 32, no. 4, pp. 120–129, 2020.

[33] F. Chollet, “Xception: deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, Honolulu, HI, USA, 2017.

[34] J. Tompson, R. Goroshin, A. Jain, Y. Le Cun, and C. Bregler, “Efficient object localization using convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 648–656, Boston, MA, USA, 2015.

[35] Y. Tian, Q. Zhang, H. Huang et al., “Aboveground biomass of typical invasive mangroves and its distribution patterns using UAV-LiDAR data in a subtropical estuary: Maoling River estuary, Guangxi, China,” *Ecological Indicators*, vol. 136, p. 108694, 2022.

[36] X. Y. Tong, G. S. Xia, Q. Lu et al., “Land-cover classification with high-resolution remote sensing images using transferable deep models,” *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] S. Lee and C. Lee, “Revisiting spatial dropout for regularizing convolutional neural networks,” *Multimedia Tools and Applications*, vol. 79, no. 45-46, pp. 34195–34207, 2020.