IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 25, 2020, accepted January 11, 2021, date of publication January 18, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052021

Pixel-Level Prediction for Ocean Remote Sensing Image Features Fusion Based on Global and Local Semantic Relations

HAO GAO[1,2,3], XUEJUN XIONG[3], LIN CAO[2], DINGFENG YU[2], GUANGBING YANG[3], AND LEI YANG[2]

[1] College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
[2] Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao 266061, China
[3] First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China

Corresponding authors: Xuejun Xiong (xiongxuejun_fio@163.com) and Lin Cao (caolin_summer@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 40406009 and Grant 41376038; in part by the National Science and Technology Major Project under Grant 2016ZX05057015, in part by the Key Research and Development Plan of Shandong Province under Grant 2019GHY112017, and in part by the Youth Innovation and Technology Support Program in Colleges and Universities of Shandong Province under Grant 2019KJN009.

ABSTRACT With the rapid development of remote-sensing imaging technology, remote-sensing images have become increasingly diverse, and people are paying more attention to ocean remote-sensing research. Because ocean remote-sensing data are complex, and the ocean environment is diverse, results will differ, even if the same target is detected at different times in the same scene. To obtain more semantic features and better pixel-level prediction capabilities, this paper proposes a pixel-level ocean remote-sensing image algorithm (GLPO-Net) that combines local and global features. First, texture features, color features, and spatial relationship features are extracted. Second, the algorithm constructs a multiscale local cross-attention mechanism strategy to obtain feature weight information in different directions to fully mine the local features of ocean remote-sensing images. Concurrently, an algorithm constructs a multiscale global cross-attention mechanism strategy to obtain global features. Then, the fusion of global features and local features is described in each submodule to obtain more representative deep features. Finally, small-sample ocean remote-sensing is described via image pixel-level prediction. The algorithm proposed in this paper has been tested with three public ocean remote-sensing datasets. The experimental results show that the proposed GLPO-Net algorithm can learn features from small samples of ocean remote-sensing images. Compared to the prediction results of other remote-sensing image algorithms, GLPO-Net exhibits better prediction capabilities.

INDEX TERMS Ocean remote sensing, deep learning, features fusion, multi-scale convolutional.

I. INTRODUCTION

With the development of remote-sensing imaging technology, remote-sensing images have become increasingly diverse. Therefore, the demand for remote-sensing technology has diversified and includes remote-sensing target detection [1]–[3]; remote sensing scene classification [4], [5]; remote sensing image semantic segmentation [6]–[8], [30], [32], [34], [39] and other tasks. In marine remote-sensing, the pixel distribution of the foreground (recognition target) is often below that of the background (sea water, etc.). Therefore, the pixel-level recognition of small samples of ocean images is typically challenging. In recent years, with the rapid development of deep learning [27], [28], [36]–[38], even a few samples can yield more semantic information, thus allowing more meaningful deep feature [29], [31], [33], [35] to be recognized. Research on deep learning algorithms in different scenarios of remote-sensing images has also yielded good results [9]–[15]. Recently, some researchers have done a lot of research on obtaining semantic information and generating more valuable features.

In terms of obtaining semantic information: Pan *et al.* [16] proposed an atrous model to obtain deep global semantic information from remote-sensing images and then used

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

Canny and morphological algorithms to enhance the deep semantic information and amplify the edge semantic information. Finally, using edges and regions, a joint feature was used to achieve pixel-level target recognition. Xu *et al.* [17] proposed a hierarchical semantic propagation framework to improve the target detection performance of remote sensing images. The core idea of their framework is that semantic information can be transmitted between different components along a network. First, spatial details and global semantic information are obtained through the pyramid network, and then the hierarchical semantic layer is used to obtain the hierarchical semantic information. Finally, the comprehensive evaluation of three sets of data sets highlights the superiority of this method. Zhang *et al.* [18] introduced a high-resolution network (HRNet) to enhance features to obtain contextual semantic information. This method uses the spatial linking method of the model to show more semantics for the low-resolution information containing more semantic information to the high-resolution information, and enhance the high-resolution information, thereby solving the positioning caused by cascading pooling. Loss of accuracy and preservation of spatial details. These models focus more on the preservation and mining of global semantic information. Various algorithms are used to obtain the global semantic information of remote-sensing images but ignore local semantic information.

In terms of feature mining. You *et al.* [19] proposed a SAGP algorithm that uses convolution to mine deep features of images, uses independent recurrent neural networks to retain the contextual semantic information of remote sensing images, and uses graph convolution algorithms to build relationships between features. Through the generation of these three features, many features are used in the model, and the best prediction result is obtained using the Populus euphratica dataset. Imbriaco *et al.* [20] proposed an image retrieval pipeline that focuses on the mining of local features, and finally aggregates all local features into a complete global feature. By comparing various verification coefficients, this strategy yields better performance than other feature extraction methods, even without combining external factors. Zhang *et al.* [21] proposed a multi-scale dense network. First, the network obtains different scale information and integrates multiscale information. This method obtains shallow and deep features. Concurrently, the three-dimensional dense link structure is used to achieve different levels of feature clustering. By considering the proportion information in a remote-sensing image, three different proportion feature maps are extracted. Finally, the algorithm obtained the best results on the five data sets called Indian Pines, Pavia University, Salinas, Botswana, and Kennedy Space Center. This research focuses on mining local semantic information. Concurrently, this research shows that multiscale convolution can obtain more local semantic information and that an independent recurrent neural network plays an active role in learning the contextual semantic information of remote-sensing images.

hlThese studies show that the semantic information of remote-sensing images plays a positive role in the final recognition. These algorithms learn global features using different strategies, and some focus on learning local features. Although the global or local semantic information of remote-sensing images can be obtained, this information is typically incomplete. For this reason, the algorithm proposed in this paper will obtain global and local semantic information concurrently. Obtaining global and local semantic information guarantees a final prediction of the deep learning model. Different algorithms produce different features, and multiple features describe ocean remote-sensing images from multiple aspects, thus providing more support for the final prediction. To better obtain the semantic information of ocean remote-sensing and mine deep image features in more detail, this paper proposes a cascading algorithm of global and local features (GLNet). This algorithm yields better predictions of ocean remote-sensing images.

The main contributions are as follows:

- To better obtain the global feature information of each pixel in an ocean remote-sensing image, we constructed a bidirectional independent recurrent neural network to integrate the semantic information of all features in each pixel and constructed a global attention mechanism as global feature distribution weight coefficients. Finally, we constructed a dilation and dense module to achieve the integration of multiple global information in different receptive fields. The purpose of this network is to further strengthen the integration of global features and eliminate many of redundant features.
- To determine the local semantic information of each pixel in the ocean remote-sensing image in more detail, we map all features into a two-dimensional space, assign multiple weights to each feature through the cross-attention mechanism, combine them with a multiscale volume product, and then fully describe the local deep features of each pixel.
- After mining the local and global features in each layer, we merge the local and global features to generate deeper features with more semantic information and ultimately enhance feature description.

The remainder of the sections in this article are summarized as follows. The second part primarily analyzes the importance of global and local features. The third part primarily describes the characteristics of the GLPO-Net algorithm and the process of feature construction. The fourth part describes the performance of the GLPO-Net algorithm experimentally. Finally, the fifth part summarizes the results of the study and provides recommendations for future research.

II. RELATED RESEARCH

To better describe the pixel-level prediction of ocean remote-sensing images, it is necessary to fully obtain the global and local features of each pixel. Concurrently, to obtain more semantic features, it is also necessary to integrate multiple features. Thus, we investigate the fusion of global features, local features and multiple features.

Research on remote sensing image recognition based on global features. To improve the recognition accuracy of small samples of hyperspectral remote sensing images, Wang *et al.* [22] proposed the MACBINet algorithm. The algorithm obtains the contextual semantic information of deep features through an independent recurrent neural network and concurrently mitigates gradient disappearance during feature training of small sample data sets. This algorithm yields the best results on three publicly available hyperspectral datasets. Shao *et al.* [2] used the MF-CNN algorithm to obtain the multiscale global features of remote-sensing images and combined high- and low-level semantic information during the learning process to categorize the pixels of thick clouds, thin clouds and cloudless areas. Finally, compared to various cloud detection methods, the best performance was obtained.

Research on remote sensing image recognition based on local features. Through research, Yuan *et al.* [23] found that the features of the last fully connected layer pay more attention to global features and ignore local features during remote-sensing image classification, reducing the classification accuracy of certain images that are more correlated with local features. Thus, a local feature rearrangement algorithm is used to emphasize the importance of local features, thus retaining deep local features. To solve the limitations of existing deep learning algorithms for extracting features from an entire image. Li *et al.* [24] proposed a new regional deep feature extraction framework that extracts regions that may contain target information from the entire image. The algorithm then uses convolutional neural networks for feature learning and finally uses improved vector local aggregation descriptors to encode local features to achieve local feature extraction. Liu *et al.* [25] discussed the local description of deep convolution, using different scales of convolution to extract convolution features, and obtaining local feature descriptors by linking these convolutions, and eliminating redundant features of local feature descriptors through PCA. Finally, this paper shows that the performance of the local feature descriptor is better than the features extracted by the fully connected layer.

The fusion of global and local features is also very important. Zhang *et al.* [26] proposed a context-aware detection network. The network first adjusts feature performance through an attention mechanism, captures their global information through the overall scene, and uses target objects to capture their local information. Through the linkage of global and local information, the context information of both global and local features can be described in more detail. Finally, the feasibility of the algorithm is verified using two public datasets. Also, Zhu *et al.* [40] proposed a multimodal image fusion method that uses different strategies to achieve complementary information between different modalities of the same image to improve the expression of the entire image. To improve the expression of local features, the author uses the sum modified Laplacian method and the steering kernel feature concurrently, which highlights the importance of joint learning of an image's global and local semantic information.

These studies show that the independent recurrent neural network plays an active role in obtaining the contextual semantic information of a feature and retains global features more accurately. Multiscale convolutional neural networks play an active role in obtaining local features and fully describe the mining of local features. Concurrently, to expand the difference between features and eliminate redundant features, we described the weight distribution of global and local features through a global attention mechanism and a cross-attention mechanism [30].

In this article, we use the global attention mechanism, the independent recurrent neural network and dilation dense network to generate global features that contain more global semantic information. We use the cross-attention mechanism and multiscale convolution to achieve local features that contain more local semantic information. We also describe the deep fusion of local features and global features by fusing the local features and global features of each step; thus, feature points can obtain more semantic information. Through the fusion of these local features and global features, ocean remote-sensing images from shallow features to deep features are retained to the largest extent. Via experiments, the feasibility and robustness of this method are shown.

III. PROPOSED FRAMEWORK

Figure 1 shows the GLPO-Net algorithm model. First, we extracted the features [7] of three subsets (beach, island and sea_ice) of ocean remote sensing images in the NWPU-RESISC45 dataset [43]. Second, these three groups of features were jointly input into the independent recurrent neural network under the global attention mechanism to obtain global features. Concurrently, these three sets of joint features were input into the multiscale convolution under the cross-attention mechanism to obtain local features. The fusion of features at each step was achieved by fusing the global and local features of each submodule. The final algorithm was classified by sigmoid, and the function predicted each remote-sensing pixel.

A. FEATURE EXTRACTION

To accurately predict the foreground category (e.g., beaches, islands, ice cubes, etc.) and background category (water) of ocean remote sensing, three types of features are extracted to represent each pixel in this study. The extracted features are as follows: 1) spatial relationship features based on weights (SRW); 2.) local texture features (LTF); and 3) HIS color features (HIS).

Spatial relationship features based on weights: We convert the R, G, and B three channels of the original ocean remote sensing image in the NWPU-RESISC45 dataset to gray value conversion., and the specific formula 1 is as follows [41]:

$$\text{Gray} = R * 0.299 + G * 0.587 + B * 0.114 \tag{1}$$

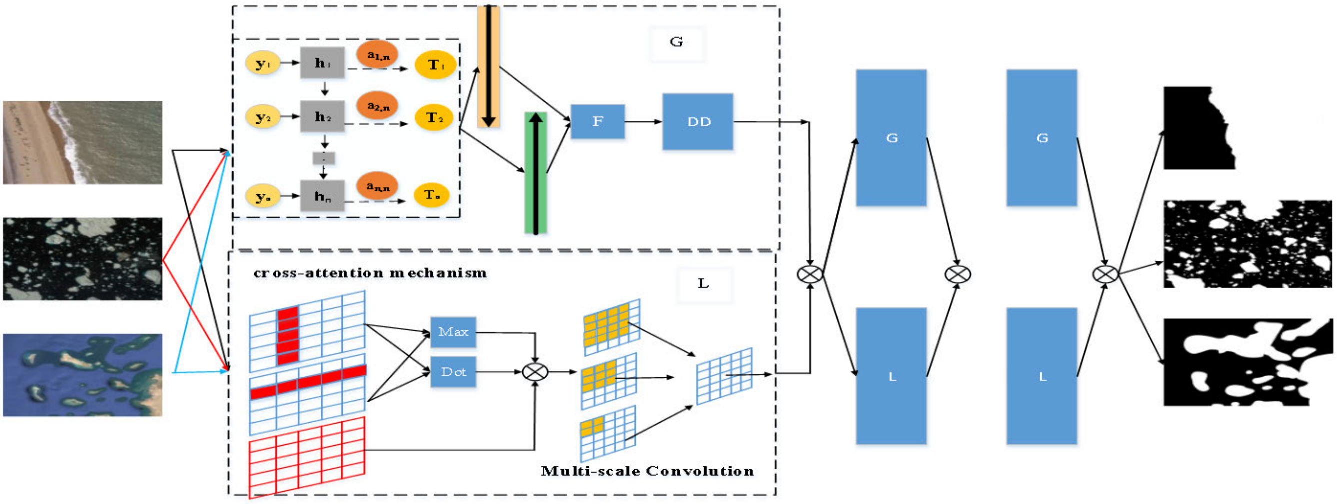

FIGURE 1. GLPO-Net algorithm model. where *G* represents the global feature extraction module; *L* represents the local feature extraction module; *Max* represents the maximum strategy; *Dot* represents the dot product strategy; *F* represents the feature fusion strategy; The downward arrow and the upward arrow represent the independent recurrent nerve semantic information in both directions of the network. ⊗ represents the integration of multiple characteristics; and *DD* represents the dilation dense module; The red matrix represents the original feature map.

where *Gray* represents the gray pixel value image that generated the pixel; *R* represents the red channel pixel value; *G* represents the green channel pixel value; *B* represents the blue channel pixel value; and Other values represent the proportion of each channel.

To expand the difference between the foreground and the background, we add a set of weight coefficients between the gray values. When the two adjacent gray values differ greatly(The gray value between them is greater than 50), the weight is set to 0; when two adjacent gray values When the gap is small, the weight is set to 1(The gray value between them is less than 50). We also set a threshold (50) to detect changes between pixels. This algorithm can be described as follows:

$$W = \text{where}\left(\left(\text{Gray}_1 > \text{Gray}_2\right), 1, 0\right) \tag{2}$$

where $Gray_1$ represents the first gray value; $Gray_2$ represents the second gray value; *where* represents the judgment of the first gray value and the second gray value; and w represents the generated weight.

Local texture features: A texture feature can describe the shape, stripes and other aspects of a predicted target in a remote-sensing image, thereby enhancing the characteristics of the predicted target shape in the image. To improve the acquisition of semantic information in surrounding pixels, we use small vector image blocks as the attributes of the central feature point. However, to reduce the image where the surrounding pixels are all central pixels, we introduce a weighted coefficient based on the reciprocal of the Euclidean theorem to assign different weights to the surrounding texture features and generate new local texture features as follows (3), as shown at the bottom of the next page: where T represents the texture feature; V represents the set of vector blocks; n represents the value representing the length of the vector block; m represents the value representing the width of the vector block; i represents the horizontal coordinate of the vector block; j represents the vertical coordinate of the vector block.

HIS color features: HIS color features are based on human vision and are expressed by hue (H), intensity (I), and saturation (S), and can express subtle changes in these visual aspects. Therefore, this paper selects the color feature of HIS as one of the most important features in remote-sensing images. The process of generating H, I, and S parameters from RGB data is described in reference [42] and is reproduced as follows:

$$H = \begin{cases} \theta \quad (B \le G) \\ 360 - \theta(B > G) \end{cases} \tag{4}$$

$$\theta = \cos^{-1}\left(\frac{\frac{1}{2}((R-G)+(R-B))}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right) \tag{5}$$

$$I = \frac{R+G+B}{3} \tag{6}$$

$$S = 1 - \frac{3}{R+G+B}\min(R, G, B) \tag{7}$$

Among them, R represents the red pixel value; G represents the green pixel value; B represents the blue pixel value; θ represents the cosine value between R, G, and B. $\le$ means less than or equal to; $>$ means "greater than"; and $+$ and $-$ are arithmetic symbols. These formulae generate new HIS three-dimensional features from RGB, and the generated vector features will be used as the input of the dual-channel deep learning algorithm.

B. GLOBAL FEATURE ACQUISITION

To obtain more accurate global semantic information, we constructed an independent recurrent neural network

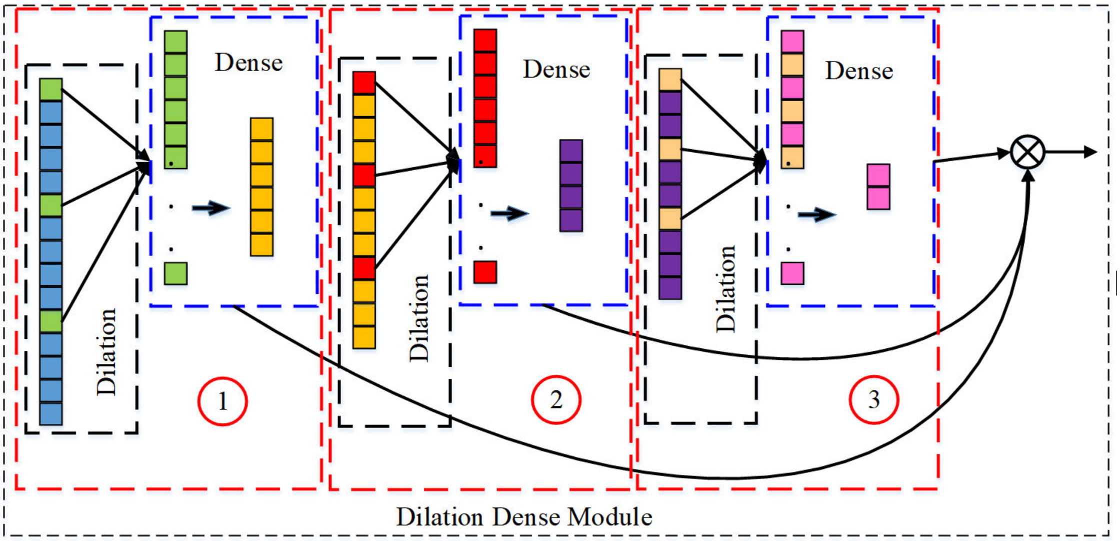

FIGURE 2. shows the dilation dense module network model. The black dashed box represents the 1-dimensional dilation convolution; the blue dashed box represents the global feature compression process; ⊗ represents the fusion strategy of different compression features. We achieved shallow to deep feature mining of global semantic information via dilated convolution. In the black dashed box, we use different expansion scales to describe the mining of global semantic information. In the blue dashed box, we use feature compression and extraction through dense layers. Finally, we use intensive strategies to fuse the global semantic information with different expansion scales. The red dashed box represents the process of extracting features by a set of dilation convolutions. Here we use three sets of dilation convolutions to mine features from shallow to deep. In "1" we set the expansion scale of the dilation convolution to 4; in "2" we set the expansion scale of the dilation convolution to 3; in "3" we set the expansion scale of the dilation convolution to 2. "Dilation" represents the dilation convolutional layer, whose purpose is to filter features; "Dense" represents the dense layer, whose purpose is to integrate features.

model based on the global attention mechanism. First, the feature vector of each pixel is input into the global attention mechanism, and each pixel is assigned a weight coefficient. Then, half of the features with larger weights are retained. Then, these features are input as new feature vectors to the independent recurrent neural network. In the network, we build a bidirectional independent recurrent neural network to obtain global semantic information in two directions. Finally, we use the normalization layer to normalize and integrate the generated semantic features. Through these algorithm, we retain the global semantic information of each pixel and improve the feature quality of the hidden layer, thereby providing a larger weight coefficient for foreground features. The global attention mechanism and independent recurrent neural network formulae are described as follows:

$$G = \mathrm{Max}_{1\sim n/2}\,(\text{ global}\left\{\sum_{j=1}^{n}\frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{n}\exp\left(e_{ik}\right)}h_{ij}\right) \quad (8)$$

$$\mathrm{Ind}_t = f\left(\omega_1 x_t + \omega_2 G_{t-1}\right) \quad (9)$$

$$\mathrm{Ind}'_t = f\left(\omega_3 x_t + \omega_5 \mathrm{Ind}_{t+1}\right) \quad (10)$$

$$O_t = g\left(\omega_4 Ind_t + \omega_6 Ind'_t\right) \quad (11)$$

where G represents half of the features after the global attention distribution weight; i represents the moment, and e represents the energy value of the ith moment; j represents the length of the feature sequence; h_ij represents the hidden state information of the j by feature vectors. Ind_t represents a feature sequence containing semantic information from left to right; Ind'_t represents the feature sequence containing semantic information from right to left; O_t represents a new feature sequence that combines left and right semantic information. $1 \sim n/2$ represents the largest half of the value in the entire feature vector; and ω represents the feature weight coefficients in different layers of the Bi independent recurrent neural network.

After improving the quality of global features, we realized further mining of global features and elimination of redundant features through the dilation dense module. In order to better express the dilation dense module I mentioned, we made a detailed description of this module through Figure 2.

$$\underset{n=\{3,5,7\}}{V} = \begin{bmatrix} \frac{1}{\sqrt{n^2+m^2}}T_{i-n,j-m} & \cdots & \frac{1}{\sqrt{n^2+m^2}}T_{i+n,j-m} \\ \cdots & 1^*\left(T_{i,j}\right) & \cdots \\ \frac{1}{\sqrt{n^2+m^2}}T_{i-n,j+m} & \cdots & \frac{1}{\sqrt{n^2+m^2}}T_{i+n,j+m} \end{bmatrix} \quad (3)$$

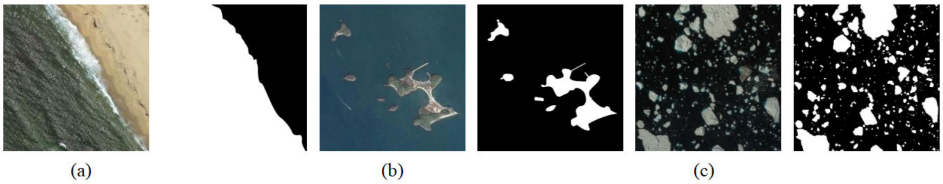

FIGURE 3. NWPU-RESISC45 data set. Including original images (left) and their ground truth (right). (a) beach_064. (c) sea_ice_060. (f) sea_ice_485.

C. LOCAL FEATURE ACQUISITION

To obtain more accurate local semantic information, we constructed a multiscale convolution submodule based on the cross-attention mechanism. First, the feature vector of each feature point is mapped onto a 2-dimensional space, and then a cross-attention mechanism is constructed to assign two weight coefficients to each feature. The difference of the weighted coefficients is expanded using the maximum value strategy and the point multiplication strategy. Then, multiscale convolution is used to generate local feature maps of different scales to obtain different local features. Multiscale local feature maps compensate for the limitations of single-scale features, use multiple scales to obtain precise positioning of foreground classes, reduce the total number of parameter calculations through convolution with a convolution kernel of 1, and improve the calculation efficiency of the algorithm.

D. EXPERIMENT ENVIRONMENT

This experiment uses the Keras deep learning library (version 2.1.5). The GPU server used was a Tesla V100 16G, and all experiments were performed in Python (version 3.6.4) in Windows 10. The pixel-level accurate prediction of ocean remote-sensing data is described through the sigmoid activation function, and the weight of the algorithm is updated through the Adam optimizer. The other parameters are as follows: the learning rate is equal to 0.0001; the range of the attenuation rate of the first-order matrix and the second-order matrix is 0.9 0.999 through the Adma optimizer.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION METHODS

We verify the proposed method using three subsets of the NWPU-RESISC45 dataset [30] to achieve pixel-level prediction of ocean remote-sensing. We use three features to express the original ocean remote-sensing image. We use 10% of the samples of each type of data as the training set and 90 % of the samples as the test set. We use precision, recall rate and the F1 score to evaluate the algorithm. The formula for each evaluation indicator is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP \times FN} \tag{13}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \ \text{Recall}}{\text{Precision} \times \ \text{Recall}} \tag{14}$$

where TP means that the original sample is positive, and the final judgment is positive;TN represents that the original sample is positive and the final judgment is negative; FP represents that the original sample is negative and the final judgment is negative; and FN means that the original sample is negative and finally judged as positive. Figure 3 shows the original images and labels of the three data sets.

B. PREDICTION RESULTS OF EACH MODEL ON DIFFERENT DATASETS

To verify the feasibility of the proposed algorithm, we selected classic algorithms (AlexNet, ResNet and DenseNet) and other more modern algorithms that have achieved better prediction results on remote-sensing datasets (AML, SAGP, and MAMC). We tested each algorithm in the same experimental environment, and the learning features of each model were the same. Table 1 shows the experimental results of each model, and Figure 4 shows a visual analysis of the prediction results of each model to highlight the differences in performance across the models.

1) RESULTS FROM THE BEACH DATASET

The beach data set concentrates the foreground (beach) and background (sea water) samples. Table 1 shows that the SAGP algorithm yields the highest precision on this data set, highlighting that the image volume algorithm plays a positive role in constructing the relationship between features, particularly when many samples of the same type are present. The proposed algorithm achieves the best recall, which shows that the proposed method predicts positive and negative samples best among the tested algorithms. Concurrently, the proposed algorithm also yields the best F1 score, showing that the proposed algorithm performs best. The precision of the AML, SAGP, and MAMC subcategories are all higher than those of the other three groups of comparative experiments, which highlights the feasibility of the attention mechanism. Figure 4 also shows that SAGP exhibits high accuracy on positive samples. Other models have marginally worse predictions on positive samples; however, the SAGP algorithm has the worst effect on negative samples. The proposed algorithm produces the least amount of noise during prediction of positive and negative samples, thus highlighting its superiority in overall prediction performance.

2) RESULTS FROM THE ISLAND DATASET

Figure 3 shows that the island areas are irregular; thus, the algorithm must have better edge processing capabilities.

TABLE 1. Experimental results of different models on different data sets through different evaluation criteria.

Methods	Beach(P)	Beach(R)	Beach(F1)	Island(P)	Island(R)	Island(F1)	Sea_ice(P)	Sea_ice(R)	Sea_ice(F1)
AlexNet	0.9836	0.9964	0.9901	0.8062	0.8795	0.8412	0.8534	0.9534	0.9099
ResNet	0.9811	0.9938	0.9874	0.8373	0.8473	0.8244	0.8657	0.9508	0.9063
DenseNet	0.9839	0.9935	0.9888	0.8380	0.8722	0.8547	0.8627	0.9679	0.9123
AML	0.9839	0.9966	0.9902	0.8111	0.8945	0.8507	0.8726	0.9544	0.9108
SAGP	**0.9953**	0.9574	0.9760	0.8393	**0.9364**	0.8852	0.8516	0.9777	0.9105
MAMC	0.9856	0.9918	0.9887	0.8373	0.8781	0.8572	0.8836	0.9511	0.9161
GLPO-Net	0.9916	**0.9978**	**0.9946**	**0.8529**	0.9286	**0.8891**	**0.8954**	**0.9804**	**0.9359**

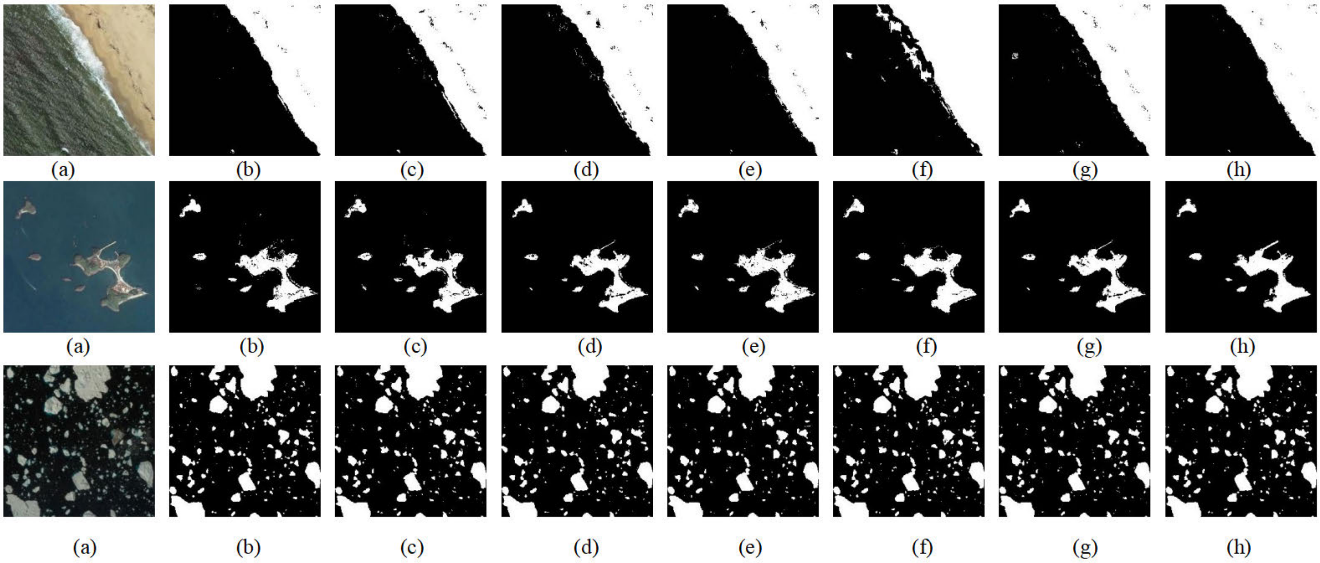

FIGURE 4. Segmentation result of beach, island, and sea_ice data set. (a) original images. (b)AlexNet. (c)ResNet. (d)DenseNet. (e)AML. (f)SAGP. (g) MAMC.

Table 1 shows that GLPO-Net yields the best prediction results with positive samples and is 1.36 percentage points higher than the second highest forecasting algorithm (SAGP). In this sample, the recall of SAGP is 0.59% higher than that of the proposed algorithm. We speculate that this result is caused by the sea being more concentrated than the islands; thus, the result of predicting the negative sample is better, while GLPO-Net's F1 score is 0.39% higher than that of SAGP, highlighting the superior overall performance of the proposed algorithm. Concurrently, DenseNet yields better accuracy when fusing deep and shallow features. The MAMC algorithm yields more features through multiple scales and can only obtain good prediction results in the prediction of irregular targets. In the second column of Figure 4, DenseNet, MAMC and the proposed GLPO-Net yield better results on edges, allowing them to predict small and prominent island boundaries more accurately. Concurrently, SAGP and GLPO-Net produced the least amount of noise for prediction in sea water.

3) RESULTS FROM THE SEA_ICE DATASET

Figure 3 shows that ice in the ocean is characterized by many samples and no concentration. Therefore, a model must obtain both global and local semantic information to perform adequately. With this data set, the proposed model achieved the best results regarding precision, recall, and F1 score. In particular, its F1 score was much higher than the other models, which was likely due to the fact that more local information was produced through multiscale convolution along with more global information through a bidirectional independent recurrent neural network and cryptographic dilation convolution. In the following experiments, we investigated the performance of each submodel of the GLPO-Net algorithm. As shown in the third column of Figure 4, the proposed algorithm exhibits better predictions, particularly for small targets

C. INFLUENCE OF THE FUNCTIONS OF THE ATTENTION MECHANISM ON THE EXPERIMENTAL RESULTS

To verify the influence of the cross-attention mechanism, its strategy and the global attention mechanism on the experimental results, we verify this mechanism's different attention mechanism modules. The specific verification process is shown in Table 2, where "MAMC_no_Horizontal" represents only the vertical attention module; "MAMC_no_Vertical" represents the horizontal attention module only; "MAMC_no_Max" represents that the cross-attention mechanism does not use the maximum strategy; "MAMC_no

TABLE 2. Experimental results of different attention mechanisms and strategies.

Methods	Beach(P)	Beach(R)	Beach(F1)	Island(P)	Island(R)	Island(F1)	Sea_ice(P)	Sea_ice(R)	Sea_ice(F1)
MAMC_no_Horizontal	0.9812	0.9934	0.9873	0.7515	0.8592	0.8017	0.8695	0.9459	0.9061
MAMC_no_Vertical	0.9796	0.9932	0.9864	0.7581	0.8440	0.7950	0.8614	0.9620	0.9089
MAMC_no_Max	0.9839	0.9968	0.9900	0.8219	0.9185	0.8675	0.8698	0.9652	0.9150
MAMC_no_Dot	0.9834	0.9947	0.9890	0.8186	0.8984	0.8567	0.8716	0.9684	0.9174
MAMC_no_Global	0.9896	0.9907	0.9901	0.8347	0.8817	0.8576	0.8828	0.9619	0.9206
GLPO-Net	**0.9916**	**0.9978**	**0.9946**	**0.8529**	**0.9286**	**0.8891**	**0.8954**	**0.9804**	**0.9359**

TABLE 3. The feasibility of each module in the GLPO-Net algorithm.

Methods	Beach(P)	Beach(R)	Beach(F1)	Island(P)	Island(R)	Island(F1)	Sea_ice(P)	Sea_ice(R)	Sea_ice(F1)
GA+CA+MC+BI	0.9883	0.9956	0.9919	0.8307	0.8779	0.8534	0.8652	0.9605	0.9103
GA+DD+MC+BI	0.9843	0.9935	0.9888	0.8438	0.9046	0.8731	0.8892	0.9641	0.9251
DD+CA+MC+BI	0.9891	0.9962	0.9926	0.8269	0.9102	0.8665	0.8642	0.9547	0.9072
GA+CA+DD+BI	0.9880	0.9946	0.9912	0.8550	0.9037	0.8514	0.8754	0.9687	0.9196
GA+DD+MC+CA	0.9857	0.9945	0.9900	0.7920	0.9159	0.8494	0.8623	0.9703	0.9131
GLPO-Net	**0.9916**	**0.9978**	**0.9946**	**0.8529**	**0.9286**	**0.8891**	**0.8954**	**0.9804**	**0.9359**

_Dot'' represents the cross-attention where the force mechanism does not use the dot product strategy; and ''MAMC_no_Global'' represents not using the global attention mechanism module. Table 2 shows the experimental results of each strategy on three subdatasets.

Table 2 shows that when only the one-way attention mechanism is used, the predicted results are worse than other attention structures. When the prediction target is smaller or more irregular, cross attention plays a stronger role. In the Island data set, the predicted precision is 1.28% and 1.61% below that of ''MAMC_no_Global'', even if there is one fewer strategy. However, in the Sea_ice data set, the proposed predicted precision is 1.12% and 2.30% below that of ''MAMC_no_Global''. The overall experimental results of the ''MAMC_no_Global'' strategy in the three datasets are lower than those of the GLPO-Net algorithm; thus, more global semantic information through the global attention mechanism can improve the accuracy of target recognition.

D. FEASIBILITY OF EACH MODULE IN THE GLPO-NET ALGORITHM

Different models describe task learning with different characteristics. In this paragraph, we describe an experimental analysis of all modules in GLPO-Net, where ''GA'' represents the global attention mechanism module; ''DD'' represents the dilation dense module; ''CA'' represents the cross-attention mechanism module; ''BI'' represents the bidirectional independent recurrent neural network; and ''MC'' represents the multiscale convolution module. The results of medical image recognition based on the pixel level are shown in Table 3.

Table 3 shows that the absence of the CA module on the Beach dataset yields the lowest accuracy in predicting positive samples. This result is likely caused by the fact that sample points are relatively concentrated, and local semantic information is more important than global semantic information. When the GA module is missing, the prediction result is higher than that of the other models; this result is verified from the negative side. Also, in the Island data set, the lack of a BI module worsens prediction; however, the prediction error on the negative samples is lowest. Therefore, the edge prediction of positive samples can be described more accurately by integrating contextual semantic information through the BI module. When the DD module is missing, the predicted probability of the worst negative sample is obtained, highlighting that the dilation dense integration module yields better learning with data that has more negative samples than positive samples; the DD module is also missing on the Sea_ice data set. The prediction results obtained by the GA module and the BI module are both poor, particularly when there are any positive samples, highlighting the need to mine and integrate global semantic information. Concurrently, the GA module is shown to play an active role in the global feature distribution weight, the BI module obtains global semantic information, and the DD module plays an active role in mining global features.

E. INFLUENCE OF DIFFERENT FEATURES ON EXPERIMENTAL RESULTS

To represent the information of each feature point, we constructed three features: SRW, LTF and HIS. In this section, we compare these three features and compare the value of each feature more intuitively using visualizations of each feature on the three sets of evaluation criteria using polar axis pie charts(As shown in Figure 5).

In this section, we use the proposed GLPO-Net algorithm to evaluate various single feature or double combination features on three sets of evaluation coefficients. These polar pie charts show that the SRW feature yields the worst pixel-level prediction but plays a critical role in the algorithm. The HIS feature contains more semantic information; thus, the Beach and Island data sets have the best predictive effect among

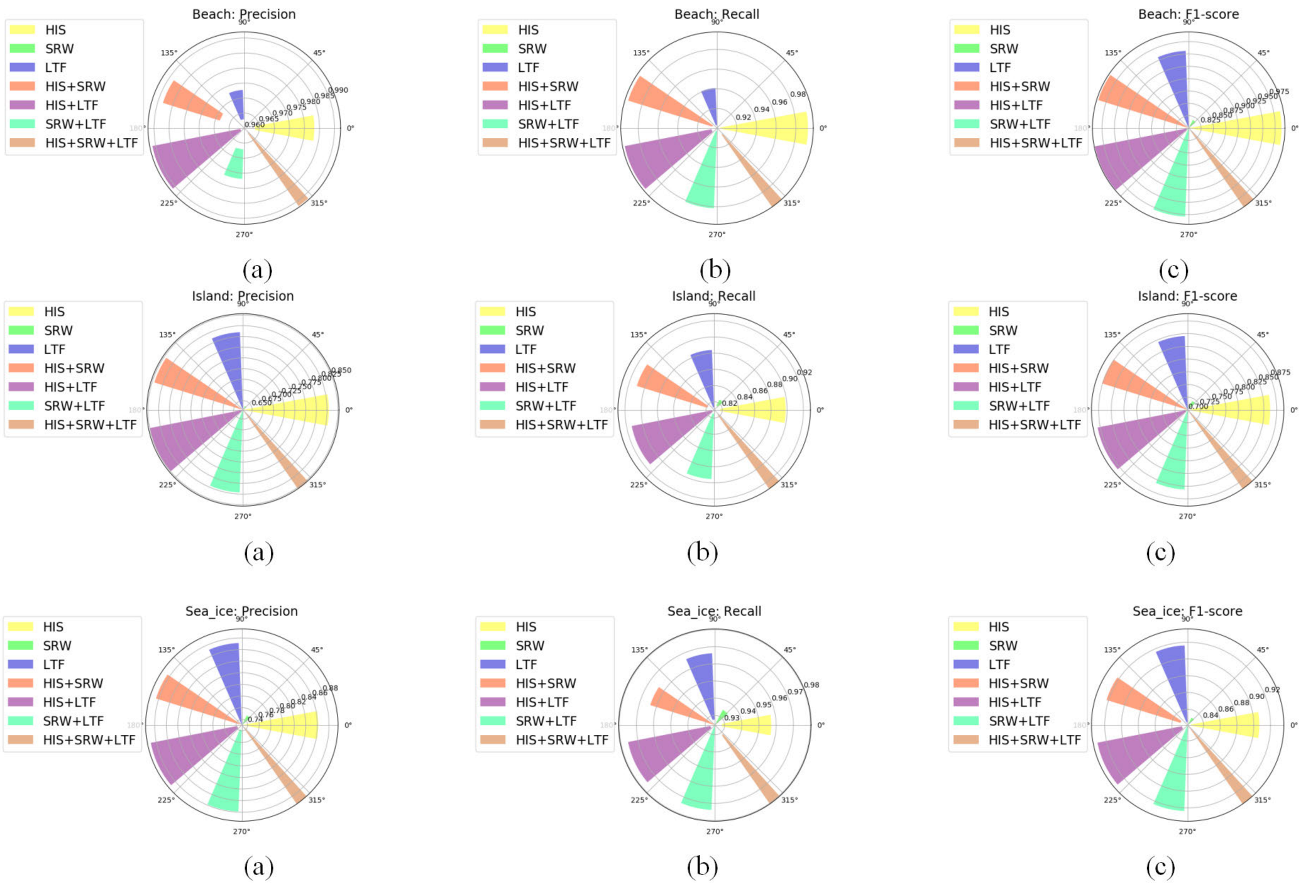

FIGURE 5. Polar axis pie chart visualization results. (a) Precision. (b) Recall. (c) F1-score.

the three features. However, the LTF feature on the Sea_ice dataset has a better predictive ability. Local features could thus be described in more detail. The fusion of any two groups of features is shown to be yield better results than that of other single-type features, which highlights that the fusion of multiple features retains more semantic information than a single feature. The SRW, LTF and HIS features we extracted from the three data sets thus all play a positive role, highlighting the feasibility of the three sets of features

V. CONCLUSION

This paper describes the mining of local features through the cross-attention mechanism model and multiscale convolution, and further expands the difference of feature information through the maximum weight strategy and the weighted point multiplication strategy. Concurrently, the global attention mechanism fuses bidirectional independent recurrent neural network, describes the extraction of global features, and eliminates many of redundant features while fully mining the global features through the dilation dense network. Finally, the proposed GLPO-Net algorithm exhibits the best performance on three public data sets, highlighting the feasibility of the proposed model.

ACKNOWLEDGMENT

The authors would like to thank the editor-in-chief, the associate editor, and the reviewers for their insightful comments and suggestions.

REFERENCES

[1] A. Stateczny, W. Kazimierski, D. Gronska-Sledz, and W. Motyl, "The empirical application of automotive 3D radar sensor for target detection for an autonomous surface Vehicle's navigation," *Remote Sens.*, vol. 11, no. 10, p. 1156, May 2019.

[2] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, Jun. 2019.

[3] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[4] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[5] X. Lu, W. Ji, X. Li, and X. Zheng, "Bidirectional adaptive feature fusion for remote sensing scene classification," *Neurocomputing*, vol. 328, pp. 135–146, Feb. 2019.

[6] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 140–152, Jan. 2020.

[7] H. Gao, L. Cao, D. Yu, X. Xiong, and M. Cao, "Semantic segmentation of marine remote sensing based on a cross direction attention mechanism," *IEEE Access*, vol. 8, pp. 142483–142494, 2020.

[8] R. Dong, X. Pan, and F. Li, ''DenseU-Net-based semantic segmentation of small objects in urban remote sensing images,'' *IEEE Access*, vol. 7, pp. 65347–65356, 2019.

[9] F. Huang, Y. Yu, and T. Feng, ''Hyperspectral remote sensing image change detection based on tensor and deep learning,'' *J. Vis. Commun. Image Represent.*, vol. 58, pp. 233–244, Jan. 2019.

[10] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, ''Deep learning in remote sensing applications: A meta-analysis and review,'' *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[11] F. A. N. Rongshuang, C. H. E. N. Yang, X. U. Qiheng, and W. A. N. G. Jingxue, ''A high-resolution remote sensing image building extraction method based on deep learning,'' *Acta Geodaetica et Cartographica Sinica*, vol. 48, no. 1, p. 34, 2019

[12] G. Chen, ''Agricultural remote sensing image cultivated land extraction technology based on deep learning,'' *Revista de la Facultad de Agronomia de la Universidad del Zulia*, vol. 36, no. 6, pp. 1–12, 2019.

[13] J. Song, S. Gao, Y. Zhu, and C. Ma, ''A survey of remote sensing image classification based on CNNs,'' *Big Earth Data*, vol. 3, no. 3, pp. 232–254, Jul. 2019.

[14] R. Fan, L. Wang, R. Feng, and Y. Zhu, ''Attention based residual network for high-resolution remote sensing imagery scene classification,'' in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2019, pp. 1346–1349.

[15] W. Li, H. Liu, Y. Wang, Z. Li, Y. Jia, and G. Gui, ''Deep learning-based classification methods for remote sensing images in urban built-up areas,'' *IEEE Access*, vol. 7, pp. 36274–36284, 2019.

[16] S. Pan, Y. Tao, C. Nie, and Y. Chong, ''PEGNet: Progressive edge guidance network for semantic segmentation of remote sensing images,'' *IEEE Geosci. Remote Sens. Lett.*, early access, Apr. 22, 2020, doi: 10.1109/LGRS.2020.2983464.

[17] C. Xu, C. Li, Z. Cui, T. Zhang, and J. Yang, ''Hierarchical semantic propagation for object detection in remote sensing imagery,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4353–4364, Jun. 2020.

[18] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, ''Multi-scale context aggregation for semantic segmentation of remote sensing images,'' *Remote Sens.*, vol. 12, no. 4, p. 701, Feb. 2020.

[19] H. You, S. Tian, L. Yu, and Y. Lv, ''Pixel-level remote sensing image recognition based on bidirectional word vectors,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1281–1293, Feb. 2020.

[20] R. Imbriaco, C. Sebastian, E. Bondarev, and P. de With, ''Aggregated deep local features for remote sensing image retrieval,'' *Remote Sens.*, vol. 11, no. 5, p. 493, Feb. 2019.

[21] C. Zhang, G. Li, and S. Du, ''Multi-scale dense networks for hyperspectral remote sensing image classification,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[22] Z. Wang, C. Zou, and X. Cui, ''Low-sample size remote sensing image recognition based on a multihead attention integration network,'' *Multimedia Tools Appl.*, vol. 79, pp. 32525–32540, Aug. 2020.

[23] Y. Yuan, J. Fang, X. Lu, and Y. Feng, ''Remote sensing image scene classification using rearranged local features,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.

[24] P. Li, P. Ren, X. Zhang, Q. Wang, X. Zhu, and L. Wang, ''Region-wise deep feature representation for remote sensing images,'' *Remote Sens.*, vol. 10, no. 6, p. 871, Jun. 2018.

[25] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, ''Exploiting convolutional neural networks with deeply local description for remote sensing image classification,'' *IEEE Access*, vol. 6, pp. 11215–11228, 2018.

[26] G. Zhang, S. Lu, and W. Zhang, ''CAD-Net: A context-aware detection network for objects in remote sensing imagery,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[27] X. Ning, K. Gong, W. Li, and L. Zhang, ''JWSAA: Joint weak saliency and attention aware for person re-identification,'' *Neurocomputing*, to be published.

[28] W. Cai and Z. Wei, ''PiiGAN: Generative adversarial networks for pluralistic image inpainting,'' *IEEE Access*, vol. 8, pp. 48451–48463, 2020, doi: 10.1109/ACCESS.2020.2979348.

[29] X. Ning, W. Li, B. Tang, and H. He, ''BULDP: Biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition,'' *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2575–2586, May 2018, doi: 10.1109/TIP.2018.2806229.

[30] W. Cai and Z. Wei, ''Remote sensing image classification based on a cross-attention mechanism and graph convolution,'' *IEEE Geosci. Remote Sens. Lett.*, early access, Oct. 1, 2020, doi: 10.1109/LGRS.2020.3026587.

[31] J. Zhou, M. Hao, D. Zhang, P. Zou, and W. Zhang, ''Fusion PSPnet image segmentation based method for multi-focus image fusion,'' *IEEE Photon. J.*, vol. 11, no. 6, pp. 1–12, Dec. 2019, doi: 10.1109/JPHOT.2019.2950949.

[32] Z. Wang, C. Zou, and W. Cai, ''Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model,'' *IEEE Access*, vol. 8, pp. 71353–71363, 2020, doi: 10.1109/ACCESS.2020.2986267.

[33] J. Zhou, D. Zhang, P. Zou, W. Zhang, and W. Zhang, ''Retinex-based Laplacian pyramid method for image defogging,'' *IEEE Access*, vol. 7, pp. 122459–122472, 2019, doi: 10.1109/ACCESS.2019.2934981.

[34] Z. Huang, Y. Zhang, Q. Li, T. Zhang, N. Sang, and H. Hong, ''Progressive dual-domain filter for enhancing and denoising optical remote-sensing images,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 759–763, May 2018, doi: 10.1109/LGRS.2018.2796604.

[35] Z. Huang, L. Huang, Q. Li, T. Zhang, and N. Sang, ''Framelet regularization for uneven intensity correction of color images with illumination and reflectance estimation,'' *Neurocomputing*, vol. 314, pp. 154–168, Nov. 2018.

[36] Z.-L. Yang, X.-Q. Guo, Z.-M. Chen, Y.-F. Huang, and Y.-J. Zhang, ''RNN-stega: Linguistic steganography based on recurrent neural networks,'' *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1280–1295, May 2019, doi: 10.1109/TIFS.2018.2871746.

[37] S. Zhang, C. Lu, S. Jiang, L. Shan, and N. N. Xiong, ''An unmanned intelligent transportation scheduling system for open-pit mine vehicles based on 5G and big data,'' *IEEE Access*, vol. 8, pp. 135524–135539, 2020, doi: 10.1109/ACCESS.2020.3011109.

[38] W. Cai and Z. Wei, ''Diversity-generated image inpainting with style extraction,''' 2019, *arXiv:1912.01834*. [Online]. Available: https://arxiv.org/abs/1912.01834

[39] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, ''TARDB-Net: Triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification,'' *Multimedia Tools Appl.*, to be published, doi: 10.1007/s11042-020-10188-x.

[40] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, ''A novel multi-modality image fusion method based on image decomposition and sparse representation,'' *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.

[41] C. Saravanan, ''Color image to grayscale image conversion,'' in *Proc. 2nd Int. Conf. Comput. Eng. Appl.*, vol. 2, Mar. 2010, pp. 196–199.

[42] V. Chernov, J. Alander, and V. Bochko, ''Integer-based accurate conversion between RGB and HSV color spaces,'' *Comput. Electr. Eng.*, vol. 46, pp. 328–337, Aug. 2015.

[43] G. Cheng, J. Han, and X. Lu, ''Remote sensing image scene classification: Benchmark and state of the art,'' *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

HAO GAO received the M.S. degree from Newcastle University, Newcastle, U.K., in 2013. He is currently pursuing the Ph.D. degree with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao, China. He has been with the Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), since 2014. His research interests include marine environmental monitoring, big data analysis, remote sensing digital image processing, and deep learning techniques.

XUEJUN XIONG received the Ph.D. degree from the Ocean University of China, Qingdao, China. He is currently a Researcher with the First Institute of Oceanography, Ministry of Natural Resources. His research interests include regional oceanography, ocean dynamics, ocean engineering, and marine survey equipment and technology.

LIN CAO received the M.S. degree from Nankai University, Tianjin, China, in 2014. She is currently with the Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences). Her research interests include digital image processing, machine vision, computational intelligence, and deep neural network.

DINGFENG YU received the M.S. and Ph.D. degrees from the University of Chinese Academy of Sciences, China, in 2013. He is currently an Associate Fellow with the Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao, China. His research interests include ocean remote sensing monitoring, ocean spectral measurement technology, and satellite remote sensing calibration and validation.

GUANGBING YANG is currently with the Key Laboratory of Marine Science and Numerical Modeling, First Institute of Oceanography. He is involved in research in acoustics, oceanography, and acoustic engineering. His current project is on Observation of the effect of bottom water temperature change on shallow marine sediment in weather process.

LEI YANG was born in Yantai, Shandong, China, in 1988. He received the B.S. degree in communication engineering and the M.S. degree in information and communication engineering from the Harbin Institute of Technology, in 2011 and 2013, respectively. From 2013 to 2016, he was a Research Assistant with The Institute of Oceanographic Instrumentation, Shandong Academy of Sciences, where he has been an Engineer since 2017. His research interests include data analysis and processing, system design, remote sensing, and array signal processing.