Ship Identification via Adjacent-Branched Saliency Filtering and Prior Representation-Based Classification

Jianming Hu, *Student Member, IEEE*, Tianjun Shi, Shikai Jiang, *Student Member, IEEE*, Xiyang Zhi, Xiaogang Sun, and Wei Zhang

***Abstract*—Ship identification in optical remote sensing images is essential for a wide range of civil and military applications, including maritime rescue, port management, and sea area surveillance. However, current studies focus mainly on ship detection or coarse-grained ship size identification, rather than fine-grained type identification. Moreover, interference from clouds and port facilities as well as complex conditions such as occlusion and shadows increase the difficulty of ship type identification. To address these problems, we propose a novel ship identification method by employing adjacent-branched saliency filtering and prior representation-based classification strategies, which achieves high-precision type recognition performance for large and medium-sized ships under complex environmental interference conditions. In the candidate region extraction stage, a multiscale feature aggregation structure that utilizes feature map fusion in adjacent layers and receptive field mining within the same extraction branch is presented, providing fine representation of the location and edge characteristics of ship targets in complex scenes. In the classification stage, the low-rank term describing interclass differences and the graph-based regularization term describing intraclass differences are added to the representation model as prior constraints, which can correctly classify ships in the presence of complex environmental interference such as occlusion and shadow. Experimental results on two high-quality ship datasets indicate that the proposed method realizes state-of-the-art identification performance compared with benchmark methods.**

***Index Terms*—Adjacent-branched saliency filtering, complex scene, optical remote sensing, prior representation, ship identification.**

I. Introduction

Maritime ship identification is of great significance for automatic fishery management, port rescue, marine traffic maintenance applications [1]. The accuracy of identification technology directly determines the safety and timeliness of military and civilian applications [2]. Visible light images have rich details and textures, obvious target structure features, and their information contained is intuitive and easy to understand, which conforms to the daily observation habits of the human eye. Therefore, the visible light remote sensing image has become an important image data source for ship detection and identification [3]. However, due to the wide-area imaging characteristics of remote sensing images, there are usually various environmental factors such as clouds and various port facilities in the image scene. Ship identification in complex optical image scenes is still a challenging task.

For ship identification in complex scenes, researchers have proposed various strategies and models [4]. Visual saliency model is one of the representative works [5]. It imitates the attention mechanism of human eyes and can quickly locate abnormal areas or points in complex scenes, having a wide prospect in massive data processing applications. Early visual saliency models applied traditional image attributes to highlight salient areas, such as contrast, background prior, and compactness characteristics. On this basis, various extraction strategies are adopted to achieve more accurate contour highlighting of objects, such as frequency domain analysis, cellular automata, random walk, and Bayesian theory. These methods have obtained high-performance object detection and identification performance for specific scenes. However, the approaches via traditional strategies have a good effect for relatively simple close-up imaging scenes, it is easy to produce a large number of false alarms when directly applied to large-scale remote sensing scenes. Therefore, it is urgent to present an effective strategy to solve this issue.

As we know, the deep learning networks represented by CNN has a strong ability to represent high-level semantic features [6], which provide an effective framework for remote sensing image ship detection and identification [7]. With the popularity of deep learning technology, the saliency methods based on deep learning network have significantly improved the performance of region extraction and ship classification. In order to make the network suitable for detecting ship targets of different scales and types, feature pyramid network (FPN) [8] is widely used in the feature extraction stage of learning-based detection models. In

Manuscript received 3 June 2024; revised 14 July 2024; accepted 26 July 2024. Date of publication 30 July 2024; date of current version 5 September 2024. This work was supported by National Natural Science Foundation of China under Grant 62305088 and China Postdoctoral Science Foundation under Grant 2023M740900. *(Corresponding author: Shikai Jiang.)*

Jianming Hu, Tianjun Shi, Shikai Jiang, Xiyang Zhi, and Wei Zhang are with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: hujianming@hit.edu.cn; 22B921004@stu.hit.edu.cn; jiangshikai@hit.edu.cn; zhixiyang@hit.edu.cn; wzhang@hit.edu.cn).

Xiaogang Sun is with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: sxg@hit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3435675

addition, many subsequent detection models have added connection channels between feature extraction branches based on the FPN to fuse and represent multilevel target features. However, the semantic gap between features of different levels is seldom considered in the fusion of the current networks. In fact, the fusion of feature maps with large resolution differences is easy to introduce unnecessary background noise, and the fusion of too many features may reduce the efficiency of the network in learning target features, leading to the nonoptimal representation of ship target features. In addition, the classification modules of the existing detection networks are rarely designed in combination with the characteristics of the remote sensing image scene, resulting in poor algorithms robustness for conditions such as object occlusion and shadow. Therefore, it is urgent to design and optimize the ship detection network structure in combination with the characteristics of image scenes.

To address these issues, a new ship identification model via adjacent-branched saliency extraction and prior representation-based classification (PRC) is proposed. Specifically, at the saliency extraction stage, we present an interactive fusion module that only merges feature maps from adjacent layers to effectively address semantic gap issues caused by significantly different resolutions, achieving spatial consistency of multiscale features. In addition, we design a multiscale information mining module that utilizes dilated convolutions with different dilation rates on the same extraction branch, enhancing the perception of suspected target regions and suppress background interference. On this basis, combining the characteristic prior knowledge of the ship targets, a ship type classifier that is robust to complex environmental disturbances such as occlusion and shadows is constructed. Experimental results on high-quality public dataset and self-built dataset demonstrate that the proposed algorithm has excellent ship type recognition accuracy in complex interference scene.

The main contributions of this work are summarized as follows:

1) A ship identification framework combining adjacent-branched saliency extraction and prior representation-based classification is proposed, which can achieve high-accuracy type recognition performance of typical ship targets in complex sea and port scenes.
2) A novel multiscale feature aggregation module is proposed, which utilizes the feature map fusion of adjacent layers and the mining of different receptive fields in the same extraction branch to accurately represent the location and edge characteristics of the ship target.
3) A prior representation-based target classification module is proposed, which can correctly classify ships in the presence of complex environmental interference such as occlusion and shadow.

II. Related Research

In this section, we first briefly introduce the development of the salient object detection technology and point out the shortcoming of the current methods. On this basis, we illustrate the solution of our method. We then review the use of typical discriminant classifiers and clarify the difference between the proposed approach and other representative classification methods.

A. Saliency Detection Model

Ship detection and identification in complex scenes has attacked considerable concern in the remote sensing field [9]. In [10], research scholar proposes a multidirectional ship detection via dynamic soft label assignment strategy. In [11], researchers improve the accurate direction capture of ships through active rotating filters and complementary reconstruction of feature maps. In [12], a foreground-aware feature map reconstruction network is presented by calculating the weighted distance of foreground weights, which can achieve accurate classification under few-shot conditions. The visual attention mechanism of the eyes is an important component of visual information processing. By suppressing negligible stimuli and quickly focusing on important regions in the image scene, it helps rapidly search and locate the target of interest in complex background, such as the sparsely distributed ships. Inspired by the human visual attention mechanism, researchers have proposed several visual saliency models for ships detection [13]. In [14], researchers apply dual mask attention based on multidirectional feature fusion, refining target features while suppressing background clutter interference. In [15], researchers design a feature fusion structure that balances the receptive fields of different backbone layers and utilizes contextual attention strategies to enhance target feature learning, ultimately achieving high confidence recognition of targets in complex scenes. These models concentrate on mining the important and valuable information in the salient regions, thus timely realizing the ship location in complex scenes.

Visual saliency models can be broadly classified into top-down methods and bottom-up methods [16] according to the human eye visual attention mechanism. The former top-down methods are task-driven, which require high-level priori information such as the specific scene, objects with clear status and unique observation condition. Such approaches belong to an advanced cognitive process with complex modeling representations, which require a large amount of computing resources. The latter bottom-up methods are driven by data, focusing on mining the feature differences between pixels and surrounding areas in terms of color, luminance, edges, etc. These methods utilize the distinct characteristics in the spatial and frequency domain at multiple scales as the basis for the final saliency results. In [17], the phase spectrum of Fourier transform with low computation complexity is applied to draw the saliency map and a homogeneous filter is followed to obtain the suspected regions. In [18], the phase saliency map combined with extended wavelet transform to improve the ability of candidate extraction in the complicates scenes. In [19], a combined saliency model is proposed with self-adaptive weights generate the regions of interest in the maritime background. In [20], a visual saliency method is constructed to prescreen the ship candidates by the statistical characteristics difference. It can be seen from the above methods that the performance of traditional visual saliency models relies heavily on the hand-crafted feature descriptors. They can only achieve promising detection results for certain special scenes, and may be less effective in complex scenes such as ports containing various artificial facilities. The rise of deep learning network makes the feature extraction no longer rely on artificially designed descriptors. The traditional operators are gradually replaced by convolutional neural networks (CNNs) with stronger feature characterization ability, further promoting the development of visual saliency detection. Almost all CNN-based saliency detection methods can be viewed as an encode–decode structure with supervised training in the form of pixel-level truth annotation and end-to-end implementation. DHSNet [21], as a typical saliency detection model, captures the salient objects from a global perspective and then refines the details through a hierarchical recurrent convolution neural network. Subsequent approaches have been developed based on the encode–decode structure, but gradually incorporate the multiscale structures, multilevel structures and attention structures [22] to address the problems of scale variant, boundary refinement, and false alarm suppression. In [23], a multilevel feature aggregation network is put forward to fuse multilevel feature maps into multiple resolutions. In [24], a channel and spatial attention mechanism (SAM) is embedded in the encode–decode architecture to align the context information among the multiscale feature maps. In [25], a pyramid attentive module and a salient edge detection module are proposed to improve the multiscale feature extraction capability and sharpen the boundary representation, respectively.

Although the above methods can enhance target features in specific applications, they ignore the semantic gap between different extraction branches in feature integration, thus limiting the optimal representation of target features to some extent. Different from other works, we propose an interactive feature fusion module that only fuses the feature maps of adjacent layers to effectively address the semantic gap caused by the resolution difference, which is guided by the channel attention mechanism (CAM) to enhance the details and position information at different levels. Moreover, we design a single feature branch information mining module. It applies several dilated convolutions with different expansion rates to enlarge the receptive field, and employs the SAM to improve the perception ability of candidate regions and suppress the background interference.

B. Classifier in Object Detection Methods

Based on the ship suspected candidates extracted in the first stage, the following stage comes to achieve ship discrimination through the classification model. The traditional methods generally utilize the feature descriptors to extract the shape, material, structure, and other general features, and then apply the classifier to realize the object classification. The most widely used classifiers are support vector machines (SVM) [26], sparse representation-based classification (SRC), and improvements based on the above models.

SVM, as a typical classifier, has a sound theoretical foundation and is suitable for processing small sample data. In [27], a spatial bag of visual words (BOVW) model is applied to describe the object characteristics with scale-invariant feature transform (SIFT) keypoints, and the final decision is made by SVM. In [28], the composite kernel support vector is proposed to improve the features fusion quality, which is composed of the shape and texture features. In [29], multiple features representation methods are combined such as multiorientation Gabor filters, fisher vector, and BOVW to achieve the optimized feature, then the SVM is adopted to give the posterior-probability estimation. SVM relies heavily on the representation ability of feature extraction operators. However, there are large intraclass differences in object characteristics affected by lighting conditions and environmental factors, which makes SVM hard to adapt the changeable task scenes. As an efficient multiclass classifier, the sparse representation (SR) based methods have high robustness and generalization. At the same time, the methods are not sensitive to occlusion and have the potential to identify multiple types of targets in complex scenes. In [30], the discriminative sparse coding model is constructed with the consideration of within-class difference and between-class difference to improve the feature representation ability. In [31], K-singular value decomposition algorithm is used to achieve the sparse coding with the SIFT descriptors. In [32], the structured SR model with low dimensions is put forward to realize a promising performance for the inshore ship detection. The above methods can only be applied to coarse-grained classification and cannot realize fine-grained recognition. Therefore, we propose a SR model that fully mine the differential and low-rank characteristics of multiclass targets. Through integrating the regularization constraints of the typical ship features, high-precision ship type discrimination can be achieved even in complex port scene images.

III. Proposed Method

A. Method Overview

As illustrated in Fig. 1, we provide the architecture diagram of the proposed method, which consists of two-stage components: adjacent-branched mining based saliency filtering (ABSF) stage and PRC stage. By applying the first stage, we mine the deep semantic information of the input images combining with the training of a large amount of data, extracting the multiscale suspected target salient regions. To suppress the false alarm generated by the saliency extraction part, we integrate the target prior information into the SR model in the second stage, realizing the ship identification with high accuracy and low false alarm rate (FAR) in complex port scenes.

B. Adjacent-Branched Mining Based Saliency Filtering

The premise step of ship identification is to extract the suspected target region, but the scale variation and various types of ships make it tough to accurately describe the target characteristics. Considering that the deep learning network has the advantage of extracting features at multiple levels, we adopt the typical FPN to obtain the target features at different scales. However, the conventional FPN usually fuses the features of all levels directly. Due to the semantic gap and structural property difference between the feature maps on each extraction branch,

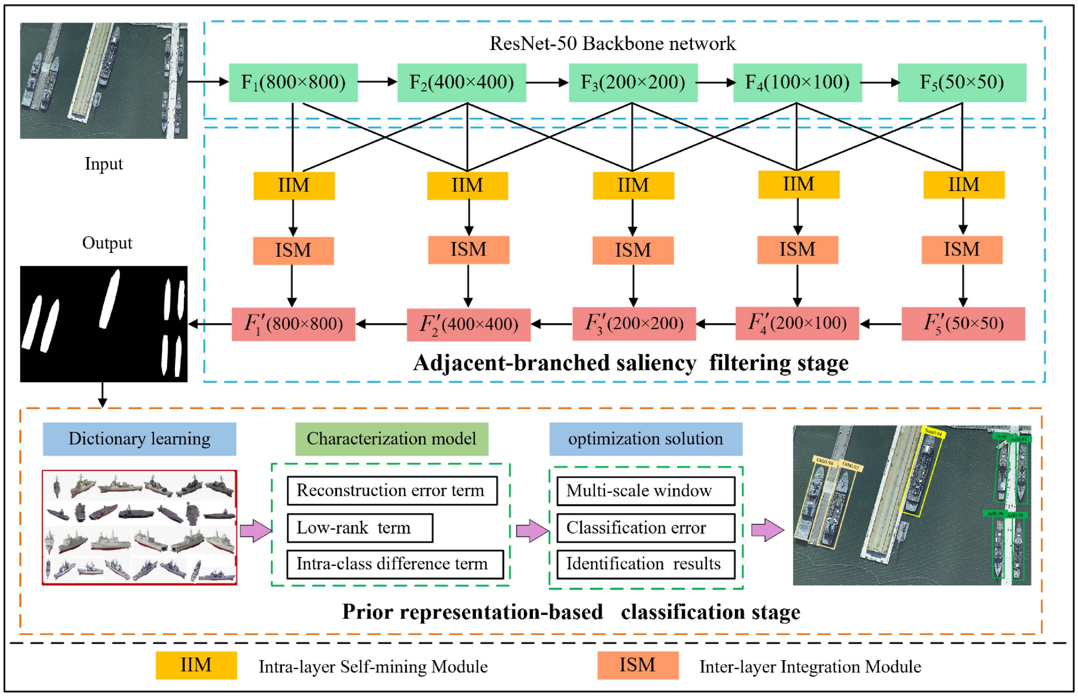


Fig. 1. Flow of the proposed method.

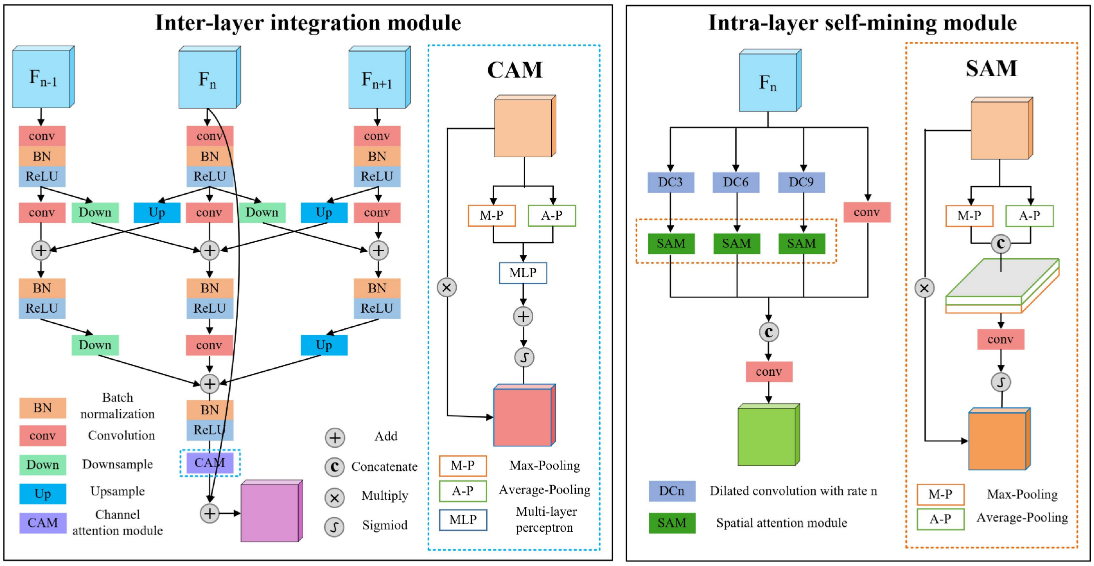


Fig. 2. Structure of two proposed modules.

the direct fusion is hard to obtain the optimal representation results. Moreover, the fusion also ignores to mine the mutual interaction between the interlayers, which reduces efficiency of information utilization.

To address the multiscale representation problem in ship identification, as shown in Fig. 2, two modules are proposed to mine the interlayer and intralayer information in the salient region extraction stage. The first module, interlayer integration module (IIM), enhances the edge details and semantic information of different feature extraction branches by fusing the feature maps of adjacent levels. The second module, intralayer self-mining module (ISM), mines the feature information of different receptive fields in the same extraction branch to better capture the scale change of the target.

For the feature maps of different levels extracted by the backbone network, due to the diversity of resolution, the edge details and semantic levels in each branch are different. Consequently, the direct cross-level fusion is obviously not the optimal representation of the ship target. Based on this consideration, the proposed IIM only integrates adjacent levels. On the one hand, this operation can reduce the computing cost and information redundancy caused by too much map fusion. On the other hand, it reduces the fusion difficulty and noise interference caused by resolution difference. The fusion between adjacent feature maps can deepen the detailed information at the shallow levels and enhance the semantic information at the deep levels.

Specifically, suppose that the feature maps of three adjacent levels extracted by the backbone network are F_{n-1}, F_n, and F_{n+1}, where $n \in [2, 3, 4]$. We first perform convolution, batch normalization, and ReLU operations on the feature maps to reduce the channel dimension

$$F_i^1 = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{conv}\left(F_i\right)\right)\right), i = n-1, n, n+1 \tag{1}$$

where conv, BN, and ReLU represent the convolution, batch normalization, and a kind of nonlinear activation process, respectively.

Then, by performing the downsampling and upsampling operations on F_{n-1}^1, F_{n+1}^1 respectively, we make the maps of two branches be the same resolution scale with F_n^1, thus the fusion of multibranch feature maps can be written as

$$F_i^2 = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{conv}\left(F_n^1\right) + \mathrm{Ds}\left(F_{n-1}^1\right) + \mathrm{Us}\left(F_{n+1}^1\right)\right)\right) \tag{2}$$

where Ds and Us represent the downsampling and upsampling operations, respectively.

Similarly, the fusion of other branches F_{n-1} and F_{n+1} can be calculated as

$$F_{i-1}^2 = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{conv}\left(F_{n-1}^1\right) + \mathrm{Us}\left(F_n^1\right)\right)\right) \tag{3}$$

$$F_{i+1}^2 = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{conv}\left(F_{n+1}^1\right) + \mathrm{Ds}\left(F_n^1\right)\right)\right). \tag{4}$$

The interaction of adjacent feature layers can improve the edge details and semantic representation in different levels. Moreover, the fusion of multilevels F_i^3 is carried out on the basis of the enhanced features. It is worth putting out, a residual learning architecture is introduced to improve the efficiency of network optimization

$$F_i^3 = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{conv}\left(F_i^2\right) + \mathrm{Ds}\left(F_{i-1}^2\right) + \mathrm{Us}\left(F_{i+1}^2\right)\right)\right) \tag{5}$$

$$F_i^4 = \mathrm{ReLU}\left(\mathrm{BN}\left(F_i^3\right) + F_i^1\right). \tag{6}$$

Since each channel of the fused feature contribute differently to the final ship recognition, inspired by CBAM [33], we apply the CAM to assign different weight coefficients along the channel dimension. The CAM automatically obtains the significance of each channel through learning, and strengthens the edge details and semantic information after reassigning channel weights. We aggregate spatial dimension information through average pooling and global pooling, and then achieve the weight vectors through shared multilayer perceptron (MLP). The final output F_i^{output} is calculated as

$$\begin{aligned} F_i^{\mathrm{output}} = \; & \mathrm{ReLU}\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(F_i^4\right)\right)\right. \\ & \left. + \mathrm{MLP}\left(\mathrm{MaxPool}\left(F_i^4\right)\right)\right). \end{aligned} \tag{7}$$

Due to the scale diversity of ships, there are differences in the salient regions size and feature representation corresponding to various targets. In order to make better use of the multilevel feature map information fused by the IIM part, we propose the ISM part to optimize the characteristics of different receptive fields. Unlike the SPP and ASPP networks which are only added at the deepest level, the ISM part processes the feature maps at all levels after fusion, so as to better mine the information with different scales and ensure the feature representation ability for diverse target types.

Specifically, suppose the input feature map in the ISM part is F_j. In order to gain a variety of receptive fields, we apply dilated convolution to extract the multiscale features. Actually, the dilated rate parameters help the conventional convolution expand the receptive field, and make the network have better feature extraction ability without extra computation complexity. Here, we use the 3 × 3 dilated convolution with the dilated rate of 3, 6, and 9, respectively. The dilated convolution with the rate of *n* is formulated as

$$F_j^{1n} = \mathrm{dconv}_n\left(F_j\right) \tag{8}$$

where dconv means the dilated convolution operation.

In fact, due to the complexity of the distribution of port scene elements, partial shadows and shore facilities may be extracted as the salient regions of the ship. Therefore, it is necessary to allocate different significant contributions to regions in different positions. Based on this consideration, we introduce the SAM [33] to better utilize the extracted multiscale features. By generating a position-weighted mask, specific regions of interest are enhanced with irrelevant background regions attenuated. We explore the spatial correlation between features by using average pooling and max pooling along the channel dimension. Then a 7 × 7 convolution is used to embed the neighboring information into the weights. The feature map obtained after the dilated convolution and the SAM operation is computed as

$$F_{tmp} = \mathrm{concat}\left(\mathrm{MaxPool}\left(F_j^{1n}\right), \mathrm{AvgPool}\left(F_j^{1n}\right)\right) \tag{9}$$

$$F_j^{2n} = \mathrm{ReLU}\left(\mathrm{conv}\left(F_{tmp}\right)\right) \tag{10}$$

where concate$(\cdot)$ represents the concatenation of the corresponding feature maps.

Finally, the features of different scales are fused by channel splicing and the 1 × 1 convolution operation. It is worth noting that we additionally introduce a 1 × 1 convolution branch to ensure the original scale information

$$F_j^{\mathrm{output}} = \mathrm{conv}\left(\mathrm{concat}\left(F_j^{23}, F_j^{26}, F_j^{29}, \mathrm{conv}\left(F_j\right)\right)\right). \tag{11}$$

After the above steps, we apply the minimum bounding rectangle to mark all potential salient regions. It should be noted that this component can generate potential target regions like the classical region proposal network (RPN), but the RPN structure is relatively simple, with only two branches that do not involve feature refinement and complementarity. Unlike it, the proposed ABSF adopts the neighboring feature fusion strategy to effectively utilize semantic information of multilevel feature maps, improving the representation ability of the target. In addition, considering that each channel of the fused features contributes differently to the final ship recognition, we introduce the attention mechanism to reassign weights to the extracted channels, enhancing the edge details of the target meanwhile suppressing the interference of background clutter. Both strategies can improve the binary classification ability of potential target regions, thus enabling more accurate segmentation of targets and backgrounds than the original RPN.

C. Prior Representation-Based Classification

After obtaining the region slices of the suspected ships, it is necessary to design a discriminant classifier to effectively distinguish different types of ships. As we know, the direct feature extraction on different slice images easily leads to insufficient feature discrimination ability, that is, the extracted features may be effective for some specific targets and difficult to distinguish other types of targets. Therefore, the idea of direct feature extraction has limitations in the multiclass ship identification tasks.

Aiming at the multiclass ship identification in complex port scene, we propose a SR based strategy to design the multiclass ship classifier. The essential idea of the proposed approach is to construct a mathematical representation model based on the sparse structure of the solution, realizing the signal decomposition under the constraint of sparse regularization. In fact, the typical SR theory is proved to be independent on a specific powerful feature. Moreover, this method shows excellent robustness to complex environmental conditions, such as occlusion and low contrast. In addition, this method is easy to introduce the prior information of the target, which is conducive to improving the accuracy of identification. Based on the above advantages of SR theory, we seek to design new regularization constraints combining with the port application scene and realize the accurate discrimination of ship types in complex environment.

The typical SR method is to find an overcomplete dictionary for the samples with ordinary dense representation through task learning (as shown in Fig. 3), and transform the samples into the form represented by a few atoms. Given a training sample set $X = [x_1, \ldots, x_i, \ldots, x_p]$, the SR on the training set is regarded as a mathematical optimization problem, which is defined as

$$\min_{B,\alpha_i} \sum_{i=1}^{p} \|x_i - B\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \tag{12}$$

where x_i represents the ith sample in the training set, B is a dictionary matrix that needs to be learned, α_i is the sparse coding coefficient, and λ is a regular parameter greater than 0. Actually, $\|x_i - B\alpha_i\|_2^2$ represents the l_2 norm which constrain the reconstruction error and $\|\alpha_i\|_1$ is the l_1 norm constraint on the coefficient α_i.

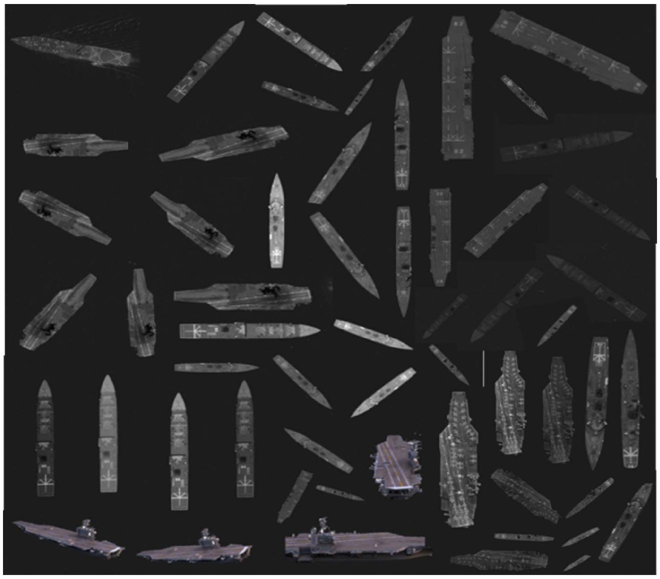


Fig. 3. Typical samples for dictionary learning.

Given a test data $\hat{x}$, according to the ordinary SR method, the form of SR is

$$\hat{\alpha} = \arg\min_{\alpha} \|\hat{x} - B\alpha\|_2^2 + \lambda\|\alpha\|_1. \tag{13}$$

The classification error on the test data is

$$e_k(\hat{x}) = \|\hat{x} - B\hat{x}_k\|_2^2 \tag{14}$$

where $\hat{x}_k$ is the coefficient vector related to the kth class target.

For multiclass target identification, especially the similarity between some ship categories is high, such as different types of aircraft carriers, the discrimination information between different types directly affects the precision of the representation model. Moreover, the ordinary reconstruction error term is described only by the l_2 norm $\|x_i - B\alpha_i\|_2^2$ directly, which may be difficult to fully describe the fitting error between data, thus it is significant to establish an improved regularization term. Consequently, we propose a data training based reconstruction error regularization term t_{err}. By constraining the projection vector via a semisupervised training strategy, we reduce the discrimination error of the projection representation model on the training set, thus improving the discrimination ability of the learned projection vector between ships and the environmental background, and at the same time improving the learning efficiency of the projection vector.

In addition, considering that the artificial facilities in the port scene have a large amount of redundant information, to reduce the complexity of calculation, we present a low-rank regularization term t_{lr} based on the local characteristics of the background of the training images. Besides, due to the possible occlusion and illumination changes in the port scene, ships of the same class may show diversified characteristics. Therefore, it is essential to consider a regularization term and learn the intraclass difference. we introduce a new discriminant regularization term t_{id}. Based on the above analysis, then the SR optimization issue can be extended to

$$\min_{B,\alpha_i} \{\lambda_1 t_{\text{err}}(x_i, B, \alpha_i) + \lambda_2 \|\alpha_i\|_1 + \lambda_3 t_{lr}(\alpha_i) + \lambda_4 t_{id}(\alpha_i)\} \tag{15}$$

where λ_1, λ_2, λ_3, and λ_4 are tradeoff parameters to adjust the impact of each item.

As for the reconstruction error term $t_{err}(x_i, B, \alpha_i)$, inspired by the research work [30], to ensure the fitting accuracy of SR, three reconstruction error constraints should be met in the reconstruction process. First, the sample x_i should be well represented by the dictionary B. Second, x_k, the sample of the kth category, can be well expressed by dictionary B_k. Third, x_t, the sample of the tth category, should not be well expressed by dictionary B_k. Therefore, the reconstruction error constraint is defined as

$$t_{eer}\,(x_i, B, \alpha_i) = \|x_i - B\alpha_i\|_F^2 + \sum_{k=1}^{N} \|x_k - B_k\alpha_k\|_F^2 + \sum_{k=1}^{N} \sum_{t=1, t\neq k}^{N} \|B_t\alpha_k\|_F^2 \tag{16}$$

where $\|\cdot\|_F^2$ is the Frobenius norm and N is the number of ship types to be distinguished. Obviously, by applying the supervised constraint of the reconstruction error, we make the representation model learn the interclass distinguishability.

As for the low-rank regularization term, we apply this constraint to the final coding coefficient α_i, making the model learn the shared atoms from different directions and reduce the influence of discriminant atoms from specific category dictionary. Nevertheless, the rank minimization issue is a NP hard problem mathematically, we then apply the trace norm $\|\alpha_i\|_*$ to solve it indirectly.

With regard to the constraint of intraclass difference, we adopt a graph-based regularization term to solve this issue

$$\min_{G} \sum_{p,q} G_{p,q} \|\alpha_{i,p} - \alpha_{i,q}\|_F^2 = \frac{1}{2}\text{trace}(\alpha^T L\alpha) \tag{17}$$

where $\alpha_{i,p}$ and $\alpha_{i,q}$ are two samples in the same class. $G_{p,q}$ represents the similarity graph of these two samples. Moreover, when $G_{p,q} \to \infty$, $\|\alpha_{i,p} - \alpha_{i,q}\|^2 \to 0$. L is the Laplacian matrix by $L = D - G$ and $D = \text{diag}(\text{sum}(G, 1))$. Through this regularization term, we can utilize the local similarity structure of samples to learn intraclass differences. Moreover, this learning can further enhance the optimization and updating of the representation model.

When dealing with the slices generated by saliency network, we need to apply the multiscale matching idea to classify each slice image. Suppose the scale set $S = [1, 2, \ldots, s]$, the learning-based representation model of different types of ships is $\psi_{i=1,\ldots,N}$.

Then, we obtain the category of test slice image by calculating the classification error

$$L_{sr,\text{class}} = \arg\min_{k} e_k(\psi, x_S). \tag{18}$$

Fig. 4. Representative samples in the considered dataset.

We define the overall loss function of the proposed network as

$$L_{\text{total}} = L_{\text{sre,class}} + L_{\text{sre,reg}} + L_{\text{sr,class}} \tag{19}$$

where $L_{\text{sre,class}}$ and $L_{\text{sre,reg}}$ represent the class prediction loss and the position regression loss in the salient region extraction stage, respectively. $L_{\text{sr,class}}$ represents the class prediction loss generated by the SR-based classification stage. By constraining the overall loss of training images, we make the proposed network learn the features of different types of ships that are conducive to class recognition.

IV. Experimental Results

A. Experimental Settings

1) Datasets Description: In order to verify the performance of the proposed method, two datasets are selected for performance analysis. The first one is the public high-quality HRSC2016 dataset [34], and the second one is a dataset we established based on typical optical satellite images. Moreover, inspired by the HRSC2016 and FGSCR [35] datasets, the scenes we selected for the second dataset images are mainly famous military ports such as Norfolk Naval Base, San Diego Naval Base, and Murmansk harbor. The second datasets contain 2212 optical remote sensing images and 3067 ship instances in total, of which 1565 images are obtained from Google Earth, and the rest are captured from Worldview-3, Jilin-1, and Pleiades satellite images. The image resolution of the two datasets is about 0.4 to 2 m and the image size is mainly distributed in 800 × 800 to 1800 × 1800 pixels. Some typical samples of the two datasets are shown in Fig. 4. Considering the number and relative uniformity of ship instances, we only label eight representative categories of large and medium-sized warships. The categories and instance numbers of selected ships are shown in Table I.

2) Evaluation Metrics: In our method, the extraction of saliency map and the classification of suspected target slices are the key steps. Therefore, we first conduct ablation experiments to evaluate the effectiveness of these two modules. The widely used precision and recall metrics are employed to measure the

TABLE I
CATEGORIES INFORMATION AND INSTANCE NUMBERS OF TYPICAL SHIPS

Ship type	Abbreviation	Ship length	Instance number
Nimitz-class aircraft carrier	Nimi	333	565
Kuznetsov-class aircraft carrier	Kuzn	306	328
Midway-class aircraft carrier	Midw	307	252
Tarawa ship	Taraw	250	260
Ticonderoga-class cruiser	Tic	173	622
Arleigh Burke-class destroy	Burke	154	624
San Antonio-class dock ship	Anto	209	420
WhidbeyIsland-class landing ship	Whid	186	104
Perry-class frigate	Perry	136	204
Submarine	Sub	170	168
Total	-	-	3547

performance of saliency model. These metrics are defined as

$$\text{precision} = \frac{|S \cap G|}{|S|} \tag{20}$$

$$\text{recall} = \frac{|S \cap G|}{|G|} \tag{21}$$

where S and G represent the predicted saliency map and ground truth.

In order to measure the overall performance of ship identification, the mean average precision (mAP) and the FAR [3] are employed to measure the overall performance of the proposed algorithm. They mainly characterize the application ability of the algorithm from the accuracy and error probability terms. The two metrics are defined as

$$\text{AP} = \frac{N_{ci}}{N_r} \times 100\%, mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{22}$$

$$\text{FAR} = \frac{N_{df}}{N_{dc}} \times 100\% \tag{23}$$

where K is the number of ship categories, N_{ci} and N_r are the correctly identified ship number and the real ship number, respectively. Similarly, N_{df} and N_{dc} are the detected false alarm number and the detected candidate number, respectively.

3) Implementation Details: All the experiments are carried out on the workstation with an Nvidia RTX 2080 GPU. The implementation of the proposed algorithm is based on PyTorch framework. The typical pretrained ResNet-50 [36] is selected as the backbone to extract different levels of features of the input image. During the training process, the epoch we set is 300 and the train batch size is 4. In our experiments, the stochastic gradient descent (SGD) strategy is applied to calculate the loss function in the training stage and update the model parameters. The learning rate starts from 0.001, and the weight decay and momentum are assigned to 0.0005 and 0.9, respectively. Besides, the training of classifier is mainly based on the slice images of FGSCR dataset. In order to improve the generalization of images, we have performed data augmentation operations on these training data such as rotation, clipping, and brightness change.

B. Performance Analysis

1) Ablation Analysis: In order to verify the necessity and effectiveness of the proposed modules, the ablation experiments are carried out. In the first group, the proposed learning-based saliency model is compared with several advanced saliency extraction methods, including the Amulet [23], MLMS [37], GRoIE [38], CFPN [39], BASNet [40], and DMT [41] methods. In the second group, the performance of the proposed SR-based classifier is compared with that of typical classification components. In the third group, ablation experiments are conducted to investigate the impact of different modules on ship recognition performance.

Fig. 5 shows the maps provided by the proposed learning-based saliency model and three typical salient region extraction methods. From the visual effect of saliency maps, we can find that the proposed method can highlight the potential areas more evenly and finely. Specifically, Amulet method effectively highlights the ships in two typical scenes with good lighting conditions in the HRSC dataset, but it is difficult to eliminate the interference of port facilities. In addition, for the scene images with shadow interference and densely docked conditions in the established dataset, it is difficult to highlight all ships. CFPN method can highlight the target with strong contrast, but it is difficult to comprehensively capture the ships with low contrast in the images. DMT method can accurately detect the ships and basically remove the interference of port facilities for typical scene images in HRSC dataset. On the second dataset, this method can effectively detect low contrast ships under cloud shadow, but it is difficult to accurately distinguish the edge contour of all ships for densely docked ships. The method we proposed correctly extracts all targets in complex scenes, and the extracted edge contour is more refined than that of several typical methods compared. This is mainly because we have adopted a multilevel feature extraction architecture, which can effectively capture multiscale target features in complex scenes. In order to more intuitively represent the processing effect of each method on two dataset images, the quantitative precision–recall curves of these saliency models are given in Fig. 6. From the comparison of the curves, we can find that our algorithm has the largest area of the precision–recall curve, which indicates the proposed algorithm achieves an optimal overall evaluation performance.

In order to verify the effectiveness of the proposed classification model, we compare its performance with six typical object classification methods. The comparison methods include: Fourier HOG with linear SVM [20], Rotation-Invariant Descriptor with Gaussian SVM [42], LBP with SVM [43], MD-DCM [44], and LSRTN [45]. The first three methods apply feature extraction combined with the classical SVM classifier to achieve target classification and false alarm suppression. The last two methods design detection operators to directly extract ship targets.

The candidate salient regions extracted by first module are used as the input of various classification algorithms, and the average precision and false detection rate are applied to evaluate the ship classification and false alarm suppression performance of various algorithms. The results of different classification models are shown in Table II. It can be found from the table that the proposed prior representation-based method can obtain the highest mAP and the lowest FAR values. The first three methods are all based on spatial domain features to extract

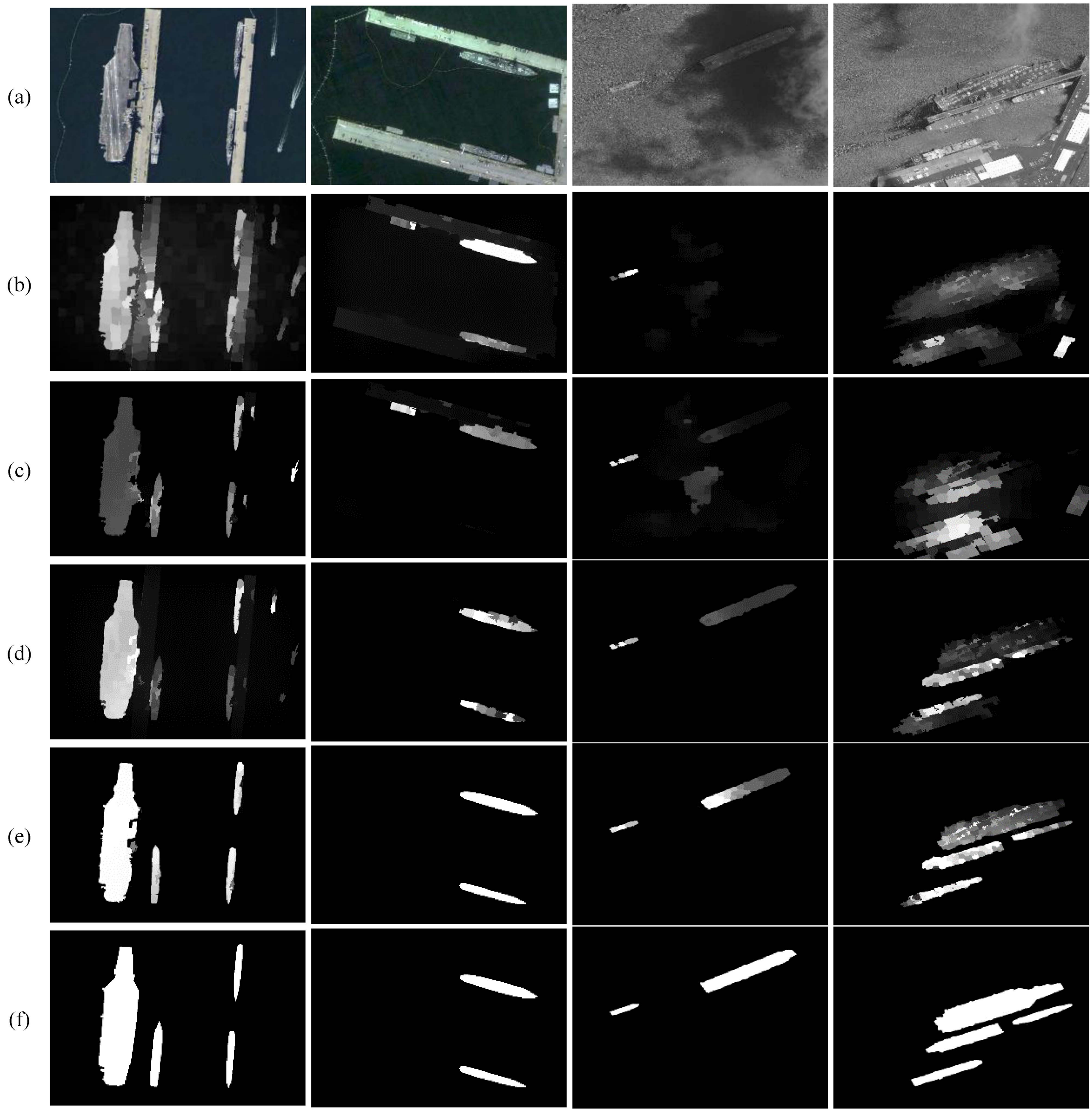
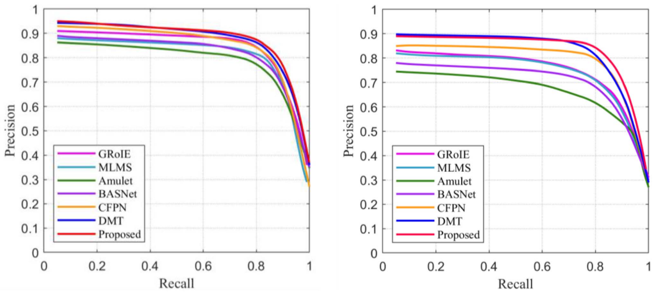


Fig. 5. Comparison results of the proposed method and several advanced saliency models on the two datasets (Columns 1–2 from the HRSC2016 dataset, Columns 3–4 from the established dataset). (a) Input image. (b) Amulet. (c) CFPN. (d) DMT. (e) Proposed method. (f) Ground truth.

candidate regions and then adopt SVM for classification. These methods have relatively poor performance in image slices for complex sea and sky scenes, mainly because their detection performance strongly depends on the precision of spatial domain feature operators. In fact, due to the failure to effectively mine the differences between similar ships and the fine edge features of targets in low contrast scenes, these methods have poor discrimination ability for ships with similar spatial distribution characteristics. Although the MDDCM method employs a deep network to learn target features, it only uses contrast features to extract targets in the spatial domain, making it difficult to adapt for ship target classification in highly contrast dynamic changing images, resulting in a relatively high FAR. The LSRTN method extracts scene features based on low-rank sparse decomposition theory, and has strong distinguishing capability for low-contrast targets. However, this model seldom considers the similarity between different ship classes, it is difficult to have high robust classification capability for ships in complex scenes such as partial cloud occlusion and shadow, limiting the target classification performance to some extent. The proposed classification method, on the one hand, characterizes the supervised reconstruction error of the salient regions of background interference



(a) (b)

Fig. 6. Precision-recall curve of different saliency extraction models on the two datasets: (a) Curve on the HRSC2016 dataset. (b) Curve on the established dataset.

TABLE II
Performance Comparison Results of Different Classification Models

Models	HRSC2016 dataset		Established dataset	
	mAP (%)	FAR (%)	mAP (%)	FAR (%)
HOG+linear SVM	43.87	19.46	47.63	17.34
RI + Gaussian SVM	48.63	16.45	54.31	15.25
LBP+SVM	57.18	13.59	63.45	11.74
MDDCM	69.74	9.42	70.36	9.64
LSRTN	77.25	7.86	72.89	6.76
Proposed	79.56	4.64	74.43	5.68

and ship targets, thus effectively characterizing the differences between the two types of elements. At the same time, it combines the regular items of the similarity graph to learn the intraclass differences between the original samples. Compared with other methods, the recognition accuracy is significantly improved and the background interference is effectively suppressed.

In order to visually and objectively illustrate the impact of the proposed modules on ship recognition performance, Table III presents the effects of sequentially adding and combining different modules on the final recognition performance. For the convenience of comparison, faster R-CNN is chosen as the basic framework. Upon encountering constraints within the Faster R-CNN detection architecture, the implementation of the original ResNet50 and ResNet101 yields mAP scores of 74.30% and 76.02%, respectively. Through the results of these mAP indicators, it can be found that incorporating the ABSF module into the faster R-CNN framework can improve the mAP to 77.36% and 78.64%, respectively. Further refinement through the integration of the PRC module on the ResNet50 extraction structure elevates the mAP to 79.56%, manifesting a substantial performance augmentation relative to the baseline detection architecture. Consequently, the aforementioned ablation studies substantiate the efficacy and utility of the introduced modules.

When the faster R-CNN detection framework is limited, the AP values obtained by configuring the original ResNet50 and ResNet101 as the multiscale extraction module are 68.52 and 69.38, respectively. From the AP measure, we can find that when we embed the multiscale dilated convolution module into the faster R-CNN detection framework, the AP value reaches 70.72. After further strengthening the characterization with the attention-based feature enhancement module, the final AP value reaches 72.68, showing a significant performance improvement compared with the original detection framework.

2) Algorithm Performance Comparison: To verify the effectiveness and robustness of the proposed model, we compare the proposed approach with nine representative object identification models on the two datasets, including typical faster R-CNN [46], YOLOv4 [47], RetinaNet [48], YOLOv7 [49], R2CNN [50], RRPN [51], R3Det [52], S2A-Net [53], ReDet [54], ROOD [55], AEDet [56], and Ofcos [57]. These methods are recognized as excellent models in the object recognition applications. It is worth noting that the first four methods are networks designed for general object recognition, and the last five are models used for ship target recognition in remote sensing images in recent years.

To clarify the application effect of different recognition methods, the faster R-CNN, R3Det, and ReDet are selected for result comparison display. The identification results of different models on the HRSC2016 dataset are given in Fig. 7. Specifically, on the HRSC2016 dataset, faster R-CNN method is robust to images containing shadow interference through deep network training. However, due to the lack of fine extraction strategies for multiscale features, it is easy to generate erroneous recognition and misses detection for docks with similar target characteristics and ships with relatively low contrast to the surrounding environment. R3Det method applies a multiscale feature skip connection strategy, which can better utilize target features with different sampling levels compared to faster R-CNN. Therefore, it effectively identifies partially occluded ship targets (as shown in the second scene of the cth row), and the overall confidence level of recognition is also improved. The ReDet method basically detects all ships, but there are certain limitations on the discrimination performance of destroyers, frigates, and cruisers with similar shapes and lengths. Our method has the best recognition performance because it considers the discrimination degree of different types of targets and focuses on the differences of ships with the same category.

For the established dataset, as shown in Fig. 8, in the first scene where ships are densely arranged, the considered three methods all fail to detect relatively small-scale submarines, and they have errors in determining the types of multiple destroyers and frigates. In contrary, despite a missed detection, the recognition quantity and accuracy of our method are still superior to other algorithms compared. In the second scene, for ships obscured by thin clouds, the three comparative methods all detect the position of the objects, but there are errors in the judgment of the type of the targets, and our method can correctly identify the ships. In the third scene with local shadows, although the considered three methods all detect multiple submarine targets, they all miss detection for targets in the shadow of the dock. Our method takes into account the mining of target multiscale information and the learning of target uniqueness features, thus correctly identifying all ship targets. The quantitative evaluation results of the two datasets are further demonstrated in Tables IV and V. Comparing the recognition results of different types of ships, it can be found that the recognition probability of submarines is relatively low, while that of aircraft carriers, destroyers, and cruisers is relatively high. This is mainly because submarines with small

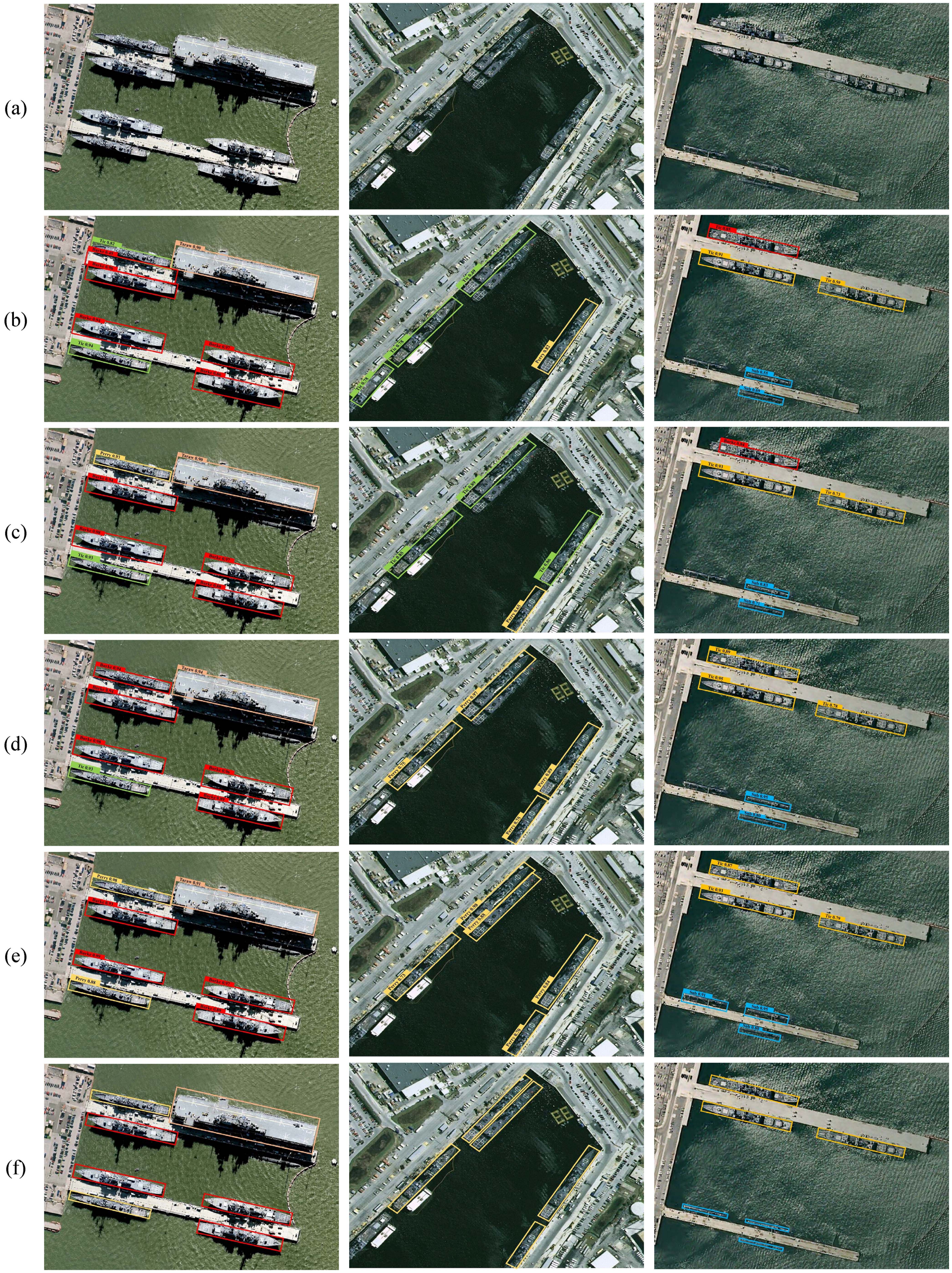


Fig. 7. Identification results obtained by our method and the baseline approaches on the HRSC2016 dataset. (a) Input image. (b) Faster R-CNN. (c) R3Det. (d) ReDet. (e) Proposed method. (f) Ground truth.

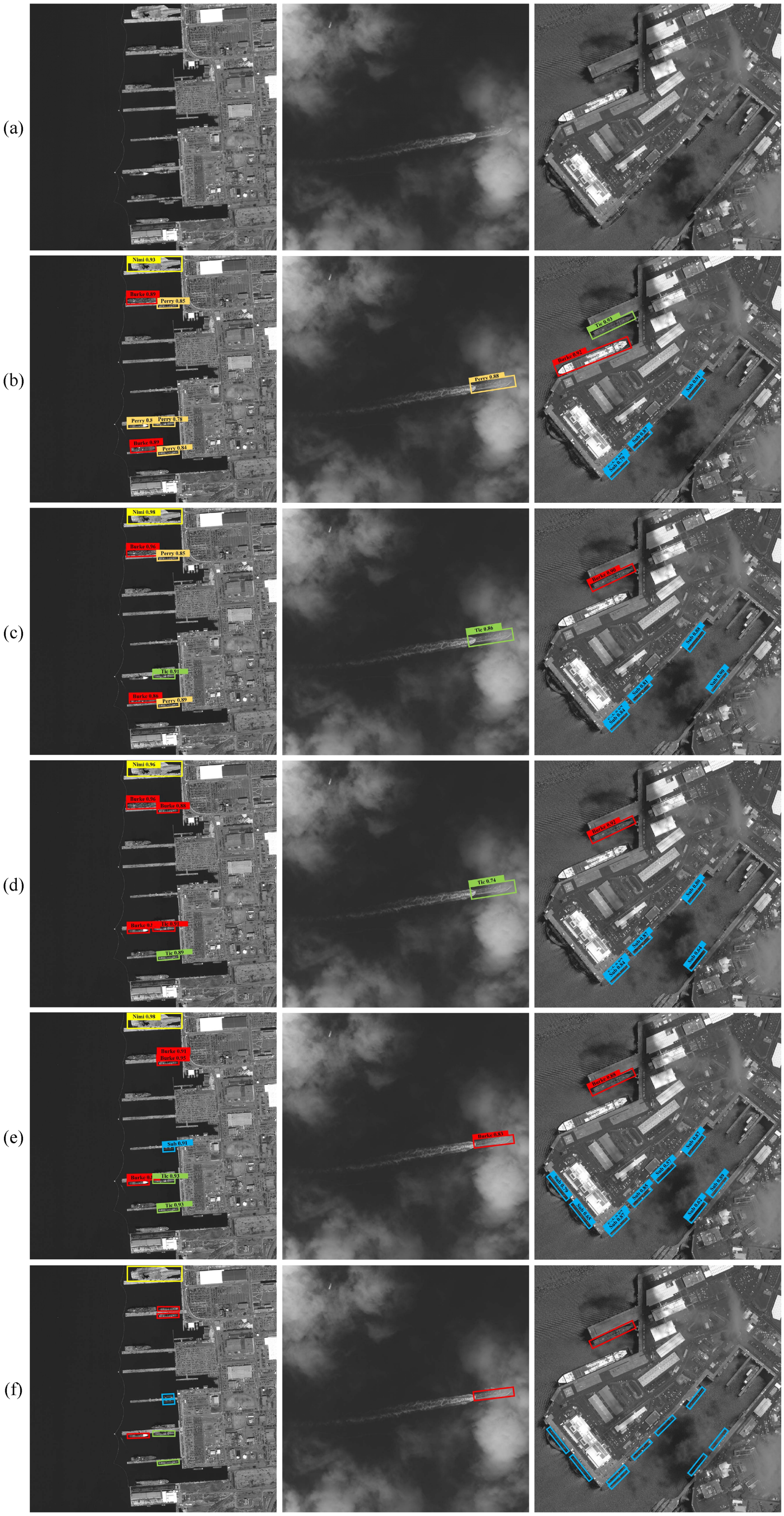


Fig. 8. Identification results obtained by our method and the baseline approaches on the established dataset. (a) Input image. (b) Faster R-CNN. (c) R3Det. (d) ReDet. (e) Proposed method. (f) Ground truth.

TABLE III
EVALUATION RESULTS OF ABLATION EXPERIMENTS ON THE HRSC DATASET

Methods	Backbone	Multi-Scale Aggregation Module	Classification Module	mAP(%)
Faster R-CNN	ResNet50	-	-	74.30
Faster R-CNN	ResNet101	-	-	76.02
Faster R-CNN	ResNet101	Proposed ABSF	-	77.36
Faster R-CNN	ResNet50	Proposed ABSF	-	78.64
Faster R-CNN	ResNet50	-	Proposed PRC	76.79
Faster R-CNN	ResNet101	-	Proposed PRC	78.49
Faster R-CNN	ResNet50	Proposed ABSF	Proposed PRC	79.56
Faster R-CNN	ResNet101	Proposed ABSF	Proposed PRC	80.28

TABLE IV
QUANTITATIVE EVALUATION RESULTS ON THE HRSC2016 DATASET

Methods	Nimi	Kuzn	Midw	Taraw	Tic	Burke	Anto	Whid	Perry	Sub	mAP(%)	FAR(%)	FPS
Faster R-CNN	73.82	92.50	71.38	74.62	79.50	81.56	79.39	61.35	72.12	56.78	74.30	12.36	19
YOLOv4	76.45	88.64	63.25	72.74	83.28	80.06	82.35	58.27	74.32	47.36	72.67	8.27	30
RetinaNet	68.60	94.35	66.72	65.47	84.52	79.02	39.36	50.36	71.15	44.28	66.38	14.82	18
YOLOv7	82.15	90.24	68.46	76.11	85.63	87.35	85.43	50.46	76.39	56.32	75.85	6.98	34
R2CNN	61.52	98.75	63.71	68.20	77.80	75.89	78.85	53.70	62.32	31.93	67.27	12.20	17
RRPN	74.45	30.56	68.52	68.32	85.43	87.58	84.53	66.84	75.12	53.78	69.51	9.65	18
R3Det	78.10	99.85	74.12	76.04	78.10	84.28	67.46	55.25	69.45	74.95	75.76	7.24	18
S2A-Net	79.24	95.34	76.34	77.5	85.21	83.83	78.45	68.21	68.43	59.32	77.19	5.35	18
ReDet	81.87	99.95	75.12	78.36	87.35	81.05	83.24	67.87	72.32	55.12	78.23	5.72	4
ROOD	78.61	97.62	73.77	75.72	79.42	84.64	70.65	60.24	70.05	62.75	75.35	7.28	19
AEDet	82.62	99.96	77.24	78.05	86.36	84.63	82.63	65.51	73.16	56.63	78.68	5.64	21
Ofcos	80.35	99.05	75.56	79.03	81.51	80.61	81.66	66.26	64.61	57.72	76.61	7.66	23
Proposed	85.42	98.45	76.25	79.64	82.73	83.36	86.64	72.37	65.54	65.16	79.56	4.64	22

TABLE V
QUANTITATIVE EVALUATION RESULTS ON THE ESTABLISHED DATASET

Methods	Nimi	Kuzn	Midw	Taraw	Tic	Burke	Anto	Whid	Perry	Sub	mAP(%)	FAR(%)	FPS
Faster R-CNN	79.36	88.49	75.12	65.60	65.87	65.98	71.81	51.21	49.76	40.54	65.35	13.56	18
YOLOv4	89.15	91.56	80.26	70.23	69.48	68.84	62.13	54.32	52.58	53.63	69.11	8.47	33
RetinaNet	85.15	92.56	86.13	65.24	67.36	61.15	58.56	43.15	45.4	39.75	64.36	12.06	19
YOLOv7	94.35	99.5	87.45	68.62	75.25	79.27	79.43	63.32	58.15	56.79	70.60	6.73	36
R2CNN	80.12	93.61	77.12	64.46	69.32	71.24	63.52	40.13	39.87	42.96	64.21	14.28	18
RRPN	76.23	88.58	70.37	65.90	67.27	75.28	63.37	47.13	50.27	58.48	66.33	13.85	18
R3Det	81.43	92.65	75.38	69.54	65.2	74.65	61.32	50.21	53.23	70.34	69.38	8.53	19
S2A-Net	87.43	88.45	80.45	72.38	67.54	83.68	80.86	54.42	55.52	54.32	72.52	5.84	19
ReDet	84.32	98.65	75.43	74.10	70.32	72.89	79.43	59.43	61.43	64.78	74.08	6.63	6
ROOD	85.62	90.35	76.86	74.51	69.15	71.66	80.11	56.26	60.38	63.24	72.81	6.72	19
AEDet	86.06	97.72	75.62	73.25	70.23	73.83	78.13	62.51	60.52	63.16	74.10	7.24	21
Ofcos	82.36	90.31	75.86	72.34	68.37	75.47	80.35	60.82	59.62	63.42	72.89	7.05	24
Proposed	79.34	96.24	72.12	74.86	71.46	74.64	81.45	66.42	63.78	64.43	74.43	5.68	24

size and similar appearance to the sea color are more difficult to identify types, while aircraft carriers and amphibious ships with large and medium size are easier to perceive. Moreover, the overall recognition accuracy of the Whidbey island-class ship is low due to the small number of training samples, which makes it difficult to fully train the network model. In addition, it can be found from the evaluation index results that our algorithm obtains the highest average accuracy and the lowest FAR on the two datasets. Based on the above analysis, it can be concluded that our method can achieve state-of-the-art performance compared to other methods in complex scenes, and has better robustness to scene elements such as shadows, occlusion, and dense arrangement.

V. CONCLUSION

In this article, we propose a ship identification method via adjacent-branched saliency extraction and prior representation-based classification, which is capable of achieving high-accuracy type recognition performance for typical ship targets in occlusion and shadow application scenes. First, a feature extraction network based on adjacent branch fusion is utilized to extract suspected salient regions, which consists of IIM and ISM module. By applying the extraction structure, feature maps with small semantic differences are fused to enhance the edge details and semantic information of multiscale ships, significantly highlighting candidate target regions with relatively less false detection. Then, a classification method based on SR learning is proposed, which can integrate the low-rank and intraclass characteristic prior into the classification strategy, effectively improving the classification performance under complex environmental interference conditions. Finally, extensive experiments are conducted on the HRSC2016 dataset and the established dataset, the superiority of the proposed method is verified through a comparison with representative identification methods.

References

[1] M. J. Er, Y. Zhang, J. Chen, and W. Gao, "Ship detection with deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 10, pp. 11825–11865, 2023.

[2] U. Kanjir, H. Greidanus, and K. Oštir, "Vessel detection and classification from spaceborne optical images: A literature survey," *Remote Sens. Environ.*, vol. 207, pp. 1–26, 2018.

[3] J. Hu, X. Zhi, S. Jiang, H. Tang, W. Zhang, and L. Bruzzone, "Supervised multi-scale attention-guided ship detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630514.

[4] J. Escorcia-Gutierrez, M. Gamarra, K. Beleño, C. Soto, and R. F. Mansour, "Intelligent deep learning-enabled autonomous small ship detection and classification model," *Comput. Elect. Eng.*, vol. 100, 2022, Art. no. 107871.

[5] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[6] T. Wu et al., "MTU-Net: Multilevel transUNet for space-based infrared tiny ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601015.

[7] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning ERA: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[9] J. Hu, X. Zhi, T. Shi, W. Zhang, Y. Cui, and S. Zhao, "PAG-YOLO: A portable attention-guided YOLO network for small ship detection," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3059.

[10] Y. Li, C. Bian, and H. Chen, "Dynamic soft label assignment for arbitrary-oriented ship detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1160–1170, 2022.

[11] Y. Li, L. Chen, W. Li, and N. Wang, "Few-shot fine-grained classification with rotation-invariant feature map complementary reconstruction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5608312.

[12] Y. Li and C. Bian, "Few-shot fine-grained ship classification with a foreground-aware feature map reconstruction network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622812.

[13] J. Hu, X. Zhi, W. Zhang, L. Ren, and L. Bruzzone, "Salient ship detection via background prior and foreground constraint in remote sensing images," *Remote Sens.*, vol. 12, no. 20, 2020, Art. no. 3370.

[14] Y. Han, X. Yang, T. Pu, and Z. Peng, "Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5612318.

[15] Y. Han, J. Liao, T. Lu, T. Pu, and Z. Peng, "KCPNet: Knowledge-driven context perception networks for ship detection in infrared imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5000219.

[16] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[17] S. Qi, J. Ma, J. Lin, Y. Li, and J. Tian, "Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1451–1455, Jul. 2015.

[18] T. Nie, B. He, G. Bi, Y. Zhang, and W. Wang, "A method of ship detection under complex background," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 6, 2017, Art. no. 159.

[19] F. Xu, J. Liu, M. Sun, D. Zeng, and X. Wang, "A hierarchical maritime target detection method for optical remote sensing imagery," *Remote Sens.*, vol. 9, no. 3, 2017, Art. no. 280.

[20] C. Dong, J. Liu, and F. Xu, "Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 400.

[21] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 678–686.

[22] Y. Ji, H. Zhang, Z. Zhang, and M. Liu, "CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances," *Inf. Sci.*, vol. 546, pp. 835–857, 2021.

[23] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[24] Y. Ji, H. Zhang, and Q. J. Wu, "Salient object detection via multi-scale attention CNN," *Neurocomputing*, vol. 322, pp. 130–140, 2018.

[25] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1448–1457.

[26] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[27] L. Zhang, L. Zhang, D. Tao, and X. Huang, "A multifeature tensor for remote-sensing target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 374–378, Mar. 2011.

[28] B. Pan, Z. Jiang, J. Wu, H. Zhang, and P. Luo, "Ship recognition based on active learning and composite kernel SVM," in *Proc. 10th Chin. Conf. Adv. Image Graph. Technol.*, Beijing, China, Jun. 19–20, 2015, pp. 198–207.

[29] L. Huang, W. Li, C. Chen, F. Zhang, and H. Lang, "Multiple features learning for ship classification in optical imagery," *Multimedia Tools Appl.*, vol. 77, pp. 13363–13389, 2018.

[30] J. Han et al., "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogrammetry Remote Sens.*, vol. 89, pp. 37–48, 2014.

[31] X. Wang, S. Shen, C. Ning, F. Huang, and H. Gao, "Multi-class remote sensing object recognition based on discriminative sparse representation," *Appl. Opt.*, vol. 55, no. 6, pp. 1381–1394, 2016.

[32] Y. Zhuang, L. Li, and H. Chen, "Small sample set inshore ship detection from VHR optical remote sensing images based on structured sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2145–2160, 2020.

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

[34] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[35] Y. Di, Z. Jiang, and H. Zhang, "A public dataset for fine-grained ship classification in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 747.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[37] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8150–8159.

[38] L. Rossi, A. Karimi, and A. Prati, "A novel region of interest extraction layer for instance segmentation," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 2203–2209.

[39] Z. Li, C. Lang, J. H. Liew, Y. Li, Q. Hou, and J. Feng, "Cross-layer feature pyramid network for salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 4587–4598, 2021.

[40] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.

[41] L. Li et al., "Discriminative co-saliency and background mining transformer for co-salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7247–7256.

[42] C. Dong, J. Liu, F. Xu, and C. Liu, "Ship detection from optical remote sensing images using multi-scale analysis and Fourier hog descriptor," *Remote Sens.*, vol. 11, no. 13, 2019, Art. no. 1529.

[43] T. Nie, X. Han, B. He, X. Li, H. Liu, and G. Bi, "Ship detection in panchromatic optical remote sensing images based on visual saliency and multi-dimensional feature description," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 152.

[44] W. Yu, H. You, P. Lv, Y. Hu, and B. Han, "A moving ship detection and tracking method based on optical remote sensing images from the geostationary satellite," *Sensors*, vol. 21, no. 22, 2021, Art. no. 7547.

[45] T. Wang, Y. Gu, and G. Gao, "Satellite video scene classification using low-rank sparse representation two-stream networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622012.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[49] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOV7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.

[50] Y. Jiang et al., "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.

[51] R. Nabati and H. Qi, "RRPN: Radar region proposal network for object detection in autonomous vehicles," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3093–3097.

[52] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.

[53] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5602511.

[54] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2786–2795.

[55] L. Hou, K. Lu, and J. Xue, "Refined one-stage oriented object detection method for remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 1545–1558, 2022.

[56] K. Zhou et al., "Arbitrary-oriented ellipse detector for ship detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7151–7162, 2023.

[57] D. Zhang, C. Wang, and Q. Fu, "OFCOS: An oriented anchor-free detector for ship detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6004005.

Jianming Hu (Student Member, IEEE) received the M.Eng. and Ph.D. degrees in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017 and 2022, respectively.

He is currently an Assistant Professor in HIT. He was a Visiting Scholar with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, in 2019. His research interests include remote sensing image processing, image characteristic analysis, sea-aero target detection and identification.

Tianjun Shi received the B.Eng. and M.Eng. degrees in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2020 and 2022, respectively. He is currently working toward the Ph.D. degree in optical engineering with the Harbin Institute of Technology, City, China.

His research interests include remote sensing image acquisition and processing, small target detection and identification.

Shikai Jiang (Student Member, IEEE) received the M.Eng. and Ph.D. degrees in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2018 and 2022, respectively.

He is currently an Assistant Professor in optical engineering with the School of Astronautics, HIT. His research interests include optical information processing, remote sensing image intelligent interpretation, and innovation space optical imaging system and application.

Xiyang Zhi received the Ph.D. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017.

He is currently a Full Professor with the HIT. His current research interests include remote sensing image acquisition and processing, optical target detection and identification.

Xiaogang Sun received the Ph.D. degree in precision instruments and machinery from the Harbin Institute of Technology (HIT), Harbin, China, in 1998.

He is currently a Full Professor at the HIT. His current research interests include infrared measurement, photoelectric testing, intelligent instruments, and medical detection.

Wei Zhang received the M.S. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 1986, and the Ph.D. degree in electronic engineering from the Tohoku Institute of Technology, Miyagi, Japan, in 2000.

He is currently a Full Professor with HIT. His research interests include remote sensing image acquisition and processing, optical system design, and automatic target detection and identification.