# Supervised Multi-scale Attention-guided Ship Detection in Optical Remote Sensing Images

Jianming Hu, Xiyang Zhi, Shikai Jiang, Hao Tang, Wei Zhang, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Ship detection in optical remote sensing images plays a significant role in a wide range of civilian and military tasks. However, it is still a challenging issue owing to complex environmental interferences and a large variety of target scales and positions. To overcome these limitations, we propose a supervised multi-scale attention-guided detection framework, which can effectively detect ships of different scales both in complex pure ocean and port scenes. Specifically, a multi-scale supervision module is first proposed to adjust the semantic consistency of different feature levels, obtaining extracted features with small semantic gaps. Next, an attention-guided module is utilized to aggregate context information from both spatial and channel dimensions by calculating map correlations, adaptively enhancing the feature representation. Moreover, to preserve the attribute and spatial relationship of the optimized features, we adopt a capsule-based module as the classifier and obtain satisfactory classification performance. Experimental results conducted on two public high-quality datasets demonstrate that the proposed method obtains state-of-the-art performance in comparison with several advanced methods.**

***Index Terms*—Ship detection, multi-scale extraction, attention-guided feature representation, capsule-based classification, optical remote sensing.**

## I. Introduction

With the development of earth observation technology, high-resolution aerial and satellite images bring great convenience for data analysis and interpretation in many maritime engineering applications such as area surveillance, port management and sea rescues. Automatic ship detection, has become a central issue in these varied security and service applications [1]. At present, mainstream remote sensing data can be summarized into two categories: synthetic aperture radar (SAR) [2] and optical images [3]. Compared with the former, optical imaging data offers merits in providing richer spectral information, clearer texture features, more intuitive structural details and lower background noise, thus becoming a popular and significant research material in the field of automatic ship detection [4].

However, accurate target detection in optical images is a challenging issue. On the one hand, due to the long-range imaging conditions and the variety of target appearances, the ships in the image show a changeable scale difference. On the other hand, owing to the wide range of remote sensing imaging scenes, it is inevitable that there will be a variety of background interferences such as islands, clouds, sea clutter and potential artificial facilities in the imaging field [5], which further increases the difficulty of efficient ship detection. Therefore, it is urgent and critical to find a detection solution that can adaptively and accurately detect targets of different scales with complex scene interferences.

In recent years, with extensive research of deep convolution networks, object detection technology has made a great breakthrough. Various outstanding deep learning networks have emerged, which greatly improve the ability of target representation, showing obvious advantages over traditional methods. Among these approaches, the feature pyramid network (FPN) [6], which is an effective multi-branch architecture, is widely employed to solve the problem of multi-level feature extraction. Generally speaking, FPN generates well-organized features composed of rich semantics and fine structural details. Nevertheless, most FPN-based studies ignore a defect caused by the design of multi-level structure of the FPN itself, that is, directly fusing features of different levels may lead to suboptimal results. Actually, since there are semantic gaps between different levels of a pyramid structure, direct fusion of these features inevitably reduces the representation effect of the multi-scale characteristics of targets. Consequently, it is necessary to add a supervision mechanism before the feature integration to narrow these gaps, so as to make full and harmonious use of the representation advantages of different feature maps. In addition, the feature pyramid generates many multi-scale feature maps, and each feature map contains global and local information of different granularity. Considering that the global features can reflect the semantic information of the input scene, and the local context of the target neighborhood can describe the relationship between the foreground and the background, both of them provide support for the subsequent target classification. Therefore, when using these numerous feature maps, finding ways to preserve the global context while enhancing the local context features is also an important issue to ensure the accurate detection of ship targets.

To address these limitations, we propose a novel multi-scale

This work was supported by the National Natural Science Foundation of China under Grant 61975043. *(Corresponding author: Xiyang Zhi.)*

J. Hu is with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: hjm1007491571@163.com).

X. Zhi is with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: zhixiyang@hit.edu.cn).

S. Jiang is with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: jsk8023@163.com).

H. Tang is is with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland (e-mail: hao.tang@vision.ee.ethz.ch).

W. Zhang is with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: wzhang@hit.edu.cn).

L. Bruzzone is with Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

ship detection network in this paper. Specifically, a supervision module is first proposed to constrain the semantic consistency of the multi-scale features extracted from the classical residual network, ensuring that the features of different levels learn similar semantic information. Then, considering that the attention mechanism can help screen out valuable candidate regions from a large number of unrelated background regions, a dual-attention module is introduced to adaptively capture the relevance of context information both from the spatial and channel dimensions, making the network learn more distinctive elements that are conducive to the subsequent accurate and robust classification. This strategy can significantly enhance the feature representation of ship targets in complex environment. Moreover, to preserve the properties and spatial relationship of the features, we adopt a capsule-based classifier and obtain satisfactory classification performance. Experimental results on two high-quality datasets demonstrate that the proposed method significantly outperforms existing detection methods.

The main contributions of this work are summarized as follows:

1) A semantic supervision module is utilized to constrain the semantic gap of different scale features, ensuring the semantic consistency of multi-scale feature representation.
2) An attention-guided module is adopted to adaptively integrate context information both from the spatial and channel aspects by calculating the map correlation of different positions and different channels, thus strengthening the multi-scale feature representation.

The remainder of this paper is organized as follows. We introduce the related works and existing problems of multi-scale ship detection in Section II. Section III illustrates the rationale and details of the proposed framework. In Section IV, we validate the effectiveness of the proposed method and present experimental results based on two public datasets. Finally, we draw the conclusions in Section V.

## II. Previous Related Research

In this section, we review previous research works about multi-scale deep supervision, attention mechanism in deep network and capsule-based network in detection model. Specifically, we briefly illustrate the ideas and shortcomings of the current typical methods for the above problems. On this basis, we illustrate the motivations and differences of the proposed method.

### A. Multi-scale Deep Supervision

Due to the randomness of the size and position of the detection target in practical application, multi-scale detection is one of the issues that cannot be ignored in a detection model [7]. Before the popularity of learning-based methods, researchers usually built image pyramids with different resolutions for hierarchical prediction. As convolutional neural networks show significant advantages in feature representation [8], multi-scale architecture design based on learning networks has developed rapidly.

Multi-scale architecture generally represents image objects at different levels of detail through different feature extraction branches. With the increase of network branches and the expansion of depth, deep supervision has become a core link in an excellent network architecture. To realize the deep supervision, an auxiliary classifier is usually added to some hidden layers of a network to constrain the overall or local performance of the network. From the perspective of the supervision method development, the deep supervision strategy was first presented as a training trick in 2014 [9], mainly solving the problems pertaining to training gradient disappearance and slow convergence speed. Furthermore, in [10] a multi-scale network structure was designed, which embeds multiple classifiers into a learning-based network and interconnects these classifiers through dense connected branches, to learn different levels of detail. In [11] the work used the probability knowledge learned by multiple auxiliary classifiers as an additional constraint to supervise the knowledge matching among branches, and finally realized the dynamic cooperation of different branches in the classification task. In [12] the supervision mechanism was applied to adjust the balance between the accuracy of different module combinations and computing resources, so as to achieve the purpose of optimal allocation. Varying from the aforementioned works, we provide new insights on applying the supervision mechanism to ensure the semantic consistency of multi-scale features, thus enhancing the feature representation of learning-based networks.

### B. Attention Mechanism in Deep Network

In recent years, attention has developed into one of the most influential concepts in deep networks. It was first proposed by simulating human cognitive habits. When people observe an image scene, they usually pay special attention to some abnormal regions. Inspired by this, researchers have designed a variety of learning networks and presented different activation functions to make the machine automatically obtain a local response to potential target areas, which greatly improves the performance of various visual tasks [13]. Actually, the attention used in deep networks is mainly a task-oriented focused module. Specifically, the network introduces the attention mechanism according to the needs of various tasks to guide the updating of the feature maps, highlighting the relevant features and simultaneously suppressing the irrelevant features.

Generally, there are two main design ideas of attention module in current learning networks. The first one is to optimize the global context information directly, and the second idea is to introduce a self-attention mechanism to improve the dependence of different dimensions. One of the representative works guided by the first idea is SENet [14], which was proposed in 2018. It attempts to employ the global-pooling operation to model channel-wise interdependencies. It is worth noting that the pattern of spatial information squeezing in this work has a profound impact on follow-up research. On this basis, CBAM [15] further presented a spatial domain correlation calculation module to selectively allocate the importance of different features. Interestingly, it innovatively compares the performance of different connection modes of the channel

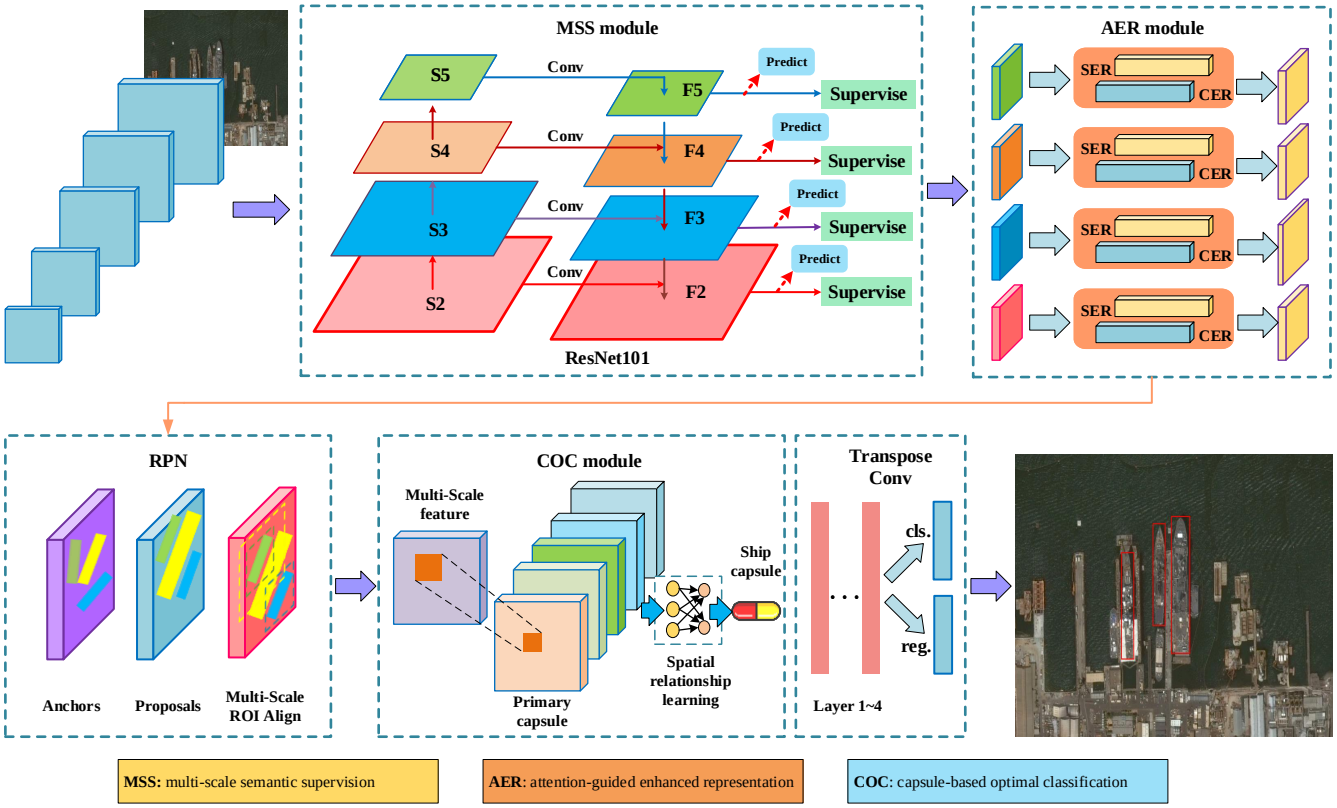



Fig. 1. Overview of the proposed framework.

dimension and the spatial dimension attention, leading a wave of attention-based parallel design. Subsequently, SK-Net [16] employed several parallel convolution kernel branches with different receptive fields to learn the weight of multi-scale features. SPA-Net [17] adopted a spatial pyramid composed of multiple adaptive average pooling parts to model the local and global context information. In terms of self-attention ideas, researchers have also created many valuable works. In [18], the self-attention mechanism was first applied to mine long-range relevance, inspiring many researchers to adaptively improve the global context. In [19], a classical non-local module was proposed to extract the global context and model long-distance feature dependencies. In 2019, the non-local idea was introduced into both the channel domain and spatial domain [20], adaptively integrating features of different receptive fields.

From the above literature on attention, we can find that the use of long-range dependencies is conducive to object classification and detection, yet the SENet-based methods mainly focuse on the information mining of local context to reallocate the weights of different channels and locations, ignoring the global dependencies. The Non-local method is redundant in the calculation of attention weight, especially in the spatial dimension. Studies [21] have proved that the long-range dependencies of many coordinate points are highly similar, which leads to a waste of computing resources. Different from existing approaches, in the spatial dimension, we provide new insights into calculating the space-point-based weight by blocks, and use the mutual mapping of small blocks in the channel dimension to calculate the reasonable weight for each block, highlighting the significant regions and avoiding the tedious calculation of directly using a non-local autocorrelation matrix. In the channel dimension, we apply the channel adjustment factor to effectively reduce the calculation amount of the channel-dimension long-range dependence relationship. The experimental results show that the effectiveness and robustness of the proposed method.

### C. Capsule Network in Detection Model

A traditional deep convolution network usually extracts features from images by using convolution kernels, and processes these generated feature maps using the pooling layer, so as to detect the same kind of objects in different images. Although deep convolution networks have achieved remarkable results in many applications, it is obvious that pooling operations cannot describe the spatial relationship between features, and the positional relationship between features has been proven to have a significant impact on target detection and recognition. In addition, the pooling operation loses information of the target to a certain extent, which weakens the network's ability to describe the target with changing direction. Aiming at these problems, in 2011 Hinton et al. [22] proposed the capsule concept as an alternative to the convolutional neural network (CNN) model. Unlike the CNN model, which transmits information in the form of scalar weight, the capsule architecture encodes and records attitude characteristics (such as accurate position, direction, etc.) in the form of vector. Moreover, the transmission of features from input to output is equivariant,

which has advantages in fine learning the part-whole relationship in the feature maps. The capsule architecture was first implemented [23] and applied to solve practical problems in 2017. On this basis, many innovative works have been proposed to optimize the performance of capsule networks in detection and recognition tasks.

For instance, in [24] the layers of the capsule network were deepened to learn the features of the input image, obtaining satisfactory results in terms of detection accuracy and false alarm rates. In [25] superpixel patches were input into the capsule network and the max-pooling operation is applied to classify objects with varying scales, orientations and occlusion conditions. In [26] an extended capsule-based architecture was presented to handle the feature preserving issue in the encoder network, showing the effectiveness and robustness in object recognition with limited training samples. Motivated by these works, we introduce the capsule network architecture to describe the spatial relationship of the extracted features, so as to enhance the perception ability of our method for ships in different positions and random directions.

## III. Proposed Method

### A. Method Overview

Compared to previous multi-scale ship detection approaches, we seek an alternative solution with more semantically consistent multi-scale extraction, stronger feature representation ability and more accurate classification performance. As illustrated in Fig. 1, the proposed framework mainly consists of three modules: multi-scale semantic supervision (MSS), attention-guided enhanced representation (AER) and capsule-based optimal classification (COC). Specifically, we obtain the supervised target features of different scales with smaller semantic gaps by using the MSS module. Then, the AER module is utilized to integrate local features with global dependency while adaptively strengthening the relationship between different interdependent channels. Finally, we apply the COC module to encode the various property parameters of different objects and use these parameters to assist the class and location prediction.

### B. Multi-scale Semantic Supervision

As we know, in most learning-based detection methods, the FPN architecture is adopted to extract the multi-scale features of input images. Specifically, the FPN architecture first employs a bottom-up pathway to generate several sampling stages, thus establishing a feature pyramid with different degrees of semantic information distributed in each level. Then these feature maps of various levels are merged by top-down up-sampling and lateral connection design. It should be emphasized that among these feature maps, those with less down-sampling processing contain more high-resolution edge details, while those with more sampling processing have lower resolution but richer semantic information levels. Actually, the semantic features contribute to describing the relative position distribution between the target region and its surrounding neighborhood and the edge details help to define the clear boundary of the target.

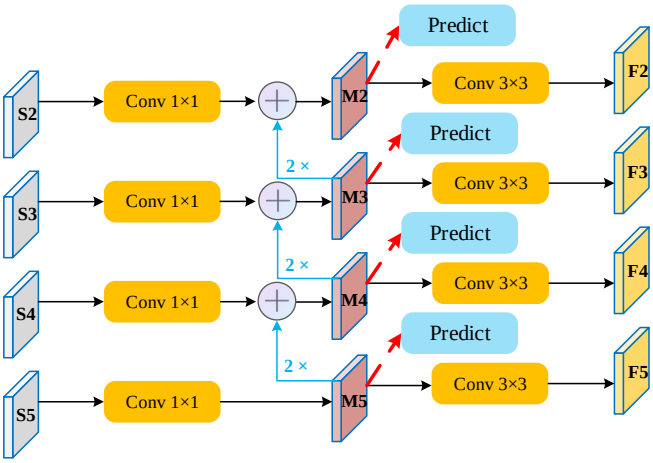



Fig. 2. Design of the MSS module.

Although the FPN structure considers different levels of features through independent prediction at each level, the merger is sub-optimal due to great semantic differences between different pyramid stages. This inspires us to find a supervision strategy aiming at narrowing the semantic gap between multi-scale features, so as to obtain a more consistent feature representation of the detection scene.

For this reason, we propose the MSS module (as illustrated in Fig. 2). It is placed before the pyramid feature fusion, which can effectively supervise these semantic gaps. The main idea of this module is to confine the loss function in the feature extraction stage within the expected range. Let us define the output of the last residual block in each level of the ResNet backbone as $S$. To reduce the computational complexity, we abandon the feature map with a stride of two pixels, thus obtaining $S_i$= $\{S_2$,$S_3$,$S_4$,$S_5\}$ corresponding to maps with strides $\{4, 8, 16, 32\}$ pixels. Similar to the typical faster R-CNN method [27], we apply the region proposal network (RPN) to extract the regions of interest (ROIs) from these maps $S_i$. Moreover, to accurately pool these ROIs into feature maps of a fixed size based on the location coordinates of these ROIs, the ROI alignment operation is employed.

In order to define the specific values of these ROIs, we select four sampled points for each ROI unit, and use bilinear interpolation to describe their coordinates. Suppose that the coordinates of the four sampled points on the boundary are $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$ and $Q_{22} = (x_2, y_2)$, respectively. Then, according to the principle of interpolation, the coordinate $f$ in floating point form at point $(x, y)$ can be expressed as:

$$f(x,y) = \frac{1}{(x_2-x_1)(y_2-y_1)}\left[x_2-x, x-x_1\right]\cdot \begin{bmatrix} f(Q_{11}) & f(Q_{12}) \\ f(Q_{21}) & f(Q_{22}) \end{bmatrix} \begin{bmatrix} y_2-y \\ y-y_1 \end{bmatrix} \quad (1)$$

After obtaining these fixed size feature maps, we establish the relationship between the original maps $S_i$ and the new feature maps $M_i = (M_2, M_3, M_4, M_5)$ as follows:

$$M_i = \begin{cases} conv(S_i) + upspl(M_{i+1}) & i\text{=2, 3, 4} \\ conv(S_i) & i\text{=5} \end{cases} \quad (2)$$

where $conv$ represents a $1 \times 1$ convolution operation, $upspl$ represents the up-sampling operation with a sampling multiple of 2. Moreover, the up-sampling operation is realized by the nearest neighbor interpolation.

Now we have obtained numerous candidate target regions of different scales. It is difficult to analyze their semantic differences directly. Consequently, we combine the prediction performance of the detection task to indirectly narrow these semantic gaps of different scale feature maps. Specifically, these feature maps after the ROI alignment process are passed into the R-CNN [28] part to generate bounding box predictions and class scores. More importantly, the prediction parameters of these maps with different scales are shared to obtain similar level semantic information extraction. In addition, to reduce the aliasing effect caused by the nearest neighbor interpolation, the feature maps $M_i$ are processed by a $3 \times 3$ convolution, and the final features $F_i$ are obtained. Then, we propose a loss function to comprehensively supervise the degree of semantic information learning.

The final loss function of semantic supervision is computed as follows:

$$\begin{aligned} L_{msc} =& \alpha(L_{cls,M}(p_M, p_{GT}) + \beta L_{loc,M}(t_M, t_{GT})) + \\ & L_{cls,F}(p_F, p_{GT}) + \beta L_{loc,F}(t_F, t_{GT}) \end{aligned} \tag{3}$$

where $\alpha$ is the index used to adjust the weight between the semantic constraint loss and the original prediction loss. The classification loss $L_{cls}$ includes cross-entropy loss calculated by the RPN part and classification loss calculated by the R-CNN, and the location regression loss $L_{loc}$ contains the anchor position deviation computed by RPN and the prediction box position deviation computed by R-CNN. $L_{cls,M}$ and $L_{loc,M}$ indicate the loss functions calculated by layers $M_i$. Similarly, $L_{cls,F}$ and $L_{loc,F}$ are the functions of layers $F_i$= $\{F_2$,$F_3$,$F_4$,$F_5\}$. $p_{GT}$ and $t_{GT}$ are real class labels and target locations, respectively. $p_M$, $p_F$ and $t_M$, $t_F$ indicate the prediction information of layers $M_i$ and $F_i$, respectively. $\beta$ is a judgment factor, which can be assigned to 1 only when the prediction on the class label is correct, otherwise, it is 0.

By using the above loss function, we optimize the parameter settings of the multi-scale extraction network based on training data, thus obtaining the multi-scale representation of the image scene with a smaller semantic gap.

### C. Attention-guided Enhanced Representation

After the processing of the MSS module, we realize a preliminary multi-scale representation of the input remote sensing scene. However, to achieve accurate detection of ship targets in complex sea surface and harbor scenes, we still need to propose new strategies to effectively distinguish ships from interferences with similar characteristics, such as islands, sea clutter and thick clouds. Actually, these interferences are likely to be similar to the real ship in size, direction, shape, color, texture and other characteristics, which poses a significant challenge to the final detection task. Inspired by [13], extracting rich global and local context information from multi-scale maps is conducive to improving the ability of distinguishing different elements in the scene. Therefore, focusing on this problem, we propose the AER module that optimizes feature representation from two domains. As illustrated in Fig. 3, the module is mainly composed of two parallel branches: the enhanced representation branch in spatial domain and the channel domain. The spatial branch is used to mine the local block based context information, while the main purpose of the channel branch is to optimize the response of a specific class by mining long-range dependencies between channels, both enhancing the differences between various element classes.

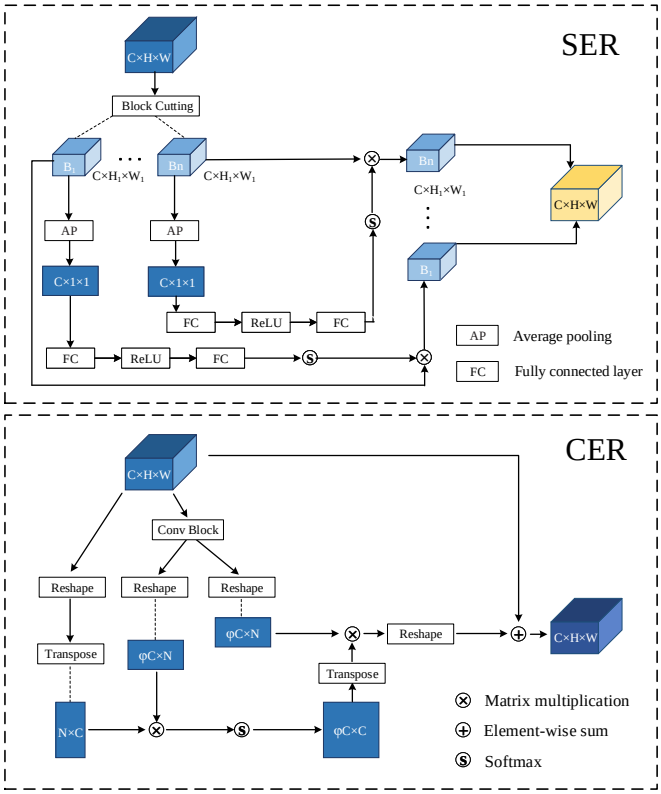



Fig. 3. Design of the AER module.

Before introducing this module, it is essential to define the input of the module. Considering that these resolutions of feature maps $F_i$ are inconsistent, we first unify all the images to the same resolution through an interpolation operation, thus obtaining maps with the same resolution. We merge all feature maps to aggregate features of multiple levels, and perform a convolution operation to change dimensions. The multi-scale feature map $F_{ms}$ is obtained as

$$F_{ms} = conv\left\{[F_2', upspl\left(F_3'\right), upspl\left(F_4'\right), upspl\left(F_5'\right)]\right\} \tag{4}$$

Then, we integrate $F_{ms}$ into different scale branches as the input of the AER module.

*Spatial Enhanced Representation*. We first illustrate the details of the spatial domain branch. Considering that different regions of large-scale remote sensing images may have different degrees of importance, and the direct calculation of point-based dependency calculation is redundant. Consequently, given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, we first employ the block cutting processing to generate many new feature layers $\{X_1, X_2, ..., X_n\} \in \mathbb{R}^{C \times H_1 \times W_1}$. $C$ is the number of channels, $H_1$ and $W_1$ are the height and width of the feature

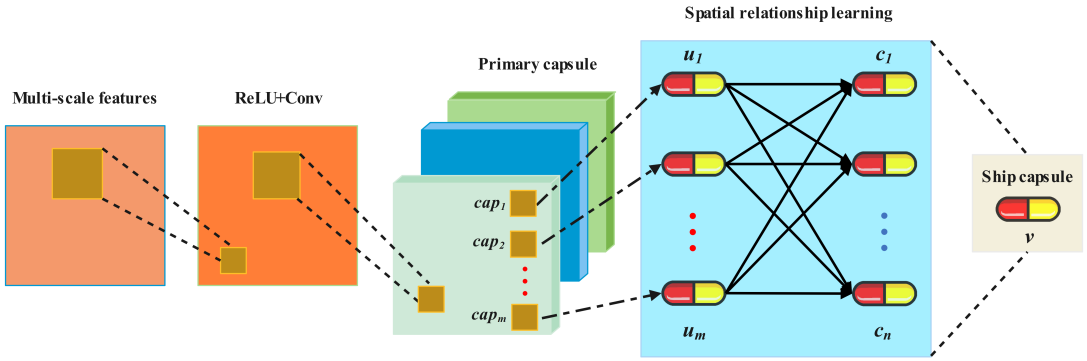



Fig. 4. Design of the COC module.

after cutting, respectively. It is worth noting that the sizes of different region blocks are set according to the ship size of the dataset images. In order to highlight the features of the ships, the size of each block should be larger than that of the largest preset anchor box. The training strategy of preset anchor box in deep learning network can refer to [4]. On the premise that the block size is larger than the preset maximum anchor box size, we divide the input image into blocks of the same size. We then apply the average pooling operation to highlight the significance of different blocks. Based on the operation, we calculate the local significance value $Y_b \in \mathbb{R}^{C\times 1\times 1}$ from each small block $X_n$. In fact, the results of different $Y_b$ also reflect the global distribution characteristics in the each feature map to a certain extent. This process can be characterized as

$$Y_b = \frac{1}{W_1 H_1}\sum_{x=1}^{W_1}\sum_{y=1}^{H_1} Q(x,y) \tag{5}$$

where $Q(x,y)$ indicates a pixel point at any channel.

In order to train accurate weights, inspired by SENet [14], we assign appropriate weights to each block through the response of multi-layer channels. For weight learning and updating, we apply two fully connected layers. The first layer compresses $C$ channels of $Y_i$ into $C/\phi$ channels. $\phi$ is the channel compression ratio,which can be optimized by gradual learning. After a ReLU operation, the second fully connected layer restores the output map to C channels. Consequently, the relationship between the final output $W_s$ and the input $Y_b$ is constructed as

$$Y_k = \mathrm{FC}(\mathrm{ReLU}(\mathrm{FC}(Y_b))), k \in [1,n] \tag{6}$$

$$W_s = (Y_1 \cdot X_1, Y_2 \cdot X_2, ..., Y_n \cdot X_n) \tag{7}$$

Obviously, each element in feature $W_s$ is a weighted sum of the feature related to the position and the original feature. Therefore, sufficient context information can be obtained, moreover, since similar features show more significant correlation, the differences between different element classes in the feature map can be highlighted more obviously.

*Channel Enhanced Representation*. Each channel map can be seen as a response to a specific feature. We calculate the interdependence between different channels and highlight the dependent feature mapping, which can realize the feature representation of the candidate area with a high suspected degree. We still take $X \in \mathbb{R}^{C\times H\times W}$ as an example of the input feature in the channel branch. Unlike the spatial branch, we only apply one convolution layer to process the input feature $X$, which adjusts the computational burden of the feature extraction. Actually, when the number of channels is large, it should not be neglected the computational burden and efficiency. Consequently, we introduce a dimension adjustment parameter $\varphi \in (0,1)$ and then obtain the new feature $X_{da} \in \mathbb{R}^{\varphi C\times H\times W}$ based on the convolution layer.

Next, we reshape $X_{da}$ to $X'_{da} \in \mathbb{R}^{\varphi C\times N}$ and $X$ to $X' \in \mathbb{R}^{C\times N}$. Through a matrix multiplication and a softmax operation, we obtain the map $Y_{da} \in \mathbb{R}^{\varphi C\times C}$. Similar to the spatial branch, the relationship between the final output $W_c$ and the input $X$ is constructed as

$$W_c = \rho \cdot \mathrm{Reshape}\left(Y_{da}^T \cdot X'_{da}\right) + X \tag{8}$$

where $\rho$ is a scale influence parameter. Similar to $\phi$, it can be optimized by gradual learning.

Finally, we combine the maps of two branches to gain the output of the module as

$$W = W_s + W_c \tag{9}$$

From the above processes, it can be inferred that the output map fuses features with strong low-resolution semantic information and features with weak high-resolution semantic information but rich spatial information.

### D. Capsule-based Optimal Classification

After the attention-guided module, we have embedded the position information into these extracted features. We then input the enhanced features into the RPN network for candidate region generation. It is worth noting that the traditional deep network usually employs multiple pooling layers to achieve the final classification. This step can maintain the spatial position invariance in the feature map, but it is difficult to capture small changes in spatial local feature, which restricts the ship detection performance in complex port scenes to a certain extent. For example, when a ship in the port rotates at a certain angle or two ships are berthed close to each other, the conventional CNN network is difficult to capture these spatial relative position changes. However, these situations are common phenomena in the port scene.

The pooling layer in the original CNN seldom takes into account the relative spatial relationship between the extracted features [23], so the distribution information of the foreground and background is discarded to a certain extent, which directly affects the final classification performance. Therefore, it is crucial to explore a strategy that combines the relative spatial relationship information with the final classification idea.

To handle this issue, we adopt the COC module (the architecture is illustrated in Fig. 4), which can retain the important state information of all features in the form of the capsule, and apply these state properties to achieve an accurate prediction of the target class. The capsule is a vector that can record the properties of a specific type of entity, including position, size, direction and so on. The COC module is mainly composed of one conventional convolutional layer, one primary capsule layer and one ship capsule layer. Given some input features, these features are first encoded into many capsules to characterize various levels of entities. In addition, in order to enhance the representation ability of encoding, we apply a simple but efficient rectified linear unit (ReLU) function to describe the nonlinear relationship. We describe the internal structure of a capsule through the transformation of input and output vectors, and the transformation relation can be described as

$$v_n = \frac{\|c_n\|^2}{1+\|c_n\|^2} \cdot \frac{c_n}{\|c_n\|} \tag{10}$$

where $v_n$ indicates the output vector of capsule n, $c_n$ indicates the input vector. It can be found that this transformation makes the long vector close to 1 while the short vector is close to 0. By controlling the length of the output vector to be between 0 and 1, we employ the length of the output vector to represent the probability value of a feature.

For the transfer rules between capsules of different layers, we establish the mapping relationship as

$$c_n = \sum_m a_{mn} u_{n|m} \tag{11}$$

$$u_{n|m} = w_{mn} u_m \tag{12}$$

where $a_{mn}$ is the coefficient that characterizes the contribution of capsule $m$ to capsule $n$, $u_{n|m}$ indicates the prediction vector input to capsule $n$. $u_m$ is the output vector of the lower capsule m, and $w_{mn}$ indicates the weight matrix connecting the two capsule layers.

Through the above steps, we establish the transmission strategy of the relative spatial relationship between different capsule layers. Considering that each capsule layer has vectors with low contribution rate, the capsule model can be pruned and optimized in the training process to reduce the redundant parameters of the classification stage. In order to prune the model, we set a pruning ratio $\gamma$ and analyze the parameters of the capsule layer to be pruned layer by layer. Suppose the capsule layer to be optimized has $l$ units. We sort them in ascending order according to their weight parameters, and eliminate the units less than a certain threshold. This process can be expressed as

$$cap_i = \begin{cases} 1 & i \geq l * \gamma \\ 0 & i < l * \gamma \end{cases} \tag{13}$$

After removing some redundant capsules with low contribution rate, we need to retrain the model to obtain the optimized weight parameters.

To effectively improve the classification performance of the capsule network, we apply the margin loss function to control the iterative process. The function $L_k$ is defined as follows:

$$\begin{aligned} L_k = & T_k \cdot \max(0, m^+ - \|u_k\|^2) + \\ & 0.5 \cdot (1 - T_k) \cdot \max(0, \|u_k\|^2 - m^-)^2 \end{aligned} \tag{14}$$

where $k$ indicates the target ship class, $T_k = 1$ only when the classification is correct, otherwise it is 0. $m^+$ and $m^-$ are the upper and lower probability thresholds that need to be set when training samples, respectively. As in [23], they are set to 0.9 and 0.1 in our experiments, respectively.

We obtain the class prediction result by inputting the output of the last capsule to a decoder network with four fully-connected layers. Moreover, to ensure the precision of the classification, we apply the widely used mean square error (MSE) index to measure the difference between the decoded image and the original image. Therefore, the overall loss function of the proposed network is defined as

$$L_{last} = L_{msc} + L_{RPN} + L_k + L_{decoder} \tag{15}$$

where $L_{RPN}$ is the loss function of the typical RPN network. For the specific calculation method, please refer to the Faster R-CNN network [27]. $L_{decoder}$ is the MSE loss computed by the decoding network.

## IV. Experimental Results

In this section, we first introduce the experimental settings, including the datasets employed, the measurement criteria and some implementation specifics. Then, ablation experiments are conducted to analyze the impact of each functional module on the final algorithm performance. Finally, the comparison results of several baseline methods and the proposed method are presented to validate the overall algorithms performance.

### A. Experimental Settings

*1) Datasets:* We have designed extensive experiments on two widely applied and high-quality datasets to demonstrate the validity of the proposed approach. It is worth noting that the first dataset mainly includes ships in natural scenes, that is, the main interferences in this dataset are islands, reefs and waves on the sea surface and thick clouds in the sky, while the other dataset mainly contains ships in port scenes, which means that there are many artificial facilities interferences in the second dataset images.

The first dataset is the Airbus ship detection challenge dataset (available from: https://www.kaggle.com/c/airbusship-detection). The Airbus dataset is regarded as having the most samples in the ship detection research community and has nearly two hundred thousand training samples and more than ten thousand testing samples. The sizes of all images are 768 $\times$ 768 pixels. Considering that the dataset is very large and most of the samples are pure background, we choose six thousand images from the dataset as the experimental sample

TABLE I
ABLATION EXPERIMENTAL RESULTS ON THE AIRBUS DATASET (IN THE FOLLOWING TABLE, $AP_{50}$ AND $AP_{75}$ REPRESENT AP VALUES WHEN THE IOU THRESHOLD IS 0.5 AND 0.7, RESPECTIVELY. THE VALUES IN BRACKETS INDICATE THE INCREASE OF AP INDEXES COMPARED WITH THE BENCHMARK ARCHITECTURE, AND THE BOLD NUMBER REPRESENTS THE HIGHEST EVALUATION INDEX OF ALL COMBINATIONS).

| Settings | MSS | AER | COC | Module combination | $AP_{50}$ (%) | $AP_{75}$ (%) | AP (%) |
|---|---|---|---|---|---|---|---|
| RetinaNet | × | × | × | None | 86.02 | 80.42 | 72.64 |
| RetinaNet | ✓ | × | × | 1 | 88.38 | 82.01 | 73.48 (+0.84) |
| RetinaNet | × | ✓ | × | 2 | 89.89 | 83.83 | 73.88 (+1.24) |
| RetinaNet | × | × | ✓ | 3 | 89.72 | 83.68 | 73.76 (+1.12) |
| RetinaNet | ✓ | ✓ | × | 1,2 | 93.08 | 85.37 | 75.96 (+3.32) |
| RetinaNet | ✓ | × | ✓ | 1,3 | 92.45 | 86.43 | 76.62 (+3.94) |
| RetinaNet | × | ✓ | ✓ | 2,3 | 93.3 | 87.54 | 77.88 (+5.84) |
| RetinaNet | ✓ | ✓ | ✓ | 1,2,3 | 94.15 | 88.72 | **80.12 (+7.48)** |

TABLE II
ABLATION EXPERIMENTAL RESULTS ON THE HRSC2016 DATASET.

| Settings | MSS | AER | COC | Module combination | $AP_{50}$ (%) | $AP_{75}$ (%) | AP (%) |
|---|---|---|---|---|---|---|---|
| RetinaNet | × | × | × | None | 90.03 | 85.35 | 80.81 |
| RetinaNet | ✓ | × | × | 1 | 91.48 | 86.27 | 82.24 (+1.43) |
| RetinaNet | × | ✓ | × | 2 | 93.06 | 88.11 | 83.42 (+2.61) |
| RetinaNet | × | × | ✓ | 3 | 92.65 | 87.87 | 83.13 (+2.32) |
| RetinaNet | ✓ | ✓ | × | 1,2 | 93.42 | 90.45 | 85.35 (+4.54) |
| RetinaNet | ✓ | × | ✓ | 1,3 | 93.23 | 89.28 | 84.36 (+3.55) |
| RetinaNet | × | ✓ | ✓ | 2,3 | 95.47 | 92.52 | 87.10 (+6.29) |
| RetinaNet | ✓ | ✓ | ✓ | 1,2,3 | 96.02 | 93.33 | **89.79 (+8.98)** |

set. In particular, the selected sample scenes cover a variety of complicated natural components, such as ocean clutter, clouds, wakes and islands.

The second dataset we employed is the HRSC2016 ship dataset [29], in which the images are gathered from six famous ports in the visible domain. The dataset contains 1061 images, and the image resolutions range from 0.4 to 2 meters. Moreover, these images vary in size and are mainly from 300 × 300 to 1500 × 900 pixels. Noting that all these images in this dataset are provided with three levels of annotations, that is, ship class, ship category and specific type. Since our method mainly focuses on the granularity of class detection, we verify the performance of the proposed method by applying the labels on the first level.

*2) Evaluation Metrics:* In the experiments, the average precision (AP), false alarm rate (FAR) and frames per second (FPS) are adopted as the metrics for quantitatively evaluating the performance of the proposed method. These metrics are the most widely employed indexes in object detection applications. The AP index is defined as

$$precision = \frac{TP}{TP + FP} \tag{16}$$

$$recall = \frac{TP}{TP + FN} \tag{17}$$

$$AP = \int_0^1 precision(recall) d(recall) \tag{18}$$

where $TP$ and $FP$ indicate true positive and false positive, respectively. Similarly, $FN$ represents false negative. The higher the AP value, the better the algorithm performance.

The false detection probability of ship targets directly determines the practical application ability of the proposed algorithm, consequently, the false alarm rate is introduced as a key indicator, which is calculated as

$$FAR = \frac{N_{fa}}{N_{dc}} \tag{19}$$

where $N_{fa}$ represents the number of detected false alarms. $N_{dc}$ indicates the total number of predicted targets. A lower FAR means a better performance.

*3) Implementation Details:* The proposed model was implemented with the PyTorch framework and was end-to-end trained on an Nvidia RTX 2080 GPU. The ResNet-101 pre-trained model was utilized as our basic feature extractor. The size of the input image was set to 800 × 800 pixels. Moreover, we applied the Adam optimizer to update the network weight after each iteration with a weight decay of 0.001. The initial learning rate was set to 0.01 and it decreased by a ratio of 0.1 after the 10th epoch. In the training process, we set the batch size to 4 and trained the model with 300 epochs. Besides, after several comparative experiments, $\alpha$ in Equation (3) is set to 0.3 to make the semantic supervision loss curve obtain excellent convergence efficiency. In order to obtain a sufficient training result, we assigned the proportion of the training set, verification set and test set to 7:2:1. Considering that the randomness of ship direction may lead to dramatic fluctuations in the intersection over union (IOU) values, we compared a variety of commonly used IOU thresholds for the fair comparison of various methods.

## B. Results and Discussion

*1) Ablation Experiments:* To measure the contribution of each proposed module, we carried out component-wise experiments on the two datasets. The high-quality RetinaNet structure [42] was adopted as the benchmark method. The proposed modules were configured on different branches of the basic architecture, and their contribution to the final results was

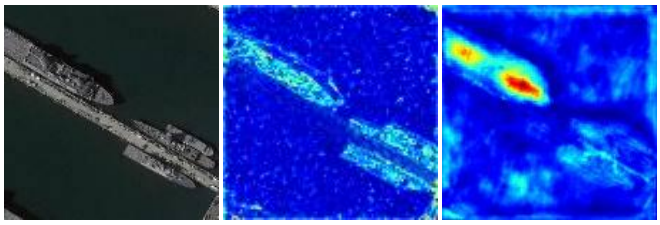


Fig. 5. Visualization results of the AER module. (a) $W_s$ of the $F_3$ branch and (b) $W_c$ of the $F_3$ branch.

TABLE III
COMPARATIVE EXPERIMENTAL RESULTS ON THE AIRBUS DATASET.

| Methods | Backbone | Input image size | AP (%) | FAR (%) | FPS |
|---|---|---|---|---|---|
| SSD [30] | VGG16 | 512×512 | 67.78 | 11.25 | 23 |
| FRIFB [31] | - | - | 67.89 | 11.04 | 5 |
| ORSIm [32] | - | - | 68.03 | 10.12 | 5 |
| YOLOv3 [33] | CSPDarknet53 | 416×416 | 68.12 | 10.42 | 28 |
| Faster R-CNN [27] | ResNet-101 | 800×600 | 69.38 | 12.02 | 4 |
| DAFA [34] | MobileNetV2 | 500×500 | 69.45 | 11.35 | 16 |
| YOLOv4 [35] | CSPDarknet53 | 608×608 | 69.64 | 8.40 | 34 |
| Mask R-CNN [36] | ResNet-101 | 512×512 | 71.45 | 10.12 | 4 |
| YOLOv5 [37] | CSPDarknet53 | 608×608 | 71.51 | 7.63 | 53 |
| HSF-Net [38] | ResNet-101 | 500×500 | 76.58 | 7.65 | 4 |
| RIE [39] | HRGANet-W48 | 800×800 | 78.06 | 10.47 | 25 |
| SSE attention [40] | DLA-34 [41] | 512×512 | 79.04 | 5.31 | 14 |
| Proposed | ResNet-101 | 800×800 | **80.12** | **4.78** | 5 |

analyzed by comparing the detection performance evaluation of these modules before and after used. It is worth noting that for a fair comparison, we uniformly applied the PASCAL VOC2007 metric [43] to measure the experimental results.

Table I reports the evaluation results of the ablation experiments on the Airbus dataset. The AP value gained by the basic RetinaNet network is 72.64%. Through the comparison of the AP values, we can see that the AP values increase by 0.84%, 1.24% and 1.12% with the addition of the MSS, AER and COC modules, respectively. This indicates that each module has a positive impact on the benchmark architecture. In addition, the configuration of the attention module shows the most obvious gain on the final result. When the three modules are configured on the benchmark network at the same time, the final AP evaluation value increases by 7.48%.

The experimental results on the HRSC2016 dataset are presented in Table II. Compared with the benchmark architecture, the AP improvements of adding the three modules in turn are 1.43%, 2.61% and 2.32%, respectively. Moreover, the combination of these three modules brings a significant increase in the AP term, which is 8.98%. On the whole, the effect of our proposed method in the second dataset is more significant than that in the first dataset. This may be because the HRSC2016 dataset mostly contains large and medium-sized ships, and their characteristics are richer than those of small ships in the Airbus dataset. Our method improves the representation of ship characteristics by configuring three modules in the basic network architecture, thus obtaining better detection results.

From the results of the two datasets, it can be found that among the three modules, the AER module has the most obvious improvement on the performance of the proposed detection method. In order to visually show the role of this module on the features of ships, as shown in Fig. 5, we

TABLE IV
COMPARATIVE EXPERIMENTAL RESULTS ON THE HRSC2016 DATASET.

| Methods | Backbone | Input image size | AP (%) | FAR (%) | FPS |
|---|---|---|---|---|---|
| SSD [30] | VGG16 | 512×512 | 76.37 | 15.14 | 17 |
| FRIFB [31] | - | - | 76.63 | 14.56 | 3 |
| ORSIm [32] | - | - | 78.40 | 13.32 | 3 |
| YOLOv3 [33] | CSPDarknet53 | 416×416 | 80.84 | 12.46 | 20 |
| Faster R-CNN [27] | ResNet-101 | 800×600 | 82.25 | 13.45 | 3 |
| DAFA [34] | MobileNetV2 | 500×500 | 82.28 | 13.30 | 12 |
| YOLOv4 [35] | CSPDarknet53 | 608×608 | 83.42 | 10.38 | 23 |
| Mask R-CNN [36] | ResNet-101 | 512×512 | 83.76 | 11.51 | 3 |
| YOLOv5 [37] | CSPDarknet53 | 608×608 | 83.93 | 9.47 | 35 |
| HSF-Net [38] | ResNet-101 | 500×500 | 85.72 | 12.24 | 3 |
| RIE [39] | HRGANet-W48 | 800×800 | 88.13 | 13.37 | 20 |
| SSE attention [40] | DLA-34 | 512×512 | 88.65 | 6.45 | 10 |
| Proposed | ResNet-101 | 800×800 | **89.79** | **6.25** | 4 |

provide the visualization results of $W_s$ and $W_c$. It can be found that for some remote sensing scenes, attention-based feature enhancement by blocks is conducive to distinguishing ships and interferences in different positions. In addition, it may be difficult to highlight all targets by directly allocating attention weights to the whole image scene. Therefore, the above evaluation results demonstrate the effectiveness and necessity of the proposed modules.

*2) Comparative Experiments:* In order to verify the overall detection performance of the proposed framework, we compared the method with several state-of-art detectors, including classical SSD [30], Faster R-CNN [27], YOLOv4 [35], HSF-Net [38] and SSE attention [40]. Since the open source codes of some methods adopt the horizontal prediction box by default, for fair comparison, all comparison methods use the horizontal annotation box uniformly. The quantitative results on the two employed datasets are illustrated in Table III and Table IV. It can be seen that the proposed method obtains the highest AP and the lowest FAR, outperforming all compared methods on both datasets. In contrast, among these compared methods, SSD gains the lowest AP. As for the false alarms, Faster R-CNN and SSD gain the highest FAR values on the two datasets, respectively. In addition, since the HRSC2016 dataset has a higher resolution than the Airbus dataset, it contains more details of ship targets, leading to a higher AP for the entire dataset. However, the artificial facilities in the port scene cause serious interferences to the ship detection, resulting in more false detection on the HRSC2016 dataset. Besides, as can be seen from the Table III and Table IV, we also compare the calculation time of different algorithms. On the whole, the typical single-stage networks, such as the SSD and the YOLO-based models, show speed advantages over the representative two-stage Faster R-CNN. The DAFA, RIE and SSE attention methods can obtain significantly faster computing speed than typical two-stage networks by using lightweight backbone networks, but they are difficult to ensure the balance of the AP and FAR indexes. On the contrary, our method pays attention to the balance between algorithm precision and false alarm, so as to better meet the needs of practical applications.

In order to intuitively compare the processing results of the two dataset images produced by different algorithms [44], we select Faster R-CNN, HSF-Net and SSE attention approaches from all compared methods for illustration. Fig. 6 shows the

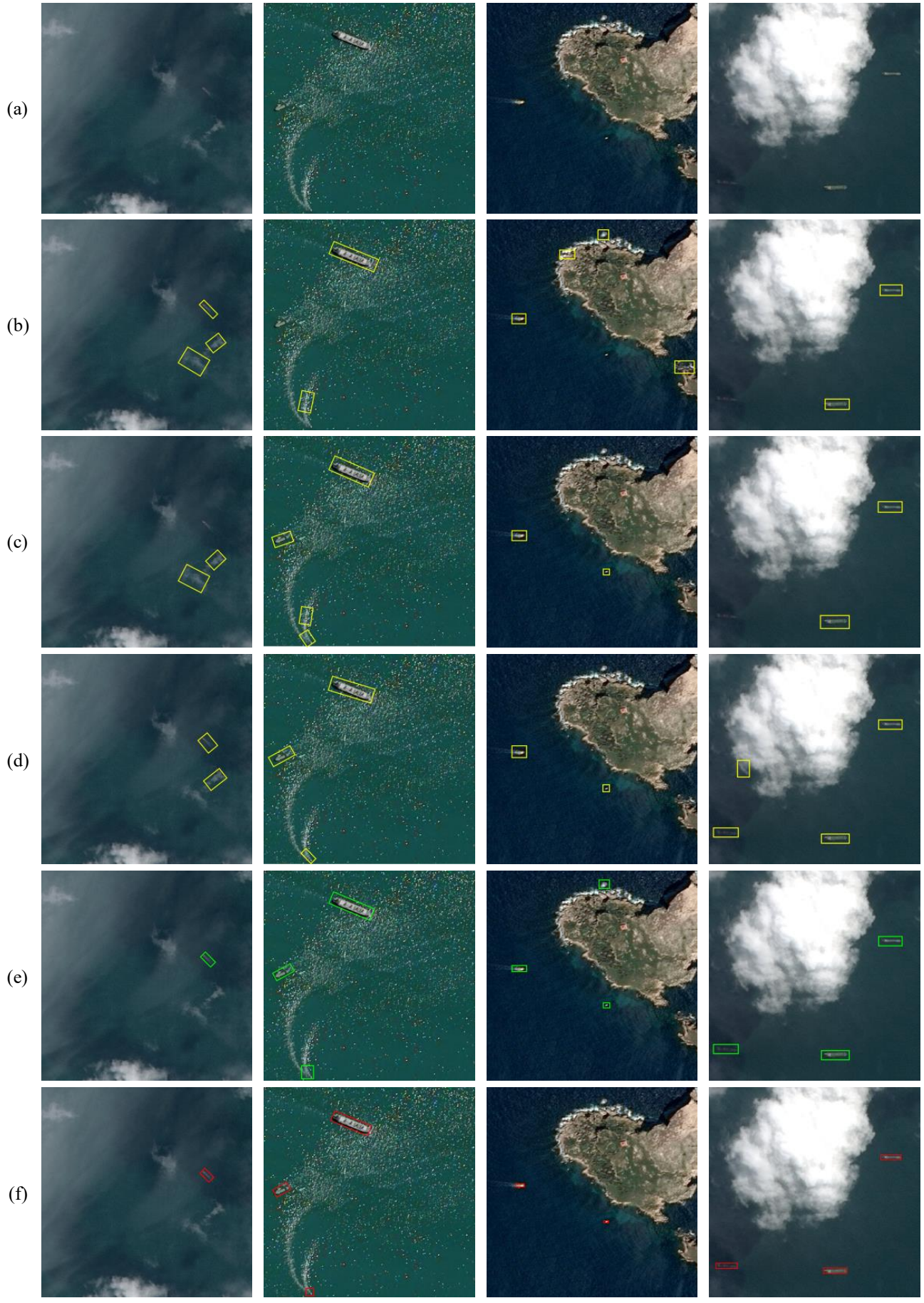



Fig. 6. Examples provided by the different considered methods on the Airbus dataset. (a) Test image, (b) Faster R-CNN, (c) HSF-Net, (d) SSE attention, (e) Proposed method and (f) Ground truth.


Fig. 7. Examples provided by the different considered methods on the HRSC2016 dataset. (a) Test image, (b) Faster R-CNN, (c) HSF-Net, (d) SSE attention, (e) Proposed method and (f) Ground truth.

application effects of these selected methods on the Airbus dataset. It can be found that the Faster R-CNN model is easily disturbed by environmental factors such as thin clouds, waves and reefs near the island, resulting in false alarms. Moreover, as can be seen from the results in the second and third columns, the Faster R-CNN method shows missed detection of small targets. This is mainly because this method predicts the candidate target regions based on the last layer of the basic extraction network. Since the information of small targets is reduced after multiple convolutions, this method has limitations in the detection performance of small targets. Particularly, the HSF-Net method can effectively detect small-scale ships in different complex scenes. This is because it adopts a hierarchical selection strategy of filter layers, which can generate the features with adaptive convolution degree for multi-scale targets. However, this method is not sensitive to targets that have a low contrast with the background (see the first and fourth columns), resulting in missed detection. In contrast, our method can still accurately detect ships even in the shadow, mainly because our attention-based module enhances the interdependence between different positions and different channels, highlighting the feature mapping of the low-contrast

target regions. In addition, the SSE attention model and the proposed method correctly detect all ships in the test samples. Actually, both methods apply attention mechanism to extract the potential target regions, but our method obtains less false alarms, mainly because our method unifies the characteristics of different scales at the semantic level and enhances the ability of feature representation, thus obtaining the training parameters more suitable for ship targets.

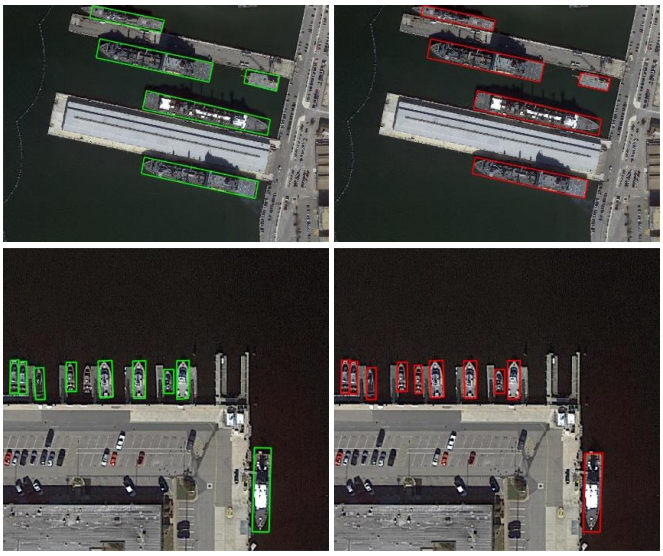
Fig. 7 provides the detection results of typical examples on the HRSC2016 dataset. The challenges of this dataset lie in the complexity of the port facilities and the possible dense distribution of ships. As can be seen from the figure, both the Faster R-CNN and HSF-Net produce false alarms in each displayed scene image. In addition, the SSE attention model and our method have a lower false detection rate than the above two methods. Besides, there is a noticeable detail in the detection results of the SSE attention approach, that is, this method easily predicts densely arranged ships as the same object. In contrast, our method accurately distinguishes two densely distributed ships, which is mainly due to the classification strategy based on the capsule packaging. The capsule-based network can completely record the direction, edge, texture and other characteristics of global and local areas, which is conducive to the accurate identification of the same object class in different forms of images. Therefore, based on these above qualitative and quantitative results, we can see that the proposed method is robust to natural scene elements and port facility interferences, outperforming all the compared techniques in terms of the AP and FAR indexes.

*3) Extended Experiments:* In order to further verify the applicability of the proposed method to other remote sensing datasets, we conduct ship detection performance verification experiments on the public and widely used DOTA-v1.0 dataset [45]. The quantitative evaluation results of the ship detection on DOTA dataset are shown in Table V, which proves the superiority of the proposed method in terms of average precision and false alarm rate. Moreover, detection results of typical examples are shown in the Fig. 8. From the image results, we can find that the proposed method has high accuracy rate for large and medium-sized ships with sparse distribution. However, as shown in Fig. 9, the proposed model still has some missed detection for the densely distributed small ships in the port scene. For visual comparison, we mark the main difference regions with the yellow dashed boxes in Fig. 9. This may require the redesign of dense blocks and transition layers to make the network pay more attention to the location distribution characteristics of small targets. To address this limitation, we plan to study more robust feature metrics to improve ship detection performance in complex scenes.

In addition, we present the model size details of different detection methods in Table V. As can be seen from the table, the model size of the proposed method is similar to that of the typical Faster R-CNN. Combining the FPS evaluation index of different detection algorithms, we can conclude that the proposed method achieve real-time performance with FPS of 5 on the premise of ensuring the best detection precision.

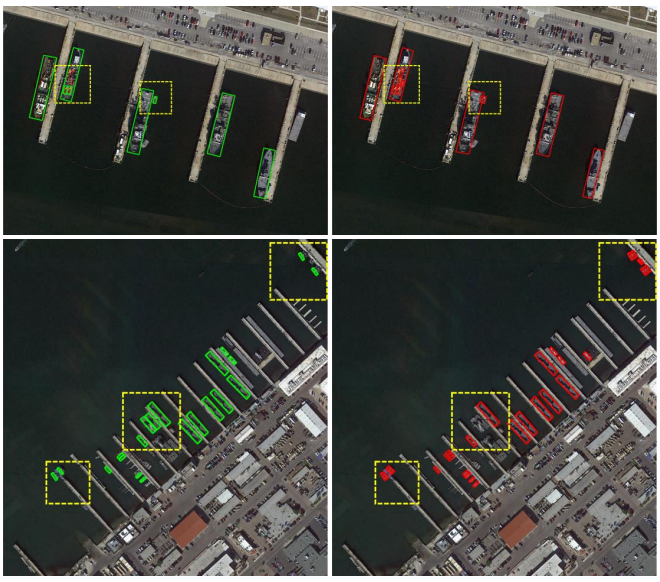
TABLE V
SHIP DETECTION RESULTS ON THE DOTA DATASET.

| Methods | Backbone | AP (%) | FAR (%) | Model size(MB) | FPS |
|---|---|---|---|---|---|
| FRIFB [31] | - | 65.21 | 19.23 | - | 5 |
| ORSIm [32] | - | 68.63 | 18.17 | - | 5 |
| SSD [30] | VGG16 | 69.54 | 16.72 | 140.3 | 23 |
| DAFA [34] | MobileNetV2 | 70.65 | 10.90 | 50.3 | 14 |
| Faster R-CNN [27] | ResNet-101 | 71.28 | 12.43 | 203.7 | 5 |
| YOLOv3 [33] | CSPDarknet53 | 71.50 | 12.30 | 235.3 | 25 |
| Mask R-CNN [36] | ResNet-101 | 72.42 | 12.65 | 233.3 | 4 |
| HSF-Net [38] | ResNet-101 | 75.21 | 10.20 | 199.8 | 6 |
| RIE [39] | HRGANet-W48 | 76.43 | 13.42 | 207.5 | 16 |
| YOLOv4 [35] | CSPDarknet53 | 76.85 | 10.02 | 289.1 | 27 |
| R-Libra R-CNN [46] | ResNet-101 | 77.42 | 7.57 | 240.4 | 5 |
| $R^3$Det [47] | ResNet-101 | 77.54 | 7.19 | 227.0 | 5 |
| YOLOv5 [37] | CSPDarknet53 | 77.69 | 9.53 | 41.2 | 42 |
| SSE attention [40] | DLA-34 [41] | 78.35 | 8.13 | 74.9 | 11 |
| $O^2$-DNet [48] | Hourglass-104 | 78.70 | 9.43 | 182.5 | 7 |
| CenterMap [49] | ResNet-101 | 79.20 | 7.75 | 351.9 | 4 |
| Proposed | ResNet-101 | **79.57** | **6.72** | 201.5 | 5 |



(a) (b)

Fig. 8. Results of the proposed algorithm on the DOTA dataset. (a) Detection results, (b) Ground truth.



(a) (b)

Fig. 9. Missing detection results of the proposed algorithm on the DOTA dataset. (a) Detection results, (b) Ground truth.

## V. Conclusion

In this paper, we propose a multi-scale attention-guided detection framework to detect ships in complex natural scenes and port conditions. Three modules are proposed to enhance the representation of effective features at different levels. Specifically, a multi-scale supervision module is first applied to handle the semantic inconsistency between features produced by different extraction layers, ensuring that these features learn similar semantic-level information. Then, to improve the ability to effectively extract targets in complex environment interferences, an attention-guided module is adopted to integrate context information from both spatial and channel dimensions by calculating the map correlation, adaptively removing redundant features and mining useful hierarchical features. Moreover, a capsule-based module is presented, which can preserve the attribute and spatial relationship between features and support the accurate classification of suspected target regions. Experimental results conducted on two public high-quality datasets indicate that the proposed method outperforms all compared methods in terms of the AP and FAR indexes.

As for the future work, considering that the false alarm rate of our current detection model is still slightly high, we first plan to extend this work in the direction of mining features that more effectively represent the characteristics of ship targets. Second, we plan to introduce the strategy of a rotating prediction box to describe the locations of ships more accurately, Third, we will further optimize strategies to achieve more detailed target information interpretation, such as type recognition and component recognition.

## References


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

[2] G. Gao, Y. Luo, K. Ouyang, and S. Zhou, "Statistical modeling of pma detector for ship detection in high-resolution dual-polarization sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4302–4313, 2016.

[3] H. Chen, T. Gao, W. Chen, Y. Zhang, and J. Zhao, "Contour refinement and eg-ght-based inshore ship detection in optical remote sensing image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8458–8478, 2019.

[4] 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 Sensing*, vol. 13, no. 16, p. 3059, 2021.

[5] S. Jiang, X. Zhi, W. Zhang, D. Wang, J. Hu, and C. Tian, "Global information transmission model-based multiobjective image inversion restoration method for space diffractive membrane imaging systems," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[7] 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 Sensing*, vol. 12, no. 20, p. 3370, 2020.

[8] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with svd networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5832–5845, 2016.

[9] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*. PMLR, 2015, pp. 562–570.

[10] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," *arXiv preprint arXiv:1703.09844*, 2017.

[11] D. Sun, A. Yao, A. Zhou, and H. Zhao, "Deeply-supervised knowledge synergy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6997–7006.

[12] L. Yao, H. Xu, W. Zhang, X. Liang, and Z. Li, "Sm-nas: structural-to-modular neural architecture search for object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 661–12 668.

[13] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[16] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[17] J. Guo, X. Ma, A. Sansom, M. McGuire, A. Kalaani, Q. Chen, S. Tang, Q. Yang, and S. Fu, "Spanet: Spatial pyramid attention network for enhanced image recognition," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[20] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[21] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[22] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *International conference on artificial neural networks*. Springer, 2011, pp. 44–51.

[23] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *arXiv preprint arXiv:1710.09829*, 2017.

[24] C. P. Schwegmann, W. Kleynhans, B. P. Salmon, L. W. Mdakane, and R. G. Meyer, "Synthetic aperture radar ship detection using capsule networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 725–728.

[25] Y. Yu, T. Gu, H. Guan, D. Li, and S. Jin, "Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 12, pp. 1894–1898, 2019.

[26] H. Ren, X. Yu, L. Zou, Y. Zhou, X. Wang, and L. Bruzzone, "Extended convolutional capsule network with application on sar automatic target recognition," *Signal Processing*, vol. 183, p. 108021, 2021.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[29] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *International conference on pattern recognition applications and methods*, vol. 2. SCITEPRESS, 2017, pp. 324–331.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[31] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 302–306, 2019.

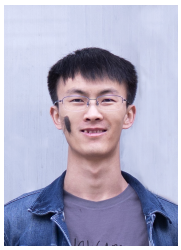
[32] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5146–5158, 2019.

[33] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

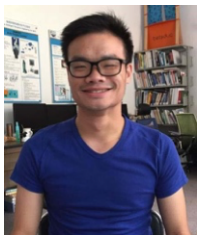
[34] X. He, S. Ma, L. He, L. Ru, and C. Wang, "Learning rotated inscribed ellipse for oriented object detection in remote sensing images," *Remote Sensing*, vol. 13, no. 18, p. 3622, 2021.
[35] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[37] B. Yan, P. Fan, X. Lei, Z. Liu, and F. Yang, "A real-time apple targets detection method for picking robot based on improved yolov5," *Remote Sensing*, vol. 13, no. 9, p. 1619, 2021.
[38] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7147–7161, 2018.
[39] F. Gao, Y. He, J. Wang, A. Hussain, and H. Zhou, "Anchor-free convolutional network with dense attention feature aggregation for ship detection in sar images," *Remote Sensing*, vol. 12, no. 16, p. 2619, 2020.
[40] Z. Cui, X. Wang, N. Liu, Z. Cao, and J. Yang, "Ship detection in large-scale sar images via spatial shuffle-group enhance attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 379–391, 2020.
[41] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.
[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
[43] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
[44] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2018.
[45] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.
[46] H. Guo, X. Yang, N. Wang, B. Song, and X. Gao, "A rotational libra r-cnn method for ship detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5772–5781, 2020.
[47] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3163–3171.
[48] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 268–279, 2020.
[49] J. Wang, W. Yang, H.-C. Li, H. Zhang, and G.-S. Xia, "Learning center probability map for detecting objects in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4307–4323, 2020.

**Jianming Hu** received the M.S. degree in optical engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2017. Currently he is pursuing the Ph.D. degree in HIT and is a visiting Ph.D. with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His research interests include remote sensing image processing, salient target detection and optical target identification.

**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 at the HIT. His current research interests including remote sensing image acquisition and processing, optical target detection and identification.



**Shikai Jiang** received the B.E. in Electronic Science and technology (2016) and the M.E. in Optical Engineering (2018) from Harbin Institute of Technology (HIT). He is currently pursuing a Ph.D. in Optical Engineering in School of Astronautics of HIT. He interested in image processing of space optical remote sensing satellite include image inversion restoration, image fusion, super-resolution, target detection and identification.



**Hao Tang** is currently a Postdoctoral with Computer Vision Lab, ETH Zurich, Switzerland. He received the master's degree from the School of Electronics and Computer Engineering, Peking University, China and the Ph.D. degree from Multimedia and Human Understanding Group, University of Trento, Italy. He was a visiting scholar in the Department of Engineering Science at the University of Oxford. His research interests are deep learning, machine learning, and their applications to computer vision.

**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 at the HIT. His current research interests including remote sensing image acquisition and processing, optical system design, automatic target detection and identification.

**Lorenzo Bruzzone** (S'95-M'98-SM'03-F'10) received the Laurea (M.S.) degree in electronic engineering (*summa cum laude*) and the Ph.D. degree in telecommunications from the University of Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications at the University of Trento, Italy, where he teaches remote sensing, radar, and digital communications. Dr. Bruzzone is the founder and the director of the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento. His current research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is the Principal Investigator of the *Radar for icy Moon exploration* (RIME) instrument in the framework of the *JUpiter ICy moons Explorer* (JUICE) mission of the European Space Agency. He is the author (or coauthor) of 215 scientific publications in referred international journals (154 in IEEE journals), more than 290 papers in conference proceedings, and 21 book chapters. He is editor/co-editor of 18 books/conference proceedings and 1 scientific book. His papers are highly cited, as proven form the total number of citations (more than 26000) and the value of the h-index (74) (source: Google Scholar). He was invited as keynote speaker in more than 30 international conferences and workshops. Since 2009 he is a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS).

Dr. Bruzzone ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. Since that he was recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award and the 2017 IEEE IGARSS Symposium Prize Paper Award. Dr. Bruzzone was a Guest Co-Editor of many Special Issues of international journals. He is the co-founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003 he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the founder of the IEEE Geoscience and Remote Sensing Magazine for which he has been Editor-in-Chief between 2013-2017. Currently he is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing. He has been Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society between 2012-2016.