Multilayer Feature Extraction Network for Military Ship Detection From High-Resolution Optical Remote Sensing Images

Peng Qin, Yulin Cai, Jia Liu, Puran Fan, and Menghao Sun

***Abstract*—Rapid and accurate detection of maritime military targets is of great significance for maintaining national defense security. Few studies have used high-resolution optical images for the detailed classification of maritime military targets. This article, inspired by EfficientDet trackers, presents a method to classify military targets on the sea from high-resolution optical remote sensing images. In the first stage, a multilayer feature extraction network is constructed to extract various features. At the same time, residual connection and dilation convolution are introduced to prevent the deep network features from disappearing. Moreover, we use multilevel attention mechanism approaches to make more effective use of multilayer features. ReLU is introduced to replace the original swish activation function to reduce the computational cost in the pretreatment stage. After this, deep feature fusion networks and prediction networks are constructed to locate and distinguish different types of ships. Different types of ships use different degrees of data expansion methods to solve the problem of sample shortage and imbalance. The multiclassification method is used to solve low classification accuracy caused by little difference between civil and military ships. Experimental results suggested that the proposed method can accurately identify multiple types of military ships.**

***Index Terms*—Attention mechanism, data enhancement, efficientdet network, feature extraction, feature fusion, military target detection, multiclassification.**

I. Introduction

THE RAPID and accurate identification of maritime military targets is crucial for understanding enemy naval vessels and military equipment on the sea. This helps the military with early warning and rapid response, thereby maintaining national defense security [1], [2]. Remote sensing imaging of ships at sea is easily affected by cloud, wave, light, etc., and if boats dock at coastal ports, their imaging will also be influenced by the background of coastal factors [3], [4]. Therefore, it is important but challenging to detect and classify maritime military targets on the sea quickly and accurately in a large range and multiple scenes.

Various types of remote sensing data are widely used to identify vessels on the sea due to their characteristics of economy and timeliness [5]–[7]. However, the resolution of some data is relatively low and vulnerable to noise [8], which brings difficulties to coastal ship identification [9]. Optical remote sensing image, with its high spatial resolution, makes up for the shortage of other remote sensing data to a great extent and, thus, is widely used [10], [11]. Still, the traditional ship classification algorithm of optical remote sensing images mostly uses manually designed features [9], [12]–[14], these features are easily affected by some clouds and waves, and the robustness is poor [15], which is not suitable for large-scale and multiscene ship target recognition [16].

In recent years, with the development of deep learning technology, in terms of scene recognition [17], image segmentation [18], or target detection [19], deep learning methods show some obvious advantages compared with traditional image recognition methods. Especially in the aspect of image target detection, some neural networks have a great improvement in accuracy and efficiency [20]. All kinds of target detection networks can be divided into one-stage and two-stage networks according to whether candidate regions are generated [21]. The one-stage network represented by single shot multibox detector (SSD) [22] and you only look once (YOLO) [23] can quickly and directly detect the entire image, but there are great limitations in accuracy [24]. To improve the accuracy, scholars tried to divide the whole image into different candidate regions and then classify and locate them. This greatly reduces the interference of negative samples, and the final model accuracy will be improved [25].

Recent studies have shown that the application of deep learning to remote sensing detection of ships at sea has achieved promising results. Networks such as R-CNN (Region proposals with CNN) network and improved Fast R-CNN can rapidly detect vessels [26], [27]. The latest research found that using a multiscale feature extraction network to extract multiscale features, the capacity to enhance model accuracies in ship extraction [28]. However, most networks only distinguish between ships and nonships, with high intraclass variation causing misclassification [16]. Multiclass methods [12], [16], [29] provide a viable solution because the results have more categories, and more

Manuscript received August 25, 2021; revised October 2, 2021; accepted October 23, 2021. Date of publication October 27, 2021; date of current version November 10, 2021. This work was supported by the National Natural Science Foundation of China National Defense Science and Industry Administration Civil Aerospace "Thirteenth Five-Year Plan" technology advance research. (*Corresponding author: Yulin Cai.*)

Yulin Cai, Jia Liu, Puran Fan, and Menghao Sun are with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: caiyl@sdust.edu.cn; liujia123@sdust.edu.cn; fanpuran@sdust.edu.cn).

Peng Qin is with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China, and also with the School of Geospatial Engineering and Science, Sun Yat-sen University, Guangzhou 510275, China (e-mail: qinpeng1995@sdust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3123080

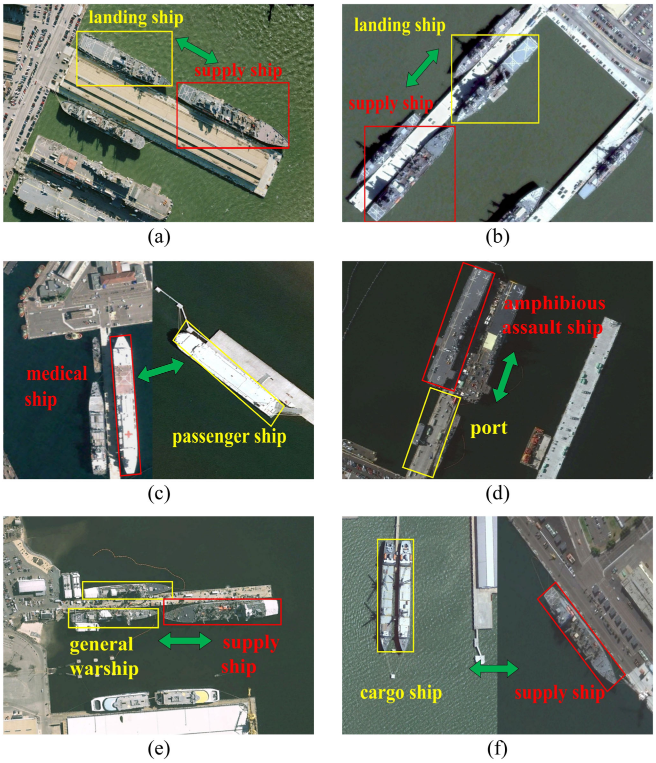


Fig. 1. Various ships with similar characteristics: (a) (b) supply and landing ship; (c) medical ship and passenger ship; (d) amphibious assault ship and port; (e) general warship and supply ship; (f) cargo ship and supply ship.

detailed ship information can be extracted. In order to improve the accuracy of these methods, some specially designed features have been created. For example, before the ship detection, the water land separation is carried out, or some other information is used to extract the hull [16]. These non-end-to-end [28] measures increase the complexity of the model, and the quality of feature extraction will also affect the final accuracy. What's more, most of the studies on ship identification are aimed at ordinary ships and the military ships are directly classified as general ships or a single category without detailed division [30], [31]. At the same time, the differences between some types of military ships are small and difficult to distinguish. As shown in Fig. 1, the characteristics of general warships, landing ships, and supply ships are very close, medical ships have similar features to passenger ships, amphibious assault ships are to be mistakenly be classified as port terminals. At the same time, supply ships are very similar to cargo ships in terms of functions and features. All of these similarities bring great difficulties to the classification of specific military ships. This is another reason why there is little research on warship classification.

Considering the abovementioned situation, this article proposes a novel method of warship recognition based on deep learning. The network can extract richer features while ensuring efficiency and accuracy, which is conducive to the detailed classification of military ships. High-resolution remote sensing images can be directly input into the network to obtain high-precision ship classification results without manually extracting other features of the image. Our contributions are as follows.

1) To accurately extract different sizes of warships, especially some large warships, in the first stage, we construct a multilayer feature extraction network based on a multilevel attention mechanism to increase feature values and enrich semantic information [4]. Additionally, residual connection [32] and dilation convolution module [33] are introduced into the deep network to avoid vanishing gradients. These features are input into the feature fusion network in the next stage [34].
2) The ReLU activation function is used to replace the swish activation function with a high computational cost at the beginning to cooperate with the use of a deep network.
3) Imbalance of the number of different samples will lead to low recognition accuracy [35]. In order to solve this problem, different degrees of data expansion are used to reduce the difference in the number of different samples.
4) We made a more detailed classification to highlight intraclass differences to avoid misclassification [16]. In addition to dividing military ships into seven classes, civilian ships are also divided into two categories.

II. Data Preparation

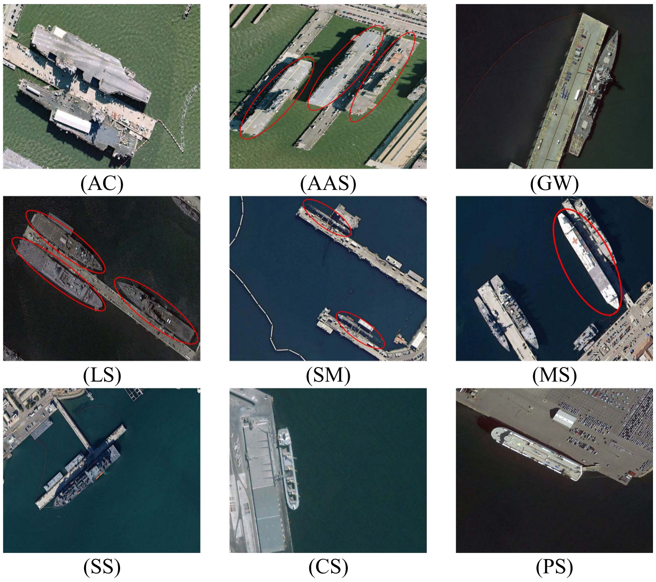
A. Data Acquisition

The data used in the experiment is the "2016 High Resolution Ship Collections" ship dataset [36], which contains 1072 Google Earth images with a resolution ranging from 0.4 m to 2 m, and most of the images are about 1200 × 800 pixels in size. Some of the images that do not contain military ships and the types of ships that cannot be visually identified are removed. After screening, a total of 777 images were selected, with 472 for training and 305 for testing. These images include about 3000 ships of various types, of which 450 are civilian ships, and the rest are military ships. The dataset contains a variety of recognition backgrounds of different times, places, sea states, and weather, which is in line with the actual classification scene.

B. Data Enhancement

In order to detect specific types of military ships, it is necessary to distinguish different types of ships in the training data set. Due to the limited image resolution and the slightly different rules for naming ships in different countries. Referring to the naming rules of the world's mainstream warships [37], we rename the vessel according to the purpose, shape, size, bridge, deck, naval gun, and apron. As illustrated in Fig. 2, military ships are divided into seven categories, including aircraft carriers (AC), amphibious assault ships (AAS), general warships (GW), landing ships (LS), submarines (berthed in ports, SM), medical ships (MS), and supply ships (SS). The civilian ships are divided into two categories, including cargo ships (CS), and passenger ships (PS).

To obtain sufficient training samples, the traditional data expansion method performs translation, rotation, and random noise addition of all data [38]. In this article, we have performed different degrees of data expansion on different samples because the number of military ships in the image is small and unevenly



(AC) (AAS) (GW)

(LS) (SM) (MS)

(SS) (CS) (PS)

Fig. 2. Specific characteristics of various ships.

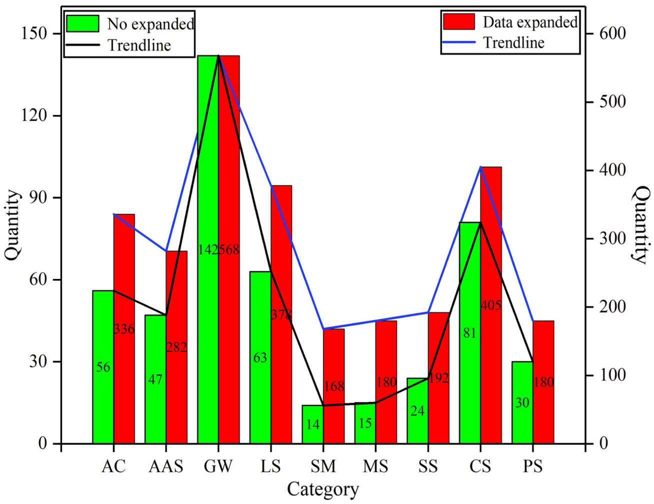


Fig. 3. Number of statistics before and after data expansion.

distributed. The original training data set contains 56 images of aircraft carriers, 47 amphibious assault ships, 142 general warships, 63 landing ships, 24 supply ships, 14 submarines, 15 medical ships, 81 cargo ships, and 30 passenger ships. The number of images containing different ships varies greatly. In addition, the orientation of different ships is also different. These factors will affect the accuracy of the classification results [35], [39]. Hence, we chose to expand the dataset by rotating the original images at different angles. The images of aircraft carrier, amphibious attack ship, general warship, landing ship, supply ship, submarine, medical ship, cargo ship, and passenger ship have been expanded by six times, six times, four times, six times, eight times, twelve times, twelve times, five times, and six times. Where the expansion of six times means the original image will be rotated by 60°, 120°, 180°, 240°, 300°. Finally, the resulting training data set contains 2689 images. Fig. 3 shows the changes in the samples of different categories of vessels before and after the expansion. The change curve of the expanded samples is flatter than that before the expansion, which means that the difference in the number of different kinds of samples is reduced to a certain extent.

Finally, different types of ships in the training dataset and the test set data are marked with rectangular boxes and labelled with text. Since warships are generally large in size, we only marked ships with a length of about 50 m or more, with other ships as background.

III. Method

EfficientDet is a two-stage new target detection network [40]. When proposed in 2020, it achieved a score of 51.0 mAP (mean Average Precision) on the common object in context dataset, which was the highest level at the time. Unlike the previous neural network methods that only change one dimension, the backbone feature extraction network of EfficientDet is a three-dimensional model scaling method, including network depth, network width, and image resolution. At the same time, the feature fusion network used in the second stage of the network uses different weights for fusion according to different scale features extracted in different stages. This not only ensures efficiency but also integrates multiscale information to extract multifeature military ships. In this article, we refer to the structure of this network to build a multilayer feature extraction network to achieve ship detection.

A. *Multilayer Feature Extraction Network*

In general, shallow networks extract more generalized information, such as the general structure of a ship, while deep networks extract more detailed information, such as the characteristics of specific different kinds of military ships [41]. But the deep network is sometimes unable to perform gradient updates. By constructing residual neural networks, the features can be extracted from deep networks [32]. Therefore, we build a residual unit based on the structure of the residual network (as shown in Fig. 4) to extract multilayer features. First, a 1×1 convolution is performed on the input features to increase the image dimension. The second step is to carry out a depthwise separable convolution [42], which has fewer convolution parameters and is beneficial to deepening the network. The third step is an attentional mechanism model, which will be described later, the output of this step, after the convolution operation, will be fused with the initial input features to form the final output. With the above operation, we construct a large residual unit, the backbone feature extraction network is to repeatedly stack these residual units into seven layers, each layer repeats these residual units 1, 2, 2, 3, 3, 4, 1 times, respectively. At the same time, the dimension of the image gradually rises from 3 to 320 dimensions.

In the process of constructing the residual unit, the number of channels of the feature map is changing, after visualized the outputs of different channels, we find that the response is different for different targets on different channels (as shown in Fig. 5, a general warship shows different responses in different

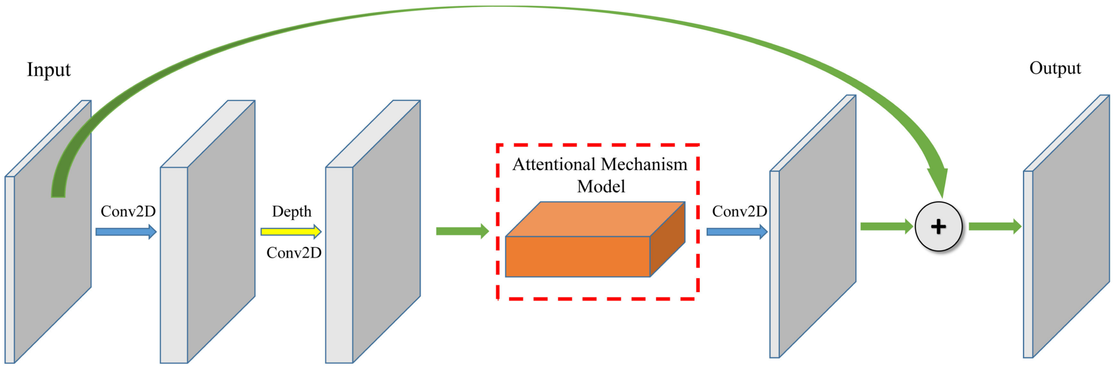


Fig. 4. Residual unit structure.

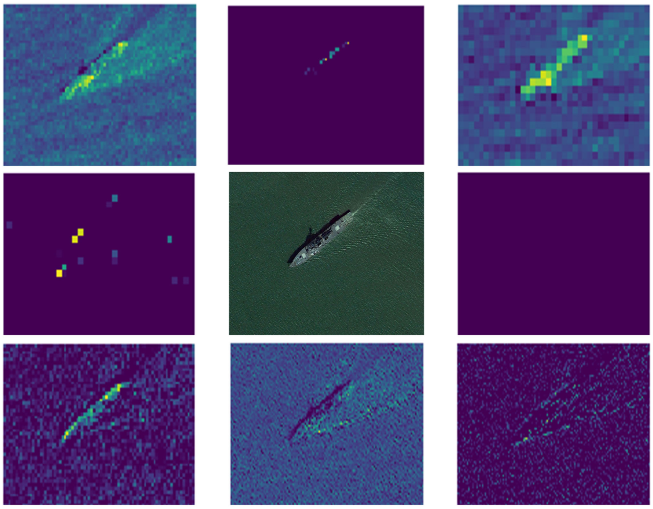


Fig. 5. Channel visualization: the original image in the middle, the rest are the results of different channel feature map visualization.

channels: some channels show a clear response to shape, some channels show a clear response to warship feature points, while some channels show little or no response). Since the accuracy of the subsequent prediction network extraction results depends on the quality of these feature mapping, we improve the accuracy of the model by increasing the focus on the high response channels. Specifically, in the backbone feature extraction network, the extracted feature maps are fed into the attention mechanism model. As shown in Fig. 6, in different attention mechanism models, firstly, feature maps with length, width and number of channels n, n, c, respectively, are converted to 1×1×c feature map after a global averaging pooling. Two convolutional layers follow these features to change the size and dimension. In the meantime, the activation function of the first operation is ReLU, and the activation function of the second operation is sigmoid. As shown in (1), after the sigmoid function processing, the input values will be mapped to the interval of 0-1. Therefore, the feature map is processed by the attention mechanism model and transformed into a c-dimensional vector with each value between 0 and 1, where c corresponds to the number of channels of the input feature map. Finally, the vector is multiplied with the initial input feature map to give different weights to the channels with different responses. It also means that give different attention to different channels to strengthen the primary information and weaken the secondary information.

$$\text{sigmoid} : f(x) = \frac{1}{1 + e^{-x}}. \tag{1}$$

We transformed the input remote sensing images into some features from shallow to deep layers through the feature extraction network. Since these features are large in number and vary in quality, we processed them using an attention mechanism to select some useful information in every residual unit. Finally, a multilayer feature extraction network with multilevel attention mechanism models is formed.

B. *Feature Fusion Network*

Since the size of various ships is not fixed, it is necessary to extract small target features based on shallow networks and large target features based on deep networks [4]. Therefore, the extracted feature information needs to be integrated. For this reason, we build a top-down feature fusion network, in which five layers of features are input from bottom to top, as shown in Fig. 4. First, the upper layer features extracted by the backbone network are up-sampling and merged with those of the low layers. The specific steps are: the top-layer (P5) features are up-sampling and merged with the fourth-layer features (P4) to obtain A4; A4 is up-sampling and fused with P3 to get A3, and A2 is obtained in the same way. Second, A2 is up-sampling and is fused with P1 to obtain B1. Third, the lower layer features are down-sampling and merged with the upper layer ones. Specifically, B1 is downsampled and merged with A2 and P2 to obtain B2, B2 is downsampled and merged with A3 and P3 features to obtain B3, and so on to obtain B4 and B5. Thus, a total of 5 layers of features, B1–B5, are output, and a complete feature fusion unit is formed, as shown in Fig. 7. The output of the previous feature fusion unit will be used as the input of the next unit. The feature fusion network is finally formed by repeating the feature fusion unit four times.

C. *Category Prediction and Box Prediction Network*

The category and position of each target in the image can be obtained by inputting the fused features into the category prediction network and the box prediction network.

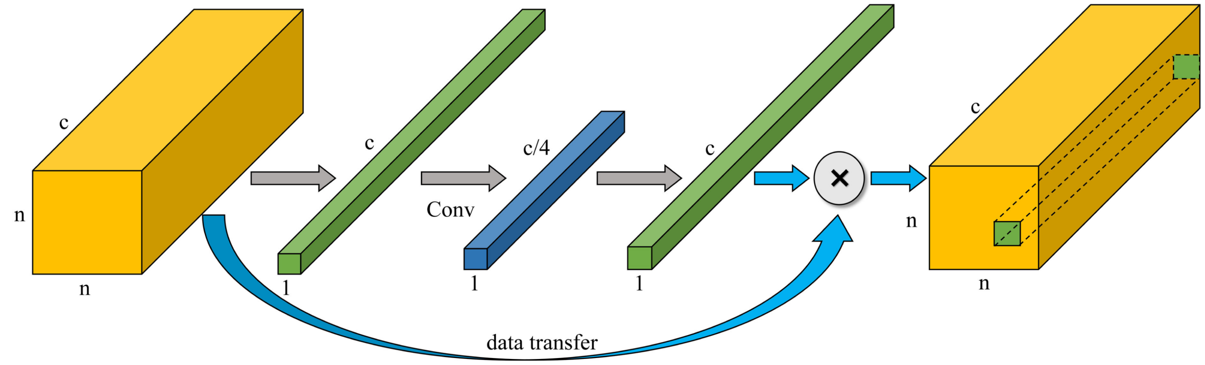


Fig. 6. Attention mechanism model.

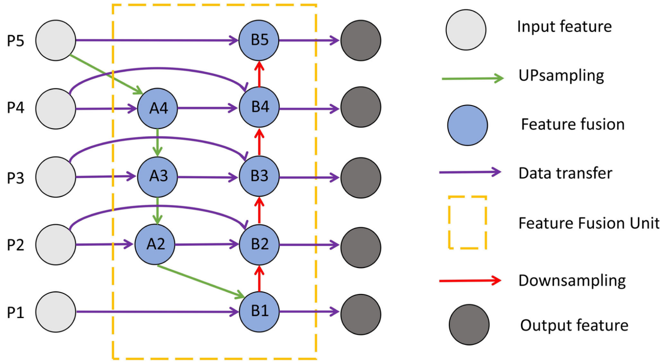


Fig. 7. Feature fusion unit.

In the category prediction network, three depthwise separable convolutions are first carried out to adjust the number of image channels to 88. Then, another depthwise separable convolution with a filter depth of 9 × 9 is performed, where the two 9 s, respectively, represent the number of preset prediction boxes and prediction categories in each candidate region. The loss function used in this network is Focal, in order to further reduce the imbalance between different samples, a balancing factor is added to the loss function [43]. The Loss function Focal is defined as

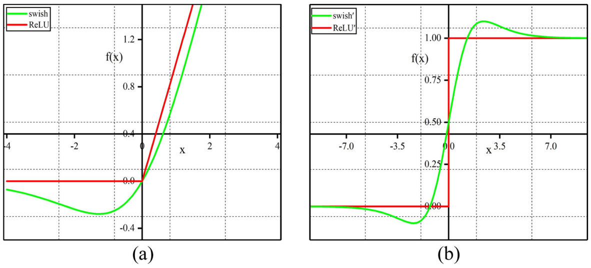
$$\mathcal{L}(\text{Focal}) = \begin{cases} -\alpha(1-y')^{\gamma}\text{log}y' & y=1 \\ -(1-\alpha)y'^{\gamma}\log(1-y') & y=0 \end{cases} \tag{2}$$

where α andγrepresent adjustment parameters, which are constants 0.25 and 2, respectively, and y' is the predicted probability value of output.

In the box prediction network, six depthwise separable convolutions are performed first with a filter depth of 88, and then another depthwise separable convolution is used to adjust the number of channels to 9×4. Here, nine represents the number of predicted boxes and four represents four adjustment parameters of each prediction box. Smooth_L1 [25] is set as the loss function in this process, which is formulated by

$$\mathcal{L}(\text{Smooth\_}L1) = \begin{cases} 0.5x^2 & |x|<1 \\ |x|-0.5 & |x|\geq 1 \end{cases} \tag{3}$$

where x is the difference between the real value and the predicted value.



(a) (b)

Fig. 8. Comparison of two different activation functions: (a) original function; (b) derivative function.

D. Other Detail Improvements

In addition to the adoption of multicomponent model and the change of data expansion method in the data preparation stage. We also made some modifications to the original EfficientDet network to improve the accuracy and efficiency of the ship information extraction, including replacing activation functions and changing network structure.

In order to extract more semantic information, it is generally necessary to continuously deepen the network by boosting the number of layers [44]. In the model pre-processing stage original EfficientDet network uses swish [40] as the activation function (4), but the calculation cost of swish is relatively high. Therefore, the ReLU activation function is selected here (5), which has less calculation and faster calculation speed. Without the influence of exponential function, the calculation process is greatly simplified. Additionally, from Fig. 8, we can see that ReLU activation function transforms the original complex features into more discrete and simpler features [45], the simpler derivative function allows the gradient to be updated more quickly. It is more conducive to distinguish the subsequent features in the preprocessing stage of the model, and the actual performance is also better.

The swish and ReLU activation function is formulated by

$$f(\text{swish}) = x \cdot \frac{1}{1+e^{-\beta x}} \tag{4}$$

$$f(\text{ReLU}) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{5}$$

where is β a constant (1.0) and x is the input.

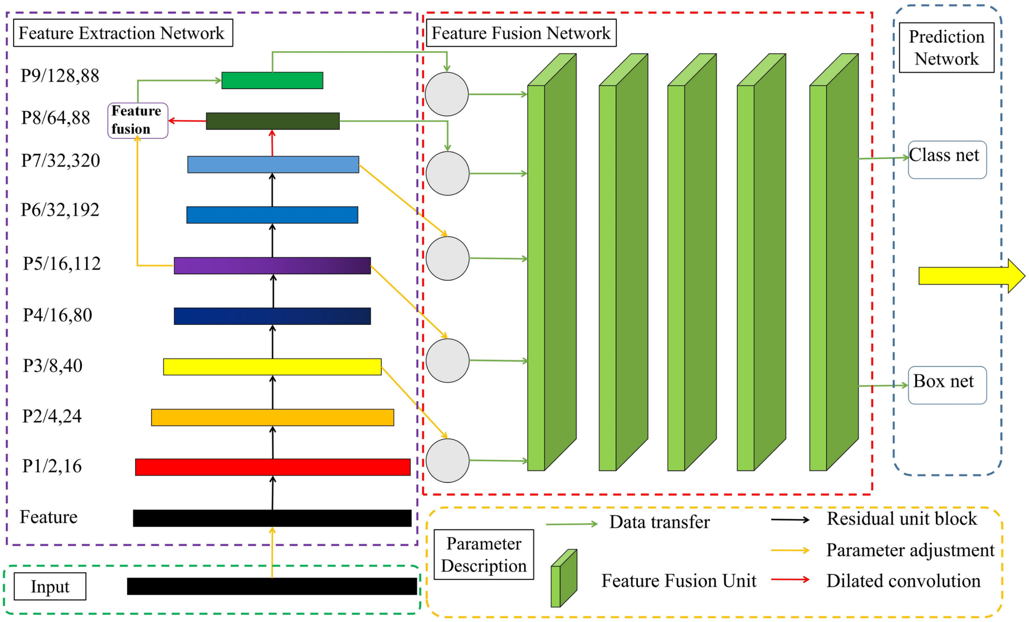


Fig. 9. Network structure of the model in this article.

The original network directly uses the backbone feature extraction network's 3–7 layer features for feature fusion. However, due to the large changes in the size of various ships and the high resolution of remote sensing images, the extracted features of different types of ships are not obvious, which is not conducive to the classification of specific types of ships. Especially for the larger size of the vessels, the deep features extracted are insufficient, so we choose to extract third, fifth, seventh layer features and input them into the fusion network of the next stage. This avoids the reuse of information, makes the extracted features more discrete, and is more conducive to the differentiation of ships of different sizes, and properly adjusts the size of each layer feature map. At the same time, on the basis of layer 7, an expansion convolution, normalization, and maximum pooling are carried out to obtain a new layer of feature P8. The purpose of using expansion convolution here is to avoid the disappearance of deep network features, improve the resolution of feature maps [46]. When we continue to use dilation convolution on this new feature, the shallow features become very limited with the deepening of the convolution layer, which is not conducive to the extraction of small-size vessels. Therefore, residual connection [32] is introduced and P8 is fused with the fifth layer feature after expansion convolution to obtain the last feature layer P9.

Prediction results will be obtained after inputting these fused features into the final category prediction and boxes prediction network. The final network structure is shown in Fig. 9, and the leftmost label is the change in the size and dimension of the feature map.

E. Accuracy Evaluation

To evaluate the integrity rate and false alarm rate of ship recognition, we take recall and precision as the accuracy evaluation criteria. The calculation formula for recall and precision is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

The higher the recall, the more ships that are correctly identified, and the higher the precision, the higher probability that the identified ship is a real ship. When counting the number of ships, if more than 90% of the area of a ship is correctly identified, it is considered as a positive sample (TP), otherwise, it is regarded as an incorrect negative sample (FN). If it is not a ship but is identified as a ship, this is a false positive sample (FP).

IV. Results

A. Model Results

After the model is trained, the test data are imported into the model to get the predicted value. Fig. 10 shows loss function curve and Precision-Recall curve for different categories in the test set. Table I is the comparison result between the predicted values with the labels of the test data, and the recall and precision of each ship category are then calculated based on Table I, as shown in Table II. After statistics, the overall recall for warships is 0.935, and the precision is 0.970 (remove the two categories of cargo ships and passenger ships). The specific identification effects of all kinds of ships are shown in Fig. 11. Results show that, in various complex scenes and multitarget scenes, the method proposed in this article has high recognition accuracy for all types of ships. It can be seen from Table II that the precision of amphibious assault ships and passenger ships is lower than that of other types of ships, possibly due to the influence of the background of nearshore ports. In addition, as can be seen from

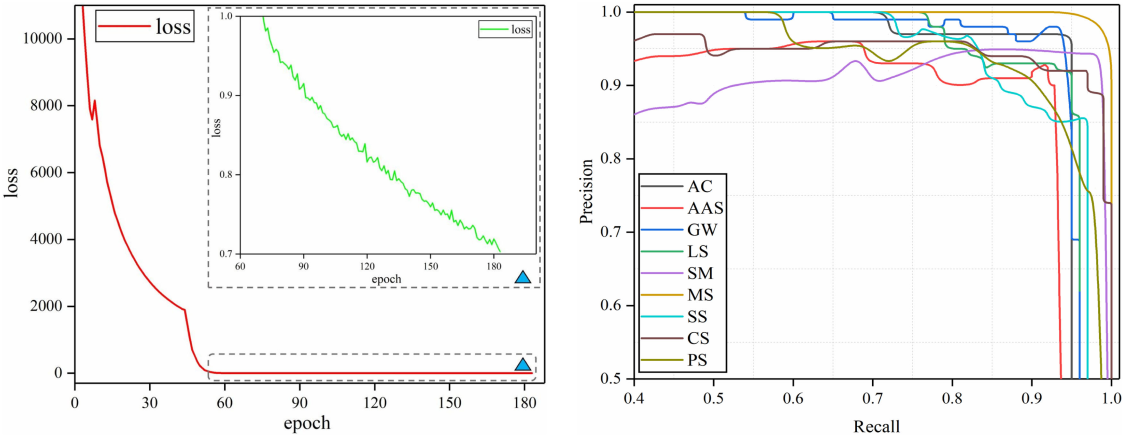


Fig. 10. Left: loss function curve; right: Recall-Precision curve.

TABLE I
VARIOUS TYPES OF SHIP IDENTIFICATION STATISTICS

No.	AC	AAS	GW	LS	SM	MS	SS	CS	PS	other
AC	39	1	0	0	0	0	0	0	0	0
AAS	0	40	0	1	0	0	1	0	0	2
GW	0	1	344	3	0	0	0	1	0	1
LS	0	0	0	73	0	0	3	0	0	1
SM	0	0	0	0	50	0	0	0	0	3
MS	0	0	0	0	0	10	0	0	0	0
SS	0	0	0	1	0	0	37	0	0	0
CS	1	0	0	0	0	0	3	77	1	1
PS	1	0	0	0	0	0	1	3	26	2
other	1	1	20	0	1	1	0	1	0	null

Horizontal axis is the true category, and the vertical axis is the predicted category.

TABLE II
ACCURACY OF VARIOUS SHIPS

NO.	AC	AAS	GW	LS	SM	MS	SS	CS	PS
Recall	0.929	0.930	0.945	0.936	0.980	0.909	0.822	0.939	0.963
Precision	0.975	0.909	0.983	0.948	0.943	1.000	0.974	0.928	0.788

Table I, supply ships are easily classified into other categories because they are similar to cargo ships and landing ships, and the number of samples is small. But the model still maintains a fairly high accuracy overall.

B. Ablation Experiments

We compared the performance of ship extraction before and after the model was improved. The comparison results are as follows

1) In the traditional ship identification research, the target ship is only labelled as a warship without specific classification. The data obtained in this way is input into the network and trained to obtain a two-class model. Experiment results show that the overall recall of the warship identified by the model is 0.973, and the precision is 0.896. Although the recall is very high, the precision is relatively low. Moreover, although there are few civilian ships, they are easily identified as warships, and the backgrounds such as ports are also easily mistaken for warships. For this reason, in the experiment, we chose a multiclassification method to classify ships, rather than simply to divide the dataset into two types: warships and nonwarships. Experiment results show that this multiclassification strategy can avoid misclassification caused by small differences between classes. After adopting this strategy, the overall recall of the warship is 0.935, and the precision is 0.970. Although the recall is slightly lower, the precision is much improved.
2) The traditional data expansion method expands the datasets of different targets by the same multiple. Following this rule, we rotate all types of ships by the same six times, and then input them into the network for model training. The final recall and precision of various types of ships are shown in Table III. Results show that the recognition effect is poor for ships with fewer samples and difficult to identify. This phenomenon is mainly caused by an imbalance in the categories of the different samples, in which the smaller number of samples (e.g., submarines and supply ships) can easily be identified with the larger

TABLE III
ACCURACY OF TRADITIONAL DATA EXPANSION METHOD

NO.	AC	AAS	GW	LS	SM	MS	SS	CS	PS
Recall	0.929	0.884	0.962	0.885	0.725	0.818	0.244	0.841	0.926
Precision	0.886	0.792	0.931	0.802	0.925	0.750	1.000	0.841	0.735

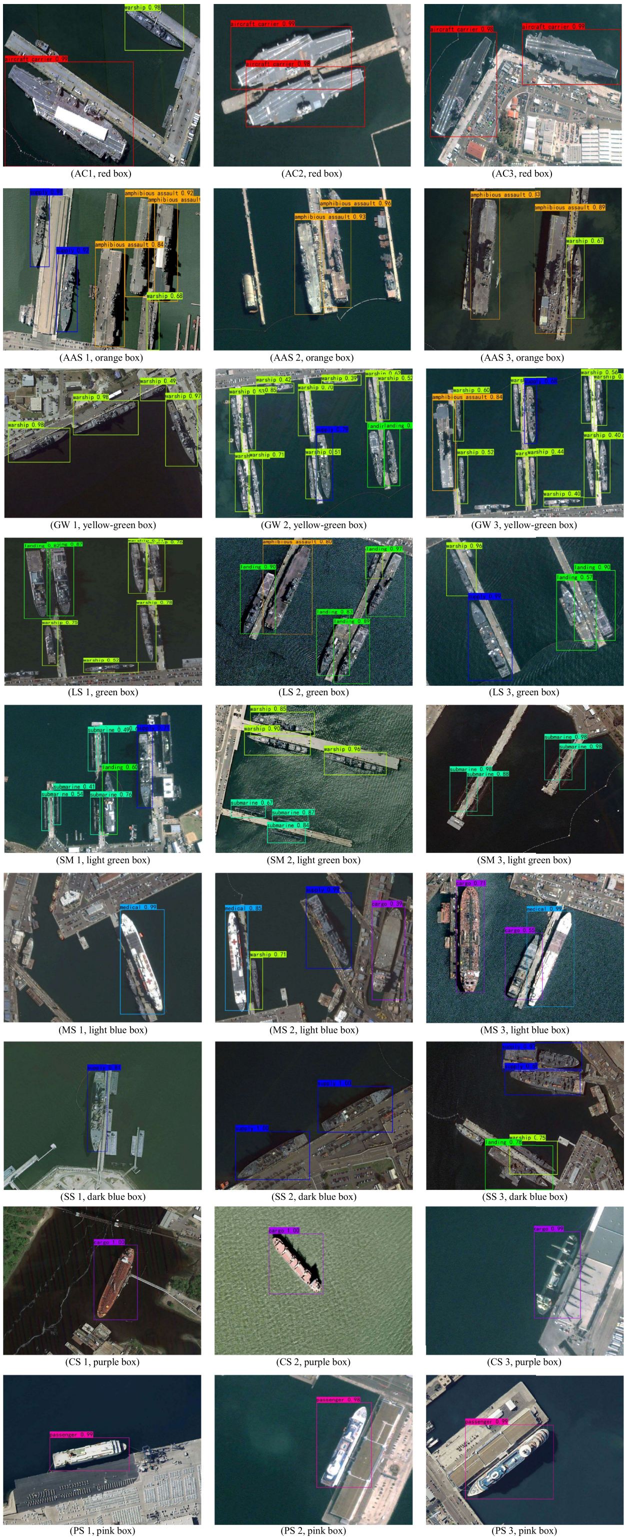

(AC1, red box) (AC2, red box) (AC3, red box)
(AAS 1, orange box) (AAS 2, orange box) (AAS 3, orange box)
(GW 1, yellow-green box) (GW 2, yellow-green box) (GW 3, yellow-green box)
(LS 1, green box) (LS 2, green box) (LS 3, green box)
(SM 1, light green box) (SM 2, light green box) (SM 3, light green box)
(MS 1, light blue box) (MS 2, light blue box) (MS 3, light blue box)
(SS 1, dark blue box) (SS 2, dark blue box) (SS 3, dark blue box)
(CS 1, purple box) (CS 2, purple box) (CS 3, purple box)
(PS 1, pink box) (PS 2, pink box) (PS 3, pink box)

Fig. 11. Test results for different types of ships.

number of other samples (e.g., general warships and cargo ships). For example, the recall of supply ships is as low as 0.244, because most of them are recognized as general warships and landing ships; the recall of submarines is 0.725, as some of them are also identified as general warships. To bridge this gap, we changed the data expansion method and carried out different times of data expansion on different types of ships, the expansion strategy is as described in Section II-B. After the expansion, the proportion of submarine pictures has increased from 2.9% to 6.2%, the proportion of supply ships has also increased somewhat, while the proportion of general warships, which are more numerous, has decreased by 30% compared to the pre-expansion period, and the proportion of cargo ships has also decreased considerably. By this method, less frequent classes were expanded more compare to more common classes. The final results show that the recognition accuracy of ships has been improved. As shown in Table II, the recall of supply ships and submarines increased to 0.822 and 0.980, respectively.

3) We chose to replace the original activation function swish with ReLU in the preprocessing stage, and compared the recognition effects of the two functions under the same other conditions. Table IV (swish) and Table II (ReLU) show the specific recognition effects of various ships using different activation functions. Results show that the overall recognition effect has been significantly improved, and the recognition accuracy of most ships has been improved. Before and after the activation function is modified, the overall recall and precision for warships increase from 0.926 and 0.938 to 0.935 and 0.970, respectively.

4) In addition to the activation function introduced before, we made many changes to the original EfficientDet network. To compare the actual results, we compared the original network with the network we used while keeping the rest of the conditions the same, and after calculating the recall of the original network for warship identification is 0.901, precision is 0.873. After compiling, Fig. 12 compares the identification results for warships under different ablation experiments.

V. DISCUSSION

In this part, we compare our method with some current state-of-the-art target detection networks, verify the portability of the model on other data, and discuss areas for improvement.

A. Quantitative Evaluation

In addition to the original EfficientDet network, we also compared our network with several other popular target

TABLE IV
ACCURACY OF SWISH ACTIVATION FUNCTION

NO.	AC	AAS	GW	LS	SM	MS	SS	CS	PS
Recall	0.905	0.884	0.951	0.885	0.961	0.909	0.822	0.927	0.926
Precision	0.927	0.809	0.980	0.920	0.875	1.000	0.841	0.916	0.862

TABLE V
COMPARISON OF RECALL OF VARIOUS METHODS

Method	AC	AAS	GW	LS	SM	MS	SS	CS	PS	All
EfficientDet	**0.952**	0.860	0.929	0.897	0.863	0.727	0.756	0.878	0.815	0.895
FRCNN	0.952	0.860	0.887	0.782	0.863	0.727	0.756	0.829	0.815	0.952
YOLO3	0.905	0.884	0.931	0.846	0.922	**0.909**	**0.844**	0.902	0.905	0.902
SSD	0.833	0.628	0.841	0.513	0.667	0.636	0.444	0.646	0.370	0.716
Ours	0.929	**0.930**	**0.945**	**0.936**	**0.980**	0.909	0.822	**0.939**	**0.963**	**0.937**

All represents the overall recall of all ships.

TABLE VI
COMPARISON OF PRECISION OF VARIOUS METHODS

Method	AC	AAS	GW	LS	SM	MS	SS	CS	PS	All
EfficientDet	0.851	0.771	0.939	0.787	0.898	0.889	0.654	0.791	0.710	0.857
FRCNN	0.816	0.685	0.910	0.763	0.830	0.571	0.447	0.624	0.524	0.816
YOLO3	0.974	0.884	0.977	0.957	0.940	1.000	0.905	0.925	**0.974**	0.952
SSD	0.875	0.659	0.887	0.727	**0.944**	0.636	0.645	0.639	0.500	0.804
Ours	**0.975**	**0.909**	**0.983**	**0.948**	0.943	**1.000**	**0.974**	**0.928**	0.788	**0.957**

All represents the overall precision of all ships.

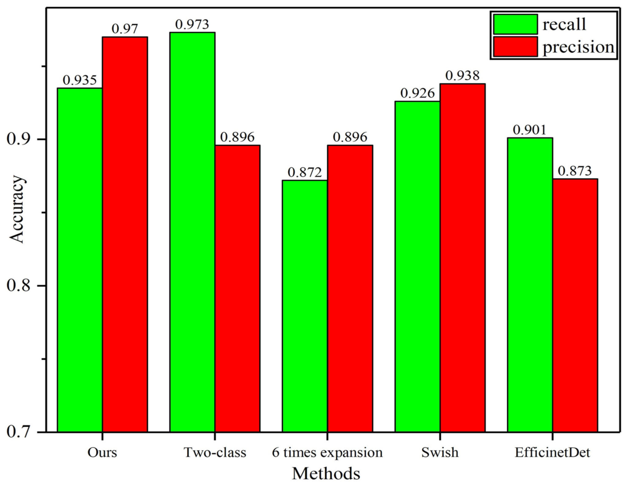


Fig. 12. Results of ablation experiments, corresponding to the methods described in (1)–(4), respectively, since the focus of this article is on military ships, the accuracy described here is only the overall accuracy after removing passenger ships and cargo ships.

detection networks, including the one-stage target detection network YOLO3, SSD, and the two-stage target detection network Fast R-CNN. We used the same training data and the same segmentation threshold to build the model, and the comparison results are shown in Tables V and VI. It can be seen that our method in this article is overall better than the original EfficientDet network, and has certain advantages in both accuracy and recognition integrity. Specifically, because the network has a larger receptive field to extract deep features, the recognition accuracy of large warships has been greatly improved. What is more, the residual connection method ensures the extraction of shallow features, so the overall recognition rate of small ships is also relatively good. Additionally, as shown in Table VI, our method also has obvious advantages compared with several widely used target detection networks. The overall recall and precision rates are generally higher than the current popular detection networks. At the same time, compared with the similar research of ship detection under the same complex conditions, its recall is 0.926 and precision is 0.953 [28], which is close to that of our proposed method, but ours can distinguish more categories.

In addition to the dataset used in this article, we also collected some ship images from “Dataset”, “DOTA”, and “Kaggle” datasets. At the same time, we also ordered some more difficult to identify maritime military targets from Google Maps images of marine military bases worldwide in the last two years to validate this model. A total of 196 warships were obtained. The final result is that the overall recall of the warship is 0.86, and the precision is 0.94. Specific recognition results are shown in Fig. 13.

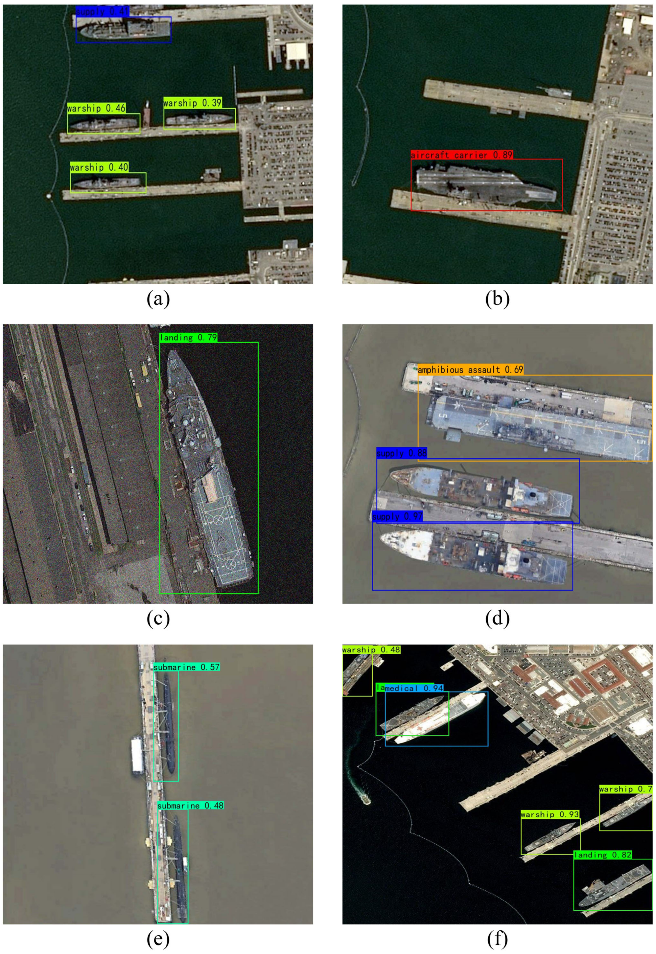


Fig. 13. Identification results of different types of ships: (a)–(c) blurred or noisy images; (d)–(f) different recognition scenarios from the original dataset or scrapped ships.

Fig. 14. Difficult to recognize scene images.

B. Potential Future Works

In the experiments conducted with the new dataset, the recognition rate of warships is relatively low. Still, in general, it can meet the demand of quickly distinguishing ship categories, which is mainly due to the large gap between the newly selected dataset and the original dataset. The poor quality of some images (as shown in Fig. 14), how to distinguish these ship categories with large gaps from the training dataset is a direction to be studied.

The following aspects of warship identification are worthy of further research and improvement: in the testing process, it is difficult to recognize features that have never appeared in the training dataset or scenes that are quite different from the training data. The domain adaptation [47] method may solve this problem, while adding some unlabeled test set data for self-training [48]. In addition, the imbalance of the training set samples will affect the accuracy of the model. The method of mixed loss function [49] may eliminate the influence of this factor by assigning different weights to samples of different orders of magnitude [50].

VI. Conclusion

This article constructed a target detection network based on a multilayer feature extraction method and realized the detection and classification of high-resolution remote sensing images of maritime military ships. The specific contributions include: In order to cooperate with the use of deep features, the residual connection, dilation convolution, and the ReLU activation function are introduced. Meanwhile, a more useful feature is extracted using the attention mechanism approach. In addition, we adjusted the structure of the original EfficientDet network to achieve the best recognition results. The recall of the overall warship is 0.935, and the precision is 0.970.

The setup of ablation experiments allow us to draw the following conclusions: the multiclassification approach used in this article can greatly improve the precision, specifically, the precision increased from the previous 0.896 to 0.970; While the different degrees of data expansion improved the recognition accuracy of less frequent samples, the recall of the supply ships and submarines improved by 0.6 and 0.26, respectively; The use of the ReLU activation function also led to some improvement in overall recognition accuracy. In addition to which, compared with the original EfficientDet network and other popular target detection networks, the improved network has a great advantage.

References

[1] A. Gambardella, F. Nunziata, and M. Migliaccio, "A physical full-resolution SAR ship detection filter," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 760–763, Oct. 2008.

[2] Y. Feng, W. Diao, X. Sun, M. Yan, and X. Gao, "Towards automated ship detection and category recognition from high-resolution aerial images," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1901.

[3] 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, Mar. 2018.

[4] J. Jiao *et al.*, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, Apr. 2018.

[5] G. H. Born, J. A. Dunne, and D. B. Lame, "Seasat mission overview," *Science*, vol. 204, no. 4400, pp. 1405–1406, 1979.

[6] M. Ballère *et al.*, "SAR data for tropical forest disturbance alerts in French Guiana: Benefit over optical imagery," *Remote Sens. Environ.*, vol. 252, Jan. 2021, Art. no. 112159.

[7] G. Gao, "A parzen-window-kernel-based CFAR algorithm for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 557–561, May 2011.

[8] T. Zhang and X. Zhang, "High-speed ship detection in SAR images based on a grid convolutional neural network," *Remote Sens.*, vol. 11, no. 10, 2019, Art. no. 1206.

[9] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[10] C. Corbane, L. Najman, E. Pecoul, L. Demagistri, and M. Petit, "A complete processing chain for ship detection using optical satellite imagery," *Int. J. Remote Sens.,* vol. 31, no. 22, pp. 5837–5854, Dec. 2010.

[11] 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.

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

[13] 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.

[14] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," *IEEE Trans. Geosci. Remote Sens.,* vol. 54, no. 10, pp. 5832–5845, Oct. 2016.

[15] Z. Li and L. Itti, "Saliency and gist features for target detection in satellite images," *IEEE Trans. Image Process.,* vol. 20, no. 7, pp. 2017–2029, Jul. 2011.

[16] L. Chen, W. Shi, C. Fan, L. Zou, and D. Deng, "A novel coarse-to-fine method of ship detection in optical remote sensing images based on a deep residual dense network," *Remote Sens.,* vol. 12, no. 19, 2020, Art. no. 3115.

[17] P. E. Carbonneau *et al.*, "Adopting deep learning methods for airborne RGB fluvial scene classification," *Remote Sens. Environ.,* vol. 251, Dec. 2020, Art. no. 112107.

[18] D. Zhang *et al.*, "A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution," *Remote Sens. Environ.,* vol. 247, Sep. 2020, Art. no. 111912.

[19] G. Ross, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2014, pp. 580–587.

[20] T. Zhang, X. Zhang, J. Shi, and S. Wei, "HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery," *ISPRS J. Photogrammetry Remote Sens.,* vol. 167, pp. 123–153, Sep. 2020.

[21] Y. Gui, X. Li, and L. Xue, "A multilayer fusion light-head detector for SAR ship detection," *Sensors,* vol. 19, no. 5, 2019, Art. no. 1124.

[22] W. Liu *et al.*, *SSD: Single Shot MultiBox Detector*. Cham, Switzerland: Springer, 2016, pp. 21–37.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[24] Z. Liu, Y. Lyu, L. Wang, and Z. Han, "Detection approach based on an improved faster RCNN for brace sleeve screws in high-speed railways," *IEEE Trans. Instrum. Meas.,* vol. 69, no. 7, pp. 4395–4403, Jul. 2020.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[26] S. Zhang, R. Wu, K. Xu, J. Wang, and W. Sun, "R-CNN-based ship detection from high resolution remote sensing imagery," *Remote Sens.,* vol. 11, no. 6, 2019, Art. no. 631.

[27] L. Qi *et al.*, "Ship target detection algorithm based on improved faster R-CNN," *Electronics,* vol. 8, no. 9, 2019, Art. no. 959.

[28] T. Tian, Z. Pan, X. Tan, and Z. Chu, "Arbitrary-oriented Inshore ship detection based on multi-scale feature fusion and contextual pooling on rotation region proposals," *Remote Sens.,* vol. 12, no. 2, 2020, Art. no. 339.

[29] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.,* vol. 10, no. 1, 2018, Art. no. 132.

[30] S. Li, Z. Zhou, B. Wang, and F. Wu, "A novel Inshore ship detection via ship head classification and body boundary determination," *IEEE Geosci. Remote Sens. Lett.,* vol. 13, no. 12, pp. 1920–1924, Dec. 2016.

[31] R. Wang, J. Li, Y. Duan, H. Cao, and Y. Zhao, "Study on the combined application of CFAR and deep learning in ship detection," *J. Indian Soc. Remote Sens.,* vol. 46, no. 9, pp. 1413–1421, Sep. 2018.

[32] 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.

[33] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 636–644.

[34] T. 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. 936–944.

[35] M. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 24, no. 4, pp. 647–660, Apr. 2013.

[36] C. Mi, C. Yaqi, L. Yafei, Z. Jing, X. Wei, and P. Jiazheng, "A new network structure for semantic segmentation of ship targets in remote sensing," in *Proc. 22th Int. Conf. Inf. Fusion*, 2019, pp. 1–8.

[37] Y. Yao, Z. Jiang, and H. Zhang, "High-resolution optical satellite image simulation of ship target in large sea scenes," *2016 IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 1241–1244.

[38] J. Jeppesen, R. Jacobsen, F. Inceoglu, and T. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.,* vol. 229, pp. 247–259, Aug. 2019.

[39] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access,* vol. 6, pp. 50839–50849, Sep. 2018.

[40] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10778–10787.

[41] Y. Qi, Y. Wang, and Y. Liu, "Object Tracking Based on Deep CNN Feature and Color Feature," *14th IEEE Int. Conf. Signal Process. (ICSP)*, pp. 469–473, 2018.

[42] S. Liu, L. Yu, and D. Zhang, "An efficient method for high-speed railway dropper fault detection based on depthwise separable convolution," *IEEE Access,* vol. 7, pp. 135678–135688, Sep. 2019.

[43] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 42, no. 2, pp. 318–327, Feb. 2020.

[44] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 730–734.

[45] X. Nie and W. X. Zheng, "Multistability and instability of neural networks with discontinuous nonmonotonic piecewise linear activation functions," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 26, no. 11, pp. 2901–2913, Nov. 2015.

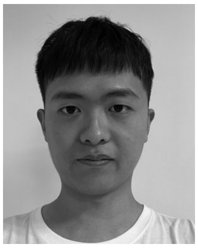
[46] Y. Chen, Y. Li, J. Wang, W. Chen, and X. Zhang, "Remote sensing image ship detection under complex sea conditions based on deep semantic segmentation," *Remote Sens.,* vol. 12, no. 4, 2020, Art. no. 625.

[47] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing,* vol. 312, pp. 135–153, Oct. 2018.

[48] Z. Yuan and C. Lin, "Research on strong constraint self-training algorithm and applied to remote sensing image classification," in *Proc. IEEE Int. Conf. Power Electron., Comput. Appl.*, Jan. 2021, pp. 981–985.

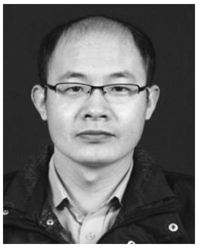
[49] V. Iglovikov, S. Seferbekov, A. Buslaev, and A. Shvets, "TernausNetV2: Fully convolutional network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 228–2284.

[50] J. Zheng *et al.*, "Growing status observation for oil palm trees using Unmanned Aerial Vehicle (UAV) images," *ISPRS J. Photogrammetry Remote Sens.,* vol. 173, pp. 95–121, Mar. 2021.



Peng Qin received the B.S. and M.S. degrees in photogrammetry and remote sensing from the Shandong University of Science and Technology, Qindao, China, in 2018 and 2021 respectively. He is currently working toward the Ph.D. degree in cartography and geographic information system with Sun Yat-sen University, Guangzhou, China.

His research interests include remote sensing and deep learning.



Yulin Cai received the B.S. degree in forestry from Shandong Agricultural University, China, in 1996, the M.S. degree in ecology and the Ph.D. degree in cartography and geographic information system from Chinese Academy of Science, Beijing, China, in 2002 and 2010 respectively, and was a visiting scholar with the University of North Carolina at Charlotte in the Department of Geography and Earth sciences from August 2014 to August 2015. He is currently an Associate Professor with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao, China. His research interests are remote sensing image processing and information extraction, as well as the application of remote sensing in environmental monitoring.

Jia liu received the B.S. degree in surveying and mapping engineering from Heilongjiang University of Technology, Heilongjiang, China in 2019. He is currently working toward the M.S. degree in surveying and mapping engineering with the Shandong University of Science and Technology, Qingdao, China.

His research interests include remote sensing of resources and environment learning.

Puran Fan received the B.S. degree in photogrammetry and remote sensing form the Shandong University of Science and Technology, Qindao, China, in 2019. She is currently working toward the M.S. degree in surveying and mapping engineering with the Shandong University of Science and Technology, Qindao, China.

Her research interests include remote sensing of resources and environment learning.

Menghao Sun received the B.S. degree in surveying and mapping engineering from the Hebei University of Technology, Tianjin, China, in 2017, and M.S. degree in surveying and mapping engineering from the Shandong University of Science and Technology, Qindao, China, in 2021.

His research interests include remote sensing of resources and environment.