LKPF-YOLO: A Small Target Ship Detection Method for Marine Wide-Area Remote Sensing Images

JUNFEI CHEN
ZHUHUA HU, Senior Member, IEEE
WEI WU
YAOCHI ZHAO
Hainan University, Haikou, China

BA HUANG
Guangzhou Marine Geological Survey, Sanya, China
China Geological Survey, Sanya, China

Ship detection based on wide-area remote sensing imagery has a wide range of applications in areas such as ship supervision and rescue at sea. However, wide-area remote sensing satellites sacrifice spatial resolution and spectral resolution to cover a larger sea area, which leads to smaller ship scales, fewer source pixels, and a lack of texture details in the images. In this paper, we propose a deep learning network, LKPF-YOLO, for detecting small-target ships in wide-area remote sensing images. For this purpose, we first create a South China Sea wide-area remote sensing dataset containing about 7600 ship instances. In order to extract features of small objects and low-contrast targets more efficiently, we design a re-parameterized large kernel module, C2Rep, to give the network a larger effective sensing field and richer gradient flow information. Finally, we design a loss function, Priori Focal Loss, based on unbalanced learning and prior knowledge, which guides the model to focus more on the training of small and difficult samples. The experimental results show that the model achieves accurate and stable small-target ship detection in wide-area remote sensing datasets. The mAP_{50} (mean Average Precision) and $mAP_{50:95}$ of the model reached 93.6% and 50.7%, which were 5.5% and 12.9% higher than the original model, respectively. In addition, the number of parameters and computation of the model are reduced by 7% and 18.7%, respectively, providing great potential for model deployment.

Received 23 May 2024; revised 15 August 2024 and 25 September 2024; accepted 6 October 2024. Date of publication 9 October 2024; date of current version 14 April 2025.

DOI. No. 10.1109/TAES.2024.3476459

Refereeing of this contribution was handled by P. Braca.

This work was supported in part by the Key Research and Development Project of Hainan Province under Grant ZDYF2022GXJS348 and Grant ZDYF2022SHFZ039, in part by the National Natural Science Foundation of China under Grant 62361024 and Grant 62161010, and in part by the Natural Science Foundation of Hainan Province under Grant 623RC446.

Authors' addresses: Junfei Chen, Zhuhua Hu, Wei Wu are with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China, E-mail: (22220854000129@hainanu.edu.cn; eagler_hu@hainanu.edu.cn; wuweido@126.com); Yaochi Zhao is with the School of Cyberspace Security, Hainan University, Haikou 570228, China, E-mail: (zhyc@hainanu.edu.cn); Ba Huang is with the Sanya Institute of South China Sea Geology, Guangzhou Marine Geological Survey, Sanya 572025, China, and also with the Academy of South China Sea Geological Science, China Geological Survey, Sanya 572025, China, E-mail: (273140684@qq.com). *(Corresponding author: Zhuhua Hu.)*

I. INTRODUCTION

As global maritime traffic continues to grow and ocean resources are extensively developed and utilized, vessels have become crucial carriers for maritime cargo transportation and important subjects for surveillance against illegal activities, such as smuggling and stowaway [1], [2]. Therefore, the detection and identification of maritime targets, such as vessels and port areas, play a vital role in both military and civilian contexts. With the advancement of satellite remote sensing technology, large amounts of maritime data can be easily obtained using satellite remote sensing techniques. Coupled with deep learning technology, these techniques can effectively aid in monitoring oceanic environments, vessels, and marine resources [3], [4]. Based on deep learning technology and satellite remote sensing technology, we have carried out relevant research in the field of ship detection. Fig. 1 is a flowchart of this study using deep learning to detect ships in wide-area remote sensing images.

At present, the object detection technology based on satellite remote sensing image is in the stage of rapid development [5]. With the continuous advancement of satellite technology and image processing algorithms, researchers have achieved remarkable results in maritime monitoring fields, such as coastline detection [6], oil spill detection [7], and ship detection [8], [9], [10]. In terms of remote sensing technology, imaging technologies, such as satellite remote sensing and autonomous aerial vehicle reconnaissance, can acquire high-resolution imagery of ground targets, providing rich image data sources for the detection and identification of targets in maritime security [11], [12]. In terms of algorithms, the application of machine learning techniques, such as deep learning and convolutional neural networks, has opened up new possibilities for the automated identification of ship targets [13], [14], [15].

However, current research on ship target detection mostly relies on high spatial resolution remote sensing satellites. These remote sensing satellites usually focus on ports, and the ships in these images are moderate in size and have rich texture details [16], [17], [18]. Apart from these high spatial resolution remote sensing satellites, there are also wide-area remote sensing satellites designed for monitoring large maritime areas. They were initially designed to increase the coverage area at the expense of other performance aspects, which results in fewer source pixels of vessels in the images and a lack of texture information. In addition, the sparse distribution of ships at sea leads to a severe imbalance between target and background samples in

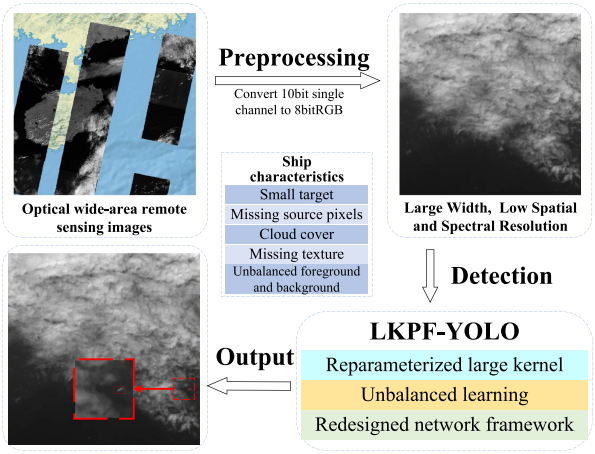


Fig. 1. We begin by creating a new dataset from the original remote sensing data using three different image conversion methods. Then, we input the augmented dataset including multiple scenarios into our model for ship detection. Finally, the output is presented by drawing the anchor frame on the input image.

the images, and the complex maritime environment makes ships more susceptible to cloud cover [19], [20].

In summary, it is a significant challenge in practical applications to address the detection difficulties arising from factors, such as sparse source pixels of ship samples, class imbalance, and lack of learnable features in wide swath satellite remote sensing images, and to enhance the robustness and accuracy of the model. This study addresses this issue by creating a set of optical remote sensing datasets for the South China Sea and proposing a target detection algorithm, large kernel and priori focal (LKPF-YOLO). This algorithm improves the detection of small target vessels under wide-area remote sensing satellites through a reparameterized large kernel structure and an unbalanced learning loss function.

The main contributions of this article can be summarized as follows.

1) In order to address the lack of wide-area remote sensing ship datasets aimed at monitoring large maritime areas, we create a set of optical remote sensing datasets for ships in the South China Sea based on the "Hainan-1" remote sensing data. This article fully utilizes the 10-bit color depth information of the original data, focusing on preserving the contrast and image details of the color value range where ships are located by marginalizing other redundant intervals through grayscale transformation. Subsequently, the original data are standardized into 8-bit RGB images, and data augmentation is implemented.
2) To tackle the issue of low detection accuracy caused by the lack of source pixels and texture details for small targets in remote sensing images, we propose a reparameterized large kernel module called C2Rep. This module possesses a larger effective receptive field and richer gradient flow information, enabling it to fully extract features and shape characteristics of small targets. This enhances the model's ability to extract features of small objects and low-contrast targets, reducing the performance degradation of the model caused by small targets or a lack of texture information. In addition, we reduce the complexity introduced by large convolutional kernels through the introduction of depthwise separable convolutions and a C3 feature extraction module.
3) Aiming at the problem of detection difficulties caused by sample category imbalance in maritime remote sensing images, we propose a priori focal loss based on unbalanced learning and prior knowledge. It first predetermines the loss weight of each sample by the distribution of the ground truth box size in the dataset and then dynamically adjusts the values of the weights according to the confidence score of the current samples in the process of model training. This ensures that the model focuses more on training on fewer and more challenging samples and addressed the problem of having too simple basis for adjustment.

The rest of this article is organized as follows. In Section II, we review relevant previous research in this field. In Section III, we provide detailed explanations of the dataset creation process and improvements to the model algorithm. We present and discuss numerical results to validate the effectiveness and improvements of the proposed algorithm in Section IV. Finally, Section V concludes this article.

II. RELATED WORK

A. Remote Sensing Dataset for Ships

In terms of maritime vessel datasets based on satellite remote sensing imagery, most of the datasets focus on high-definition near-shore port images, and there are relatively few wide-area remote sensing imagery for the purpose of monitoring a wide area of the sea. HRSC-2016 [21] was extracted from six important ports in Google Earth [22] for ship detection and classification, and the dataset contains 2976 and more than 25 types of ships. However, most of them are large ships docked in the harbor. The ShipRSImageNet dataset [23] and VHRShips dataset [24] include diverse ship types, different ship sizes, multiple nearshore locations, and different data acquisition conditions, in particular ShipRSImageNet is one of the largest remote sensing datasets for ship detection. But again, they are mostly high-resolution port images that are better suited for detection of ship types than small target detection.

DOTA [25] and SODA [26] datasets contain aerial images from different sensors and platforms, which are clear and varied and contain tens of thousands of ship targets. However, only the densely distributed docked ships near the shore do not meet the demand for ship detection over a wide sea area. The LEVIR [27] dataset consists of a large number of high-resolution images of 800 × 600 pixels and 0.2–1.0

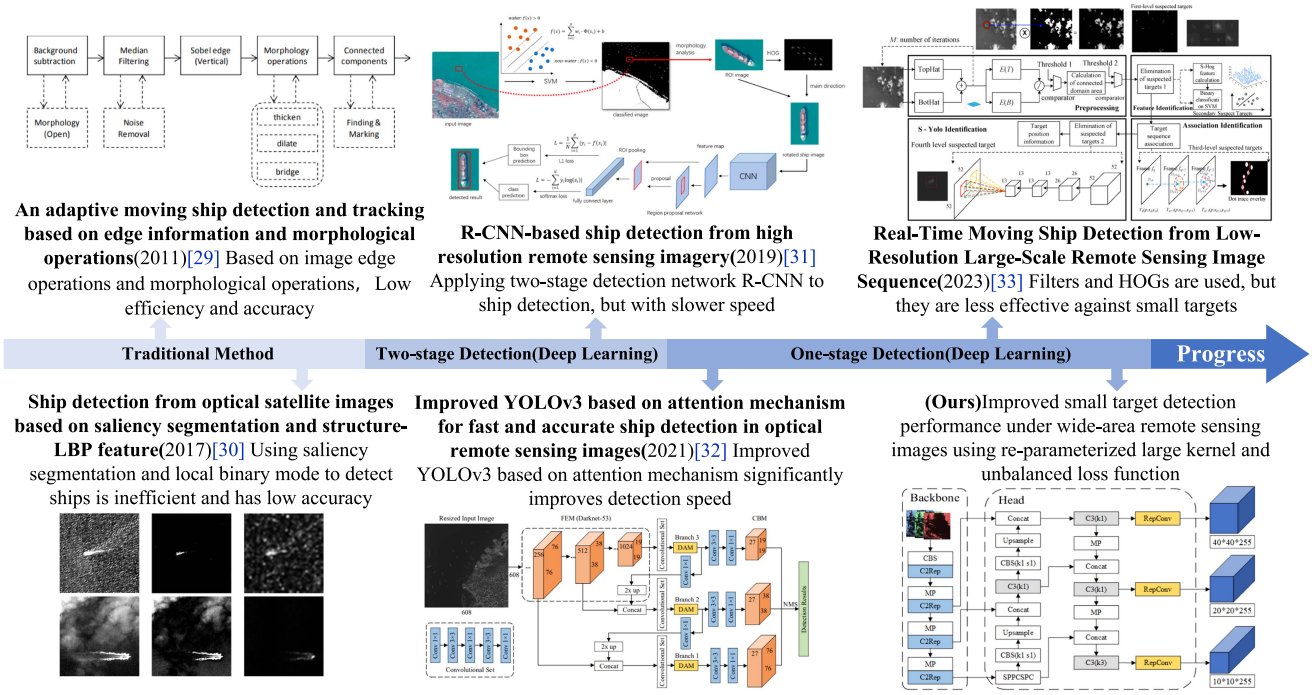


Fig. 2. Introduction of classical algorithms for ship target detection. Initially, there were only traditional methods of extracting features by hand. Then, a two-stage detection algorithm based on deep learning appeared, which improved the detection efficiency. Now, the mainstream is the one-stage detection algorithm, which greatly improves the real time.

m pixels with 3025 ship instances. The DIOR [28] dataset consists of 23 463 remotely sensed images and 190 288 target instances with a spatial resolution of 0.5–30 m. The instances in these two datasets have very high resolution, and they are all general remote sensing datasets, not specific to ships.

Therefore, most of the above datasets focus on high-definition remote sensing images of near-shore ports. They are quite different from wide-area remote sensing images with ultra-large width, complex maritime environments, low spatial resolution, and low spectral resolution.

B. Ship Target Detection

The target detection algorithms for maritime vessels have experienced a long period of development, as shown in Fig. 2. On traditional methods, Arshad et al. [29] proposed a new motion ship detection method based on correlated image edge operations and morphological operations, which achieved good results. Yang et al. [30] used saliency segmentation and local binary patterns to detect ships, but the speed and accuracy were unsatisfactory.

In terms of deep learning methods, Zhang et al. [31] applied the two-stage detection network R-convolutional neural network (CNN) to ship detection, which dramatically improved the accuracy and robustness compared with the traditional methods, but the detection speed was slow. Chen et al. [32] proposed the improved YOLOv3 network incorporating the attention mechanism, which dramatically improved the detection speed. Yang et al. [8] proposed a rotation dense feature pyramid network to solve the problem caused by narrow ship width, which can achieve excellent results in the task of detecting densely arrayed ships. However, the high false alarm rate makes it more likely to cause misjudgment in the scene of cloud interference. Su et al. [9] designed a new feature extraction network, DCNDarknet25, to realize the rapid detection of ships in large-size remote sensing images by means of a method without sliding windows. Although it can detect large-size images at one time, it cannot achieve ideal results in the face of small targets. The above algorithms can achieve better results on high-definition remote sensing image datasets, but they are less effective in the face of tiny ships and low-resolution images because the feature extraction process for small targets is not optimized.

Facing low-resolution remote sensing images, Yu et al. [33] proposed a real-time detection algorithm for moving targets in low-resolution wide-area remote sensing images, which improves the detection accuracy of low-resolution images through multiple detection. But the efficiency of the model is low and not well adapted due to multiple identifications and screenings. Therefore, there is a lack of efficient detection models in the field of ship target detection that have good adaptability, can fully extract small target features, and can adapt to the characteristics of wide-area remote sensing images.

C. Unbalanced Learning Methods

Class imbalance causes the model to be more inclined to predict the majority class, which results in a lower prediction accuracy for the minority class. And in the case of extreme imbalance, the model can achieve high accuracy by simply predicting all the samples as the majority class, but

this model is worthless in practical application. Zhang [34] used a deep cost-sensitive neural network to assign a different cost to each class, which makes the model focus more on the classification of a few classes. Jiang et al. [35] improved the traditional cross-entropy loss function definition by making the model cost-sensitive to increase the weight of a few categories. Huang et al. [36] and Khan et al. [37] made the model focus on samples of fewer classes by redefining new parameters.

All of the above studies improve the classification performance by increasing the weights of a few categories, but the number of samples of a category in image target detection is not a good representation of the difficulty level of the category and does not reflect the quality of the prediction. On the image task, Li et al. [38] proposed quality focal loss (QFL) to judge the difficulty of the current sample by the confidence score and use it to adjust the loss weight. However, this method is based on a single judgment and is easily affected by whether the confidence score is accurate or not. At the beginning of training, the model does not fully learn the characteristics of the sample, so the confidence score at that time is not reliable. This can easily lead to the model obtaining an unreliable loss weight at the initial stage.

III. PROPOSED METHODS

In recent years, there has been a gradual increase in the number of smuggling and stowaway vessels in the waters around Hainan. Remote sensing images are needed to monitor a wide range of sea areas and to strengthen the means of target detection in ship control. Based on this practical problem, this study uses the images obtained from the "Hainan-1" wide-area remote sensing satellite to produce a new set of ship datasets. Furthermore, we employ YOLOv7 [39], which offers a balanced performance in terms of accuracy, speed, and stability, as the base network. Based on this, we propose an object detection algorithm for small ships in wide-area remote sensing images, named LKPF-YOLO.

A. Marine Remote Sensing Datasets

At present, wide-width remote sensing data are scarce, and most of the published data are concentrated in high-resolution optical satellites or synthetic aperture radar (SAR) satellites. Therefore, most studies focus more on object detection with high-resolution small-range images, such as the detection of ships in ports. The lack of wide remote sensing datasets limits the rapid development of deep learning algorithms in the field of small target ship detection to some extent. Based on this, we create a dataset of remote sensing images of ships in the South China Sea based on the remote sensing images of "Hainan-1."

"Hainan-1" was launched in 2022 with a dual-line array wide-format camera. Its width is 110 km east to west and about 1900 km south to north. Its single original remote sensing image is a 10-bit grayscale image with a resolution of 28 000 × 28 000, and the resolution is about 5 m.

We analyze the composition of the original image data, and use the histogram to represent the distribution of pixel values. As shown on the left-hand side of Fig. 3, the values of the pixels in the original image are distributed between 0 and 1024, but this distribution is not uniform. Most of the pixel values are small, and only a few are close to the maximum value. If only the common mix–max normalization method is used, the image will lose a lot of dark details.

Considering the versatility and diversity of the data, we use three ways to sample the original data: bright, dark, and moderate. This results in a richer and more diverse dataset, allowing the model to learn more details about light and dark features. The three conversion modes are shown in Fig. 3.

First, as shown in Method A in Fig. 3, we process the 16-bit image using the 8-bit standard. That is, pixels with values in the interval [0, 255] are read normally, and all pixels with values greater than 255 are assigned a value of 255. Next, as shown in Method B in Fig. 3, we first remove the unused data intervals and then linearly transform the actual used intervals of pixel values. Finally, as shown in Method C in Fig. 3, we first assign the 10% of pixels with larger values directly to 255 and then linearly transform the values of the remaining 90% of pixels to the [0, 255] interval. Since most of the original pixel values are small, the overall contrast of the image can be improved by adjusting a small number of larger pixel values. The results of the three different preprocessing methods are shown in Fig. 4.

The above three preprocessed images have different detail information, such as contrast, brightness, and fine-grained features. It can be seen that the unobscured ship in Method A is more obvious, but the image is almost devoid of bright details and is greatly affected by cloud occlusion. Method B is less affected by cloud masking, but the overall brightness is darker, and the ships are not obvious. Method C is more balanced and more in line with the human eye. After preprocessing, we use LabelMe [40] software to label the images. After integrating the three kinds of images, we get a brand set of remote sensing datasets of small-target ships at sea containing 4027 images of 1024×1024 pixel size, and the dataset has a total of 7602 ship instances.

It is worth mentioning that the original remote sensing data are in the 10-bit format, while the converted universal image is in the 8-bit format (8bitRGB). So, Methods A–C all sampled about a quarter of the data from the original image according to their different methods. This means that they each have a subset of data and details that the other methods do not, and these different details form a complementary way to help the model learn more diverse features. We will demonstrate this in the ablation experiments.

B. Revisiting YOLOv7

YOLOv7 is a sequel to the YOLO series, it has a balanced accuracy, speed, and stability, and has shown

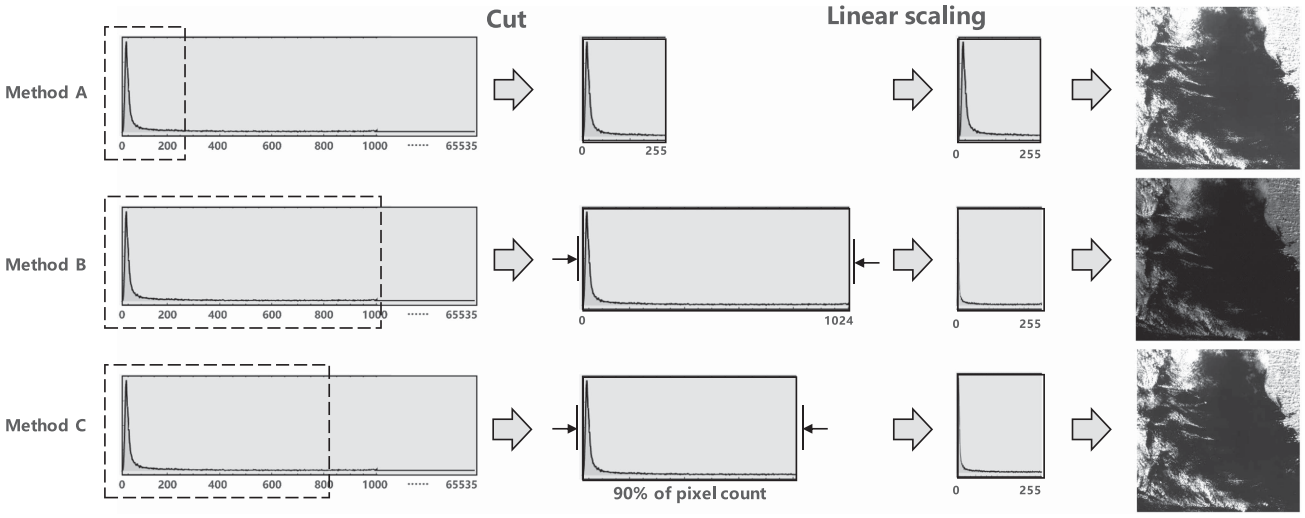


Fig. 3. Three image processing methods are used in this dataset. We sampled the original data in three ways, with the final effect being bright, dark, and moderate, so as to obtain a richer and more diverse dataset. On the left-hand side of the image is the histogram of the pixel value distribution of the original remote sensing data. The *x*-axis represents the pixel value, and the *y*-axis represents the number of pixels. The raw data use only the first ten bits in the 16-bit format, and the values of the pixels are clustered around the smaller values.

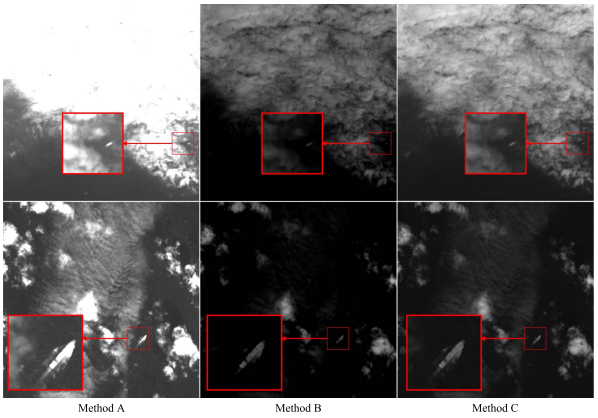


Fig. 4. Images are converted using three methods. Method A is brighter, Method B is darker, and Method C is moderate. They each have a piece of information that the other results do not.

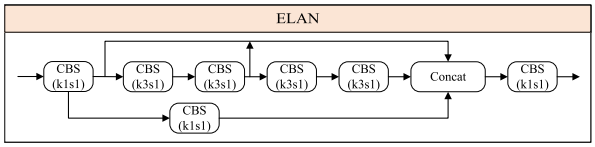


Fig. 5. ELAN module structure diagram. Its main function is to extract features.

excellent performance in both general purpose target detection and remote sensing target detection. It is mainly composed of the backbone part to extract features, the neck part to fuse features, and the head part to output results. YOLOv7 uses several innovative modules to improve performance, including efficient layer aggregation networks (ELAN) modules and SPPCSPC modules.

The function of the SPPCSPC module is to increase the receptive field and adapt the algorithm to the input of different sizes. ELAN is an efficient feature extraction module designed to improve the model's receptive field (the size of the input region corresponding to the feature map) and feature representation. It combines multiple feature layers of different scales by counterpoint addition and Concat operation (splicing multiple arrays) to make full use of the feature information of different levels. The ELAN module structure diagram is shown in Fig. 5.

Among them, conv, batch normalization and sigmoid linear unit (CBS) is mainly responsible for feature extraction, which is composed of three continuous modules: Conv, batch normalization, and activation function sigmoid linear unit (SiLU). The parameter k refers to the size of the convolution kernel and s refers to the step size of the convolution.

C. Improvement of Network Structure

YOLOv7 completes downsampling by using multiple CBS modules to gradually reduce the size of the feature map [41]. However, the small target samples have fewer source pixels, and the features in them are difficult to be learned by the deeper feature maps after the image has been drastically downsampled by the model.

In this study, the LKPF-YOLO framework is proposed to cope with the lack of source pixels and texture details of small targets by improving the model's sensitivity to shape and allowing the model to obtain richer gradient flow information. The network structure diagram of this framework is shown in Fig. 6. First, the C2Rep module proposed in this study is introduced into YOLOv7, which has a larger effective receptive field than the ELAN structure and pays more attention to the shape features in the image, avoiding a significant degradation of the model's performance due to the lack of texture information in the image. Second, the framework uses a more lightweight C3 module to replace the ELAN structure in the head part to reduce

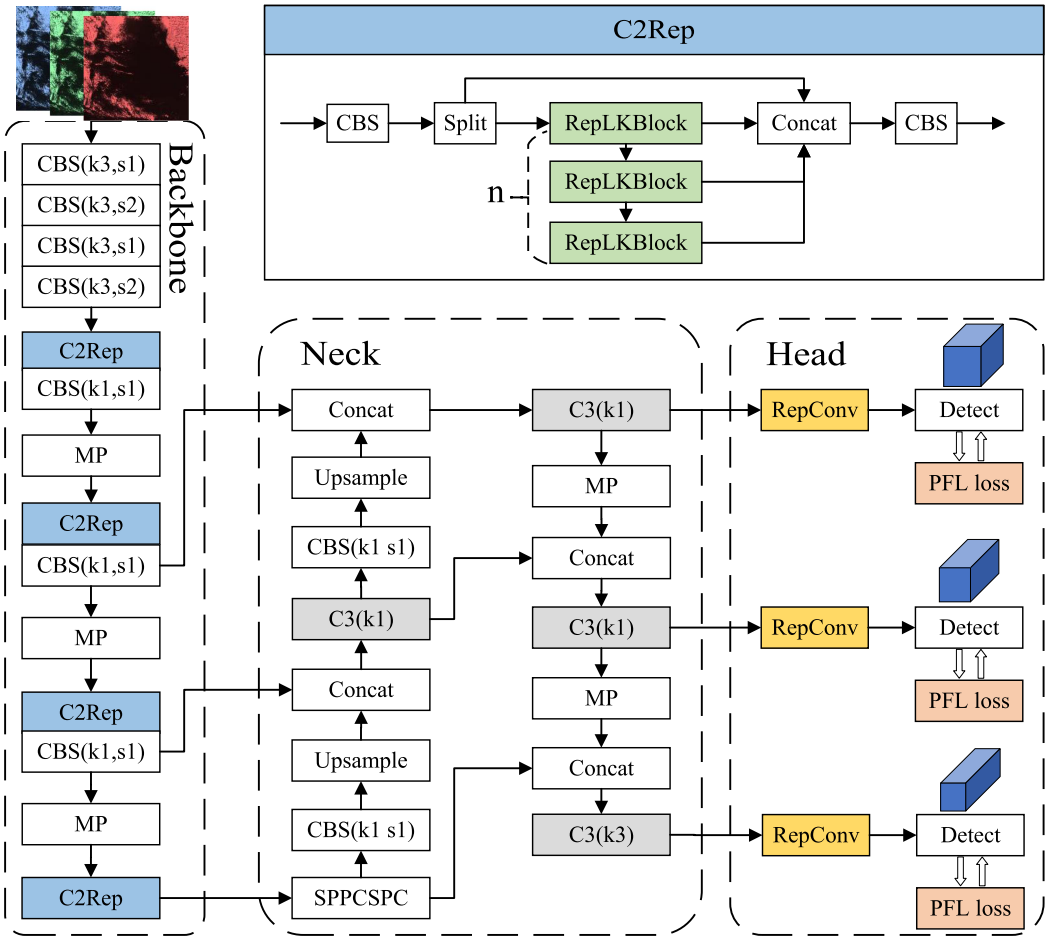


Fig. 6. Schematic diagram of LKPF-YIOLO network architecture. The colored squares in the diagram are the modules we introduced into the network. We mainly introduce the C2Rep module and PFL loss function proposed in this article to improve the detection ability of small targets.

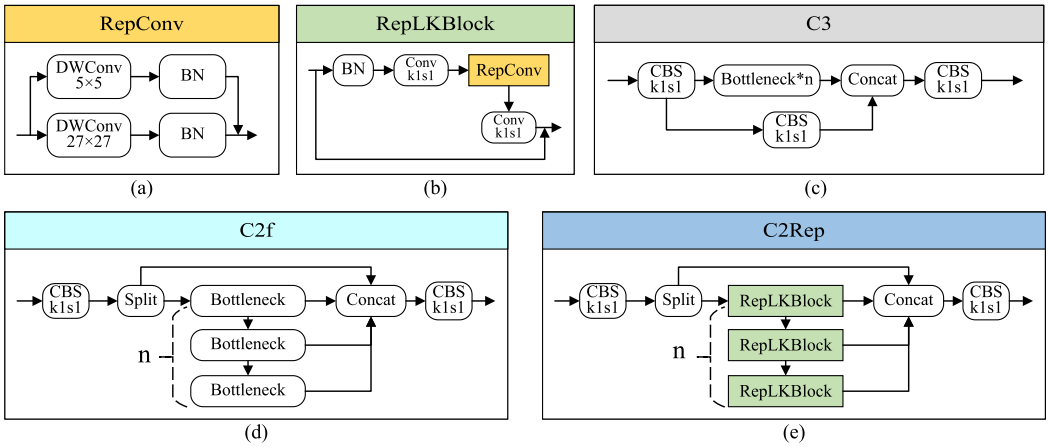


Fig. 7. Structure diagram of each module used in this article. (a) and (b) Components that provide convolution capability. (c)–(e) Full feature extraction modules.

the high number of parameters that a large convolutional kernel brings to the model. Finally, the model uses the priori focal loss (PFL) proposed in this article as a loss function to dynamically adjust the loss weights with the prior knowledge and confidence score of the current samples, so that the model pays more attention to the difficult samples.

RepConv [42] can provide competitive performance in extracting features of small targets, and its structure is shown in Fig. 7(a). It mainly consists of two depthwise separable convolution [43] of different sizes, so that each convolution kernel computes only one feature map of the corresponding channel. Thus, the large convolution kernel can be used to avoid the large number of parameters. RepLK

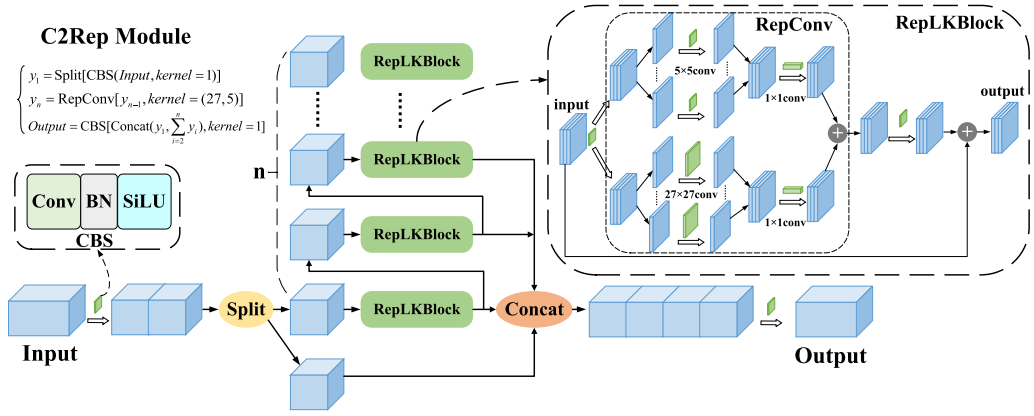


Fig. 8. C2Rep module structure diagram. The module stacks multiple RepLK blocks using residual connections, aiming to improve the model's ability to extract tiny features.

Block [42] is a large nuclear convolution structure based on RepConv, as shown in Fig. 7(b). It consists of batch normalization, two ordinary convolution, a RepConv, and residual join, which can produce a larger effective receptive field and higher shape deviation.

The C3 module [see Fig. 7(c)] is another feature extraction structure similar to ELAN, which improves the feature extraction ability of the model through the design of splitting, fusion, and partial jump connections. The bottleneck module first reduces the number of channels in the feature graph through convolution (dimensionality reduction), and then restores the number of channels (dimensionality increase), thus reducing the computation and parameter number. C2f is a feature extraction structure designed with reference to the ideas of ELAN and C3, which enables the model to obtain more abundant gradient flow information while ensuring lightweight. The structure is shown in Fig. 7(d). However, in the traditional C2f module, the limited sensory field of ordinary convolution may lead to poor accuracy in small target detection. Therefore, we propose an improved method. We introduce the RepLK block in the C2f module to improve the detection of small targets by effectively capturing the features of small targets with a reparameterized large convolutional kernel.

The C2Rep module proposed in this article is shown in Figs. 7(e) and 8. Specifically, we introduce RepLK blocks in each feature layer in the C2f structure to achieve finer feature extraction and fusion. There are multiple stacked RepLK blocks in the C2Rep module, which act as the main branching gradient module, also known as the residual module, to obtain a larger effective sensory field by reparameterizing the large convolution and thus adequately extracting small target features. This design allows the model to obtain more contextual and high-resolution information at the same time by performing operations, such as convolution, up-sampling, and channel splicing of feature maps at different levels, which improves the model's ability to detect small objects and low-contrast targets. Overall, C2rep reduces memory consumption and computational cost while increasing kernel size. It has a larger sensory field and more gradient flow information than the traditional CNN convolution, which enables the network to better extract fine-grained features when facing pixel-level image processing tasks.

In this study, C2Rep is introduced in the original YOLOv7 network to replace the ELAN structure. The C2rep structure can be represented by the following equation:

$$y_1 = \text{Split}[\text{CBS}(X, \text{kernel} = 1)] \tag{1}$$

$$y_n = \text{RepConv}(y_{n-1}, \text{kernel} = 27) \tag{2}$$

$$Z = \text{CBS}\left[\text{Concat}\left(y_1, \sum_{i=2}^{n} y_i\right), \text{kernel} = 1\right] \tag{3}$$

where X is the input of the module and Z is the output of the module. Equation (1) is the result of the Split operation after the first convolution of the input X; (2) is the respective output of the stacked RepLK blocks in the module, and the kernel size we chose is 27, which is more balanced in performance and complexity; and (3) fuses all the gradient flows [outputs of (1) and (2)] through concatenation and convolution operations to get the final output Z.

D. Improvement of Loss Function

On the ship detection task, deep learning models usually define positive and negative anchors in terms of Intersection over Union (IoU) scores, but this tends to amplify the imbalance problem of ship samples at sea, which leads to degraded detection performance [44]. This is because ships tend to be sparsely distributed in maritime remote sensing imagery, and there is a severe frequency imbalance between the ship target class and the marine background class. This leads to the fact that only a rather small number of a priori frames will be used to detect and fit positive samples, while most of the a priori frames will only be able to match negative samples. This problem is ameliorated by the proposal of QFL, which determines the difficulty of the current sample in terms of the confidence score and adjusts the loss weights in this way.

QFL is improved from focal loss (FL) [45], which enables the model to obtain better localization and classification performance and helps the model to better perform the imbalance learning task. The following equation is the binary cross-entropy loss function:

$$\mathrm{CE}(p_t) = -\log(p_t),\ p_t = \begin{cases} p, & \text{if } y=1 \\ 1-p, & \text{otherwise} \end{cases} \tag{4}$$

where y is the label value of the sample, which is positive when $y = 1$ and negative, otherwise. And p is the probability that the model predicts a particular sample to be a positive sample. If we set the parameters α and $1-\alpha$ to balance the weights of positive and negative samples, then (4) can be changed to the following form:

$$\mathrm{CE}(p_t) = -\alpha_t \log(p_t),\ \alpha_t = \begin{cases} \alpha, & \text{if } y=1 \\ 1-\alpha, & \text{otherwise.} \end{cases} \tag{5}$$

Since it is necessary to be able to dynamically recognize the difficult and easy samples and thus adjust the weights, it is necessary to set a coefficient $(1-p_t)^\gamma$ for the above equation. Then, (5) will take the following form, which is FL:

$$\mathrm{FL}(p) = \begin{cases} -\alpha(1-p)^\gamma \log(p), & \text{if } y=1 \\ -(1-\alpha)p^\gamma \log(1-p), & \text{otherwise.} \end{cases} \tag{6}$$

QFL extends FL by making it capable of handling continuous labeled values while retaining the advantages of FL for the category imbalance problem. Equation (1) extends the cross-entropy part of FL to its full form. Equation (2) generalizes the scaling factor for each sample to the absolute value between the predicted and actual values. This leads to a specific formula for QFL

$$\mathrm{QFL}(\sigma) = -|y-\sigma|^\beta[(1-y)\log(1-\sigma) + y\log(\sigma)] \tag{7}$$

where y is the quality labeling from 0 to 1, and σ is the prediction. Note that the global minimum solution of QFL is $\sigma = y$, so that the cross-entropy part becomes the full cross-entropy. QFL makes a single judgment about the difficulty of a sample, using only the confidence score and no other basis, which can cause the model to focus too much on some ambiguous samples.

In small target detection scenarios, the difficult samples are usually those with small ground truth box areas. The smaller the ground truth box, the less source pixel and texture information the sample has, and the more difficult the detection is. Taking the dataset created in Section III-A of this article as an example, the distribution of the ground truth box area is shown in Fig. 9.

The ground truth box area of this dataset is mainly concentrated around a center value (about 250). In the detection task of small targets, we can consider those with ground truth box areas smaller than the center value as small targets (difficult samples), those near the center value as normal-sized targets (normal samples), and those much higher than the center value as large targets (simple samples).

Based on the above ideas, this article fuses the a priori knowledge of the sample distribution situation with

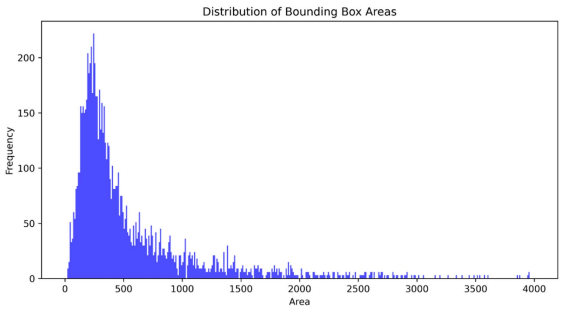


Fig. 9. Distribution of ground truth box area in the dataset. The horizontal coordinate is the area of the box, the vertical coordinate is the number of distributions, and the step size is 10. It can be found that the area of most boxes is small, and mainly distributed in the vicinity of 250 (pixels).

the dynamic imbalance learning method to propose a loss function, PFL. Specifically, we introduce a factor on prior bounding box knowledge into the QFL to assist the model in determining the loss weights, which focuses on the distribution of the dataset and the ground truth box area of the current sample. The factor first counts the distribution of the ground truth box area in the dataset, then calculates that center value in the most centralized part of the data, i.e., the one obtained when the mean absolute deviation is the smallest, and finally determines a preliminary loss weight based on the relationship between the current sample truth box area value and the center value. The factor is shown in the following:

$$k + e^{-\frac{x}{c}},\ \mathrm{k} \in (1-\mathrm{e}^{-1}, 1) \tag{8}$$

where k is a constant for adjusting the cutoff point for adding or subtracting weights, and it can get the best effect when it is set to 0.8 as presented in the ablation experiment of this study. c is the center value of the one where the ground truth box area is most concentrated, which is usually obtained when the mean absolute deviation is the smallest.

We put all the calculated ground truth box area values x_n into a set $X = \{x_1, x_2, x_3, \ldots x_n\}$. We need to obtain the value of the most centralized distribution of the elements in the set X. Here, we use the mean absolute deviation as a measure of "centralization." That is, when c is the mean value, it will minimize the mean absolute deviation of the set X, as shown in the following equation:

$$c = \mathrm{argmin}\mathrm{MAD}(u),\ \mathrm{MAD}(u) = \frac{1}{n}\sum_{i=1}^{n}|x_i - u|. \tag{9}$$

We introduce this factor into the QFL to obtain the improved loss function PFL, as shown in the following equation:

$$\begin{aligned}\mathrm{PFL}(\sigma) = &-(k+e^{-\frac{x}{c}})|y-\sigma|^\beta[(1-y)\log(1-\sigma) \\ &+ y\log(\sigma)].\end{aligned} \tag{10}$$

The loss function used in YOLOv7 consists of a weighted sum of three losses, objectness loss ($\mathrm{Loss}_{\mathrm{obj}}$), classification loss ($\mathrm{Loss}_{\mathrm{cls}}$), and localization loss ($\mathrm{Loss}_{\mathrm{box}}$),

TABLE I
Main Experimental Hardware and Software Parameters, as Well as the Settings During Model Training

Environment	Configuration	Hyp Options	Setting
System	Ubuntu 18.04.1	Input Size	640 × 640
CPU	Intel Xeon 6132	lr0	0.01
GPU	Nvidia Tesla V100	lrf	0.01
GPU Driver	CUDA 12.0	Weight Decay	0.001
DL Framework	PyTorch 1.18	Batch Size	16
Language	python 3.11	Epochs	300

as shown in the following equation:

$$\text{Loss} = a \times \text{Loss}_{\text{obj}} + b \times \text{Loss}_{\text{cls}} + c \times \text{Loss}_{\text{box}} \quad (11)$$

where a, b, and c are the weighted shares of the three losses, and the binary cross entropy (BCE) loss measure is used for both objectness loss and classification loss. The BCE formulas are as follows:

$$\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (12)$$

where N is the number of categories, y is the true label, and p is the predicted value.

Localization loss is then measured by CIoU. CIoU also incorporates the overlap area between the predicted and ground truth box, the centroid distance, and the aspect ratio into the calculation, as shown in the following equation:

$$\text{Loss}_{\text{box}} = \text{CIoU}_{\text{box}} = 1 - \left(\text{IoU} - \frac{\rho^2}{c^2} - \alpha v\right) \quad (13)$$

where ρ is the distance between the centroids of boxes A and B, c is the diagonal length of the minimum enclosing rectangle of boxes A and B, v is the aspect ratio similarity of boxes A and B, and α is the influence factor of v.

Finally, this article introduces PFL into the loss function of YOLOv7. The specific operation is to add PFL as a factor to objectness loss and classification loss, and adjust their weights according to the ground truth box area and confidence score of the current samples, while localization loss as a geometric loss is kept in its original form.

Thus, by combining (10)–(13), the loss function that incorporates PFL into YOLOv7 is obtained

$$\text{Loss} = \underbrace{a \times \text{PFL}_{\text{obj}} \times \text{BCE}_{\text{obj}}}_{\text{Objectness Loss}} + \underbrace{b \times \text{PFL}_{\text{cls}} \times \text{BCE}_{\text{cls}}}_{\text{Classification Loss}} + \underbrace{c \times \text{CIoU}_{\text{box}}}_{\text{Localization Loss}}. \quad (14)$$

IV. EXPERIMENTS

A. Experimental Setting

We use the remote sensing dataset of ships at sea created in this article to evaluate the effectiveness of the algorithm. The ratio of training set: test set: validation set as 8:1:1 is used to divide the data in this experiment. The experimental setting and training parameters for this study are given in Table I.

B. Evaluation Metrics

For the binary classification problem, category A is referred to as positive and category B is referred to as negative, and the classifier's correct prediction is denoted as true and incorrect prediction is denoted as false. Their combination forms the four base elements of the confusion matrix as true positive (TP), false negative (FN), true negative (TN), and false positive (FP).

Precision refers to the ratio of the target detection model judging the example to be a positive class and predicting it correctly to all positive classes, as shown in (15). Recall denotes the ratio of the model predicting the example to be a positive class and predicting it correctly to all predictions to be a positive class, as shown in (16)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (15)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (16)$$

When the confidence score is different, we also get different precision and recall. If we draw a graph with confidence score as the independent variable and precision and recall as the dependent variables, then we get the precision–recall curve (PR curve). AP in fact refers to the area under the PR curve and mAP is the average of the AP values of all the classes. Where mAP_{50} specifically refers to calculating the AP of all images in each class when IoU is set to 0.5, and then averaging over all classes. And $\text{mAP}_{50:95}$ indicates the average mAP when IoU is taken from 0.5 to 0.95 in steps of 0.05. The formulas of mAP_{50} and $\text{mAP}_{50:95}$ are shown in the following, respectively:

$$\text{mAP}_{50} = \frac{\sum_{i=1}^{k}\text{AP}_i}{k}, \text{IoU} \in [0.5] \quad (17)$$

$$\text{mAP}_{50:95} = \frac{\sum_{\text{IoU}=0.5}^{0.95}\text{mAP}}{k}, \text{step} = [0.5]. \quad (18)$$

In addition, we use the frames per second (FPS) metric as a measure of the model's real-time performance. It represents the number of frames the model processes per second, and the higher the FPS, the better the real-time performance of the model.

C. Results and Analysis

In order to verify the effectiveness of the proposed algorithm, the improved network, the original YOLOv7 [39], YOLOv7x [39], YOLOv7&QFL [38], SSD [46], Faster RCNN [47], YOLOv5n [48], and YOLOv8n [49] are trained and tested on the dataset in this article, respectively. The performance evaluation metrics were precision, recall, mAP_{50}, $\text{mAP}_{50:95}$, and FPS as described above.

The mAP variation curves for each network are shown in Fig. 10. It can be seen that with the increase in epoch number, the mAP of this article's method gradually tends to stabilize, and the overall curve is higher than that of other network models. In addition, we notice that although the curve of YOLOv7&QFL network is higher than that of YOLOv7 at the end, its convergence speed at the beginning

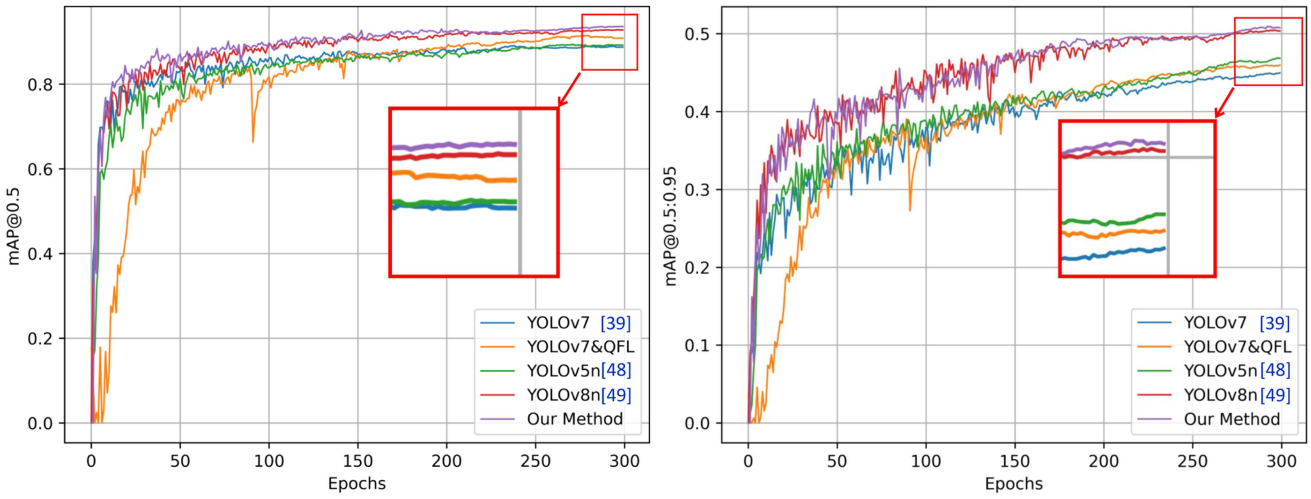


Fig. 10. mAP change curves of different models. Our model has excellent performance both in the speed of indicator improvement and the stability of indicator change.

TABLE II
Compares the Performance of Each Network

Networks	Precision (%)	Recall (%)	mAP_{50} (%)	$mAP_{50:95}$ (%)	FPS
SSD [46]	71.8	67.7	69.8	36.2	45
Faster R-CNN [47]	72.3	68.4	71.3	37.1	38
YOLOv5n [48]	86.5	83.5	89.2	46.8	126
YOLOv8n [49]	89.1	87.6	92.8	49.5	122
YOLOv7x [39]	88.2	85.1	89.1	45.5	95
YOLOv7&QFL [38]	88.6	86.7	90.9	45.9	131
YOLOv7 [39]	87.5	84.5	88.7	44.9	131
LKPF-YOLO	**90.5**	**89.2**	**93.6**	**50.7**	**135**

The higher these indicators are, the better the results, and the best score for each indicator has been highlighted in bold.

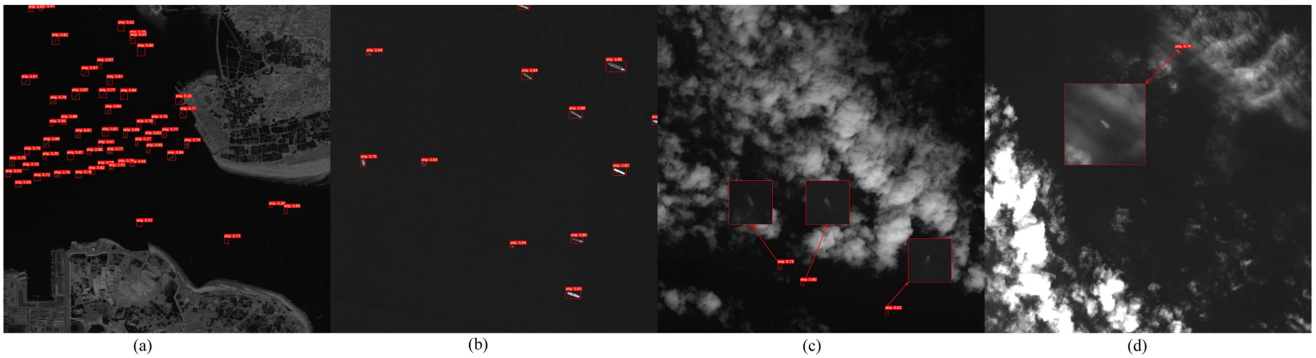
of training is significantly worse than that of the other networks, and the fluctuation amplitude is larger. The high index obtained by YOLOv7&QFL at the end of training is because QFL enables the model to focus on the difficult samples, thus achieving better performance. However, at the beginning of the training, the model has not fully learned the features of the samples, and the confidence score at this time is not too reliable, leading to the poor results of YOLOv7&QFL in the early part of the training. The LKPF-YOLO, which incorporates prior knowledge, solves this problem better. It achieved a high convergence rate at the early stage of training, and the curves did not show excessive oscillations.

Table II demonstrates the performance metrics of each network, and the model in this study has better performance compared to several mainstream models. The performance metrics (precision, recall, mAP_{50}, and $mAP_{50:95}$) of the improved network reach 90.5%, 89.2%, 93.6%, and 50.7%, which are about 3.4%, 5.6%, 5.5%, and 12.9% compared with the original network, respectively. Meanwhile, the recall metric of this article's model achieves a greater advantage over the precision metric compared to other models, which indicates that the large convolutional kernel used in this study makes the model less likely to miss small or texture-less targets. For the FPS metric, there is not much of a gap between the YOLO series. But our model still achieves the best performance, and its FPS reaches 135, which is sufficient to complete the detection task of remote sensing images.

In addition, we demonstrate the recognition effect of this article's model in complex scenes, as shown in Fig. 11. It can be seen that the improved network has a better detection effect, whether facing scenarios of dense ships, tiny ships, multiscale ships, or ships obscured by clouds.

For further comparison, we show and compare the detection effect of this article's method with other methods under the same picture, as shown in Fig. 12.

As can be seen from the comparisons in Fig. 12, the network proposed in this article can achieve accurate recognition of ships in the image both in the face of the dense ship scene and the multiscale scene, while the other models all have some omissions or misdetections. Both YOLOv7 and YOLOv7&QFL miss the detection of small ships and ships that are close to each other in the dense ship scenario; in the harbor scenario, both of them misidentify the long strip of dock as a ship; and in the cloudy scenario, YOLOv7&QFL also misidentify the blocky cloud as a ship. We notice that the long strip of dock and the blocky cloud are both elongated in shape, but they are still quite different from ships. Ships are rugby ball shaped, docks are rectangular, and clouds are irregularly shaped. This suggests that the misclassification of YOLOv7 and YOLOv7&QFL is not entirely caused by the lack of texture information in the images but also by the insensitivity of the model to the shape of small targets. For YOLOv5n and YOLOv8n, they miss detecting many small ships. This is due to the fact that the images are downsampled by the network many times, losing the feature information of the small targets.



(a) (b) (c) (d)

Fig. 11. Performance of the LKPF-YOLO network in different scenarios. Our model accurately detects ships in all scenarios. (a) Dense ship scene. (b) Multi-scale ship scenarios. (c) Tiny ship scene. (d) Ship obscured scene

TABLE III
Four Weights Were Tested on Four Datasets, Respectively

Result	DatasetA		DatasetB		DatasetC		DatasetD	
	AP1	AP2	AP1	AP2	AP1	AP2	AP1	AP2
WeightA	92.8	40.8	91.2	43.2	87.9	45.3	90.4	42.3
WeightB	91.6	43.8	90.6	44.1	87.7	42.3	89.9	43.2
WeightC	**95.3**	48.8	92.8	46.8	86.6	42.6	92.0	46.1
WeightD	93.9	**51.2**	**95.1**	**50.7**	**93.6**	**49.1**	**93.6**	**50.7**

Where WeightX refers to the results obtained by training on DatasetX using the method proposed in this study. AP1 refers to $mAP_{50}(\%)$ and AP2 refers to $mAP_{50:95}(\%)$.
The best score for each indicator has been highlighted in bold.

TABLE IV
Comparison of Various Performance Indicators of PFL With Different Values of k (On YOLOv7)

Value of k	Precision(%)	Recall(%)	$mAP_{50}(\%)$	$mAP_{50:95}(\%)$
0.70	88.0	87.8	91.9	48.5
0.75	**88.5**	88.0	92.2	48.7
0.80	88.4	**88.3**	**92.5**	**48.8**
0.85	87.8	87.9	92.1	48.4
0.90	86.8	87.7	92.0	48.2

The higher these indicators are, the better the results, and the best score for each indicator has been highlighted in bold.

TABLE V
Results of Ablation Experiments Using Different Module Combinations

ELAN	C2Rep	QFL	PFL	Precision (%)	Recall (%)	mAP_{50} (%)	$mAP_{50:95}$ (%)
√				87.5	84.5	88.7	44.9
√		√		88.6	86.7	90.9	45.9
√			√	88.4	88.3	92.5	48.8
	√			85.8	88.1	91.3	47.1
	√	√		87.1	88.5	92.2	48.3
	√		√	**90.5**	**89.2**	**93.6**	**50.7**

ELAN is the original module of YOLOv7, we replaced it with C2Rep. QFL is an improved method of loss function, PFL is our proposed loss function. The higher these indicators are, the better the results, and the best score for each indicator has been highlighted in bold.

D. Ablation Experiments

We design an ablation experiment to demonstrate the superiority of diverse data sets. We get DatasetA, DatasetB, and DatasetC by the three image conversion methods in Section III-A, and call their collection DatasetD. Then, we train the model proposed in this article on them, respectively, and the weights of result obtained by training are called WeightA, WeightB, WeightC, and WeightD, respectively. Finally, we test the four weights on each of the four datasets and compared their mAP_{50} and $mAP_{50:95}$ metrics. The results are given in Table III.

It can be seen from Table III that WeightD achieves the best results on the mAP_{50} metric except DatasetA. On the $mAP_{50:95}$ metric, WeightD achieved the best results on all four datasets. This means that the model trained on the set has the best generalization and performance, and it can achieve better performance on almost all subsets. Therefore, the richer and more diverse datasets in this article can be used to make the model achieve more competitive and generalized results.

For the value of parameter k in PFL (10), we introduce PFL into YOLOv7 and conduct a comparative experiment by changing the value of k. The results of the experiment are given in Table IV. It can be seen that the best comprehensive performance can be obtained in our model when k is 0.8. It is only slightly lower on the precision indicator than when k is 0.75.

In order to verify the effectiveness of the C2Rep module and PFL proposed in this article, ablation experiments were conducted by comparing the effects of combining different modules, and the results of the experiments are given in Table V, where the ELAN module is the feature extraction module used in the original YOLOv7.

As can be seen from the table, the introduction of both the C2Rep module and PFL effectively improves the performance of the network. The network has effectively improved both recall and mAP after incorporating only the C2Rep module, and only precision has decreased. This indicates that the large convolutional kernel in the module plays an effective role, i.e., it is able to fully extract the features of small targets, thus reducing the omission of small targets. After the network continues to introduce PFL, all the metrics are improved, and finally it is ahead of the original network in all the key metrics.

There are multiple stacked RepLK blocks in the C2Rep module proposed in this article, which produce different effects depending on the number of stacks, and there are four locations where the C2Rep module is used throughout the LKPF-YOLO network. We conducted experiments using

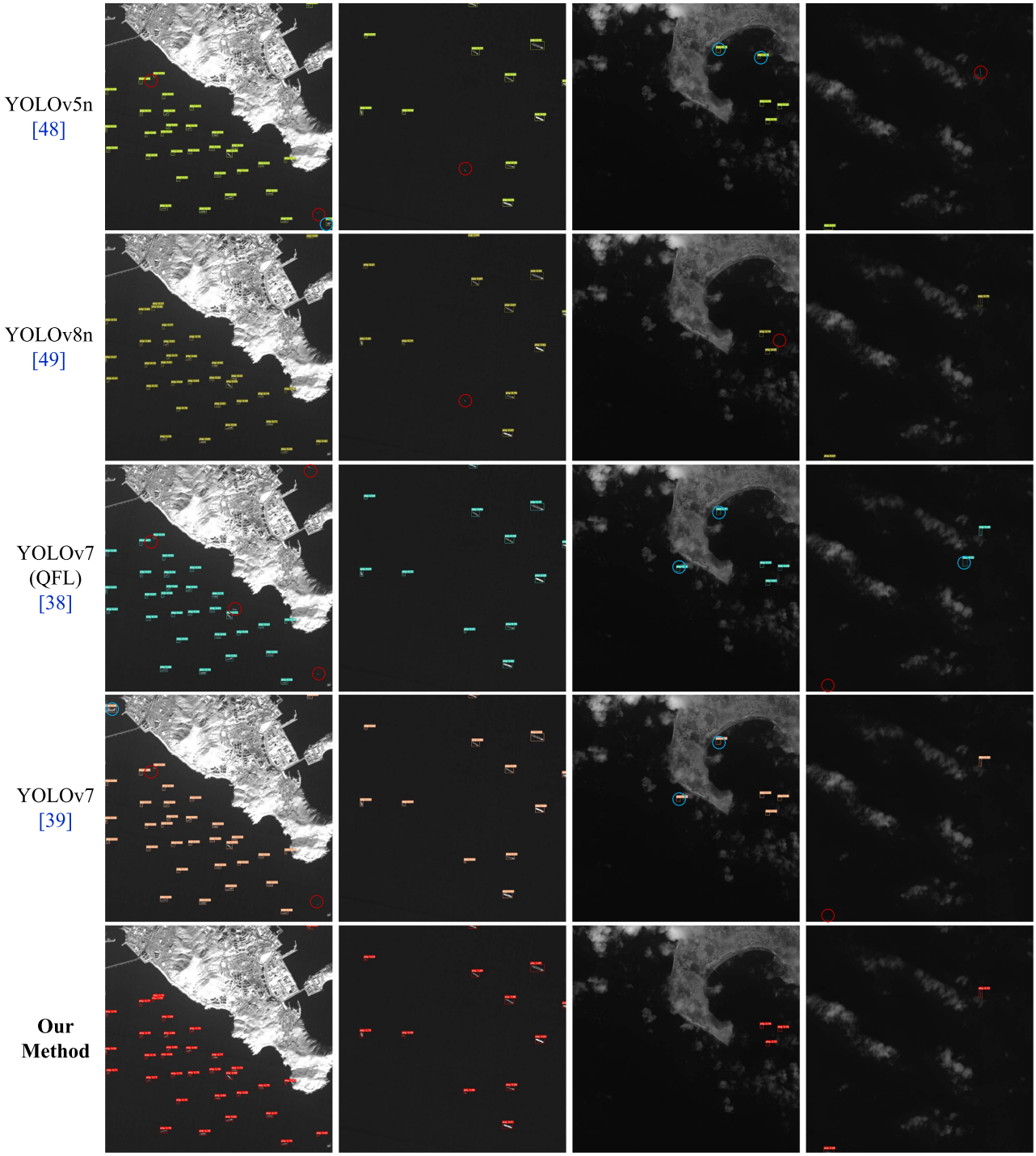


Fig. 12. Comparison of the prediction effect of LPKF-YOLO and other networks. The boxes in the figure indicate that the target is considered by the model to be a ship. The red circles indicate that a ship has been missed, and the blue circles indicate that other objects have been incorrectly identified as ships.

the original YOLOv7 loss function and adjusted only the number of times C2Rep modules were stacked at these four locations, and the results are given in Table VI.

Layers in the table denote the number of layers of the network structure; parameters is the number of parameters to be trained in the model training, and GFLOPs is the number of floating-point operations, which can be used as a measure of the model size (space complexity). As can be seen in Table VI, the parameters and GFLOPs decrease in the improved network despite the increase in the number of layers. This is due to the use of depthwise separable convolution in the C2Rep module and the introduction of the more lightweight C3 module in the head part, which can effectively reduce the number of parameters while maintaining the performance of the network. The complexity of LKPF-YOLO increases with the number of module stacks,

TABLE VI
Comparison of the Number of Times Different Modules are Stacked

Stack quantity	Layers	Parameters(M)	GFLOPs	$mAP_{50}(\%)$
YOLOv7	415	37.2	105.1	88.7
1-1-3-3	493	34.3	81.2	88
1-3-3-1	493	31.7	83.1	90.8
1-3-3-3	529	34.6	85.4	**91.3**
3-3-3-3	565	34.7	92.2	89
3-6-6-3	673	36.6	102.9	88.8

More stacks result in a higher parameter count, but performance does not necessarily increase.
The best score for each indicator has been highlighted in bold.

TABLE VII
Experiments of Each Model on SSDD Dataset

Networks	Precision(%)	Recall(%)	$mAP_{50}(\%)$	$mAP_{50:95}(\%)$
SSD [46]	85.3	81.6	89.3	45.4
Faster R-CNN [47]	81.6	92.5	89.6	46.7
YOLOv5n [48]	87.8	88.8	94.1	59.5
YOLOv8n [49]	94.8	**95**	97.2	62.1
YOLOv7 [39]	86.1	88.3	93.6	58.2
LKPF-YOLO	**95.3**	94.6	**97.9**	**64.9**

The higher these indicators are, the better the results, and the best score for each indicator has been highlighted in bold.

TABLE VIII
Experiments of Each Model on DIOR Dataset

Networks	Precision(%)	Recall(%)	$mAP_{50}(\%)$	$mAP_{50:95}(\%)$
SSD [46]	83.5	73.6	77.2	48.8
Faster R-CNN [47]	79.8	78.2	76.4	47.5
YOLOv5n [48]	80.2	82.5	82.7	58.1
YOLOv8n [49]	83.4	82.1	86.5	**65.4**
YOLOv7 [39]	79.1	**83.9**	83.5	58.7
LKPF-YOLO	**91.1**	82.7	**88.7**	64.1

The higher these indicators are, the better the results, and the best score for each indicator has been highlighted in bold.

and the performance index shows a rising and then decreasing trend. The best performance is achieved when the number of stacks of C2Rep in four positions is "1-3-3-3." In this article, the model with a module stacking number of "1-3-3-3" is used as the final model, and its parameters and GFLOPs are reduced by 7% and 18.7%, respectively, compared to the original network.

E. Generalization Experiment

As a generalization experiment, we train and test these models on DIOR [28] and SSDD [50] datasets. SSDD is a public dataset dedicated to ship target detection in SAR images. DIOR is a color optical remote sensing imagery dataset containing 20 categories with richer scenes and more categories, and most of them have small target samples. The experimental results are given in Tables VII and Table VIII.

The experimental results show that the mAP_{50} metrics of the proposed model in this article are all ahead of other networks and obtain the best overall detection performance. This means that the advantages of the proposed model in this article are not only confined to single-sample categories of grayscale remote sensing images but also obtain competitive performance on other multisample categories and multicolor channels of images.

F. Discussions

From Figs. 11 and 12, it is evident that our proposed method achieves precise detection of dense vessels, small vessels, and vessels obscured by clouds and fog. During experimentation, we observe that in images with numerous blocky clouds, all tested object detection algorithms exhibited extremely high FP rates. This is because the size and shape of small cloud patches closely resemble vessels, and in wide-area remote sensing images, objects lack texture details, making it difficult for the network to distinguish between vessels and small cloud patches. The C2Rep module proposed in this article focuses more on the shape of the target and does not rely as much on texture as other network models. Therefore, it does not produce misclassification as other networks do when facing small irregular clouds, but it is still difficult to differentiate when facing elliptical cloud clumps whose shapes are very similar to ships. These clouds are similar in shape to ships, but their edges are not as smooth as those of ships. So using methods that focus on edge features might improve this.

In addition, our method can detect vessels partially obscured by thin clouds and fog, but it remains ineffective in the presence of thick cloud cover. This limitation is due to the fact that optical remote sensing does not have the penetration capability similar to SAR, which makes it difficult to detect obscured targets and realize all-weather monitoring.

V. CONCLUSION

In this study, to address the need for small target detection in wide-area remote sensing images, we propose a model called LKPF-YOLO based on a large kernel module and unbalanced learning. This model tackles the detection challenges arising from small vessel scales and sparse distributions in wide-area satellite remote sensing images, achieving high-accuracy detection in complex scenarios, such as dense vessels, sparse vessels, small target vessels, and vessels obscured by clouds. Through comparisons with other models on the SSDD and DIOR datasets, our model is not limited to single-class single-channel maritime remote sensing datasets and it performs well across multiple classes and scenes, with the main metric mAP_{50} achieving 97.9% and 88.7% on the two datasets, respectively, surpassing YOLOv5n, YOLOv7, and YOLOv8n. These results indicate the adaptability of our model.

In the future, we will deepen the study of loss function. This study addresses the problem of detecting small targets under wide-area remote sensing imagery and therefore introduces a priori knowledge based on the idea that "the smaller the target, the harder it is to detect." However, this a priori knowledge may lose its effect in other more complicated scenarios, because not all targets in all scenarios are "the smaller, the harder." Therefore, we hope that an unbalanced loss function that covers more scenarios can be developed in the future.

REFERENCES

[1] W. Wu, X. Li, Z. Hu, and X. Liu, "Ship detection and recognition based on improved YOLOv7," *Comput., Mater. Continua*, vol. 76, no. 1, pp. 489–498, 2023.

[2] W. Wu et al., "Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image," *PLoS One*, vol. 16, no. 10, 2021, Art. no. e0259283.

[3] W. Wang, X. Zhang, W. Sun, and M. Huang, "A novel method of ship detection under cloud interference for optical remote sensing images," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3731.

[4] L. Xu, Z. Hu, C. Zhang, and W. Wu, "Remote sensing image segmentation of mariculture cage using ensemble learning strategy," *Appl. Sci.*, vol. 12, no. 16, 2022, Art. no. 8234.

[5] G. Soldi et al., "Space-based global maritime surveillance. Part II: Artificial intelligence and data fusion techniques," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 9, pp. 30–42, Sep. 2021.

[6] W. Jing, B. Cui, Y. Lu, and L. Huang, "BS-Net: Using joint-learning boundary and segmentation network for coastline extraction from remote sensing images," *Remote Sens. Lett.*, vol. 12, no. 12, pp. 1260–1268, 2021.

[7] S. T. Yekeen, A.-L. Balogun, and K. B. W. Yusof, "A novel deep learning instance segmentation model for automated marine oil spill detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 190–200, 2020.

[8] 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.

[9] N. Su, Z. Huang, Y. Yan, C. Zhao, and S. Zhou, "Detect larger at once: Large-area remote-sensing image arbitrary-oriented ship detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[10] X. Fan, Z. Hu, Y. Zhao, J. Chen, T. Wei, and Z. Huang, "A small ship object detection method for satellite remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11886–11898, 2024.

[11] B. Zhang et al., "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.

[12] C. Yu, Y. Liu, X. Xia, Z. Hu, and S. Fu, "Precise segmentation of remote sensing cage images based on SegNet and voting mechanism," *Appl. Eng. Agriculture*, vol. 38, no. 3, pp. 573–581, 2022.

[13] M. Elmikaty and T. Stathaki, "Car detection in aerial images of dense urban areas," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 1, pp. 51–63, Feb. 2018.

[14] J. Lu, Q. Hu, R. Zhu, and H. Jia, "A high-resolution remote sensing image registration method combining object and point features," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 5, pp. 4196–4213, Oct. 2022.

[15] C. Yu et al., "Segmentation and density statistics of mariculture cages from remote sensing images using mask R-CNN," *Inf. Process. Agriculture*, vol. 9, no. 3, pp. 417–430, 2022.

[16] X. Song and W. Yu, "Processing video-SAR data with the fast backprojection method," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 2838–2848, Dec. 2016.

[17] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 50–65, 2021.

[18] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.

[19] D. Barretta, L. M. Millefiori, and P. Braca, "Analytical classification performance analysis of machine-learning-based ship detection from optical satellite imagery," in *Proc. 2024-2024 IEEE Int. Geosci. Remote Sens. Symp.*, 2024, pp. 8861–8865.

[20] C. Qin, X. Wang, Y. Liu, and G. Li, "A novel end-to-end transformer network for small scale ship detection in SAR images," in *Proc. 2024-2024 IEEE Int. Geosci. Remote Sens. Symp.*, 2024, pp. 8158–8162.

[21] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. pattern Recognit. Appl. Methods*, vol. 2. SciTePress, 2017, pp. 324–331.

[22] L. Yu and P. Gong, "Google Earth as a virtual globe tool for Earth science applications at the global scale: Progress and perspectives," *Int. J. Remote Sens.*, vol. 33, no. 12, pp. 3966–3986, 2012.

[23] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "ShipRSImage Net: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8458–8472, 2021.

[24] S. Kizilkaya, U. Alganci, and E. Sertel, "Vhrships: An extensive benchmark dataset for scalable deep learning-based ship detection applications," *ISPRS Int. J. Geo- Inf.*, vol. 11, no. 8, 2022, Art. no. 445.

[25] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2018, pp. 3974–3983.

[26] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, and J. Han, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 29, 2023, doi: 10.1109/TPAMI.2023.3290594.

[27] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.

[28] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[29] N. Arshad, K.-S. Moon, and J.-N. Kim, "An adaptive moving ship detection and tracking based on edge information and morphological operations," in *Int. Conf. Graphic Image Process. (ICGIP)* vol. 8285, pp. 474–479, 2011.

[30] F. Yang, Q. Xu, and B. Li, "Ship detection from optical satellite images based on saliency segmentation and structure-LBP feature," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 602–606, May 2017.

[31] 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.

[32] L. Chen, W. Shi, and D. Deng, "Improved YOLOv3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 660.

[33] J. Yu, D. Huang, X. Shi, W. Li, and X. Wang, "Real-time moving ship detection from low-resolution large-scale remote sensing image sequence," *Appl. Sci.*, vol. 13, no. 4, 2023, Art. no. 2584.

[34] X.-L. Zhang, "Speech separation by cost-sensitive deep learning," in *Proc. 2017 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2017, pp. 159–162.

[35] J. Jiang et al., "Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network," *Biomed. Eng. online*, vol. 16, no. 1, pp. 1–20, 2017.

[36] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5375–5384.

[37] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[38] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21002–21012, 2020.

[39] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.
[40] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, pp. 157–173, 2008.
[41] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Procedia Comput. Sci.*, vol. 199, pp. 1066–1073, 2022.
[42] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11963–11975.
[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
[44] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2458–2466.
[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[46] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Computer Vision–ECCV 2016: 14th Euro. Conf.*, Amsterdam, The Netherlands, 2016, pp. 21–37.
[47] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[48] G. Jocher, K. Nishimura, and T. Mineeva, "Yolov5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5
[49] G. Jocher, A. Chaurasia, and J. Qiu, "Yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[50] T. Zhang et al., "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, 2021, Art. no. 3690.

Junfei Chen received the B.Eng. degree majored in communication engineering in 2018 from the School of Information and Communication Engineering, Hainan University, Haikou, China, where he is currently working toward the M.Eng. degree in artificial intelligence with the School of Information and Communication Engineering.

His main research interests include image object detection and digital image processing.

Zhuhua Hu (Senior Member, IEEE) received the B.Eng. degree in biological and agricultural engineering and the M.Eng. degree in computer software and theory from Jilin University, Changchun, China, in 2002 and 2005, respectively, and the Ph.D. degree in information and communication engineering from the Hainan University, Haikou, China, in 2019.

From 2005 to 2006, he was a Software Engineer with the Ningbo BIRD Research Institute of China. From 2006 to 2007, he was a Software Engineer with the Nanjing Research Institute of ZTE. From 2007 to 2009, he was a Minister of the Software Department with Shanghai Aoxun Information Technology Co., Ltd. Since 2020, he has been a Professor and Doctorial Tutor with the School of Information and Communication Engineering, Hainan University. He is currently a high-level talent in Hainan Province. He led the "multimodal information intelligent processing and decision control" innovation team, and has authored or coauthored more than 110 academic papers in journals, such as TMM, TIP, TVT, TAES, TGRS, TITS, TCAD, OE, and COMPAG, authorized 13 patents, and hosted more than ten large-scale commercial projects that have been successfully implemented. His current research interests include artificial intelligence, and signal and information processing.

Dr. Hu is a CCF Senior Member. He acted as the Reviewer for IEEE INTERNET OF THINGS JOURNAL, *Ocean Engineering*, IEEE-ACM TRANSACTIONS ON NETWORKING, Engineering Applications of Artificial Intelligence, *IMAVIS-Image and Vision Computing*, TETC-IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, ICASSP2023&2024, and ICME 2024.

Wei Wu received the B.Eng. degree in electronic information engineering from the Beijing University of Chemical Technology, Beijing, China, in 2003, and the M.Eng. degree in software engineering from Tsinghua University, Beijing, in 2008. He is currently working toward the Ph.D. degree in communication and information engineering with the School of Information and Communication Engineering, Hainan University, Haikou, China.

His main research interests include digital image processing and computer systems.

Yaochi Zhao received the M.S. degree in pattern recognition and intelligent system from Central South University, Changsha, China, in 2005, and the Ph.D. degree in computer application technology from Tianjin University, Tianjin, China, in 2023.

She worked with Ningbo BIRD Research Institute and Shanghai Wingtech Communication Company, Ltd., for three years. Later, she was engaged in teaching and research work with the College of Information Science & Technology, Hainan University, Haikou, China. She is currently an Associate Professor with the School of Cyberspace Security, Hainan University. Her current research interests include image processing and computer vision.

Ba Huang received the M.Eng. degree in communication and information systems from the Changsha University of Science and Technology, Changsha, China, in 2012.

She was responsible for the Hainan Provincial Key R&D Project—the "Precision Intelligent Joint Control Integration" combat system and demonstration application subproject for offshore vessels, serving as the person in charge of the subproject. She was in charge of extracting remote sensing data information to create corresponding products and identifying maritime vessels. She has more than ten years of extensive experience in Big Data analysis and satellite remote sensing applications. Her current research interests include data analysis, maritime target identification, and more.