frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 23 January 2023
DOI 10.3389/fmars.2022.1058401

OPEN ACCESS

EDITED BY
Xuemin Cheng,
Tsinghua University, China

REVIEWED BY
Carlos Pérez-Collazo,
University of Vigo, Spain
Ning Wang,
Dalian Maritime University, China

*CORRESPONDENCE
Jiucai Jin
jinjiucai@fio.org.cn

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 30 September 2022
ACCEPTED 28 November 2022
PUBLISHED 23 January 2023

CITATION
Zhang J, Jin J, Ma Y and Ren P (2023) Lightweight object detection algorithm based on YOLOv5 for unmanned surface vehicles.
Front. Mar. Sci. 9:1058401.
doi: 10.3389/fmars.2022.1058401

Lightweight object detection algorithm based on YOLOv5 for unmanned surface vehicles

Jialin Zhang[1,2], Jiucai Jin[1]*, Yi Ma[1] and Peng Ren[2]

[1]First Institute of Oceanography, Ministry of Natural Resources, Qingdao, China, [2]College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao, China

Visual detection technology is essential for an unmanned surface vehicle (USV) to perceive the surrounding environment; it can determine the spatial position and category of the object, which provides important environmental information for path planning and collision prevention of the USV. During a close-in reconnaissance mission, it is necessary for a USV to swiftly navigate in a complex maritime environment. Therefore, an object detection algorithm used in USVs should have high detection s peed and accuracy. In this paper, a YOLOv5 lightweight object detection algorithm using a Ghost module and Transformer is proposed for USVs. Firstly, in the backbone network, the original convolution operation in YOLOv5 is upgraded by convolution stacking with depth-wise convolution in the Ghost module. Secondly, to exalt feature extraction without deepening the network depth, we propose integrating the Transformer at the end of the backbone network and Feature Pyramid Network structure in the YOLOv5, which can improve the ability of feature expression. Lastly, the proposed algorithm and six other deep learning algorithms were tested on ship datasets. The results show that the average accuracy of the proposed algorithm is higher than that of the other six algorithms. In particular, in comparison with the original YOLOv5 model, the model size of the proposed algorithm is reduced to 12.24 M, the frames per second reached 138, the detection accuracy was improved by 1.3%, and the mean of average precision (0.5) reached 96.6% (from 95.3%). In the verification experiment, the proposed algorithm was tested on the ship video collected by the "JiuHang 750" USV under different marine environments. The test results show that the proposed algorithm has a significantly improved detection accuracy compared with other lightweight detection algorithms.

KEYWORDS

object detection, USV, ghost model, lightweight, YOLO

1 Introduction

In recent years, unmanned surface vehicle (USV) technology has developed rapidly, and USVs are widely used in maritime safety tasks, such as orderly and complex patrols, reconnaissance, and detection and tracking of specific objects. Traditional ship detection and tracking systems typically employ radar or AIS (Vesecky et al., 2009; Dzvonkovskaya and Rohling, 2010; Vesecky et al., 2010; Sermi et al., 2013). However, the radar has a relatively long scanning period and slow detection speed. It cannot distinguish between specific types of objects, and hence false and missed detections easily occur. Information collected by AIS can be intentionally turned off by ships, which sometimes results in AIS unreliability. The existing ship detection methods are based on vision; they not only have a long detection range but also have high resolution and object detailing. The traditional detection methods based on vision are mainly Mean-shift (Liu et al., 2013) and HOG-SVM (Xu and Liu, 2016). Their characteristic is that they mainly rely on a single shallow feature to complete the ship detection task. However, these features are easily affected by the ship's appearance, shape, and complex environment, resulting in poor robustness. With the rapid development of the visual field, visual object detection based on deep learning has become a popular research topic. Object detection algorithms based on deep learning have broad application prospects in the marine environment (Chen et al., 2021; Wang et al., 2022); nevertheless, their applications have not been fully valued until now (Mittal et al., 2022). For example, object detection can be used to perceive the surrounding environment. The object's orientation and image information plays an important role in path planning, collision avoidance, and object monitoring of a USV. At present, an object detection algorithm based on deep learning can more accurately classify and detect object positions. However, it has high requirements for the vision-based processing system of the USV; moreover, speed and accuracy of the object detection algorithm are also major challenges.

In this study, we propose a lightweight object detection network based on the You-Only-Look-Once-v5 (YOLOv5) to obtain fast detection speed and high accuracy for USVs. The object detection performance in a complex environment has been improved. The proposed network has reduced detection time and improvements in terms of anchor boxes, backbone, and feature pyramid network (FPN) structure. We obtained a set of anchor boxes through the K-means clustering method to adopt to the ship's characteristics. The Ghost module upgraded the convolution (Conv) in the backbone to reduce the network detection time. The Transformer is integrated into the cross stage partial network (CSPNet) of the backbone and FPN structure to achieve more useful feature extraction. The proposed network is composed of these simple but effective modules, thus balancing detection speed and accuracy well. Figure 1 shows the detailed flowchart of our training model. Lastly, the experimental results demonstrate its excellent performance on the task of detecting ship objects.

The contributions of this study include the following:

- We obtain a new set of anchor boxes to adapt to the structural characteristics; i.e., the width of the ship is longer than the height used by the K-means clustering algorithm on the ship dataset.
- A combination of Conv stacking with depth-wise Conv in the Ghost module was adopted to structure the backbone feature extraction in YOLOv5. In comparison with the original Conv, the Ghost module has better computing efficiency, which not only reduces the model training and detection times but also improves accuracy.
- We integrated the Transformer into the end of the backbone and FPN structure in the YOLOv5 network, which can improve the feature expression ability and enhance the detection accuracy without deepening the network depth.
- The proposed algorithm has achieved a good balance between detection accuracy and speed. In the actual marine environment testing process, our algorithm obtains a high accuracy rate and is found to be robust in the sea fog environment.

The remainder of this paper is organized as follows. In *Section 2*, we show the data augmentation and related work. We describe our approach in *Section 3*. The experimental results performance and discussion are presented in *Section 4*. In *Section 5*, we summarize this work.

2 Related work

2.1 Data augmentation

The purpose of data augmentation is to generate more training samples based on existing datasets. The method of data augmentation is to randomly transform the local or global features of the images, and its role is to improve the robustness and generalization ability of our trained model. In certain special circumstances, highlighting, blurring, and occlusion were encountered in the future detection process of our model. Therefore, the hue, saturation, and value have been adjusted in the model training process. With regard to the geometric distortion of the image, certain operations are performed, i.e., rotation, horizontal and vertical translation, scaling, and shearing of the image. In addition, there are some special data enhancement methods, such as Mixup (Zhang et al., 2017) and Mosaic (Bochkovskiy et al., 2020). In the Mixup data

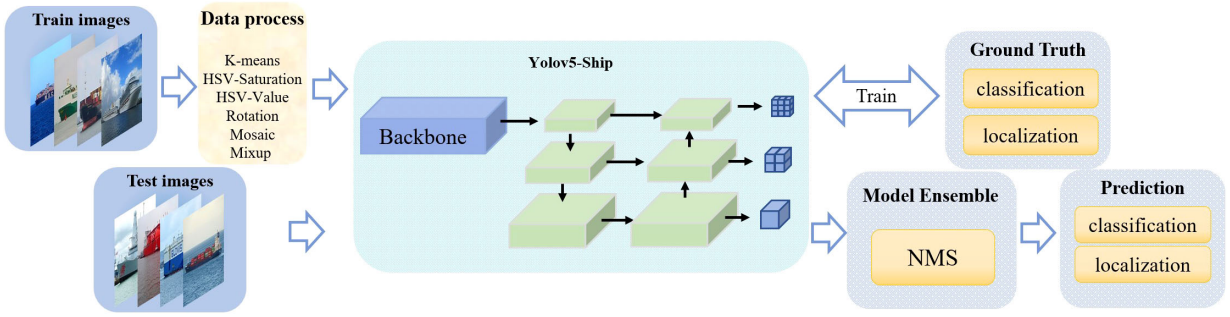

FIGURE 1
Detailed flowchart of the proposed model training.

enhancement method, new sample-label data are generated by adding two image sample-label data pairs in proportion. In the Mosaic data enhancement method, a new picture is generated using four pictures through random reduction, cropping, and arrangement. In this paper, we used a combination of Mixup, Mosaic, and traditional data augmentation methods.

2.2 Visual object detection based on deep learning

In recent years, visual detection technology has made great progress, particularly detection methods that are based on deep learning. The deep learning-based object detection algorithms are mainly divided into two types—two-stage and one-stage. The first step of a two-stage object detection algorithm is to generate a position box by generating a region proposal that can extract features; then, the second step is to perform category prediction. It has high accuracy but slow speed; thus, it is not suitable for real-time object detection like Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2015). A one-stage object detection algorithm performs classification and bounding box regression while generating candidate boxes and has fast speed but less accuracy; hence, it is suitable for real-time object detection like SSD (Liu et al., 2016) and YOLOv3 (Redmon and Farhadi, 2018). High object detection speed is essential for a USV platform; therefore, one-stage object detection algorithms are more suitable.

In the case of maritime object detection, many scholars have investigated from sea–skyline detection to ship detection. Bai et al. (2021) proposed a sea–skyline detection method based on local Otsu segmentation and Hough transform. Later, the monopole object detection method was introduced for ship detection, which reduces a certain amount of interference and calculations, and it optimizes the accuracy and speed of ship detection. Chen et al. (2021) proposed an integrated ship detection framework based on an image segmentation method for edge detection. The Canny edge detector and Gaussian filter are used to detect the edges of ships in the image, suppress the edges related to the background, and, finally, connect them to form the outline of the ship; the method achieved an effect of 32 fps. In ship detection methods based on deep learning, Gupta et al. (2021) proposed a classification method for ship detection based on support vector machines (SVMs) and convolutional neural networks (CNNs). First, the feature package is used to deal with diverse features of different types of ships, and then the CNN is used for feature extraction. Finally, 2,700 images are used for training, and the accuracy rate of their model reaches 91.04%. Zou et al. (2019) improved a maritime object detection method based on Faster R-CNN. The ResNet-50 network is replaced by the VGG16 network. The results show that the recognition and detection effect of small ships was significantly improved. Zou et al. (2020) proposed an improved SSD algorithm based on the MobileNetV2 CNN that is used in ship detection and identification. The results show that the SSD_MobileNetV2 algorithm has better performance for ship images. Shi and Suo (2018) proposed a ship detection algorithm based on an improved visual attention model. Firstly, the wavelet transform (WT) is used for feature extraction; secondly, the improved Gabor filter and deep multifaceted transformers (DMT) algorithm are used to obtain the directional and edge texture features of the image. The final test demonstrated high detection accuracy and good real-time performance. For the existing ship detection algorithms based on deep learning, it is difficult to simultaneously obtain good detection accuracy and real-time performance.

2.3 Ship detection based on YOLO

Since the YOLO algorithm was published, it has been widely studied because of its good computational efficiency and detection accuracy. Lee et al. (2018) applied the YOLOv2 algorithm to ship detection and classification. In comparison with other machine learning algorithms, their model has better robustness and scalability. Li and Qiao (2021) proposed a ship

detection and tracking algorithm based on YOLOv3. They used a graph matching algorithm and Kalman filter to achieve object matching and tracking, which solves the problems of object occlusion and label switching. Jie et al. (2021) improved YOLOv3 for ship detection and tracking in inland waterways; the K-means clustering algorithm was used to improve the anchor boxes, and it was improved by taking the single softmax classifier and introducing the Soft-NMS algorithm. Their algorithm could enhance the safety of inland navigation and prevent collisions and accidents. Zhang et al. (2020) improved a maritime object detection algorithm based on YOLOv3. They proposed an E-CIoU loss function for bounding box regression, and the improved method accelerated the convergence speed and improved the detection accuracy. Liu and Li, (2021) studied ship statistics in waterway videos. To realize automatic detection and tracking by YOLOv3, they designed a self-correcting network combining regression-based direction judgment and object counting method with variable time window. The results show that their algorithm can achieve automatic analysis and statistical data extraction in waterways videos. Sun et al. (2021) optimized the backbone network CSPDarkNet of YOLOv4 for application in an auxiliary intelligent ship navigation system. They added a receptive field block module, and the FPN of YOLOv4 was improved by combining the Transformer mechanism. Their algorithm improves the inference speed and detection accuracy. Liu et al. (2021) improved the USV maritime environment perception ability using an improved YOLOv4 object detection algorithm. The reverse depth-wise separable convolution (RDSC) was applied to the backbone and FPN structures of YOLOv4, which reduced the number of parameters of the network and improved the accuracy by 1.78% compared with the original model. Thus, the algorithm has a small network size and better performance in terms of detection speed.

In summary, the ship detection methods are mostly difficult to apply on USVs because of limited computing resources and detection speed. Thus far, the problems of accuracy and speed of maritime object detection have not been resolved. In comparison with traditional object detection algorithms, the deep learning-based object detection algorithm has good accuracy rate, but slow detection speed. Therefore, this study focuses on improving an object detection algorithm based on YOLOv5 to solve the problems of real-time performance and accuracy of the maritime ship detection algorithm applied to the USV platform.

3 Methods

The maritime object detection includes two tasks, i.e., classification and positioning of ships. A robust object detection algorithm should not only consider the detection speed, but also consider the complex environmental scenarios. In the field of object detection, the YOLO object detection algorithm performs well in various environments, such as changes in illumination in a complex sea environment, and recognition of distant small targets in the sea. The fifth version YOLO object detection algorithm has been developed, and its efficiency is very good.

YOLOv5 has high performance in terms of detection speed and accuracy. According to the depth and width of the network, it is divided into four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The basic network of the four versions is similar. The structure of YOLOv5 is mainly composed of the input, backbone, Neck, and Prediction. At the input, we perform data augmentation operations, such as Mixup and Mosaic, which can enrich the ship dataset and improve the detection efficiency of small objects. Feature maps of different scales are extracted at the backbone network. The FPN and path aggregation network (PANet) at the Neck strengthen the feature fusion ability. The FPN transfers high-level semantic features in a top-down manner, and the PANet transfers low-level strong localization features in a bottom-up manner after the FPN. The final output is the prediction of the network, and the prediction uses the non-maximum suppression (NMS) algorithm to filter the object boxes. Then, we make predictions on the image features, generate bounding boxes and predict classes.

In this study, we examine the ability of the USV to detect and classify an object quickly. We used YOLOv5 as the base network and improved it. The architecture of the improved YOLOv5 is shown in Figure 2.

3.1 Anchor box calculation

In object detection tasks, choosing suitable anchor boxes can significantly improve the speed and accuracy of object detection. Anchor boxes are boxes presented by a fixed aspect ratio in YOLO, which is used to predict the category and position offset of the bounding box. The default anchor boxes are generated in the MS COCO and VOC datasets. The COCO and VOC datasets have 80 and 20 classes, respectively, but ships are only one of their classes. Therefore, the default anchor boxes are not fully applicable to the objects in the ship dataset. To adapt the structural characteristics of the width of the ship being longer than the height of the ship, we used the K-means clustering algorithm on the ship dataset to obtain a set of anchor boxes. The clustering results for the ship dataset labels are shown in Figure 3. The steps to implement the Algorithm 1 are described as follows.

Input:
A ground truth label dataset: $S = \{x_1, x_2, x_3 \ldots x_m\}$
The number of cluster centers: k
Output:
A group of anchor boxes: $\{c_1, c_2, c_3 \ldots c_k\}$

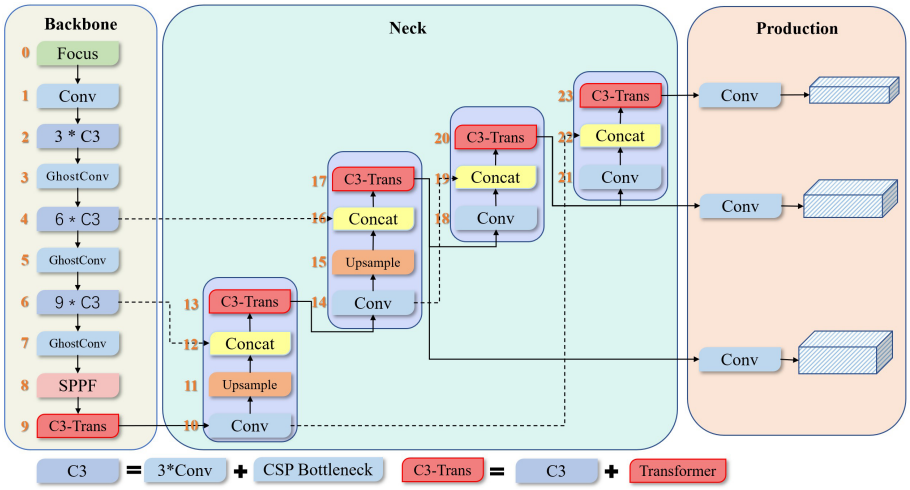

FIGURE 2
Improved YOLOv5 network structure proposed in this paper.

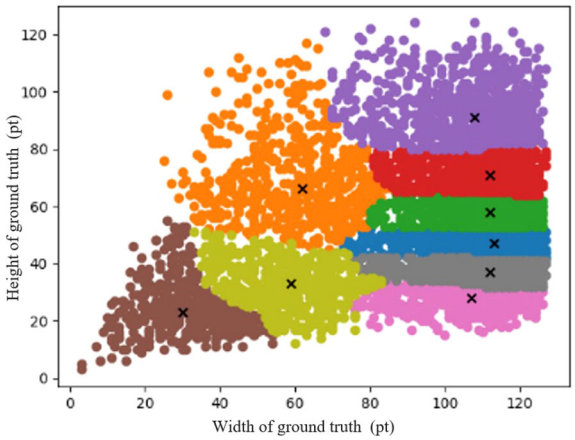

FIGURE 3
Result of ship dataset using K-means clustering. The *x*-coordinate is the width of the ground truth bounding box and the *y*-coordinate is the height of the ground truth bounding box.

Procedure:
First, select randomly nine boxes of ground truth labels from the ship dataset as the cluster centers;
for $i = 1, 2 \ldots k$ **do**
REPEAT
for $j = 1, 2, 3 \ldots m$ **do**
Calculate the distance between x_j and each cluster center $\{c_1, c_2, c_3 \ldots c_k\}$ $d_{ji} = ||x_i - c_i||_2$;
Return each label x_j to cluster centers c_i with the closest distance; Update the cluster center c_i for each class in each cluster $c_i = \Sigma_{x \in c_i} \frac{x_i}{|c_i|}$;
end for
UNTIL Cluster centers no longer change.

ALGORITHM 1
Pseudocode of K-means clustering algorithm for anchor boxes.

Finally, nine sets of adaptive anchor boxes are generated using the K-means clustering algorithm, i.e., (29,23), (58,31), (109,30), (62,60), (112,39), (114,50), (78,89), (112,65), and (112, 87). The anchor boxes of the clustering algorithm can effectively accelerate the convergence speed of the network and effectively improve the gradient descent problem in the training process.

3.2 Ghost model

There are limitations regarding the memory and computing resources of embedded industrial computers in USVs; therefore, the key to ship detection on an USV is to find a lightweight detection model that can balance detection accuracy and computational complexity. CNNs are usually composed of many convolution kernel operations, which will result in large computational cost. During model training, many redundant feature maps will be generated, as shown in Figure 4. Redundant feature maps not only have high similarity but also greatly increase computational complexity. To reduce the computational load of the model and raise the detection speed, an efficient architecture and high-performance GhostNet (Han et al., 2020) structure are adopted.

The detailed structure of the Conv and Ghost model is shown in Figure 5. Figure 5A shows the Conv operaton. A given input is

defined as $X \in \mathbb{R}^{c \times h \times w}$, where c is the number of channels of the input; h and w are the height and width of the input data, respectively. The n feature maps are generated through ordinary convolution that can be expressed as $Y = X * f + b$where $Y \in \mathbb{R}^{c \times k \times k \times n}$ is the output feature map with n channels, and * is the convolution operation; f denotes the convolution filter of this layer, b is the bias term, and $k \times k$ is the size of the convolution kernel f. The value of the floating point of operations (FLOPs) can be expressed as $n \cdot h \cdot w \cdot c \cdot k \cdot k$. Owing to the large values of n and c, the usual parameters of the model are very large. The Ghost model comprises Conv and depth-wise Conv with less parameters and computations. The Ghost model first obtains the necessary feature map of half channel of the input features through Conv. These necessary feature maps are used to perform the depth-wise Conv that can obtain similar feature maps of the necessary feature maps. Finally, the two parts of the feature maps from Conv and depth-wise Conv are spliced. The schematic diagram of the Ghost module is shown in Figure 5B. Specifically, we used the primary convolution $Y' = X * f'$ generate m feature maps $Y' \in \mathbb{R}^{h' \times w' \times m}$. To obtain the required n feature maps, the following cheap operations are used for each intrinsic feature in Y':

$$y_{ij} = \Phi_{i,j}(y'_i), \forall\ i = 1, 2, \ldots m, j = 1, 2, \ldots \ldots s \tag{1}$$

where y'_i is the ith intrinsic feature map in Y' and $\Phi_{i,j}$ is the depth-wise Conv operation to generate the jth (except the last one) Ghost feature map y_{ij}; y'_i can obtain one or more feature maps. The last $\Phi_{i,s}$ is the identity mapping to preserve the intrinsic feature map as shown in Figure 5B. We can obtain $n = m \cdot s$ feature maps for $Y = [y_{11}, y_{12} \ldots y_{ms}]$, which are taken as the output of the Ghost module. The value of the Ghost module FLOPs can be expressed as $\frac{n}{s} \cdot h \cdot w \cdot c \cdot k \cdot k + \frac{n}{s}(s-1) \cdot h \cdot w \cdot k \cdot k$. The operations $\Phi_{i,j}$ are convoluted on one channel. One convolution kernel of ordinary convolution is convoluted on every channel. The computational cost of the depth-wise Conv operation is much lower than that of the ordinary convolution.

The original convolution operation in the YOLOv5 backbone network is upgraded to Conv stacking with depth-wise Conv in the Ghost module, which can raise the operation speed and reduce the number of parameters of the model.

3.3 Transformer encoder block

In the case of ship detection, the classification result of the model can be affected because of the high similarity of ship features. Generally, an image contains rich visual information, such as the object and background information. The key is to fully mine the information in the sample and solve the problem of low accuracy. The Transformer's (Vaswani et al., 2017; Zhu et al., 2020) self-attention mechanism is used to learn the association between the foreground and background in the sample, so that the model can focus on the key areas for detection. The Transformer can improve the detection accuracy of objects. First, the Transformer constructed the sample features into sequence form and added positional encoding. Then, the self-attention mechanism of the Transformer model was used to learn the association between each feature block and assigned different attention to each feature block. Lastly, the original feature sequences are fused, and each feature block in the sequence can contain useful

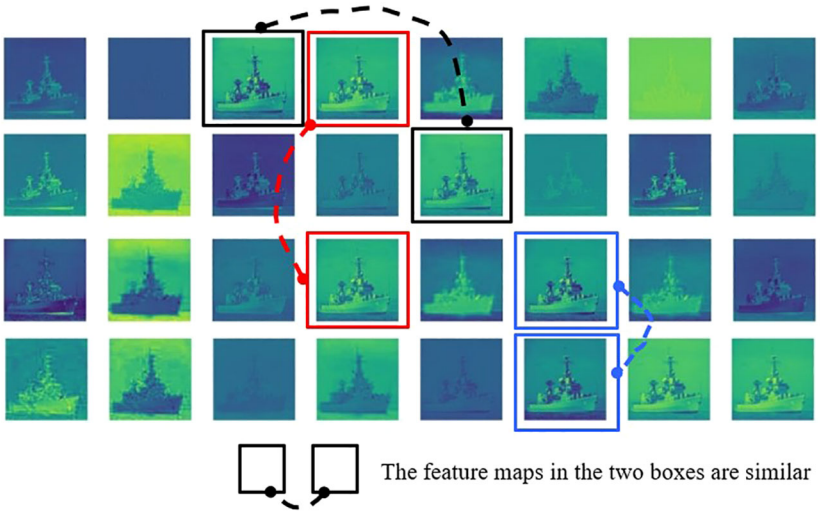

FIGURE 4
Redundant feature maps generated by original convolution.

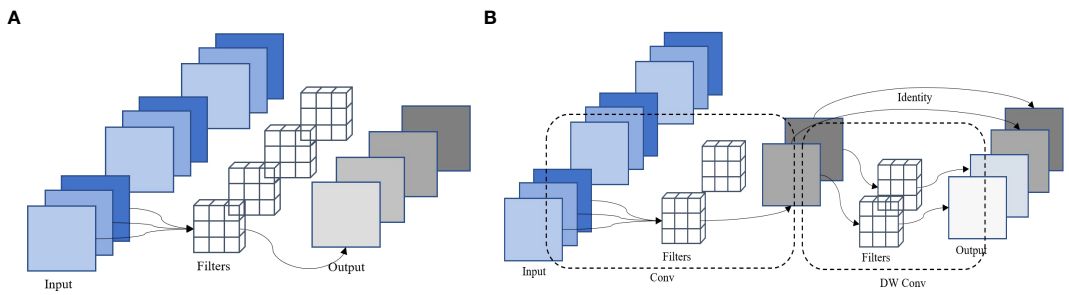

FIGURE 5
Conv and Ghost module structure diagrams **(A)** The Conv operaton, **(B)** The Ghost module.

information for detection in other feature blocks. These operations can enhance the feature expression ability of training samples and improve the accuracy of classification and detection.

The Transformer encoder comprises L layers of alternating Multihead Self-Attention (MSA) and Multilayer Perceptron (MLP) modules. The model structure of Transformer is shown in Figure 6. Therefore, the output $Z_l$if layer l based on the Transformer encoder is:

$$Z'_l = MSA(LN\ (Z_{l-1})) + Z_{l-1} \quad (2)$$

$$Z_l = MLP\ (LN\ (Z'_l)) + Z'_l \quad (3)$$

where $l = \{1, 2,\ldots, L\}$ represents the number of layers, $LN(\cdot)$ presents the layer normalization operation, and Z'_l represents the output of the Lth layer of the MSA. The final output (hidden feature) of the Transformer encoder is $Z_L \in \mathbb{R}^{N\times P\times P}$.

To improve the detection accuracy of the network without deepening the network depth, we focused on the fusion of multilayer features on the PANet and optimization of the feature transfer on the FPN structure. High-quality feature map upsampling and forward transfer were obtained, and the interference of the underlying feature background was reduced. The Transformer was integrated into YOLOv5, which could improve the feature expression. The Transformer was taken into the end of the backbone structure and CSPnet module of the FPN structure. The spatial areas of low-level features were weighted by the salient target position information contained in the attention map, which highlighted the salient regions of the low-level features and suppressed the interference of the background. Thus, it could be more conducive to the identification and classification of ships.

The Transformer could guide the model's attention to reliable and useful channels, while reducing the impact of unreliable and useless background channels. Based on the YOLOv5 model, we integrated the Transformer block at the end of its backbone and Neck networks. Because the resolution of the images at the end of the backbone network was relatively low, applying the Transformer module on the low-resolution feature maps could reduce the additional computational cost.

4 Experiment

4.1 Datasets

In marine transportation, there are generally five basic types of vessels, namely, cargo ships, general cargo ships, carrier ships, bulk carriers, and oil tankers. In addition, there are other types of ships, such as ro-ro, reefer, barge, and liquified natural gas carrier. Among them, cargo, carrier, and cruise ships account for 60%–70% of global ships (Electronic Quality Shipping

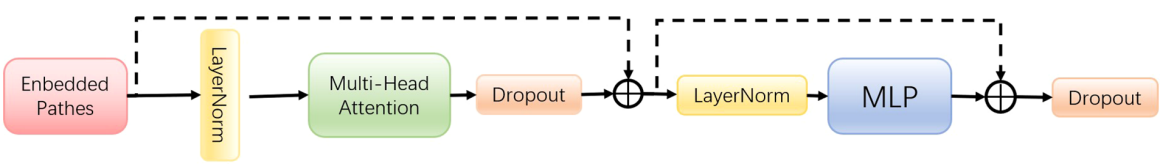

FIGURE 6
Transformer encoder architecture.

Information System, 2020). Therefore, we selected a ship dataset, which can be found on Kaggle (Jain, 2021). It includes five different ship types: cargo, military, carrier, cruise, and oil tanker. Additionally, the dataset comprises 7,604 ship images, including 1,853 cargo ships, 916 warships, 829 transport ships, 1,281 cruise ships, and 1,062 tankers. Figure 7 shows sample images that were randomly selected from ship datasets.

The "JiuHang750" USV is designed and fabricated to detect and trace ships and is used as our research platform. The USV was equipped with the three-light photoelectric platform, which comprises a 30× continuous zoom high-definition visible light camera, an 80-mm uncooled infrared thermal imager, and a 5-km laser rangefinder. The visible light camera can achieve 30× optical zoom and output video images with a 1,920 × 1,080 resolution; the stabilization accuracy of the photoelectric platform reaches 0.5 mrad, the rotation range can reach 360°, and the pitch angle can reach 70° up and down. Based on this optoelectronic platform, the "JiuHang750" USV collected images in the areas of Yellow Sea to test the detection ability of the algorithm in the maritime environment in October and December 2021 and February 2022. The video screenshots are shown in Figure 8.

4.2 Experimental environment and parameters

To ensure experimental consistency, all experiments in this study were carried out under the same hardware platform and software framework. All models used an NVIDIA RTX2080Ti GPU (11 GB) for training and testing. The operating system was CentOS 7, the test framework was PyTorch1.9.0, and the CUDA version 10.2 was the parallel computing framework. The networks were trained for 200 epochs.

4.3 Analysis of results

4.3.1 Comparison with other object detection algorithms

In this section, we evaluate the performance of the proposed improved YOLOv5 algorithm. Multiple evaluation indicators were used to evaluate the performance of the different object detection algorithms, including Average Precision (AP), Precision (P), Recall (R), and F1-score. The mean average

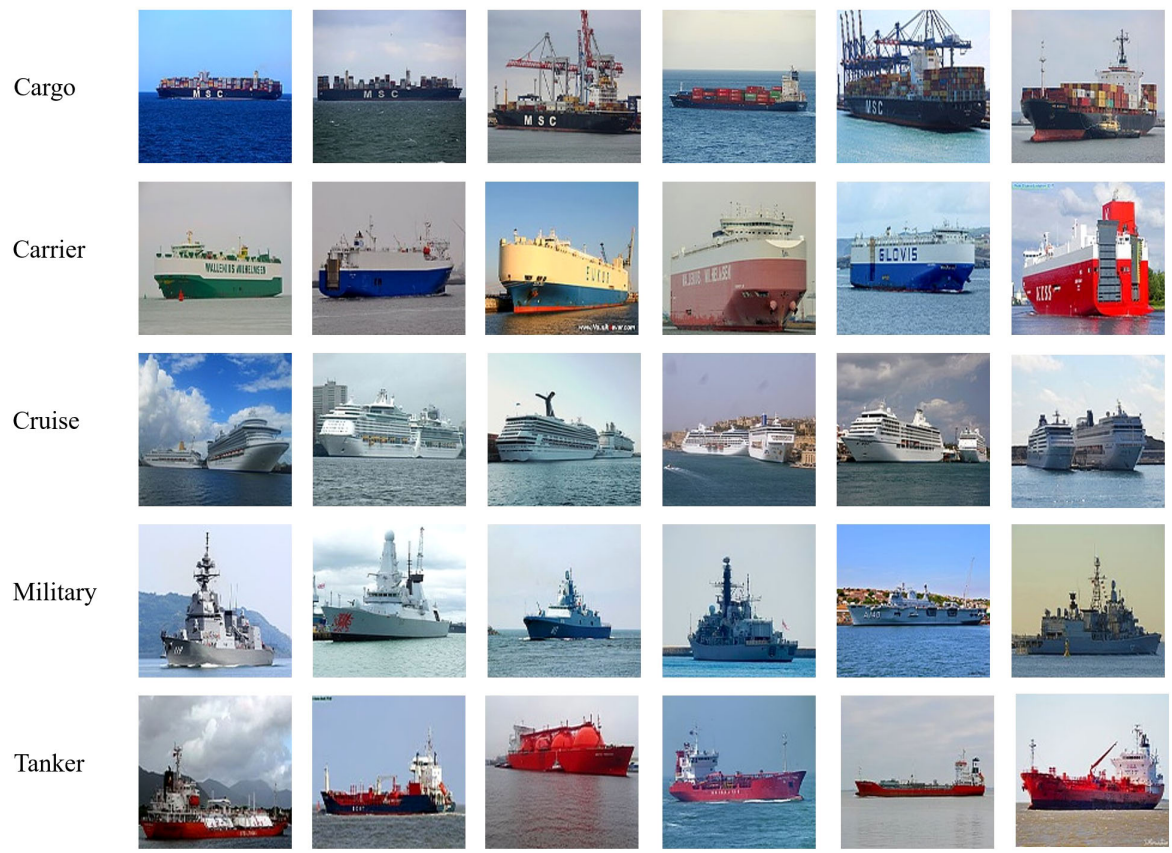

FIGURE 7
Randomly selected sample images from the dataset.

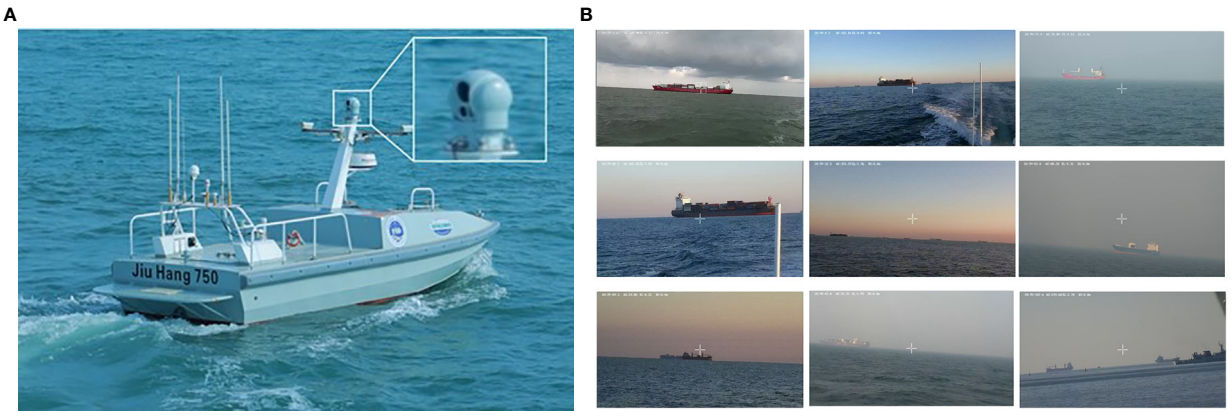

FIGURE 8
(A) "JiuHang750" USV and **(B)** its video images collected under different weather conditions.

precision (mAP) was adopted to evaluate the accuracy of the object detection algorithms. P was adopted to measure the algorithm classification accuracy, and R was used to measure the recall ability of the algorithm detection. The F1-score can consider both P and R. The frames per second (FPS) is an important indicator to evaluate the speed of a target detection algorithm, which indicates the number of frames per second processed by the detection algorithm. The calculation formulas are presented as follows:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$AP = \int_0^1 PR \cdot dR \tag{6}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{8}$$

where P represents the precision rate, R represents the recall rate, TP represents the situation where the prediction and label are both ships, and FP represents the situation where the prediction is a ship but the label is the background; FN represents the situation where the prediction is the background but the label is the ship. n represents the number of classes.

Four deep learning and two lightweight algorithms were used to compare with the proposed algorithm, including SSD, YOLOv3, YOLOv4, YOLOv5, YOLOv3-tiny, and YOLOv4-tiny. The specific test results in Table 1 show that the proposed algorithm achieves the best results between detection speed and accuracy, and its detection precision is better than SSD, YOLOv3, YOLOv4, YOLOv3-tiny, and YOLOv4-tiny. The ship detection precision of our study is 0.7% and 1.5% higher than that of YOLOv3 and YOLOv4, respectively, and 28.8% and 43.9% higher than that of YOLOv3-tiny and YOLOv4-tiny, respectively. The FPS value of our algorithm was 138. The detection speed of our algorithm is faster than that of SSD,

TABLE 1 Performance comparison of SSD, YOLOv3, YOLOv3-tiny, YOLOv4, YOLOv4-tiny, YOLOv5 and the proposed algorithm in the ship dataset.

Methods	mAP0.5 (%)	mAP@0.5:0.95 (%)	P (%)	R (%)	F1 (%)	Model size (M)	FPS
SSD	95.2	72.1	81.3	85.7	83.4	92.6M	83
YOLOv3	95.9	77.3	95.1	94.8	94.9	117M	54
YOLOv3-tiny	72.6	31.4	67.0	72.4	69.6	16.6M	149
YOLOv4	93.5	77.5	81.2	**96.4**	88.1	488M	26
YOLOv4-tiny	88.9	63.9	51.9	91.5	66.23	45M	98
YOLOv5	95.3	70.9	**95.8**	94.5	95.1	13.61M	131
Ours	**96.6**	**79.2**	**95.8**	94.7	**95.2**	**12.24M**	138

The bolded areas inside the table represent the best performance.

YOLOv3, YOLOv4, YOLOv4-tiny, and YOLOv5. The results show that the detection algorithm of the proposed algorithm achieves optimal results between speed and accuracy. Therefore, the ship detection algorithm of our study is suitable for application to USVs.

Figure 9 shows the Precision–Recall (P–R) curves of YOLOv3, YOLOv4, YOLOv5, and the proposed algorithm. The P–R curves represent the predictions of the test set samples as positive samples under different thresholds, and different precision and recall rates are obtained. The larger the area enclosed by the P–R curve with the coordinate axis, the better the precision and recall of the detection algorithm. After comparison, it can be seen that the area enclosed by the algorithm in this study is larger than that of other object detection algorithms. Hence, the algorithm in this paper is better than the three algorithms of YOLOv3, YOLOv4, and YOLOv5 in terms of detection performance.

4.4 Comparison of actual test results of USV

To test the detection effect of the proposed algorithm in an actual maritime environment, we conducted several maritime experiments in the Yellow Sea near Qingdao to detect and classify ships. Figure 10 shows the detection results of the proposed algorithm and lightweight models YOLOv3-tiny, YOLOv4-tiny, and YOLOv5 on images collected by the

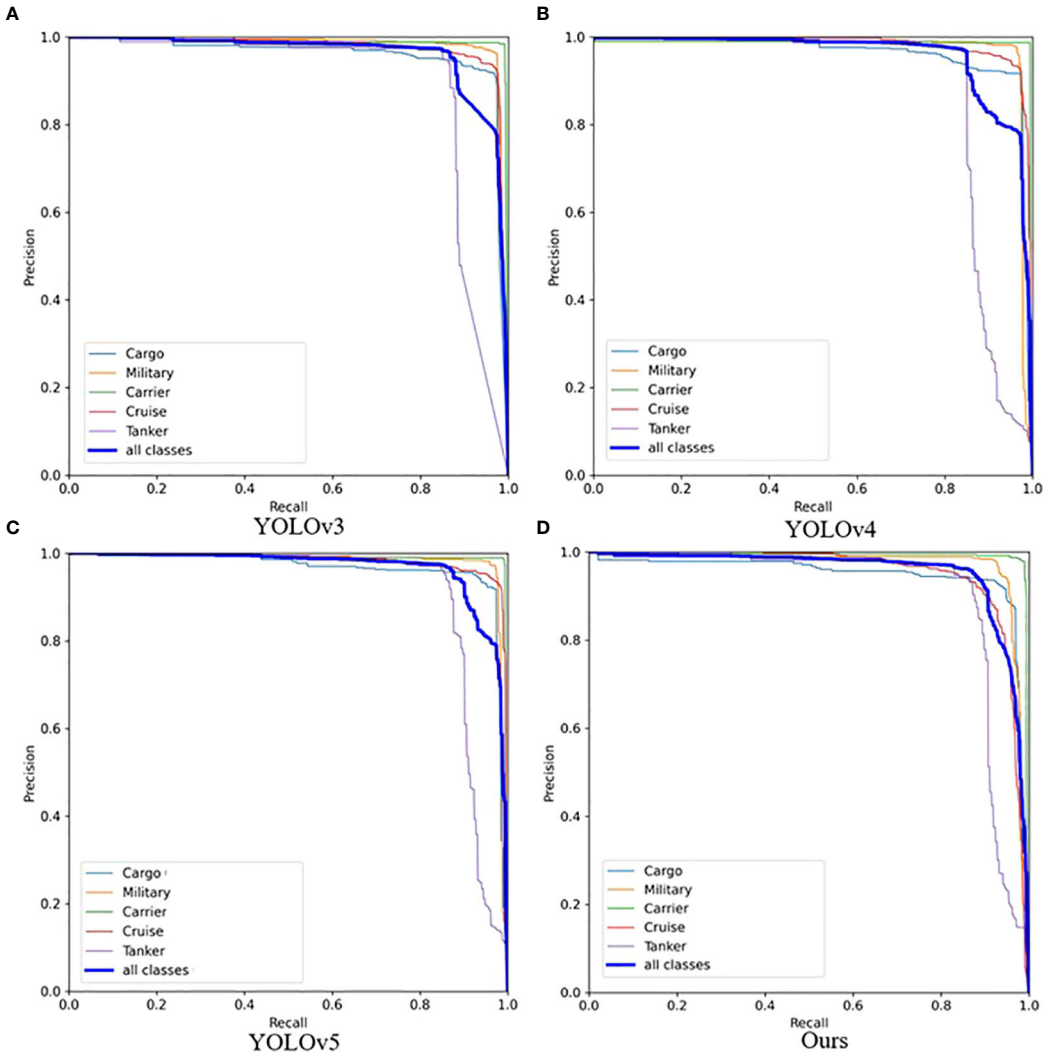

FIGURE 9
Precision-Recall (P-R) curves of different object detection algorithms on the ship dataset (IoU = 0.5). **(A)** is from YOLOv3, **(B)** is from YOLOv4, **(C)** is from YOLOv5 and **(D)** is from our proposed algorithms.

"JiuHang750" USV. The results show that the proposed algorithm has the best detection performance in the actual maritime environment. Each column presents the original image and the detection results of YOLOv3-tiny, YOLOv4-tiny, Yolov5, and the proposed algorithm from left to right. The first row shows a ship clearly. Although YOLOv4-tiny detects the object, the detection box is significantly smaller than the actual position of the ship in the image. In the second row, we show the image of a ship that is far away from the ship and has wake waves. YOLOv4-tiny recognizes the waves as a ship object, and the detection accuracy of the proposed algorithm is significantly higher than that of other detection algorithms. The third row shows the ship image under the swing of the USV. YOLOv3-tiny and YOLOv4-tiny also detect the ship object, but the detection box is inconsistent with the actual position of the ship in the image; additionally, YOLOv5 does not detect the ship object. The fourth row shows the image of the ship under dark clouds; all algorithms detect the ship object, but YOLOv4-tiny splits one ship object into two different objects. Furthermore, the accuracy of the proposed algorithm is significantly higher than that of other detection algorithms. The fifth and sixth rows show the ship image in the case of sea fog. Two images do not detect the ship object of YOLOv3-tiny and YOLOv4-tiny, and the detection accuracy is also low; however, the accuracy rate of the ship object detected by the proposed algorithm is higher.

4.5 Ablation experiments

To further evaluate the effectiveness of the proposed algorithm and each module, ablation experiments were designed, and Table 2 presents the results. Experiment 1 is set as the benchmark, which demonstrates the performance of YOLOv5s without any modification. Then, we replaced the original anchor boxes in experiment 2. In experiment 3, we added the Ghost module to the backbone structure. In experiment 4, we included the attention mechanism in the Neck network structure.

The results show that the mAP increased by 0.11% in experiment 2 after replacing the original anchor boxes. The original Conv operation in the backbone was replaced by Conv stacking with depth-wise Conv in the Ghost module in experiment 3. Compared with the results achieved by YOLOv5s, the mAP increased by 0.14% and the size of the model reduced by 1.45 M. In experiment 4, we integrated the Transformer into the end of the backbone network and FPN structure, and the mAP increased by 0.43%. These results show

FIGURE 10
Detection results of different object detection algorithms in various environments collected by "JiuHang750" USV.

TABLE 2 The results of the ablation experiment.

Experiment	Anchor boxes	Ghost module	Transformer	mAP0.5 (%)	Size (M)	FPS
1				94.80	13.61	131
2	✓			94.91	13.61	131
3		✓		95.06	**12.16**	**140**
4			✓	95.23	13.60	133
5	✓	✓	✓	**96.6**	12.24	138

The bolded areas inside the table represent the best performance.

that the addition of the two modules can improve the detection ability of the algorithm.

5 Conclusions

In this study, an object detection algorithm is improved based on the YOLOv5 model for USVs. First, based on the shape characteristics of ships, the K-means algorithm was used to optimize the initial value of the anchor boxes. Second, the Ghost module was added to the backbone, thus reducing the size of the network and improving detection efficiency. Third, we integrated the Transformer at the end of the backbone and Neck structures in the YOLOv5 network, thereby improving the model's attention to reliable and useful features. Finally, we conducted experiments to verify the accuracy of the proposed algorithm and its effectiveness in real-time detection tasks. In comparison with other deep learning object detection algorithms, the results show that the proposed algorithm achieves a mAP of 96.6%. Our model size is the smallest among all other algorithms used for comparison and only reaches 12.24 M. The detection results in different maritime environments are also significantly better than those of other detection algorithms. Additionally, our algorithm has obtained good detection results in the sea fog environment. Furthermore, the proposed algorithm was applied to the vision system of the "JiuHang750" USV and successfully realized the identification and classification of the surrounding ships of the USV.

Sea images are easily affected by weather and lighting, resulting in unclear objects on images; thus, feature extraction of objects can become difficult. In future research, we can resolve this problem by focusing on the hardware technology for image acquisition, image stabilization, and other aspects. In addition, the dataset used in this study is small in terms of size, and it is necessary to collect more photos of objects on the sea, and especially pictures at different times and light conditions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

JZ conceived, planned, and performed the designs and drafted this paper. YM and PR provided guidance and reviewed this paper. JJ provided the design ideas and edited this paper. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China (grant number 2021YFC3101101) and the National Key Research and Development Program of China (grant number 2017YFC1405203).

Acknowledgments

The authors would like to thank China University of Petroleum (East China) for technical support and all the members of our team for their contribution to the sea experiment of the USV.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bai, Y., Lei, S., and Liu, L. (2021). "The ship object detection based on Sea-Sky-Line," in *2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE)*, IEEE. 456–460. doi: 10.1109/CACRE52464.2021.9501

Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint* arXiv:2004.10934. doi: 10.48550/arXiv.2004.10934

Chen, X., Ling, J., Wang, S., Yang, Y., Luo, L., and Yan, Y. (2021). Ship detection from coastal surveillance videos *via* an ensemble canny-gaussian-morphology framework. *J. Navigation* 74, 1252–1266. doi: 10.1017/S037346332100

Chen, T., Wang, N., Wang, R., Zhao, H., and Zhang, G. (2021). One-stage CNN detector-based benthonic organisms detection with limited training dataset. *Neural Networks* 144, 247–259. doi: 10.1016/j.neunet.2021.08.014

Dzvonkovskaya, A., and Rohling, H. (2010). "Cargo ship RCS estimation based on HF radar measurements," in *11-th International Radar Symposium*. IEEE. 1–4.

Electronic Quality Shipping Information System (2020) *The world merchant fleet in 2020*. Available at: https://www.equasis.org (Accessed May 17, 2000).

Girshick, R. (2015). "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*. IEEE. 1440–1448.

Gupta, V., Gupta, M., and Singla, P. (2021). Ship detection from highly cluttered images using convolutional neural network. *Wireless Pers. Commun.* 121, 287–305. doi: 10.1007/s11277-021-08635-5

Han, K., Wang, Y., Tian, Q., Guo, J., and Xu, C. (2020). *Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE. 1580–1589.

Jain, A. (2021) *Game of deep learning: Ship datasets*. Available at: https://www.kaggle.com/datasets/arpitjain007/game-of-deep-learning-ship-datasets (Accessed June 24, 2021).

Jie, Y., Leonidas, L., Mumtaz, F., and Ali, M. (2021). Ship detection and tracking in inland waterways using improved YOLOv3 and deep SORT. *Symmetry* 13(2), 308. doi: 10.3390/sym130203

Lee, S.-J., Roh, M.-I., Lee, H.-W., Ha, J.-S., and Woo, I.-G. (2018). "Image-based ship detection and classification for unmanned surface vehicle using real-time object detection neural networks," in *The 28th International Ocean and Polar Engineering Conference (OnePetro)*. 726–730.

Li, G., and Qiao, Y. (2021). A ship object detection and tracking algorithm based on graph matching. In *Journal of Physics: Conference Series (IOP Publishing)*, vol. 1873 (1), 012056. doi: 10.1088/1742-6596/1873/1/012056

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "Ssd: Single shot multibox detector," in *European Conference on computer vision* (Springer), 21–37. doi: 10.1007/978-3-319-46448-0_2

Liu, C., and Li, J. (2021). Self-correction ship tracking and counting with variable time window based on YOLOv3. *Complexity* 2021, 1–9. doi: 10.1155/2021/7428927

Liu, T., Pang, B., Zhang, L., Yang, W., and Sun, X. (2021). Sea Surface object detection algorithm based on YOLO v4 fused with reverse depthwise separable convolution (RDSC) for USV. *J. Mar. Sci. Eng.* 9, 753. doi: 10.3390/jmse9070

Liu, Z., Zhou, F., Bai, X., and Yu, X. (2013). Automatic detection of ship object and motion direction in visual images. *Int. J. Electron.* 100, 94–111. doi: 10.1080/00207217.2012.687188

Mittal, S., Srivastava, S., and Jayanth, J. P. (2022). "A survey of deep learning techniques for underwater image classification," in *IEEE Transactions on Neural Networks and Learning Systems*. IEEE. 1–15. doi: 10.1109/TNNLS.2022.3143887

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint* arXiv:1804.02767. doi: 10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28.

Sermi, F., Mugnai, C., Cuccoli, F., and Facheris, L. (2013). "Analysis of the radar coverage provided by a maritime radar network of Co-operative vessels based on real AIS data," in *2013 European Radar Conference* IEEE., vol. 2013. 251–254.

Shi, G., and Suo, J. (2018). "Ship target detection based on visual attention," in *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE. 1–4.

Sun, X., Liu, T., Yu, X., and Pang, B. (2021). Unmanned surface vessel visual object detection under all-weather conditions with optimized feature fusion network in YOLOv4. *J. Intelligent Robotic Syst.* 103, 1–16. doi: 10.1007/s10846-021-01499-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems*. 5998–6008

Vesecky, J. F., Laws, K. E., and Paduan, J. D. (2009). "Using HF surface wave radar and the ship automatic identification system (AIS) to monitor coastal vessels," in *2009 IEEE International Geoscience and Remote Sensing Symposium* IEEE., Vol. 3. III–761-III-764. doi: 10.1109/IGARSS.2009.5417876

Vesecky, J. F., Laws, K. E., and Paduan, J. D. (2010). "A system trade model for the monitoring of coastal vessels using HF surface wave radar and ship automatic identification systems (AIS)," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 3414–3417. doi: 10.1109/IGARSS.2010.5650279

Wang, N., Wang, Y., and Er, M. J. (2022). Review on deep learning techniques for marine object recognition: Architectures and algorithms. *IEEE. Control Eng. Pract.* 118. doi: 10.1016/j.conengprac.2020.104458

Xu, F., and Liu, J. (2016). Ship detection and extraction using visual saliency and histogram of oriented gradient. *Optoelectronics Lett.* 12, 473–477. doi: 10.1007/s11801-016-6179-y

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). Mixup: Beyond empirical risk minimization. *arXiv preprint*. arXiv:1710.09412. doi: 10.48550/arXiv.1710.09412

Zhang, Y., Wu, S., Liu, Z., Yang, Y., Zhu, D., and Chen, Q. (2020). "A real-time detection USV algorithm based on bounding box regression," in *Journal of Physics: Conference Series*, Vol. 1544. 012022 (IOP Publishing). doi: 10.1088/1742-6596/1544/1/012022

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint* arXiv:2010.04159. doi: 10.48550/arXiv.2010.04159

Zou, J., Yuan, W., and Yu, M. (2019). "Maritime object detection of intelligent ship based on faster r-CNN," in *2019 Chinese Automation Congress (CAC)*. IEEE. 4113–4117. doi: 10.1109/CAC48633.2019

Zou, Y., Zhao, L., Qin, S., Pan, M., and Li, Z. (2020). "Ship object detection and identification based on SSD_MobilenetV2," in *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*. IEEE. 1676–1680.