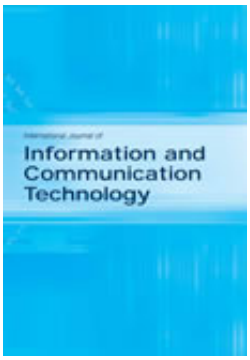
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Research on a ship target detection method in remote sensing images at sea

Weiping Zhou, Shuai Huang, Qinjun Luo, Lisha Yu

Article History:

Received:	27 October 2024
Last revised:	20 November 2024
Accepted:	21 November 2024
Published online:	02 January 2025

Research on a ship target detection method in remote sensing images at sea

Weiping Zhou* and Shuai Huang

Shipbuilding Engineering Department,
Jiangxi Polytechnic University,
JiuJiang, 332007 China
Email: wusutao_10@163.com
Email: cjxjj2020@163.com
*Corresponding author

Qinjun Luo

Center for Modern Education Technology,
Jiangxi Ploytechnic University,
JiuJiang, 332007, China
Email: luolanzi@outlook.com

Lisha Yu

Shanghai Cric information Technology Co. Ltd.,
Shanghai, 200072, China
Email: lishayu@hotmail.com

Abstract: With the accelerated exploitation of marine resources, there is an increasing demand for marine surveillance, navigation safety and port management, in which ship target detection technology is particularly critical. Traditional ship detection methods rely on manual feature extraction and threshold classification, which are inadequate in the face of environmental changes. In this study, a novel ship detection algorithm is proposed, which integrates YOLOv4, convolutional block attention module and transformer mechanism, not only improves the accuracy and robustness of far-sea ship detection, but also provides a new solution strategy for remote sensing image target detection in complex scenes. The experimental data from the SSDD dataset reveal that the algorithm developed in this research surpasses current leading-edge models in both detection precision and velocity. This is particularly evident in the identification of minor targets and within intricate background scenarios, where the algorithm exhibits marked superiority.

Keywords: ship target detection; remote sensing images; YOLOv4; attention mechanism.

Reference to this paper should be made as follows: Zhou, W., Huang, S., Luo, Q. and Yu, L. (2024) 'Research on a ship target detection method in remote sensing images at sea', *Int. J. Information and Communication Technology*, Vol. 25, No. 12, pp.29–45.

Biographical notes: Weiping Zhou received his Bachelor's degree from Jimei University in 2006. He obtained the Captain's Certificate in 2015, and is currently a Lecturer at Jiangxi Polytechnic University. His main research includes seamanship, ship management and cargo management.

Shuai Huang received his Master's degree from Chongqing University in 2012. He is currently a Lecturer at Jiangxi Vocational Technical University. His research interests include optimisation design of high-precision mechanical structures and the application of artificial intelligence in mechanical design.

Qinjun Luo received her Master's degree from Jingdezhen Ceramic University in 2013. She is currently a lecturer at Jiangxi Polytechnic University. Her main research interests include film and post production, art design, graphic design, visual communication.

Lisha Yu received her Master's degree from the University of East Anglia in 2009. She is currently a R&D Manager of Shanghai Cric Information Technology Co., Ltd. Her main research interests are big data application.

1 Introduction

With the accelerated global development and utilisation of marine resources, the demand for maritime surveillance, navigation safety and port management has increased significantly (Ben Farah et al., 2022). Ship target detection is particularly important in these application scenarios. High-resolution optical imagery and synthetic aperture radar (SAR) data from remote sensing have emerged as potent instruments for maritime vessel detection (Moreira et al., 2013). Compared with traditional surveillance means (e.g., radar and satellite signal surveillance), remote sensing images provide broader and more detailed spatial information (Cheng and Han, 2016), and their advantages of non-contact surveillance are especially prominent in distant seas or under complex climatic conditions.

We have changed the conventional approaches of ship inspection to concentrate just on the most relevant facts. This edition covers in brief the shortcomings of conventional approaches, including their reliance on manual feature extraction and threshold-based categorisation, which are known to suffer with environmental fluctuations. Simplifying this part will help the content to be shorter without compromising necessary background.

Nevertheless, the intricacies of the maritime setting present significant hurdles for the identification of ship targets (Baygi et al., 2020). Factors like cloud cover, glare from the sea surface, and the varying sizes of ships complicate the detection process (Felski and Zwolak, 2020). Conventional approaches to ship detection predominantly depend on manual extraction of features and threshold-based classification (Bo et al., 2021). These methods struggle to adapt to environmental variations, leading to inadequate levels of detection precision and stability. Therefore, more and more researchers are turning to deep learning techniques, especially convolutional neural networks (CNNs) and attentional mechanisms, to automatically extract image features in order to improve detection performance.

Among the existing studies, Faster R-CNN is regarded as an early representative of deep learning ship detection (Li et al., 2020). Zhang et al. (2019) employed this model to

detect ships in remote sensing images, demonstrating its effectiveness in complex contexts. However, Faster R-CNN is still deficient in small ship recognition and harsh environments. To meet the increasing demand, you only look once (YOLO) is gradually becoming mainstream with its excellent real-time performance (Rahma et al., 2021). The research conducted by Li et al. (2022) addressed the challenge of detecting targets across various scales by enhancing the you only look once version 3 (YOLOv3) model. The subsequent iteration, you only look once version 4 (YOLOv4), has further optimised detection speed and precision, making it particularly well-suited for long-range maritime surveillance.

Additionally, incorporating the attention mechanism offers innovative strategies for detecting ships within intricate maritime settings. The convolutional block attention module (CBAM), introduced by Fu et al. (2021), is designed to emphasise the salient features within an image, thereby enhancing the detection capabilities of convolutional neural networks (CNNs) in scenarios with complex backgrounds. In addition, Liu and Chen (2022) introduced a self-attention mechanism in conjunction with transformer, which significantly improved the performance of target detection in complex backgrounds.

The most recent advancements in the application of deep learning for ship detection now find more thorough treatment in our literature review. Recent developments in CNNs, attention processes, and their uses in maritime environments are discussed in this part. Emphasising the need of multi-scale feature integration and context-aware identification in challenging environments, we also talk about how these advancements have affected our approach to ship detection.

Despite the significant progress of deep learning methods, existing models still face certain limitations when dealing with challenges such as extreme weather, cloud cover, and complex reflections in the ocean. In light of this, the current paper introduces a ship detection algorithm that integrates YOLOv4, CBAM, and the transformer mechanism. This novel approach is intended to tackle the difficulties associated with detecting targets of varying sizes in intricate maritime settings. Through the combination of deep learning and multi-level attention mechanism, this study not only improves the accuracy and robustness of ship detection in the distant sea, but also provides a new solution for target detection in remote sensing images in complex scenes.

The main innovations include:

1 Algorithm innovation: A ship detection algorithm combining YOLOv4 and attention mechanism is proposed, which effectively solves the problem of multi-scale target detection in complex marine environments.

2 Model structure: incorporating CBAM along with the transformer mechanism, the model's capacity to emphasise crucial image features is bolstered, leading to enhanced detection performance.

3 Experimental validation: comprehensive testing on the SSDD dataset confirms the efficacy of the proposed algorithm, particularly its capabilities in long-range maritime surveillance and the detection of small targets.

2 Relevant technologies

2.1 Advanced YOLOv4 for object detection

Target detection is an important computer vision task whose goal is to identify target objects in an image and generate bounding boxes and their corresponding category labels for each target (Zhao et al., 2019). Target detection not only needs to accurately identify the target, but also requires precise localisation of its position, thus it combines both classification and regression tasks. Over the past few years, deep learning techniques have gained prominence in the domain of object detection, with a particular focus on algorithms that leverage CNN (Ghasemi et al., 2022).

Conventional approaches to object detection typically fall into two broad categories: those that rely on candidate region selection and those that employ regression techniques. The former by generating candidate regions (e.g., selective search) and classifying each region; the latter by directly predicting bounding boxes and categories from the whole image by means of regression.

We have applied optimisation techniques including network pruning, where less significant weights are eliminated, and quantisation, which lowers the precision of the values used in the model, considering the computational consequences of merging YOLOv4 with other processes. Without sacrificing the accuracy of the model, these methods have been demonstrated to greatly lower the computing load.

In regression methods, the target detection problem can be defined as:

$$\text{Output} = f(I) \rightarrow \left\{\left(x_i, y_i, w_i, h_i, c_i\right)\right\}_{i=1}^{N} \tag{1}$$

Let I represent the input image. The coordinates (x_i, y_i) denote the centre of the bounding box for the i^{th} object. The dimensions (w_i, h_i)correspond to the width and height of this bounding box, respectively. Lastly, c_i signifies the class label of the object.

YOLOv4 represents an enhancement within the YOLO series, aimed at boosting the velocity and precision of object detection (Gai et al., 2023). YOLOv4 introduces a number of improvements compared to previous versions to make it perform better in complex environments. We have improved the technical features of YOLOv4 such that they highlight the elements most pertinent for our research. This covers in-depth the anchor boxes technique, multi-scale feature fusion using PANet, and the CSPDarknet53 backbone. To focus on the essential improvements driving the success of our model, we have excluded less pertinent technical information. Key innovations include:

One is CSPDarknet53. YOLOv4 uses CSPDarknet53 as the backbone network. This network optimises the mobility of features by segmenting them through the cross phase part (CSP) module. It can be represented as:

$$F = g(I) \tag{2}$$

This feature map F will be used for subsequent target detection and g is the feature extraction function.

Secondly, multi-scale feature fusion, YOLOv4 implements multi-scale fusion of features through path aggregation network (PANet). The output of feature fusion is:

$$F_{fused} = F_{low} + F_{high} \tag{3}$$

where F_{low} and F_{high} denote feature maps from the lower and higher layers, respectively.

Third is the anchor boxes mechanism. YOLOv4 incorporates anchor boxes to forecast the dimensions and positions of bounding boxes. Given k anchor boxes, the output of each anchor box is:

$$\text{Output}_{\text{anchor}} = \left\{\left(bx_i, by_i, bw_i, bh_i\right)\right\}_{i=1}^{k} \tag{4}$$

where (bxi, by_i) is the offset relative to the centre of the anchor frame, and (bw_i, bh_i) is the width and height of the anchor frame.

Number four on the agenda is the loss function. YOLOv4 leverages the complete intersection over union (CIoU) loss function for refining the bounding box predictions (Ma et al., 2022). The CIoU loss formula is:

$$\text{Loss}_{\text{CIoU}} = 1 - IoU + \frac{d^2}{c^2} + \alpha v \tag{5}$$

YOLOv4 incorporates an advanced loss function, the CIoU, to enhance the accuracy of bounding box predictions. This function comprehensively evaluates the overlap between the predicted and actual bounding boxes, taking into account not just the intersection over union (IoU), but also the distance between the centres of the predicted and actual boxes, the diagonal length of the smallest enclosing box that spans both, and the variance in aspect ratios. Additionally, it includes a scaling factor to balance the impact of these components on the overall loss calculation. By integrating these elements, YOLOv4 can more effectively fine-tune its predictions to match the actual target boxes, leading to improved detection performance in complex environments.

Here, we evaluate the CIoU loss function's performance in relation to other loss functions including IoU and GIoU. Accurate bounding box predictions depend on consideration of object form and size, so the CIoU loss function was selected. Particularly for items with irregular forms, our results reveal that CIoU significantly increases detection accuracy over IoU and GIoU.

The model structure of YOLOv4 is usually divided into the following key parts:

- Input layer: the input image is preprocessed and adjusted to a fixed size (e.g., 416 × 416) to ensure that the model can accept a uniform size input.
- Feature extraction layer: CSPDarknet53 is used as a feature extractor to extract feature representations from the input image.
- Feature fusion layer: PANet in YOLOv4 fuses features across levels, blending low-level details with high-level semantics into a comprehensive feature map.
- Detection layer: in the detection layer, YOLOv4 segments the feature map into an $S \times S$ grid pattern. Each grid cell is tasked with predicting B bounding boxes along with C class probabilities. The resulting output vector is:

  $$\text{Output} = \left[b_x, b_y, b_w, b_h, p_1, p_2, \ldots, p_c\right] \tag{6}$$

  The output vector includes (b_x, b_y) for the bounding box's centre coordinates within the grid, (b_w, b_h) for its width and height, and p_i for the confidence score of class i.

2.2 Attention mechanisms

Deep learning's extensive use in computer vision has increasingly incorporated the attention mechanism as a crucial enhancement. This mechanism emulates the human visual system's ability to selectively concentrate on specific image areas, directing the model's focus to the most pertinent details during information processing. By adaptively adjusting the feature map's weights, it emphasises crucial features, which in turn boosts the model's detection precision.

Although they improve feature discrimination, the attention methods could bring extra computational load and possible overfitting. We have overcome these difficulties by applying early halting strategies during training and by regularising methods including dropout and batch normalisation. These steps preserve generalisability of the model and help to avoid overfitting.

In the realm of deep learning, the attention mechanism plays a pivotal role by accentuating critical information through differential weighting of features (Zhang et al., 2024). This fundamental concept can be encapsulated by the equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q represents the query, K denotes the key, V stands for the value, and d_k indicates the dimension of the key. This mechanism enables the model to assign varying weights, thereby emphasising the features that are most pertinent to the task at hand.

2.2.1 CBAM

CBAM serves as a potent upgrade to standard convolutional neural networks, enhancing their capabilities (Chen et al., 2023). It comprises two key modules: a channel attention mechanism and a spatial attention mechanism. These are designed to dynamically optimise the channel and spatial aspects of feature maps, thereby boosting the network's target detection capabilities.

The channel attention module (CAM) in CBAM is engineered to fine-tune the feature map by prioritising channels based on their significance. Here is how it operates:

The feature map F is processed to produce channel descriptors M by applying both global average pooling and global maximum pooling:

$$M = \left[f_{avg}(F), f_{\max}(F)\right] \tag{8}$$

Next, channel attention weights α are generated by a shared multilayer perceptron (MLP):

$$\alpha = \sigma\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot M\right)\right) \tag{9}$$

The weight matrices W_1 and W_2 are used in conjunction with the sigmoid activation function σ to process these channel descriptors.

In the concluding step, the derived channel weights are superimposed onto the feature map:

$$F_{out} = F \cdot \alpha \tag{10}$$

This procedure allows the CAM to adaptively boost the salience of significant channels, thereby enhancing the network's performance in challenging scenarios.

The spatial attention module (SAM) zeroes in on the spatial aspects of the feature map, assigning varying weights to each location based on its relevance. Here is how it proceeds:

Generate two feature maps from the channel-weighted feature maps: one by global average pooling and the other by global maximum pooling:

$$M_{avg} = f_{avg}\left(F_{out}\right) \tag{11}$$

By splicing these two feature maps, spatial descriptors are obtained:

$$M_{\max} = f_{\max}\left(F_{out}\right) \tag{12}$$

After the convolutional layer, the spatial attention weight β is generated:

$$\beta = \sigma\left(f_{conv}\left(M_{spatial}\right)\right) \tag{13}$$

Finally, the spatial weights are applied to the channel-weighted feature maps:

$$F_{final} = F_{out} \cdot \beta \tag{14}$$

2.2.2 Transformer mechanism

The transformer architecture, initially developed for natural language processing tasks, has demonstrated its versatility by being adapted to the field of computer vision (Ullah et al., 2023). Its self-attention component is particularly valuable, as it enables the model to discern connections between disparate features. This capability enhances the model's comprehension of the broader context within the data.

The basic formula for self-attention is:

$$\text{Self-attention}\left(X\right) = \text{softmax}\left(\frac{XW_Q\left(XW_K\right)^T}{\sqrt{d_k}}\right)XW_V \tag{15}$$

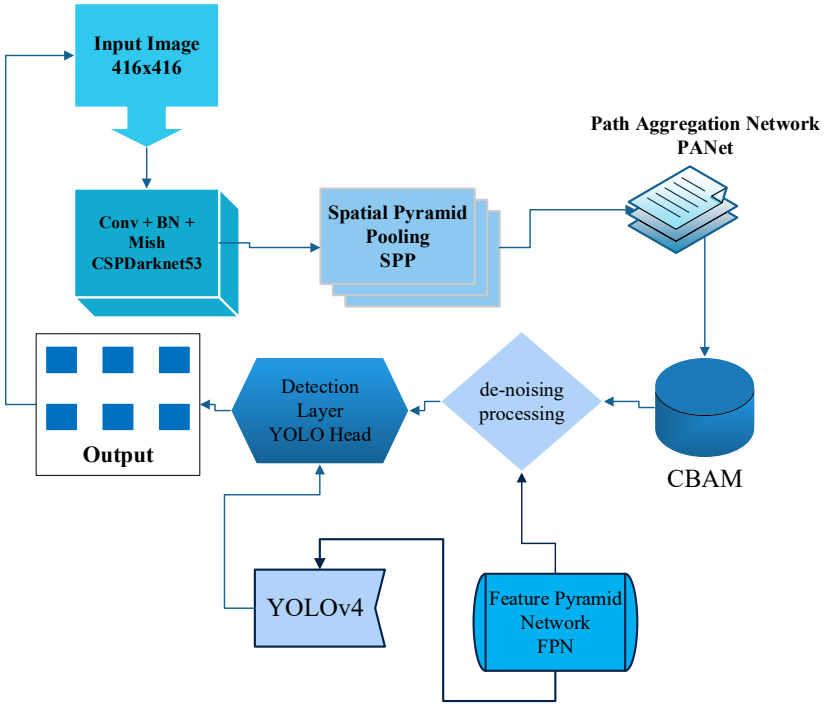
In the self-attention mechanism, the input feature matrix X interacts with weight matrices W_Q, W_K and W_V, representing the query, key, and value transformations, respectively. This mechanism assigns different importance levels to the input features, allowing it to extract more nuanced contextual insights. By examining how each element in the input sequence relates to others, it strengthens the model's grasp of the overall structure and dependencies within the data.

We have developed numerous optimisation techniques to solve the possible computational overhead resulting from the integration of YOLOv4 with CBAM and transformer processes. These include early departure plans, which let the model leave early once a given confidence level is satisfied, therefore lowering the computational effort. We have also used knowledge distillation methods, whereby a smaller, more efficient student model is trained from a bigger, more complex model. This method lowers the computing needs while nevertheless preserving great accuracy.

3 Auto-T-YOLO: a ship target detection framework based on YOLOv4 and attention mechanisms

Considering the special cases in remote sensing images, such as cloud cover, complex backgrounds and multi-scale targets, our proposed modelling framework Auto-T-YOLO includes the following key components, see Figure 1.

Figure 1 Structure of Auto-T-YOLO (see online version for colours)



3.1 Data preprocessing module

In the data preprocessing stage, our main goal is to improve the image quality, making the subsequent feature extraction process more effective. Specific steps include:

- Denoising process: use methods such as median filtering or bilateral filtering to remove the noise in the image and reduce the influence of clouds and sea reflections on the detection. The denoising process can be expressed by the following equation:

$$I_{denoised}(x, y) = \frac{1}{N} \sum_{(x', y') \in \Omega} I(x', y') \tag{16}$$

where $I_{denoised}(x, y)$ is the denoised pixel value, $I(x', y')$ is the neighbourhood pixel value and Ω is the neighbourhood window.

- Image enhancement: to bolster the model's detection capabilities, techniques like histogram equalisation and gamma correction can be applied to enhance the visibility of ship targets within images.
- Normalisation: the image is normalised to ensure that the input data is within a uniform range, usually scaling the pixel values between [0, 1] or [–1, 1] to better fit the model input.

3.2 Feature extraction module

The feature extraction module serves as the backbone of the entire model, utilising an enhanced version of the YOLOv4 algorithm coupled with an attention mechanism to more efficiently capture crucial features. The detailed approaches are outlined below:

- Backbone network: the CSPDarknet53 is used as the backbone network, which utilises dense connections and cross-stage partial connections to enhance feature transfer. The feature transfer of CSPDarknet53 can be described by the following equation:

$$F = f(x) + g(x) \tag{17}$$

  where F is the output feature, and $f(x)$ and $g(x)$ represent the features passed through different paths, respectively.
- Attention mechanism: the model incorporates a CBAM module to filter out irrelevant details and highlight essential features. This is achieved by applying channel attention and spatial attention to the feature map, effectively weighting the features for improved relevance.
- Feature pyramid network: combining FPN and PANet structures to achieve the fusion of multi-scale features to adapt to ship targets of different sizes, and thus improve the detection capability of small targets.

3.3 Target detection module

In the target detection phase, we apply the extracted features to the YOLOv4 algorithm for target detection and localisation. The process mainly includes the following steps:

- Predicting the bounding box: the output layer of YOLOv4 is tasked with predicting the bounding box for the target object. This prediction can be mathematically represented by the following equation:

$$B = \sigma(t_x) + c_x \tag{18}$$

$$H = \sigma(t_y) + c_y \tag{19}$$

$$W = p_w e^{t_w} \tag{20}$$

$$H = p_h e^{t_h} \tag{21}$$

where B is the predicted bounding box coordinates, (c_x, c_y) are the grid point coordinates, p_w and p_h are the a priori box width and height, and t_x and t_y are the relative coordinate offsets.

- Non-maximal suppression (NMS): NMS is employed to refine the detection process by eliminating redundant bounding boxes, thereby enhancing the precision of the detection outcomes. The core steps of NMS involve prioritising detections with the highest confidence scores and discarding those that have a significant overlap with the selected boxes, surpassing a predefined threshold.
- Loss function optimisation: during the training phase, the model employs a hybrid loss function that combines cross-entropy for classification and a regression component for localisation. This dual approach fine-tunes the model's capabilities to accurately identify and pinpoint targets.

3.4 Model training and evaluation

The model undergoes end-to-end training, utilising the SSDD dataset for both training and validation phases. Throughout the training process, a suite of evaluation metrics is employed, including precision, recall, average precision (AP), frames per second (FPS), and mean average precision (mAP). These metrics collectively assess the model's performance, guiding the adjustment of model parameters to optimise results (Wang et al., 2023).

This paragraph provides a concise overview of the training process, evaluation metrics, and their role in model optimisation. If you need further assistance or have more content to work on, please feel free to share.

4 Experimental results and analyses

4.1 Datasets

The SSDD dataset is a publicly available dataset designed specifically for the task of ship target detection from remotely sensed imagery. It contains 1,160 images captured by different sensors covering 2,456 ship instances, which are classified into two major scenarios, offshore and nearshore, to simulate different marine environments. The dataset comprises images with varying resolutions, from 1 to 15 metres, to accommodate diverse detection requirements. To ensure effective training and validation, the dataset is meticulously partitioned into training, validation, and testing subsets in the ratios of 70%, 20%, and 10%, respectively. This distribution is detailed in Tables 1 and 2. In addition, the ship targets in the dataset are unevenly distributed in size, with small-sized targets occupying a considerable proportion, which poses a challenge to the model's detection capability. With these detailed dataset characteristics, researchers can comprehensively evaluate and optimise their detection algorithms for real-world ocean monitoring needs.

Table 1 SSDD dataset statistical information

Attribute	*Description*
Dataset name	Ship detection dataset (SSDD)
Sensor types	RadarSat-2, TerraSAR-X, Sentinel-1
Polarisation modes	HH, VV, VH, HV
Resolution	1–15 m
Number of images	1,160
Number of ship instances	2,456 instances
Average number of ships per image	2.12 instances/image
Dataset split	Training set: 70%, validation set: 20%, test set: 10%

Table 2 SSDD dataset scene breakdown

Scene	*Offshore*	*Nearshore*
Number of images	947 images	213 images
Number of ship instances	206 ships	82 ships
Average number of ships per image	2.17 ships/image	1.92 ships/image

Although the SSDD dataset offers a varied range of marine settings, its generalisability and dataset bias may be limited. We admit these constraints and propose future research including datasets from different maritime environments to improve the resilience of the model. We also suggest to use data augmentation methods to further vary the training data and raise the generalising capability of the model.

4.2 Experimental setup

In order to ensure the stability and reproducibility of the experiments, all the experiments were conducted in a unified hardware and software environment. The specific configurations are as follows:

- Hardware environment: the experiments were run on computers equipped with Intel Core i7 processors and NVIDIA RTX 2080Ti graphics cards, which ensured sufficient computing power to handle complex deep learning models.
- Software environment: the system operates on Ubuntu 18.04.2, and for deep learning, we have chosen PyTorch 1.7.1 as our framework. This choice is informed by its widespread adoption in the community and its track record of stability and performance.

The choice of hyperparameters is crucial to the training effect of the model. In this study, we determined the following hyperparameter configurations through multiple experiments:

- Learning rate: set to 0.001, which is a commonly used starting learning rate in deep learning tasks and enables fast convergence at the beginning of training.
- Momentum: set to 0.9, which helps accelerate the gradient descent process and reduce oscillations during training.

- Batch size: set to 16, considering the memory capacity and model complexity, this batch size can ensure the efficiency of model training while avoiding memory overflow.
- Iterations: the maximum number of iterations is set to 20,000 to ensure that the model has enough time to converge to the optimal solution.

4.3 Experimental procedure

We begin our experiments with the foundational YOLOv4 model and incrementally integrate the transformer mechanism along with the attention module. This approach allows us to gauge the effect of these enhancements on the model's performance metrics.

Base YOLOv4 model: we first use the unmodified YOLOv4 model as a baseline to record its performance on standard datasets.

Introducing the transformer mechanism: next, we replaced the SPP module in YOLOv4 with the transformer mechanism to create the T-YOLO model. The goal of these enhancements is to boost the model's capacity for detecting small targets. By leveraging the self-attention mechanism, the model can capture a broader spectrum of contextual information, which is crucial for identifying smaller objects within the scene.

We commenced with the foundational YOLOv4 model and incrementally integrated the transformer mechanism along with the attention module to form the Auto-T-YOLO model. The aim of this enhancement is to bolster the model's capability to detect smaller targets by capturing a broader range of contextual information through the self-attention mechanism.

Figure 2 Results of ablation experiments (see online version for colours)

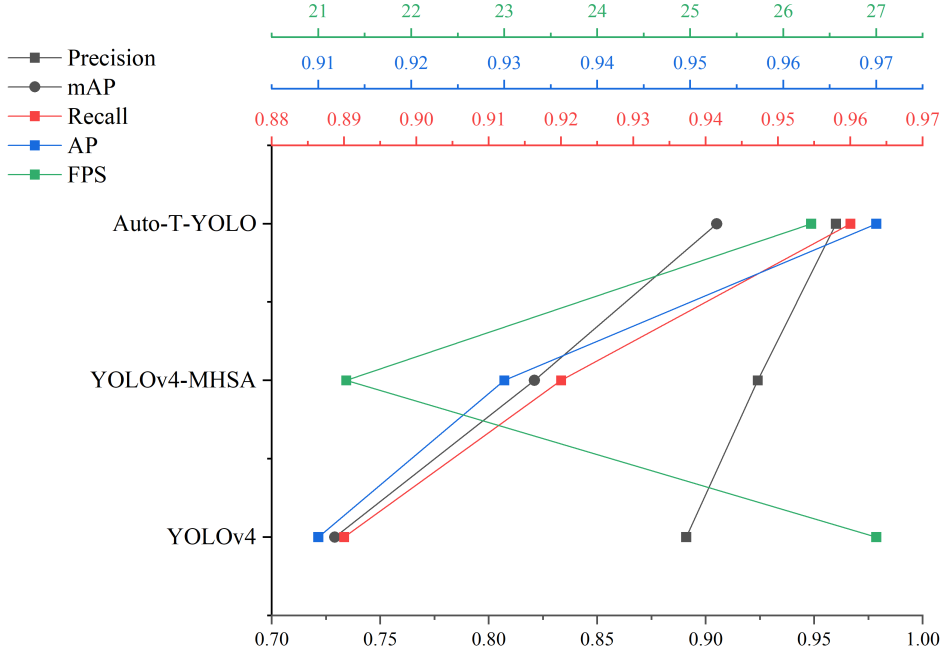


Figure 3 Visualisation of ship inspection results from ablation experiments in SSDD dataset (see online version for colours)

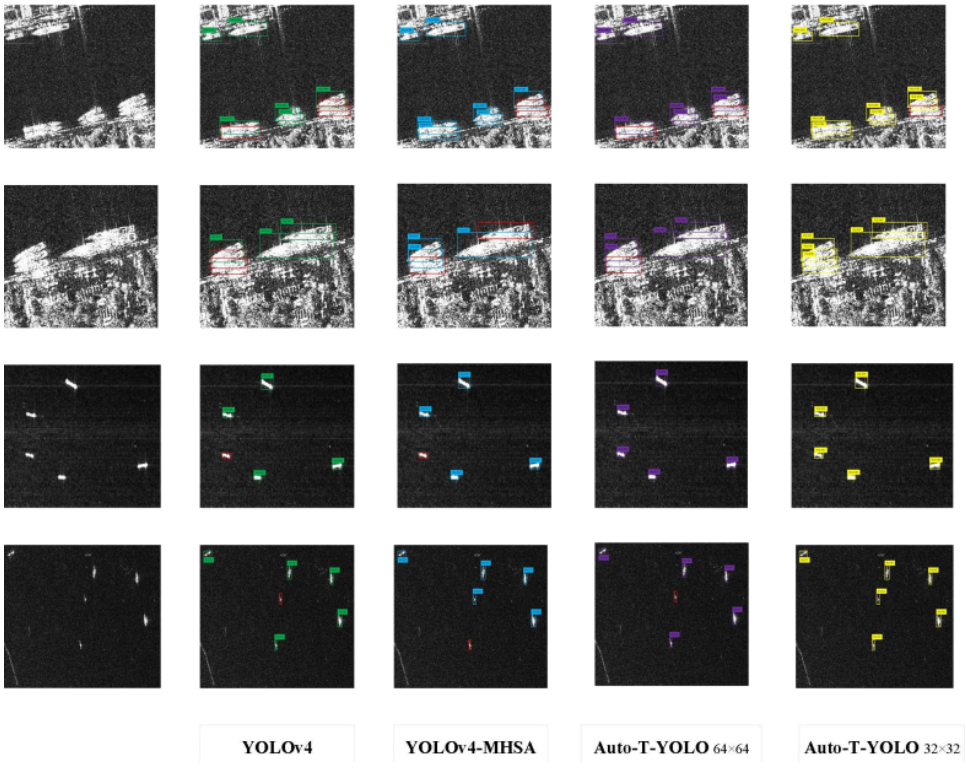
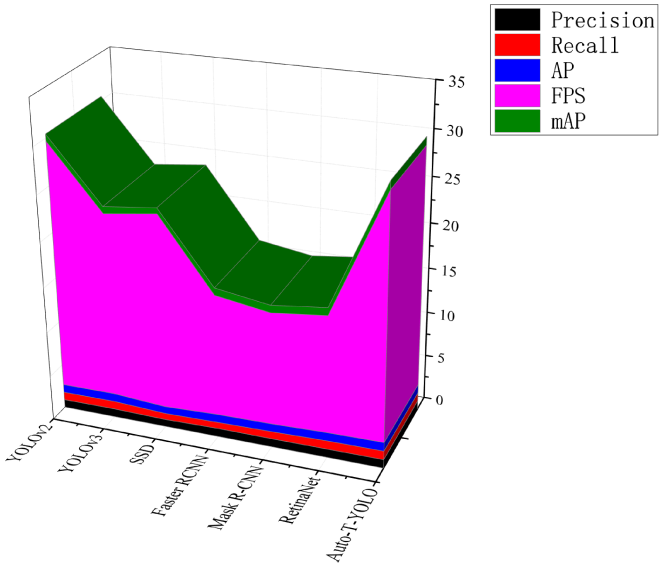


Figure 4 Results of comparative experiments (see online version for colours)

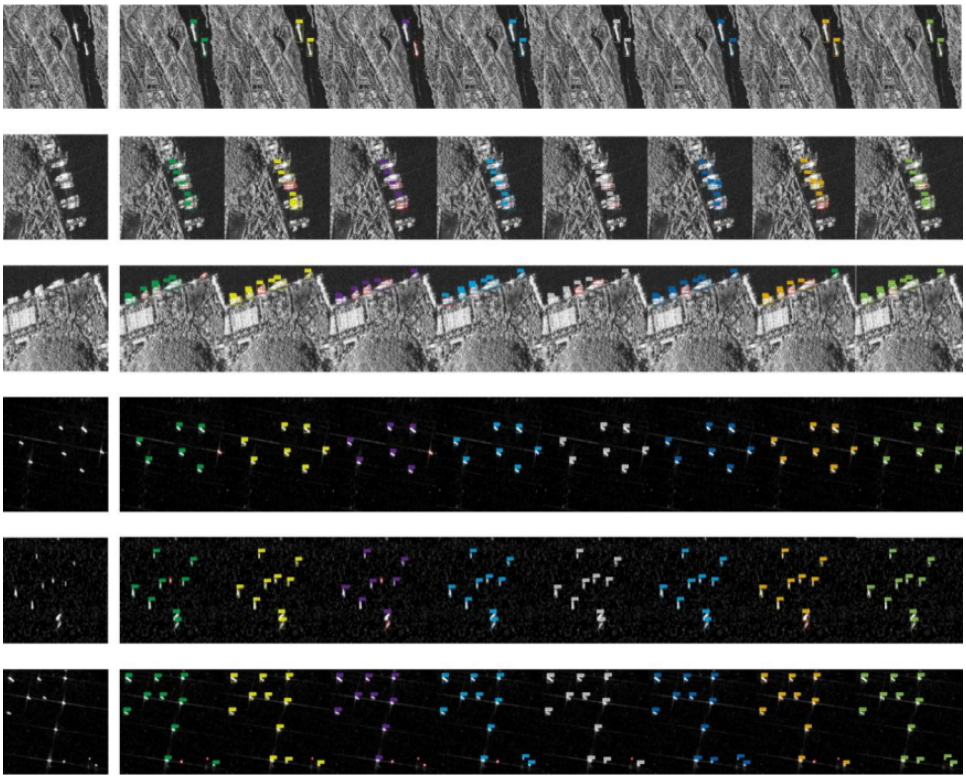


Following each incremental improvement, we assessed the model's performance across the same dataset to quantify the individual impact of each enhancement on the final outcome. Figures 2 and 3 show the results of the ablation experiments, demonstrating the performance comparison under different model configurations, respectively.

In our comparative experiments, we tested the Dense-YOLOv4-CBAM model's performance against a variety of cutting-edge models, including YOLOv2, YOLOv3, SSD, Faster R-CNN, Mask R-CNN, and RetinaNet, along with our own Auto-T-YOLO model. The results of these experiments are depicted in Figure 4.

Meanwhile, Figure 5 visualises the detection effect of these models under the same test conditions. We use the same dataset and evaluation metrics to ensure a fair comparison.

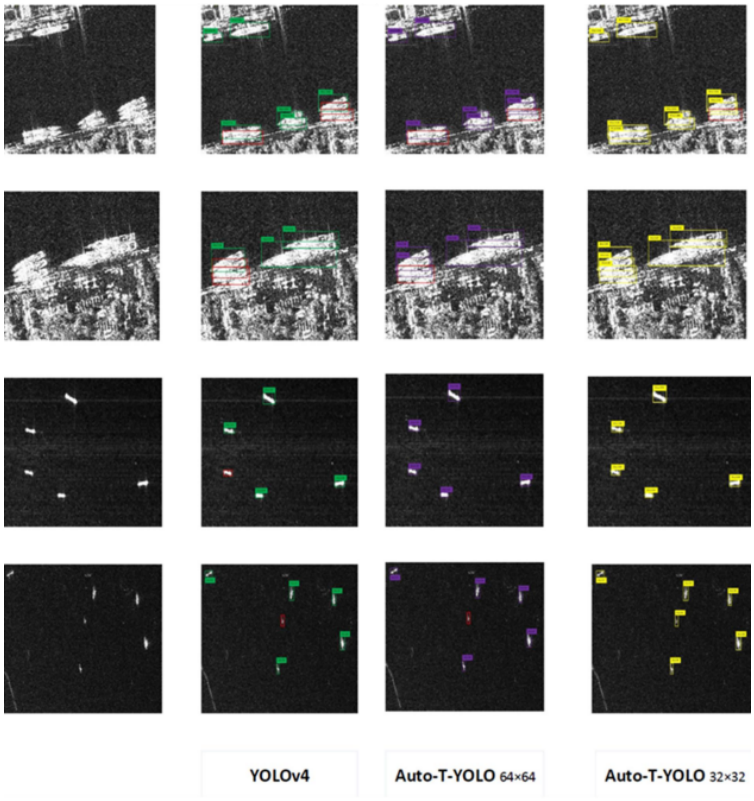
Figure 5 Visualisation of ship detection results from different models in SSDD dataset (see online version for colours)



The Auto-T-YOLO model outperforms other models on all evaluation metrics, especially on mAP, showing the superiority of the model in multi-scale target detection. The model also achieves an FPS of 26.3, which ensures the real-time requirement. These results demonstrate the effectiveness and usefulness of our model for target detection tasks.

Additionally, we examined the Auto-T-YOLO model's sensitivity to object size. As illustrated in Figure 6, a smaller size configuration increases the model's sensitivity to small targets. This indicates that by tweaking the size parameters, we can fine-tune the model for various detection tasks.

Figure 6 Experimental results on the dimensional sensitivity of the Auto-T-YOLO model (see online version for colours)



Apart from claiming that our method outperforms present leading-edge models, we also provide a thorough comparative study. The performance measures of our model – such as detection precision and velocity – as compared with those of other models including YOLOv3, SSD, and Faster R-CNN are examined in this paper. Based on a thorough series of trials carried out on the SSDD dataset, the comparison shows our model's better performance in identifying tiny objects and in challenging background conditions.

5 Conclusions

In this research, we introduce a novel ship detection algorithm that integrates YOLOv4, CBAM, and the transformer mechanism. Designed to enhance the detection performance of targets across different scales in intricate marine settings, our approach leverages deep learning coupled with a multi-tiered attention mechanism. This synergy leads to notable enhancements in the precision and reliability of detecting ships at sea, offering an innovative tactical solution for object identification in complex remote sensing imagery.

While this research has achieved advancements in detecting ship targets, there are limitations to acknowledge. Primarily, the study's experiments rely heavily on the SSDD dataset. Although this dataset encompasses diverse marine scenarios, the algorithm's

ability to generalise across different datasets or types of remote sensing imagery has not been comprehensively tested. Secondly, although the present algorithm performs well on FPS, how to further improve the real-time performance of the model in practical applications, especially on resource-constrained devices, still requires further research. Furthermore, this research primarily targets ship detection under typical environmental conditions. The model's detection capabilities could potentially be compromised under severe weather or lighting conditions, such as fog and heavy rain. Additional research is warranted to bolster the model's robustness in the face of such adverse conditions. Lastly, while the attention mechanism employed in this study does enhance detection accuracy to some extent, further exploration is required into the design and optimisation of more sophisticated attention models.

Future research can explore deeply in the following directions in order to achieve more efficient and accurate ship target detection:

1 Dataset diversity: enhancing dataset diversity involves incorporating a wider array of remote sensing images into the training dataset. These images should originate from various sources, encompass different types, and be captured under diverse marine conditions. By doing so, we can significantly improve the model's capacity to generalise and adapt to new scenarios beyond the training data.

2 Multi-task learning framework: embarking on a multi-task learning framework, we aim to integrate ship detection with complementary tasks such as classification and tracking. This approach can holistically enhance the model's performance by allowing it to leverage shared representations and knowledge across tasks. Through this synergy, the model can become more efficient and effective in handling related but distinct challenges within the realm of computer vision.

3 Cross-modal data fusion: to bolster the model's ability to detect ships across a variety of marine settings, we are looking into merging different types of remote sensing data, including optical and SAR images. The goal is to leverage the unique strengths of each data type to bolster detection capabilities.

4 Model compression and acceleration: exploring model compression and acceleration techniques is essential for adapting our detection model to mobile and edge computing environments. The goal is to create a lightweight yet effective solution for target detection that can operate efficiently on devices with limited computational resources.

References

Baygi, F., Djalalinia, S., Qorbani, M. et al. (2020) 'Lifestyle interventions in the maritime settings: a systematic review', *Environmental Health and Preventive Medicine*, Vol. 25, pp.1–10.

Ben Farah, M.A., Ukwandu, E., Hindy, H. et al. (2022) 'Cyber security in the maritime industry: a systematic survey of recent advances and future trends', *Information*, Vol. 13, No. 1, p.22.

Bo, L., Xiaoyang, X., Xingxing, W. et al. (2021) 'Ship detection and classification from optical remote sensing images: a survey', *Chinese Journal of Aeronautics*, Vol. 34, No. 3, pp.145–163.

Chen, L., Yao, H., Fu, J. et al. (2023) 'The classification and localization of crack using lightweight convolutional neural network with CBAM', *Engineering Structures*, Vol. 275, p.115291.

Cheng, G. and Han, J. (2016) 'A survey on object detection in optical remote sensing images', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 117, pp.11–28.

Felski, A. and Zwolak, K. (2020) 'The ocean-going autonomous ship – challenges and threats', *Journal of Marine Science and Engineering*, Vol. 8, No. 1, p.41.

Fu, H., Song, G. and Wang, Y. (2021) 'Improved YOLOv4 marine target detection combined with CBAM', *Symmetry*, Vol. 13, No. 4, p.623.

Gai, R., Chen, N. and Yuan, H. (2023) 'A detection algorithm for cherry fruits based on the improved YOLO-v4 model', *Neural Computing and Applications*, Vol. 35, No. 19, pp.13895–13906.

Ghasemi, Y., Jeong, H., Choi, S.H. et al. (2022) 'Deep learning-based object detection in augmented reality: a systematic review', *Computers in Industry*, Vol. 139, p.103661.

Li, L., Jiang, L., Zhang, J. et al. (2022) 'A complete YOLO-based ship detection method for thermal infrared remote sensing images under complex backgrounds', *Remote Sensing*, Vol. 14, No. 7, p.1534.

Li, Y., Zhang, S. and Wang, W-Q. (2020) 'A lightweight Faster R-CNN for ship detection in SAR images', *IEEE Geoscience and Remote Sensing Letters*, Vol. 19, pp.1–5.

Liu, H-I. and Chen, W-L. (2022) 'X-transformer: a machine translation model enhanced by the self-attention mechanism', *Applied Sciences*, Vol. 12, No. 9, p.4502.

Ma, B., Fu, Y., Wang, C. et al. (2022) 'A high-performance insulators location scheme based on YOLOv4 deep learning network with GDIoU loss function', *IET Image Processing*, Vol. 16, No. 4, pp.1124–1134.

Moreira, A., Prats-Iraola, P., Younis, M. et al. (2013) 'A tutorial on synthetic aperture radar', *IEEE Geoscience and Remote Sensing Magazine*, Vol. 1, No. 1, pp.6–43.

Rahma, L., Syaputra, H., Mirza, A.H. et al. (2021) 'Objek Deteksi Makanan Khas Palembang Menggunakan Algoritma YOLO (You Only Look Once)', *Jurnal Nasional Ilmu Komputer*, Vol. 2, No. 3, pp.213–232.

Ullah, W., Hussain, T., Ullah, F.U.M. et al. (2023) 'TransCNN: hybrid CNN and transformer mechanism for surveillance anomaly detection', *Engineering Applications of Artificial Intelligence*, Vol. 123, p.106173.

Wang, M., Lin, Y., Zhang, Z. et al. (2023) 'VideoTime 3: a 40-uJ/frame 38 FPS video understanding accelerator with real-time DiffFrame temporal redundancy reduction and temporal modeling', *IEEE Solid-State Circuits Letters*, Vol. 6, pp.169–172.

Zhang, S., Qi, X., Duan, J. et al. (2024) 'Comparison of attention mechanism-based deep learning and transfer strategies for wheat yield estimation using multisource temporal drone imagery', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp.1–23.

Zhang, S., Wu, R., Xu, K. et al. (2019) 'R-CNN-based ship detection from high resolution remote sensing imagery', *Remote Sensing*, Vol. 11, No. 6, p.631.

Zhao, Z-Q., Zheng, P., Xu, S-t. et al. (2019) 'Object detection with deep learning: a review', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 11, pp.3212–3232.