technologies

Article

Exploring Attention Placement in YOLOv5 for Ship Detection in Infrared Maritime Scenes

Ruian Zhu [1,†], Junchao Zhang [2,†], Degui Yang [2], Dongbo Zhao [3], Jiashu Chen [4] and Zhengliang Zhu [1,*]

[1] School of Ocean Information Engineering, Fujian Provincial Key Laboratory of Oceanic Information Perception and Intelligent Processing, Jimei University, Xiamen 361021, China ; pianzhizra@163.com
[2] School of Automation, Center South University, Changsha 410086, China; junchaozhang@csu.edu.cn (J.Z.); degui.yang@csu.edu.cn (D.Y.)
[3] Fujian Provincial Key Laboratory of Coast and Island Management Technology Study, Fujian Institute of Oceanography, Xiamen 361013, China; zdb@fjio.net
[4] Department of Mechanical Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China; jiashu.chen@polyu.edu.hk
* Correspondence: zhengliangzhu2024@126.com
† These authors contributed equally to this work.

Abstract

With the rapid expansion of global maritime transportation, infrared ship detection has become increasingly critical for ensuring navigational safety, enhancing maritime monitoring, and supporting environmental protection. To address the limitations of conventional methods in handling small-scale targets and complex background interference, in this paper, we propose an improved approach by embedding the convolutional block attention module (CBAM) into different components of the YOLOv5 architecture. Specifically, three enhanced models are constructed: the YOLOv5n-H (CBAM embedded in the head), the YOLOv5n-N (CBAM embedded in the neck), and the YOLOv5n-HN (CBAM embedded in both the neck and head). The comprehensive experiments are conducted on a publicly available infrared ship dataset to evaluate the impact of attention placement on detection performance. The results demonstrate that the YOLOv5n-HN achieves the best overall performance, attaining the mAP@0.5 of 86.83%, significantly improving the detection of medium- and large-scale maritime targets. The YOLOv5n-N exhibits superior performance for small-scale target detection. Furthermore, the incorporation of the attention mechanism substantially enhances the model's robustness against background clutter and its discriminative capacity. This work offers practical guidance for the development of lightweight and robust infrared ship detection models.

Keywords: YOLOv5; attention mechanism; ship detection; small object detection

check for updates

Academic Editor: Pedro Antonio Gutiérrez

Received: 24 July 2025
Revised: 21 August 2025
Accepted: 27 August 2025
Published: 1 September 2025

Citation: Zhu, R.; Zhang, J.; Yang, D.; Zhao, D.; Chen, J.; Zhu, Z. Exploring Attention Placement in YOLOv5 for Ship Detection in Infrared Maritime Scenes . *Technologies* **2025**, *13*, 391. https://doi.org/10.3390/technologies13090391

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid expansion of global maritime transportation and the widespread deployment of unmanned surface vehicles (USVs) and intelligent maritime systems, the number and diversity of vessels operating at sea have grown substantially [1]. This surge has introduced new challenges to maritime traffic management and safety monitoring. Against this backdrop, achieving efficient and accurate ship detection has become a critical technological imperative for ensuring navigational safety and enhancing management effectiveness—particularly in complex littoral environments. Currently, the primary technological approaches for maritime ship detection can be broadly categorized into four

classes: satellite remote sensing, navigational radar, visible-light imaging, and infrared vision [1]. While satellite imagery offers wide-area coverage, large-scale data acquisition, and rapid transmission, its limited spatial resolution and transmission latency impede real-time, high-precision detection, restricting its effectiveness in fine-grained maritime target recognition [2,3]. Navigational radar provides robust all-weather sensing capabilities but typically yields sparse point or strip-based outputs, lacking the resolution needed to determine target dimensions or categories accurately [4,5]. Visible-light imaging systems, known for their high spatial resolution, are capable of capturing detailed visual features of ships and have become a prominent modality for object detection [6,7]. Nevertheless, their performance is highly susceptible to environmental interference such as sea surface reflections, fog, and low illumination, leading to significant degradation under nighttime or adverse weather conditions. In contrast, infrared imaging—although generally inferior to visible-light systems in spatial resolution—operates on fundamentally different principles. By capturing thermal radiation emitted from objects, infrared systems enable passive sensing based on temperature differentials. This allows effective detection of heat-emitting targets, such as ship engines or human bodies [8]. Therefore, infrared vision exhibits strong robustness under low-light or adverse environmental conditions, making it particularly well-suited for ship detection tasks in complex coastal scenarios and offering considerable potential for practical deployment.

Traditional infrared ship detection methods are primarily based on modeling image characteristics, reflecting classical object detection paradigms in which the goal is to achieve foreground–background separation through handcrafted modeling techniques. These approaches are generally categorized into two types: spatial-domain filtering methods and transform-domain filtering methods [9,10]. Spatial-domain techniques, such as median and mean filtering, operate by smoothing local pixel neighborhoods to suppress background noise and enhance target saliency. In contrast, transform-domain methods utilize frequency or multi-scale analysis to amplify target signals while preserving edge information, thereby improving the signal-to-noise ratio (SNR) and enhancing detection robustness. In practical implementations, Mo et al. proposed a nighttime infrared ship detection method that integrates local gray-level dynamic range saliency mapping with dual-channel image separation, effectively enhancing target contrast while suppressing background clutter, which enables precise extraction of small-scale maritime targets [11]. Li et al. developed an approach that fuses temporal fluctuation features with spatial structural characteristics, exploiting the observation that infrared ship targets exhibit temporal stability, while sunlight-induced reflections vary with ocean surface undulations [12]. By incorporating target shape priors, this method achieves accurate detection even under strong solar reflection conditions. Despite their interpretability and ease of engineering deployment, traditional infrared detection methods generally suffer from limitations such as low detection accuracy, slow processing speed, and poor model adaptability. These methods often rely on handcrafted priors tailored to specific background conditions, which hampers their generalization capabilities. As a result, their feature extraction and classification performance tend to degrade in complex or highly dynamic infrared scenes, making them inadequate for the demands of real-time and robust detection in high-variability maritime environments.

In recent years, this rapid advancement of deep learning has spurred a growing interest in its application to infrared ship detection, aiming to enhance both detection accuracy and robustness. Compared with traditional methods based on handcrafted features, deep learning-based approaches offer superior capabilities in automatic feature representation, enabling effective adaptation to complex backgrounds and highly variable target appearances. As a result, these methods have led to significant performance gains in chal-

lenging infrared detection scenarios. Currently, deep learning-based infrared ship detection methods can be broadly classified into two categories: two-stage detectors and one-stage detectors. Representative two-stage frameworks include R-CNN [13], Fast R-CNN [14], and Faster R-CNN [15], which adopt a two-step paradigm wherein region proposals are first generated and subsequently refined through classification and bounding-box regression. While these methods generally achieve high detection accuracy, their relatively low inference speed hinders real-time applicability in dynamic maritime environments. In contrast, one-stage detectors such as SSD [16], RetinaNet [17], and the YOLO family [18,19] adopt an end-to-end design that treats object detection as a unified regression problem, jointly predicting class probabilities and bounding box coordinates. These models typically offer significantly faster inference speeds, making them well-suited for real-time applications. Although their detection accuracy may fall slightly short of certain two-stage methods, one-stage detectors have demonstrated strong overall performance across a range of infrared detection tasks. With the continued integration of deep learning into infrared imaging, recent studies have incorporated attention mechanisms into YOLO architecture to improve the detection of small maritime targets. For instance, Indah Monisa et al. proposed a one-stage infrared ship detection framework enhanced by a coordinate attention module, effectively boosting the model's ability to localize small targets in cluttered maritime environments characterized by low resolution, low SNR, and high background complexity [20]. Their method demonstrated improved accuracy and robustness in detecting small infrared targets. Building on YOLOv5, Jing Ye et al. introduced a detection framework that integrates attention mechanisms and multi-scale feature enhancement [21]. By incorporating high-resolution shallow feature extraction, noise suppression, and contextual information fusion, the model significantly improves the representational capacity and resilience to interference in small-target detection. Similarly, Lize Miao et al. proposed the convolutional block attention module (CBAM)-YOLOv5, which embeds a convolutional attention module and a scale-bias factor into the YOLOv5 architecture [22]. This design enhances the model's ability to extract salient features from small infrared targets and improves bounding box regression precision. The proposed method effectively addresses the challenges of low contrast, poor resolution, and small object detectability inherent to infrared maritime imagery, leading to notable improvements in detection performance. With the increasing complexity of maritime environments and the demand for reliable detection, the YOLOv5 has proven effective for infrared ship recognition due to its end-to-end design and efficient feature extraction. Incorporating the attention mechanism further improves sensitivity to small or low-texture targets, thereby enhancing overall performance under challenging infrared imaging conditions.

Motivated by these considerations, this work examines the integration of the attention mechanism into the YOLOv5 for infrared maritime ship detection. Specifically, three enhanced model variants are developed by embedding the CBAM into the neck, the detection head, and both modules simultaneously. Comparative experiments on the publicly available infrared ship dataset are then conducted to systematically assess the influence of attention placement on key performance metrics. The results demonstrate the respective advantages and trade-offs of each deployment strategy, offering both theoretical insights and practical guidance for designing lightweight yet robust infrared ship detection models.

The remainder of this paper is organized as follows: Section 2 presents a detailed description of the proposed model's system architecture, with a focus on the integration of the attention mechanism into various modules of the YOLOv5. Section 3 is dedicated to the presentation of experimental results and the comparative analysis of various attention mechanism integration strategies applied to infrared ship detection. Section 4 concludes this paper and discusses potential directions for future research.

2. Proposed Improved YOLOv5

2.1. Overview of the YOLOv5 Architecture

The YOLOv5 is a representative one-stage object detection framework, comprising four major components: the input module, the backbone network, the neck network, and the prediction head [23,24]. The overall architecture of YOLOv5 (version 6.1) is illustrated in Figure 1.

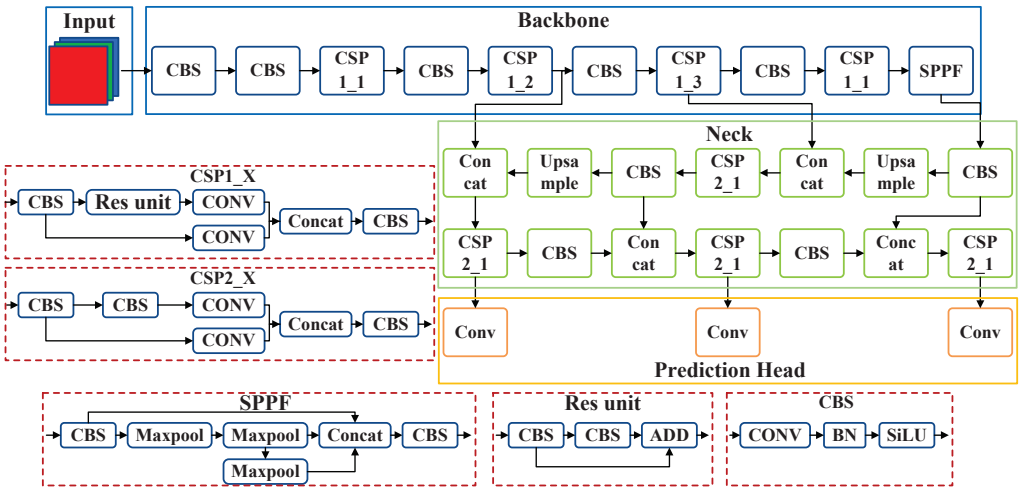


Figure 1. The architecture of the YOLOv5 (version 6.1).

In the input module, the YOLOv5 performs data pre-processing and augmentation to improve training robustness and generalization. Specifically, Mosaic data augmentation is employed, which combines four images into one by scaling, arranging, and cropping, thereby enriching the diversity of training samples. Subsequently, input images are resized proportionally and padded with the minimum amount of black borders to standardize their dimensions, ensuring consistent input to the model. In addition, the YOLOv5 adaptively initializes anchor boxes during the early stages of training. The model computes the intersection over union (IoU) between predicted boxes and ground-truth annotations to evaluate localization error. This error is then minimized through backpropagation, allowing the anchor box dimensions to be iteratively optimized, which in turn reduces the loss function and enhances overall model performance.

The backbone network is primarily responsible for feature extraction and can be viewed as an efficient convolutional architecture designed to capture multi-scale image details. It consists of several key components [1,25], including the CBS blocks, CSP structures, and the SPPF module, as illustrated in the Figure 1. These input images are first processed through a series of CBS blocks, comprising convolution layers, batch normalization, and activation functions, to extract increasingly abstract and expressive feature maps. The SPPF module aggregates multi-scale contextual information via parallel max-pooling operations at varying spatial scales, followed by feature concatenation and convolution, thereby enhancing spatial sensitivity and robustness to object scale variations.

The neck network is designed for feature fusion and primarily consists of the feature pyramid network (FPN) and path aggregation network (PANet) [26,27]. This module employs both upsampling and downsampling operations to effectively integrate multi-level features extracted from the backbone. By integrating channel-wise and spatial information, the module builds a multi-scale feature pyramid, which substantially improves the detection performance for objects of diverse sizes.

In the prediction head, the YOLOv5 performs object classification and localization by generating bounding boxes at three different feature map scales. Each scale predicts

multiple anchor boxes, with each box encoding class probabilities and spatial coordinates. During inference, candidate bounding boxes are generated by regressing from anchor boxes using the predicted offsets and scale ratios. A confidence threshold is then applied to filter out low-confidence detections, followed by non-maximum suppression (NMS) to eliminate redundant boxes and yield the final detection results.

In the context of infrared ship target detection, selecting an appropriate lightweight detection architecture is essential to achieve a balance among detection accuracy, inference speed, and model complexity. The YOLOv5 family offers multiple model variants—namely, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x—to accommodate diverse computational constraints and application requirements. Among them, the YOLOv5n is the most lightweight configuration, featuring the smallest network depth and width [28]. It significantly reduces model size and computational overhead while maintaining acceptable detection performance, making it particularly suitable for deployment on resource-constrained edge devices. In this paper, the YOLOv5n (version 6.1) is selected as the baseline model, based on a comprehensive evaluation of its performance stability, widespread adoption, and active community support. All subsequent improvements and experimental investigations are conducted on top of this architecture. Although recent iterations of the YOLO family, such as YOLOv8-YOLOv11, have demonstrated competitive performance in diverse applications [29]. We adopt the YOLOv5n (version 6.1) in this study because it offers a favorable trade-off among detection accuracy, computational efficiency, and ease of deployment on resource-constrained maritime platforms. Its extensive adoption and stable tooling ecosystem also facilitates reproducibility and comparison with prior work. Importantly, our attention-placement investigation is architecture-agnostic in spirit: the same design can be ported to newer YOLO model with minimal changes. We therefore consider the integration and evaluation of the YOLOv8-YOLOv11 as a natural extension for future work.

2.2. YOLOv5n Enhanced with Attention Mechanism

Given the strong performance of the YOLOv5n in small object detection tasks, we explore the integration of the attention mechanism into different components of the model to enhance infrared ship detection performance. As a key technique in deep learning, the attention mechanism dynamically modulates the focus of feature representations, guiding the model to concentrate on more discriminative target regions while suppressing irrelevant background information. Considering that the YOLOv5n consists of three major components—the backbone, neck, and prediction head—each contributing differently to semantic abstraction and spatial detail representation, the placement of the attention module plays a critical role in optimizing detection performance. Therefore, designing effective attention mechanism integration strategies tailored to specific stages of the model is essential for improving the detection performance of maritime infrared ships.

The attention mechanism can be categorized based on the dimension of focus into channel attention mechanism, spatial attention mechanism, and the hybrid mechanism that combines both. These mechanisms apply learnable weighting strategies to guide the model toward more discriminative feature information, thereby enhancing feature representation and improving object detection performance. For example, the channel attention mechanism is designed to assess the relative importance of different feature channels, while the spatial attention mechanism focuses on localizing critical regions within the image. By leveraging these mechanisms, the model can adaptively refine its feature maps, emphasizing content relevant to the target while suppressing irrelevant or redundant background information.

The channel attention module (CAM) enhances the expressiveness of feature maps by modeling inter-channel dependencies and assigning adaptive weights to each channel. This enables the model to amplify informative feature channels while suppressing irrelevant or redundant ones. The overall processing pipeline of the CAM is illustrated in Figure 2. Given an input feature map $\mathbf{F}$ of size $H \times W \times C$, where H, W, and C denote the height, width, and the number of channels, respectively. To generate channel-wise descriptors, the global average pooling and global max pooling are independently applied along the spatial dimensions. This yields two channel-wise descriptor vectors of size $1 \times 1 \times C$, denotes as $\mathbf{F}^c_{avg}$ and $\mathbf{F}^c_{max}$, respectively. These two descriptors are then passed via a shared multi-layer perceptron (MLP), which learns nonlinear channel relationships. The outputs of the two branches are subsequently combined via element-wise addition to produce a one-dimensional channel attention vector. This vector is normalized using a sigmoid activation function to obtain the final attention weights for each channel. Finally, the attention weights are applied to the original feature map via element-wise multiplication along the channel dimension, resulting in a reweighted feature map $M_c(\mathbf{F})$, where informative channels are enhanced and redundant ones are suppressed. The detailed computation of the CAM can be formulated as follows,

$$\begin{aligned} M_c(\mathbf{F}) &= \delta(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathbf{F}))) \\ &= \delta(\mathbf{W}_1(\mathbf{W}_0\mathbf{F}^c_{avg}) + \mathbf{W}_1(\mathbf{W}_0\mathbf{F}^c_{max})) \end{aligned}, \tag{1}$$

where the $\mathrm{AvgPool}(\cdot)$ and $\mathrm{MaxPool}(\cdot)$ represent global average pooling and global max pooling operations, respectively. The $\delta(\cdot)$ denotes the sigmoid activation function, and $M_c(\mathbf{F})$ represents the final channel attention map.

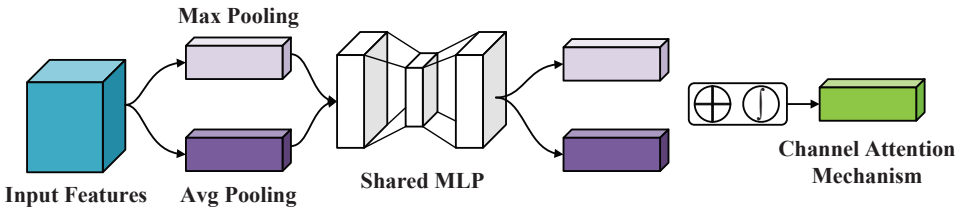


Figure 2. The structure of CAM.

The spatial attention module (SAM) aims to capture the salient regions of a feature map in the spatial dimension by analyzing the importance of different spatial locations. This guides the network to focus more precisely on target areas, thereby improving detection accuracy. Structurally, the SAM is typically applied after the channel attention module, taking as input the channel-refined feature map. Figure 3 illustrates the overall processing pipeline of the SAM. Considering the input feature map $\mathbf{F}$ of size $H \times W \times C$, the SAM first performs global max pooling and global average pooling along the channel dimension, producing spatial maps: $\mathbf{F}^s_{max}$ and $\mathbf{F}^s_{avg}$, both of size $H \times W$. These two maps are then concatenated along the channel axis to form a feature map of size $H \times W \times 2$. This concatenated map is passed through a convolutional layer with a kernel size of 7×7, followed by a sigmoid activation function, yielding the spatial attention map $M_s(\mathbf{F})$. Finally, the spatial attention map is applied to the input feature map via element-wise multiplication (broadcast along the channel dimension), thereby enhancing the spatially salient regions of the feature representation. The entire computation process of the SAM can be formulated as follows,

$$\begin{aligned} M_s(\mathbf{F}) &= \delta(f^{7\times7}([\mathrm{AvgPool}(\mathbf{F}); \mathrm{MaxPool}(\mathbf{F})])) \\ &= \delta(f^{7\times7}([\mathbf{F}^s_{avg}; \mathbf{F}^s_{max}])) \end{aligned}, \tag{2}$$

where $f^{7\times7}(\cdot)$ denotes the convolutional operation with the kernel size of 7×7. $[\cdot;\cdot]$ represents channel-wise concatenation, and $\sigma(\cdot)$ is the sigmoid activation function.

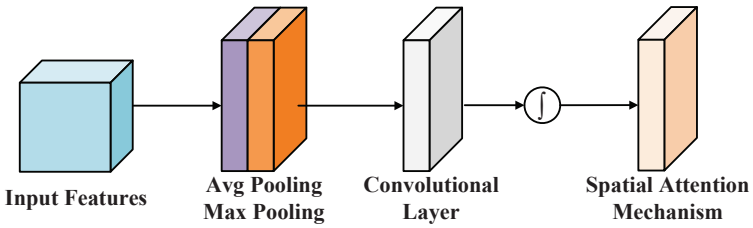


Figure 3. The structure of SAM.

Among various hybrid attention mechanisms, the CBAM has emerged as a lightweight and effective design and has been widely applied in object detection and related vision tasks. As illustrated in Figure 4, the CBAM consists of a channel attention module followed sequentially by a spatial attention module. Notably, the CBAM exhibits strong generality and plug-and-play flexibility, making it suitable for multi-layer integration within diverse deep neural network architectures. By jointly modeling both channel-wise and spatial attention, the CBAM enables the network to better focus on informative features while suppressing irrelevant information, thereby enhancing the discriminative capacity of feature representations. Based on these advantages, we adopt the CBAM as the representative attention module and integrate it into the YOLOv5n. We systematically investigate the effect of attention placement on detection performance in the context of infrared ship detection.

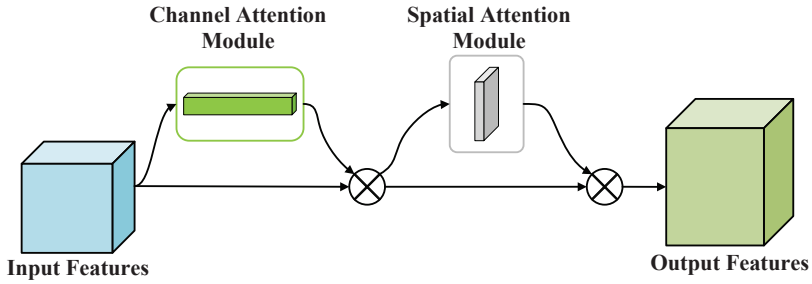


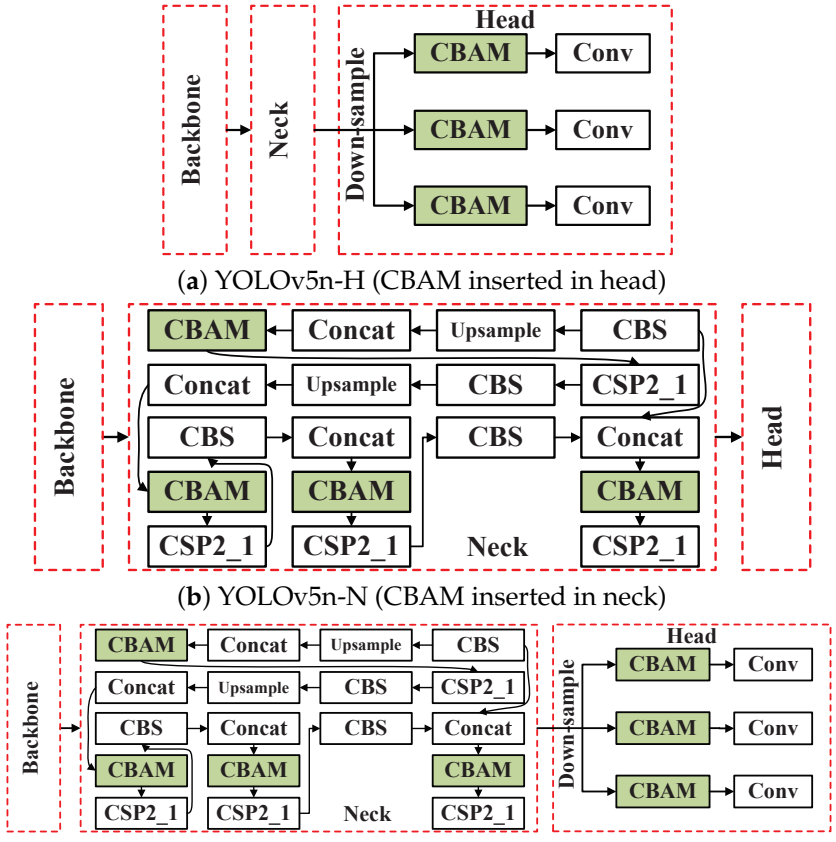
Figure 4. The structure of CBAM.

As shown in Figure 5, the CBAM is embedded into three different configurations: the neck module (YOLOv5n-N), the prediction head (YOLOv5n-H), and both the neck and head simultaneously (YOLOv5n-HN). Through comparative experiments, the impact of the attention mechanism at the feature fusion stage and output prediction stage is evaluated, providing insights into their respective contributions to detection accuracy and robustness. It is worth noting that the motivation for systematically exploring the placement of the attention mechanism is primarily based on the following considerations:

- Neck Position: Enhancing early-stage attention to small target features. In infrared imagery, small maritime objects—such as the canoe and fishing boat—tend to exhibit sparse texture and diminished saliency following deep downsampling, making them susceptible to being overwhelmed by background clutter. This indicates that embedding the CBAM into the neck network enhances the sensitivity of the model to small target during the multi-scale feature fusion stage. Specifically, the spatial attention mechanism emphasizes the spatial locations of potential targets, while the channel attention mechanism strengthens semantic channels associated with thermal emissions and structural contours of ships. This integration improves the effectiveness of early feature perception and enhances the model's ability to preserve discriminative information relevant to small objects.

- Head Position: Improving late-stage suppression of background noise. Infrared maritime scenes often exhibit complex backgrounds, with interference from waves, fog, and strong reflections, which can mislead the detection process. To address this challenge, integrating the CBAM into the prediction head enables further refinement of the candidate feature maps during the output stage. By selectively compressing and reweighting features, the model effectively suppresses background noise and enhances the response to target regions, thereby improving robustness and discriminative capability in the final prediction phase.
- Neck+Head Joint Integration: Enabling multi-stage collaborative optimization. Incorporating attention mechanism into both the neck and prediction head is theoretically advantageous, as it allows for the simultaneous enhancement of small target representations in the early stages and suppression of background interference in the later stages, potentially yielding a synergistic optimization effect. Nevertheless, the use of dual attention modules may also introduce feature redundancy or interference, thereby affecting inference efficiency and training stability. To evaluate the effectiveness of this joint strategy, a variant model named YOLOv5n-HN is constructed. Extensive experiments are conducted under various infrared maritime scenarios to assess whether this configuration offers generalized performance gains or if its benefits are limited to specific task settings.

Through the design of the three aforementioned model variants, this work not only investigates the effectiveness of the attention mechanism itself but also emphasizes a comparative analysis of its deployment at different stages of the network. The goal is to provide deeper insight into the optimal placement strategy of the attention module for infrared maritime ship detection.



(**a**) YOLOv5n-H (CBAM inserted in head)

(**b**) YOLOv5n-N (CBAM inserted in neck)

(**c**) YOLOv5n-HN (CBAM inserted in both neck and head)

Figure 5. Architectural diagrams of three YOLOv5n model variants enhanced with CBAM.

3. Experiment Results Analysis

3.1. Dataset Description and Analysis

In this section, we utilize the publicly available infrared maritime ship image dataset provided by Yantai Raytron Technology Co., Ltd. (Yantai, China). The dataset presents both unique advantages and potential challenges across multiple dimensions. It comprises a total of 8403 images, offering sufficient scale to support the parameter optimization and iterative training requirements of deep learning models. The dataset includes annotations for seven ship categories, such as liner, bulk carrier, and others. While the category set is relatively diverse, the distribution of samples is significantly imbalanced, as illustrated in Figure 6 For instance, the number of fishing boat instances reaches 8146, whereas container ships are represented by only 622 samples. Such an imbalance in data distribution may impede the model from effectively learning discriminative features for minority classes, thereby increasing the risk of missed detections or misclassifications during inference, particularly for underrepresented categories.

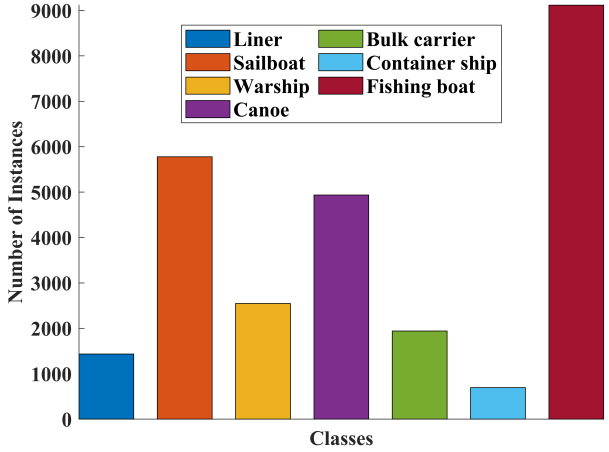


Figure 6. Class distribution of annotated ship types in the infrared dataset.

In terms of scene coverage, the dataset leverages infrared imaging technology to overcome the limitations of conventional optical sensors under challenging environmental conditions. The collected images encompass a variety of representative maritime scenarios, including open-sea navigation, port docking, and nearshore operations. Furthermore, the dataset includes infrared imagery captured under varying times of day, sea states, and illumination conditions, thereby enhancing its environmental diversity. The inclusion of multi-resolution imagery, which ranges from 384×288 to 1280×1024, further enhances the dataset, enabling models to learn discriminative ship features across varying spatial scales and thereby improving generalization in complex maritime environments. Nonetheless, infrared imaging is inherently constrained by low texture richness and reduced image contrast, which significantly increases the difficulty of detecting small ships or targets with weak thermal signatures. Additionally, due to the structural diversity of ships, the annotated bounding boxes in the dataset exhibit a wide range of aspect ratios. Some targets display extreme elongation along either the vertical or horizontal axis. The compatibility between anchor boxes and ground-truth boxes is largely dependent on the distribution of these aspect ratios. As shown in Figure 7, the normalized aspect ratio distribution is concentrated in the upper-left and lower-left quadrants, which aligns with the geometric characteristics commonly observed in maritime targets.

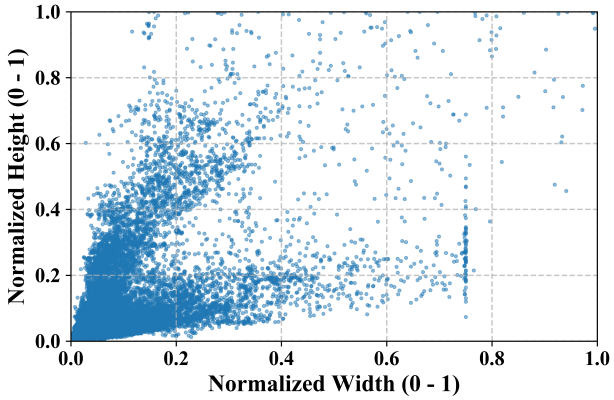


Figure 7. Distribution of normalized bounding box width and height in the infrared ship dataset.

3.2. Training Strategy and Implementation Details

The experiments were conducted on a Linux-based system (Ubuntu 22.04.1 LTS) using the PyTorch deep learning framework (version 1.13.1 with CUDA 11.7). The model training was performed on the Autodl platform. The input image size for the network was set to 640 × 640. The detailed hardware configuration used for all experiments is summarized in Table 1. The YOLOv5n was adopted as the baseline detection model. The dataset (8402 images) was split into training, validation, and test sets. First, 90% of the dataset was allocated for model development and 10% for testing. The 90% portion was then further divided into training and validation subsets at a 9:1 ratio. Consequently, the final split consisted of 81% training (6804 images), 9% validation (757 images), and 10% test (841 images).

Corresponding, Figure 8 illustrates the overall pipeline of the proposed model. During the training stage, infrared images from the training set are first subjected to data augmentation, followed by model selection and initialization. The model is then optimized via forward propagation and error backpropagation until the performance metrics satisfy the predefined criteria, yielding the final trained model. In the testing stage, infrared images from the test set are processed using the trained model to perform ship detection, and the resulting predictions are evaluated to assess detection performance.

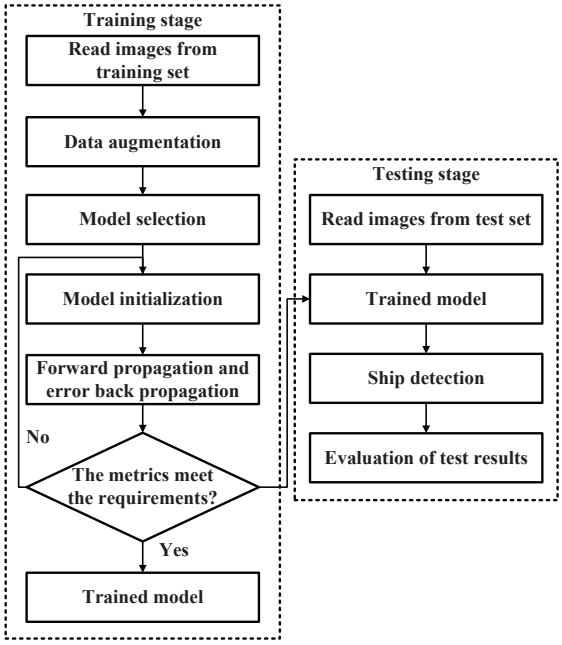


Figure 8. Overall pipeline of the proposed model, including the training stage and the testing stage.

Table 1. Experimental environment configuration.

Component	Specification
CPU	22vCPU AMD EPYC 7T83 64-Core Processor
GPU	vGPU-32GB (32 GB)
Memory	90 GB
CUDA	11.7 (compatible with PyTorch 1.13.1)
Operation System	Linux (Ubuntu 22.04.1 LTS, Chongqing, China)
Deep Learning Framework	PyTorch (1.13.1)

3.2.1. Hyperparameter Setting

According to the configuration summarized in Table 2, stochastic gradient descent (SGD) was employed as the optimizer for model parameter updates, with a momentum coefficient set to 0.937, an initial learning rate of 0.01, and a weight decay factor of 0.0005. Regarding hyperparameter settings, the batch size was configured to the maximum value permitted by the available memory, which was 32. The learning rate ranged from a maximum of 0.01 to a minimum of 0.0001. All models were trained for a total of 300 epochs, and performance was directly assessed on the designated test set. To address the class imbalance issue, data augmentation techniques, including Mosaic and MixUp, were employed to increase sample diversity. These strategies are particularly beneficial for improving the generalization capability of the model with respect to minority classes. The probability of applying Mosaic probability was set to 0.5, indicating a 50% chance of applying Mosaic at each training step. The same probability setting was used for MixUp. Additionally, the Special Auto Ratio was configured to 0.7, meaning that Mosaic augmentation was applied only during the first 70% of the training epochs. During inference, a fixed confidence threshold of 0.5 was applied to filter out low-confidence detections, following the default configuration in YOLOv5n. This setting balances false positives and missed detections under low contrast and complex backgrounds. And the anchor boxes were initialized using YOLOv5n's K-means–based adaptive anchor calculation, which was fitted to the infrared ship dataset to match the target scale distribution.

Table 2. Training hyperparameter settings.

Hyperparameter	Value
Batch Size	32
Epoch	300
Learning Rate	0.01
Mosaic Probability	0.5
MixUp Probability	0.5
Optimizer	SGD
Momentum	0.937
Special Auto Ratio	0.7

3.2.2. Loss Function Analysis

Figure 9 presents the training and validation loss curves under four models: the baseline YOLOv5n, YOLOv5n-N, YOLOv5n-H, and YOLOv5n-HN. It should be mentioned that the loss values are presented on a logarithmic scale to highlight subtle differences. As illustrated in the figure, all models exhibit high initial loss values in both training and validation sets due to the lack of convergence during the early training stages. This is attributed to the models not yet having learned effective feature representations from the training data. As training progresses, the loss values decrease rapidly. In the case of the original YOLOv5n, the loss begins to stabilize after approximately 50 epochs, eventually converging

to around 0.108. The three attention-enhanced variants follow similar convergence trends, reaching stable loss values within a comparable number of epochs. Notably, throughout the training process, the training and validation loss curves for all three attention-mechanism-integrated models converge closely, with no significant divergence between them. It means that all four models exhibit good generalization capability and maintain stable detection performance on previously unseen data.

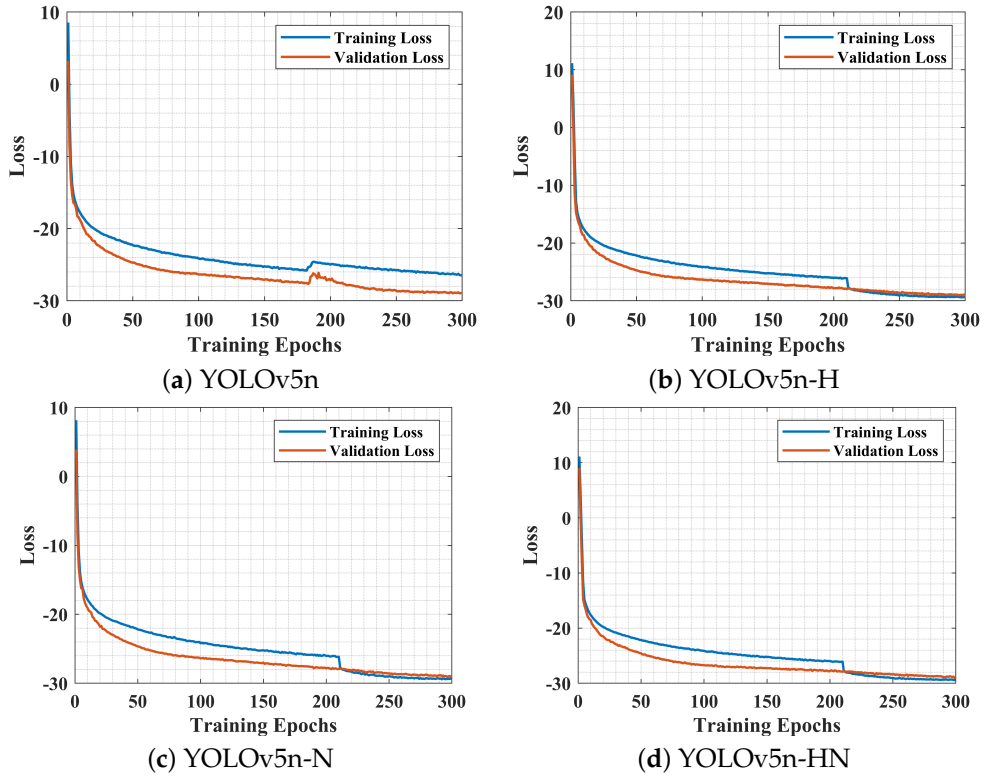


Figure 9. Training curves of YOLOv5n model variants.

3.2.3. mAP Metric Analysis

Figure 10 shows the evolution of mean Average Precision (mAP) across training epochs for all model variants. It should be noted that, to generate the performance curves in Figure 10, the models were saved every 10 training epochs, and each saved version was subsequently evaluated on the test set for mAP computation. In the early training stages, the mAP values remain relatively low across all models, indicating that the networks had not yet effectively learned discriminative target features due to parameter under-convergence. As training progresses, the mAP increases rapidly—particularly between epochs 0 and 100—demonstrating that this phase corresponds to the most active period of feature learning, during which detection accuracy improves substantially. Between epochs 100 and 150, the rate of mAP growth begins to decelerate, suggesting that the model is gradually approaching convergence. By epoch 150, all three attention-based models achieve the mAP of approximately 0.77, suggesting that the models have learned most of the dominant target features, although room for further improvement remains. Beyond this point, the mAP continues to rise but with diminishing gains. Near epoch 300, the mAP curves stabilize and converge around 0.87, indicating that the models have reached saturation in terms of detection capability and that training has effectively converged. A comparison of the curves reveals that both the YOLOv5n-H and YOLOv5n-N, which incorporate attention mechanisms, mostly outperform the original YOLOv5n in terms of mAP throughout the training process. This suggests that the integration of the attention mechanism contributes positively to the enhancement of detection accuracy.

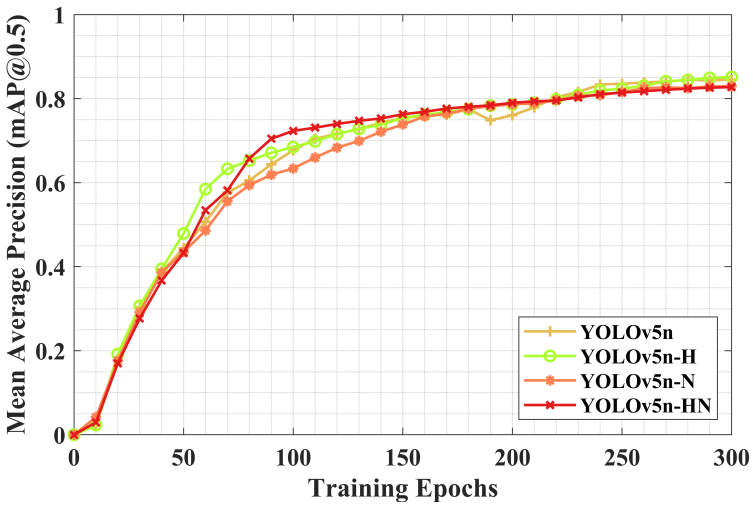


Figure 10. Comparison of mean Average Precision (mAP@0.5) across YOLOv5n with different structural modifications.

3.3. Performance Evaluation and Comparative Analysis

3.3.1. Metric-Based Comparison

In this section, model performance is evaluated using five core metrics: mAP, Average Precision (AP), F1 score, recall rate, and precision. All evaluations are conducted under a fixed confidence threshold of 0.5. The models under comparison include the original YOLOv5n and three attention-mechanism-enhanced model variants: YOLOv5n-N, YOLOv5n-H, and YOLOv5n-HN. The evaluation focuses on various categories of ship targets within the validation set. The subsequent analysis integrates both quantitative results and visualized detection outputs to assess the performance improvements brought by attention mechanisms in the context of infrared ship detection.

As shown in Table 3, in terms of the mAP@0.5 metric, the YOLOv5n-HN achieves the best overall performance with a score of 86.83%, surpassing the other three models, whose scores range from 85.95% to 86.37%. This result indicates that jointly embedding attention mechanisms into both the neck and prediction head can more effectively enhance feature representation, thereby significantly improving overall detection performance. For the mAP@0.5:0.95 metric, the YOLOv5n-H achieves a slightly higher score of 51.04% compared to 50.00% for the YOLOv5n-HN. This observation suggests that the combined attention configuration may still require further refinement to achieve a more balanced precision across varying IoU thresholds. Regarding precision, all three model variants exhibit notable improvements over the original YOLOv5n. The YOLOv5n-N achieves the highest precision at 93.49%, followed closely by the YOLOv5n-H at 93.41% and the YOLOv5n-HN at 93.13%. These results demonstrate that the attention mechanism is highly effective in reducing false positives and enhancing target discriminability. In contrast, the highest recall rate is obtained by the original YOLOv5n at 80.24%, outperforming the attention-integrated variants. This indicates that although the attention mechanism improves detection accuracy, it may also suppress certain target responses, leading to increased missed detection. Therefore, a better trade-off between precision and completeness is still needed to enhance the overall robustness of the detection system.

Table 3. Average performance of the YOLOv5n model variants on the test dataset.

Model Variant	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)
YOLOv5n	86.20	50.33	91.62	80.24
YOLOv5n-H	86.37	51.04	93.41	76.50
YOLOv5n-N	85.95	50.43	93.49	75.39
YOLOv5n-HN	86.83	50.00	93.13	76.13

Table 4 provides a comparative analysis of the AP across different ship categories under various model configurations. Taking the canoe class as a representative small-object category, the YOLOv5n-N achieves the highest AP at 74.30%, outperforming YOLOv5n-H (73.21%), YOLOv5n-HN (71.89%), and the original YOLOv5n (71.78%). This result highlights the advantage of embedding the attention mechanism in the neck, which enhances small-object feature representation and consequently improves the localization and classification accuracy for small-scale targets. Although the YOLOv5n-HN shows slightly weaker performance than the YOLOv5n-N in canoe detection, it excels in detecting large targets such as warships, where it outperforms all other models. This suggests that joint attention deployment in both the neck and head facilitates more effective capture of the high-level semantic features associated with large objects. For medium-scale targets such as liners and sailboats, the YOLOv5n-HN achieves AP values that fall between those of the YOLOv5n-H and YOLOv5n-N, demonstrating its adaptability across multiple object scales. These findings collectively indicate that joint attention integration offers a balanced performance profile, which is particularly advantageous in scenarios involving diverse target sizes.

Table 4. AP of the YOLOv5n model variants on different ship categories.

Model Variant	Bulk Carrier (%)	Canoe (%)	Container Ship (%)	Fishing Boat (%)	Liner (%)	Sailboat (%)	Warship (%)
YOLOv5n	95.18	71.78	95.95	80.51	86.75	85.46	98.79
YOLOv5n-H	95.19	73.21	94.36	79.10	86.93	85.31	98.84
YOLOv5n-N	94.71	74.30	91.68	78.52	85.22	86.98	98.63
YOLOv5n-HN	94.12	71.89	91.64	78.96	86.59	85.58	99.02

Note. AP values are reported to two decimal places (four significant digits) to highlight performance differences among model variants. These results are probability-based estimates influenced by sample size and class distribution, and should be interpreted as approximate rather than exact values.

As shown in Tables 5 and 6, the three model variants, i.e., YOLOv5n-H, YOLOv5n-N, and YOLOv5n-HN, demonstrate distinct advantages in detecting ships across different scales. Notable performance differences are observed across different object categories. In terms of precision, all three enhanced models outperform the baseline YOLOv5n in categories such as canoe and bulk carrier. For instance, the YOLOv5n-N achieves the highest precision of 88.41% for the canoe class, which is attributed to the attention module embedded in the neck, as it effectively enhances the capability of the model to capture weak and fine-grained features during the process of multi-scale fusion. The YOLOv5n-HN follows with a precision of 86.87%, reflecting the synergistic effect of attention applied to both the neck and prediction head. Despite this improvement in precision, the recall rates for the YOLOv5n-N (57.54%) and YOLOv5n-HN (55.60%) remain lower than 59.91% achieved by the original YOLOv5n. This suggests that while the attention mechanism improves discriminative ability, it may also suppress ambiguous features during compression, leading to missed detection. It also underlines the persistent challenge of balancing precision and recall in small object detection.

Table 5. Precision (%) of the YOLOv5n model variants on different ship categories.

Model Variant	Bulk Carrier (%)	Canoe (%)	Container Ship (%)	Fishing Boat (%)	Liner (%)	Sailboat (%)	Warship (%)
YOLOv5n	96.97	80.12	95.77	89.02	86.75	94.16	94.81
YOLOv5n-H	96.39	85.90	98.51	88.70	95.83	93.55	95.74
YOLOv5n-N	96.39	88.41	98.48	87.97	94.26	93.18	95.05
YOLOv5n-HN	96.23	86.87	98.48	88.19	94.35	92.32	95.47

Note. Precision values are reported to two decimal places (four significant digits) to allow meaningful comparison across categories and models. Since these metrics depend on the finite test dataset, they are subject to statistical variability and should be regarded as estimates rather than absolute measurements.

Table 6. Recall rate (%) of the YOLOv5n model variants on different ship categories.

Model Variant	Bulk Carrier (%)	Canoe (%)	Container Ship (%)	Fishing Boat (%)	Liner (%)	Sailboat (%)	Warship (%)
YOLOv5n	87.91	59.91	93.15	68.42	79.63	74.80	97.86
YOLOv5n-H	87.91	57.76	90.41	63.79	70.99	68.08	96.43
YOLOv5n-N	87.91	57.54	89.04	63.17	70.99	70.58	96.07
YOLOv5n-HN	84.07	55.60	89.04	64.51	78.22	69.64	97.86

Note. Recall rate values are presented to two decimal places (four significant digits) for consistency with other metrics. As probability estimates derived from limited data samples, they inherently contain uncertainty and should be interpreted as approximate indicators of detection performance.

For the medium-scale targets such as the fishing boats, the YOLOv5n-HN achieves a recall rate of 64.51%, surpassing the 63.17% achieved by the YOLOv5n-N, and attains a precision of 88.19%, which closely approximates the 88.70% yielded by the YOLOv5n-H. These findings suggest that the joint integration of attention mechanisms in both the neck and head modules facilitates a more effective trade-off between reducing false positives and alleviating missed detections. This improvement stems from the synergistic effect of attention mechanisms in the head and neck modules, where the former enhances classification precision and the latter reinforces multi-scale feature integration, jointly improving detection reliability for medium-scale maritime targets.

For large and thermally prominent targets such as the warships, the YOLOv5n-HN performs exceptionally well, achieving a precision of 95.47% and a recall rate of 97.86%. This high level of accuracy is attributed to the complementary functions of the attention mechanisms integrated into both the neck and head modules. Specifically, the head-based attention mechanism sharpens semantic discrimination during classification, while the neck-based attention mechanism enhances multi-scale spatial feature fusion. Their synergy enables precise localization and robust recognition, thereby supporting reliable detection in complex infrared maritime environments.

As shown in Table 7, the F1 score further substantiates the impact of different attention mechanism integration strategies on performance. For the warship category, representing large-scale targets, the YOLOv5n-HN achieves the highest F1 score of 0.97, indicating superior performance among all compared models. This result demonstrates that the collaborative embedding of attention mechanisms in both the neck and head not only improves classification accuracy but also enhances feature completeness. The synergy between precision and recall thus leads to stable and high-quality detection for large maritime objects. In detecting the fishing boats, the representative medium-scale targets, the YOLOv5n-HN attains an F1 score of 0.75, slightly outperforming the YOLOv5n-N. This indicates that the joint deployment of attention mechanisms also benefits the detection of medium-scale targets, with head-applied attention enhancing semantic discrimination during classification and neck-applied attention improving mid-level feature fusion. Their

cooperation enables a balanced trade-off between precision and recall in complex maritime scenes, thereby improving overall detection reliability.

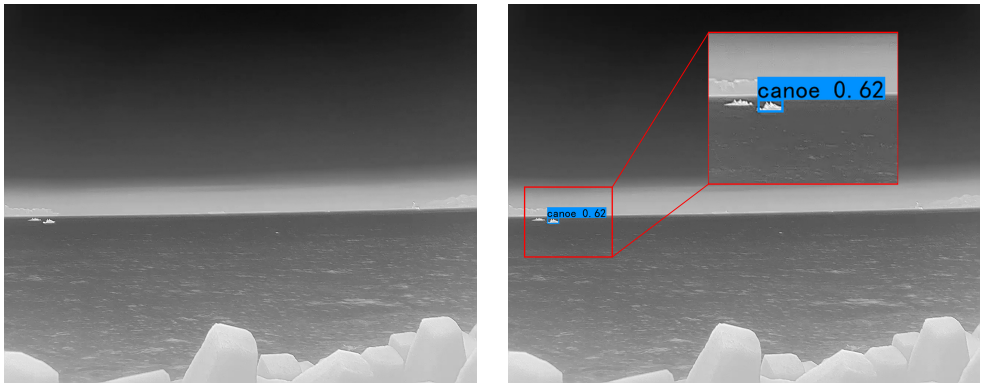
Table 7. F1 score of the YOLOv5n model variants on different ship categories.

Model Variant	Bulk Carrier	Canoe	Container Ship	Fishing Boat	Liner	Sailboat	Warship
YOLOv5n	0.92	0.69	0.94	0.77	0.86	0.82	0.96
YOLOv5n-H	0.92	0.69	0.94	0.74	0.82	0.79	0.96
YOLOv5n-N	0.92	0.70	0.94	0.74	0.81	0.80	0.96
YOLOv5n-HN	0.90	0.68	0.94	0.75	0.82	0.79	0.97

However, for small-scale targets such as the canoes, the YOLOv5n-N achieves a higher F1 score than the YOLOv5n-HN. This indicates that attention embedded solely in the neck may be more effective in enhancing fine-grained feature representations for small-object detection. The inferior performance of the joint strategy may be attributed to the low pixel ratio, blurred boundaries, and high background interference commonly associated with small infrared targets. Although the collaborative attention mechanism demonstrates clear benefits for medium and large targets, its effectiveness in small-object detection remains limited due to insufficient feature expressiveness.

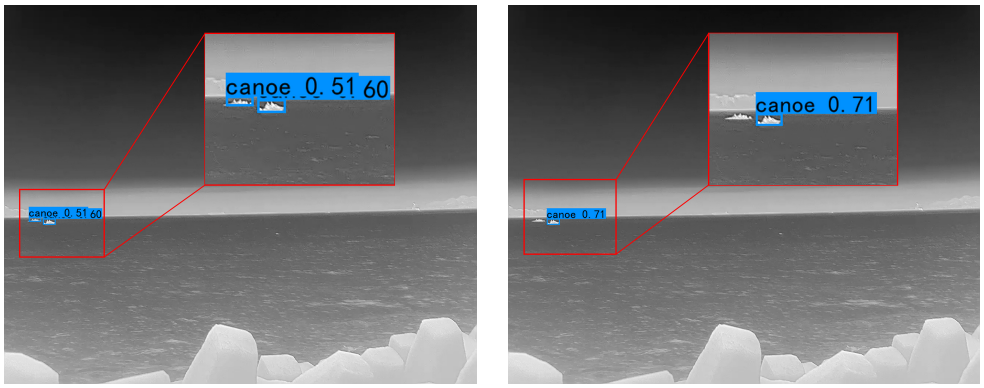
3.3.2. Case-Level Analysis

To gain deeper insights into the practical performance of each model variant, the case-level analysis is conducted using three representative infrared maritime scenarios, as illustrated in Figures 11–13. These scenarios are carefully selected to reflect diverse operational challenges, including the detection of small objects under low contrast, the distribution of dense targets in narrow waterways, and the identification of large objects under strong thermal reflections. By comparing the detection outcomes of the baseline YOLOv5n and its three attention-enhanced counterparts. In this section, we provide the qualitative complement to the quantitative evaluation in Section 3.3.1. The visual results allow for a more nuanced understanding of how the attention mechanism influences detection robustness, confidence, and boundary localization in complex infrared conditions.



(**a**) YOLOv5n (**b**) YOLOv5n-H

Figure 11. *Cont.*



(**c**) YOLOv5n-N (**d**) YOLOv5n-HN

Figure 11. Detection results of the YOLOv5n model variants under Case 1. (The red box was manually added by the authors to highlight the search region for readability and does not represent the raw bounding box output of the detector).

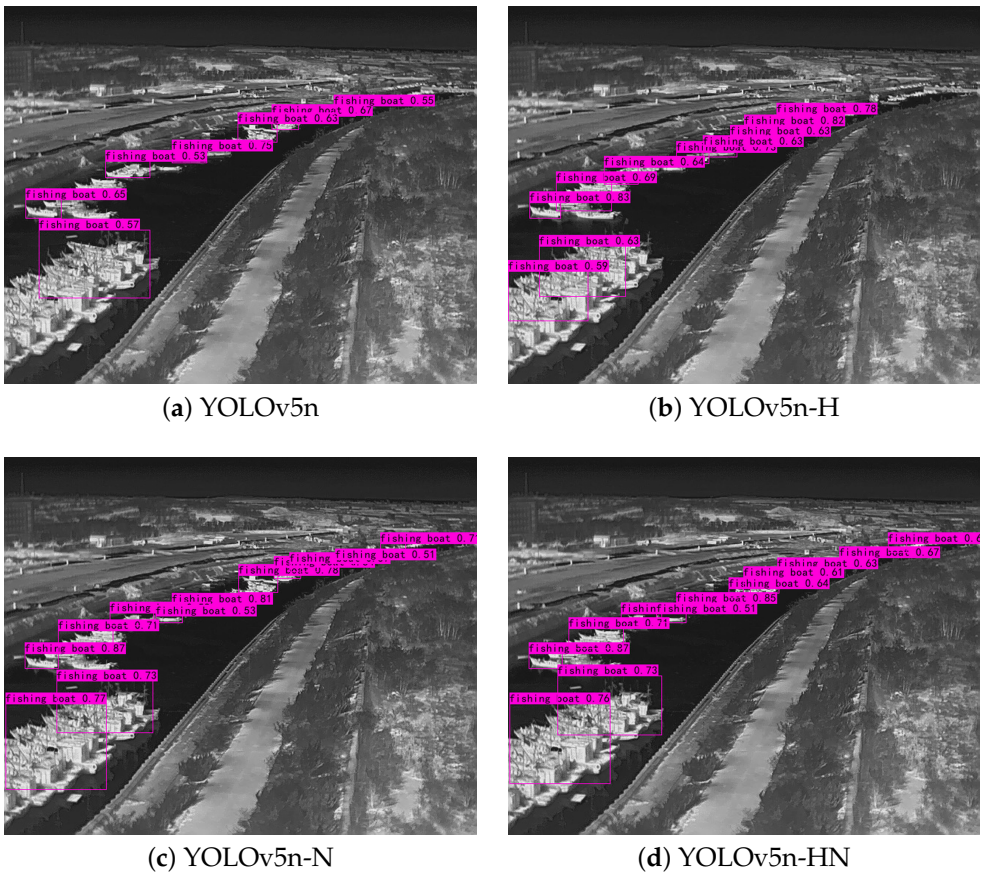
Case 1: Small-Scale Target Detection

As illustrated in Figure 11, the YOLOv5n fails to detect the canoe target, which occupies less than 5% of the image area. Due to insufficient confidence (namely, below the 0.5 threshold), no valid bounding box is generated, and the existing low-confidence outputs indicate the limited sensitivity of the model to small-scale features in cluttered infrared scenes. By introducing an attention mechanism into the detection head, the YOLOv5n-H demonstrates improved localization accuracy for the canoe, with the confidence score increasing to 0.62. Additionally, background interference, such as wave textures, is effectively suppressed, reducing false positives. Nevertheless, while target localization is enhanced, some small targets remain undetected, and the predicted bounding boxes still exhibit boundary overlap with the background. This indicates that the head-level attention mechanism improves discrimination between foreground and background but remains sub-optimal under severe noise or low-saliency conditions.

In contrast, the YOLOv5n-N, which incorporates the attention mechanism into the neck module, exhibits superior performance in densely packed small-object scenarios. It successfully detects two closely located canoe instances, demonstrating strong target separation. This is primarily attributed to the spatial attention mechanism's ability to amplify inter-object differences, thereby avoiding the cross-object bounding box fusion issue observed in YOLOv5n-H. Furthermore, the YOLOv5n-HN, with the attention mechanisms jointly embedded in both neck and head modules, achieves higher confidence and more precise localization in certain cases. This suggests that coordinated attention integration holds promise for improving detection quality. Yet the accompanying increase in model complexity and parameter count may introduce redundancy when handling extremely small targets, potentially resulting in suppressed activation or unstable predictions. This reveals a trade-off between expressive power and effective feature compression, which could be mitigated by dynamic or learnable attention allocation strategies.

Future work may consider incorporating learnable attention weighting schemes to dynamically regulate enhancement strength and timing based on contextual cues. Architecturally, the neck module in the YOLOv5n serves as the core of multi-scale feature aggregation. Therefore, embedding attention at this stage facilitates the coordinated reinforcement of both low-level spatial details and high-level semantic information, which proves particularly advantageous for small-object detection. In contrast to the head-level attention mechanism that refines outputs at the final prediction stage, the neck-level attention mechanism improves feature quality earlier in the pipeline, supplying more discriminative

inputs to downstream branches. This architectural characteristic underscores the unique effectiveness of neck-based attention in enhancing subtle target representation.



(**a**) YOLOv5n (**b**) YOLOv5n-H

(**c**) YOLOv5n-N (**d**) YOLOv5n-HN

Figure 12. Detection results of the YOLOv5n model variants under Case 2.

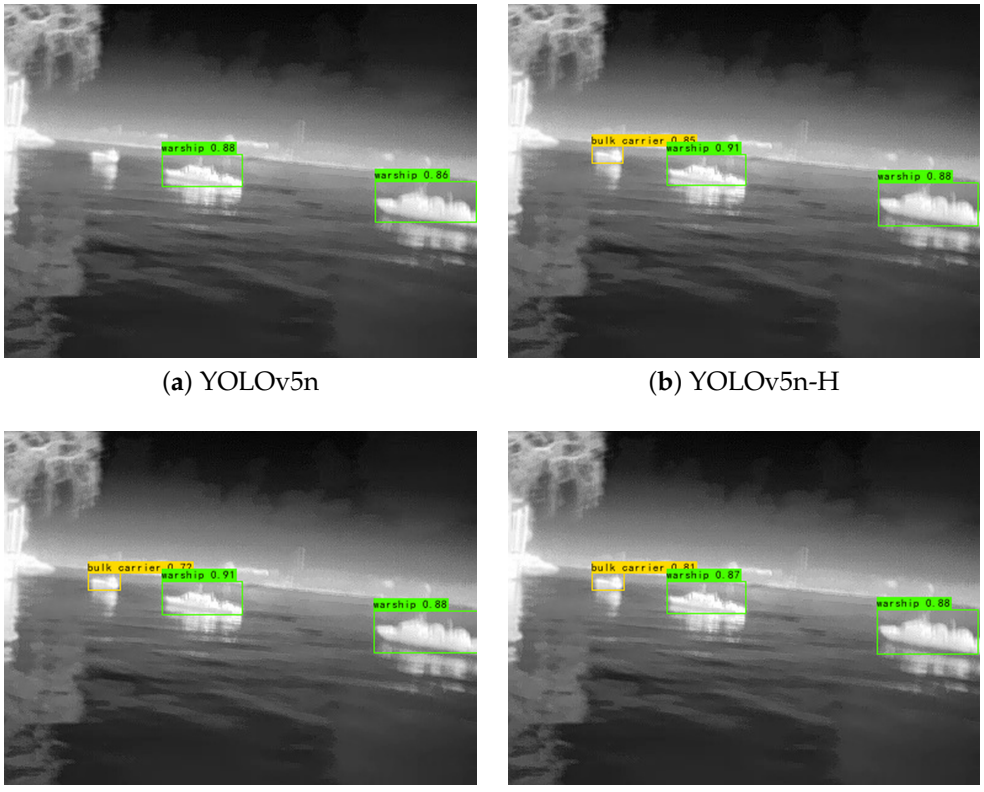
Case 2: Dense Medium-Scale Target Detection

As illustrated in Figure 12, the detection results for densely distributed fishing boats in a dockside scenario vary significantly across different YOLOv5n configurations. The YOLOv5n achieves broad coverage, successfully identifying most visible targets and demonstrating its fundamental detection capability. This strength is offset by dispersed confidence scores, particularly for distant and smaller ships, and by bounding boxes with imprecise boundaries, which indicate limited localization accuracy. This indicates that the absence of the attention mechanism restricts the model's effectiveness in discriminating small targets from complex background interference, resulting in compromised precision and increased misclassification risks.

In contrast, the YOLOv5n-H exhibits improved detection precision. Its confidence scores are more concentrated, with most fishing boats assigned significantly higher values. This suggests that the head-level attention mechanism effectively enhances classification confidence and boundary discrimination. Nevertheless, as this mechanism is positioned at the final stage of detection, it cannot compensate for information loss from earlier feature extraction processes, resulting in reduced detection coverage and residual background-induced false positives. Alternatively, the YOLOv5n-N enhances multi-scale feature fusion at earlier stages, resulting in tighter bounding boxes and substantially improved suppression of background noise, thereby enabling more accurate localization of small vessels. Although these improvements are evident, the model still struggles with extremely small or distant vessels, and a certain level of missed detections persists.

A detailed analysis of the detection results on the densely distributed fishing boats in the lower-left corner of Figure 12 is presented below. The YOLOv5n produces a single bounding box with a confidence of 0.57, failing to separate the overlapping hulls. The YOLOv5n-H generates two bounding boxes with confidence scores of 0.59 and 0.63, indicating improved boundary discrimination in the prediction head, though still with modest reliability. The YOLOv5n-N successfully separates the overlapping ships into two bounding boxes with higher confidence levels of 0.73 and 0.77, demonstrating greater resilience to feature confusion due to enhanced multi-scale fusion in the neck. The YOLOv5n-HN also detects two bounding boxes, with confidence scores of 0.76 and 0.73, yielding a performance comparable to the YOLOv5n-N. These findings suggest that neck-level attention is the primary driver of robust separation under dense and partially occluded conditions, whereas joint neck–head integration achieves similar robustness at the expense of slightly higher computational complexity.

Among the evaluated models, the YOLOv5n-HN, which integrates attention mechanisms in both the neck and head, achieves a more balanced trade-off between confidence consistency and detection coverage. The resulting bounding boxes exhibit high spatial alignment with target boundaries and minimal background interference, reflecting superior detection accuracy and robustness. Nonetheless, the use of dual attention modules introduces additional parameters and computational overhead, potentially impacting inference speed. Therefore, future research should explore the lightweight attention mechanisms to maintain detection performance while simultaneously addressing computational efficiency for practical, real-time applications.



(**a**) YOLOv5n (**b**) YOLOv5n-H

(**c**) YOLOv5n-N (**d**) YOLOv5n-HN

Figure 13. Detection results of the YOLOv5n model variants under Case 3.

Case 3: Large-Scale and High-Thermal Target Detection

As illustrated in Figure 13, significant performance differences emerge among the evaluated models under complex maritime conditions. Specifically, the baseline YOLOv5n detects only the warship and entirely fails to recognize the bulk carrier, highlighting

its limitations in extracting discriminative features for small or low-texture targets in conditions of intense sea-surface reflections. The YOLOv5n-H significantly enhances detection performance by successfully identifying the bulk carrier with a high confidence score of 0.85 and further raising the warship detection confidence to 0.91. This improvement demonstrates the effectiveness of head-level attention in amplifying high-level semantic representations, thereby increasing sensitivity to marginal or low-frequency targets.

In contrast, the YOLOv5n-N detects the bulk carrier at a lower confidence level of 0.72 while maintaining high confidence at 0.91 for the warship. While this model variant demonstrates stronger suppression of background noise and produces more compact bounding boxes, its lack of semantic-level enhancement leads to sub-optimal representation of rare-class features, resulting in slightly lower confidence for small-sample targets compared to the YOLOv5n-H. Combining attention mechanisms in both neck and head modules, the YOLOv5n-HN achieves confidence scores of 0.81 for the bulk carrier and 0.87 for the warship, striking a balanced compromise between spatial refinement and semantic enhancement. Nevertheless, the slightly reduced confidence in certain cases suggests that stacked attention modules may disproportionately favor strong-signal targets, inadvertently suppressing weaker signals from underrepresented categories.

In summary, the attention mechanism generally improves overall detection performance but may introduce unintended class bias. The original YOLOv5n exhibits no such bias but struggles with small-object recognition. The YOLOv5n-H better improves detection confidence for rare or marginal targets, whereas the YOLOv5n-N shows strength in background noise suppression but offers limited semantic enhancement. Although the YOLOv5n-HN combines the advantages of both head and neck modules, it still displays residual attention bias toward dominant targets. Future research may explore dynamic attention allocation schemes based on category-aware feature perception to achieve balanced cross-class detection performance, particularly for small infrared targets prevalent in maritime surveillance scenarios.

4. Conclusions

In this paper, we present a systematic investigation into the effect of the attention mechanism at different locations within the YOLOv5n for infrared maritime ship detection. The findings confirm that attention integration significantly enhances both the precision and robustness of detection models. The main conclusions are summarized as follows:

- The YOLOv5n-HN, which integrates attention modules into both the neck and head, delivers the most comprehensive performance, effectively fusing semantic and spatial representations for improved detection of medium- and large-scale targets.
- The YOLOv5n-N demonstrates clear advantages in small-object detection, suggesting that attention integration within the neck module is particularly beneficial for capturing fine-grained features.
- The attention mechanism enhances background suppression and target saliency. However, the simultaneous deployment of multiple attention modules may lead to overfitting or feature redundancy.

While this study focuses on infrared-based ship detection, other spectral bands provide complementary advantages: visible imagery delivers finer spatial detail under adequate illumination, whereas ultraviolet sensing enhances the radiometric contrast between sea and sky. Therefore, integrating ultraviolet–visible–infrared modalities offers a principled approach to achieving robustness against illumination variability, background clutter, and adverse weather. Incorporating such multimodal cues into deep learning–based detectors has the potential to improve both detection accuracy and out-of-distribution generalization in complex maritime scenes. Realizing these benefits will require the development of

the adaptive and learnable attention mechanism, as well as the judicious placement of the attention module to balance representational power with computational efficiency. Collectively, these advances are expected to yield detectors that are not only precise but also capable of real-time operation in practical maritime environments.

In addition, a further promising research direction lies in addressing dataset-related limitations, particularly class imbalance. Designing strategies to mitigate skewed data distributions could substantially enhance the detection of minority classes while reducing the risk of missed detections or misclassifications in real-world maritime applications.

Author Contributions: Conceptualization, R.Z., J.Z. and Z.Z.; Methodology, R.Z., J.Z., Z.Z. and D.Y.; Software, R.Z., J.Z. and Z.Z.; Validation, R.Z., Z.Z. and D.Y.; Formal analysis, R.Z., D.Y. and D.Z.; Investigation, R.Z., J.Z. and D.Z.; Resources, J.Z., D.Y. and Z.Z.; Data curation, R.Z., J.Z. and Z.Z.; Writing—original draft preparation, R.Z., J.Z. and Z.Z.; Writing—review and editing, R.Z., D.Y., D.Z. and J.C.; Visualization, R.Z. and Z.Z.; Supervision, J.Z., D.Y., J.C. and Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Xiamen under Grant 3502Z202571045; in part by the Natural Science Foundation of Fujian Province under Grant 2025J08203; in part by the Scientific Research Start-up Fund of Jimei University under Grant ZQ2024076; in part by the Science and Technology Special Envoy Project of Fujian Province under Grant B2025104; in part by the Open Funding of State Key Laboratory of Intelligent Coal Mining and Strata Control under Grant SKLIS202404; in part by the Fundamental Research Foundation of National Key Laboratory of Automatic Target Recognition under WDZC20255290209; in part by the National Natural Science Foundation of China under Grant 62471497; in part by the Research on the Application of Ecological Technologies in Seawall Engineering under Grant B21063.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, X.; Liu, M.; Yin, Y. Infrared Ship Detection in Complex Nearshore Scenes Based on Improved YOLOv5s. *Sensors* **2025**, *25*, 3979. [CrossRef] [PubMed]
2. Fan, X.; Hu, Z. A Small-Ship Object Detection Method for Satellite Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11886–11898. [CrossRef]
3. Xiong, W.; Xiong, Z.; Cui, Y.; Huang, L.; Yang, R. An Interpretable Fusion Siamese Network for Multi-Modality Remote Sensing Ship Image Retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2696–2712. [CrossRef]
4. Huang, L.F.; Liu, C.G.; Wang, H.G.; Zhu, Q.L.; Zhang, L.J.; Han, J.; Zhang, Y.S.; Wang, Q.N. Experimental Analysis of Atmospheric Ducts and Navigation Radar Over-the-Horizon Detection. *Remote Sens.* **2022**, *14*, 2588. [CrossRef]
5. Chen, X.; Mu, X.; Guan, J.; Liu, N.; Zhou, W. Marine Target Detection Based on Marine-Faster R-CNN for Navigation Radar Plane Position Indicator Images. *Front. Inf. Technol. Electron. Eng.* **2022**, *23*, 630–643. [CrossRef]
6. Huang, Q.; Sun, H.; Wang, Y.; Yuan, Y.; Guo, X.; Gao, Q. Ship Detection Based on YOLO Algorithm for Visible Images. *Iet Image Process.* **2024**, *18*, 481–492. [CrossRef]
7. Zhao, P.; Yu, X.; Chen, Z.; Liang, Y. A Real-Time Ship Detector via a Common Camera. *J. Mar. Sci. Eng.* **2022**, *10*, 1043. [CrossRef]
8. Zhang, T.; Shen, H.; Rehman, S.U.; Liu, Z.; Li, Y.; Rehman, O.U. Two-Stage Domain Adaptation for Infrared Ship Target Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4208315. [CrossRef]
9. Zhou, A.; Xie, W.; Pei, J. Background Modeling in the Fourier Domain for Maritime Infrared Target Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 2634–2649. [CrossRef]
10. Zhou, A.; Xie, W.; Pei, J. Infrared Maritime Target Detection Using the High Order Statistic Filtering in Fractional Fourier Domain. *Infrared Phys. Technol.* **2018**, *91*, 123–136. [CrossRef]

11. Mo, W.; Pei, J. Nighttime Infrared Ship Target Detection Based on Two-channel Image Separation Combined with Saliency Mapping of Local Grayscale Dynamic Range. *Infrared Phys. Technol.* **2022**, *127*, 104416. [CrossRef]
12. Li, L.; Liu, G.; Li, Z.; Ding, Z.; Qin, T. Infrared Ship Detection Based on Time Fluctuation Feature and Space Structure Feature in Sun-Glint Scene. *Infrared Phys. Technol.* **2021**, *115*, 103693. [CrossRef]
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
14. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37
. [CrossRef]
17. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2018**, arXiv:1708.02002. [CrossRef]
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
19. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
20. Firdiantika, I.M.; Kim, S. One-Stage Infrared Ships Detection with Attention Mechanism. In Proceedings of the 2023 23rd International Conference on Control, Automation and Systems (ICCAS), Yeosu, Republic of Korea, 17–20 October 2023; pp. 448–451. [CrossRef]
21. Ye, J.; Yuan, Z.; Qian, C.; Li, X. CAA-YOLO: Combined-Attention-Augmented YOLO for Infrared Ocean Ships Detection. *Sensors* **2022**, *22*, 3782. [CrossRef] [PubMed]
22. Miao, L.; Li, N.; Zhou, M.; Zhou, H. CBAM-Yolov5: Improved Yolov5 Based on Attention Model for Infrared Ship Detection. In Proceedings of the International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2021), Harbin, China, 24–26 December 2021;
SPIE: Bellingham, WA, USA, 2022; p. 33.
[CrossRef]
23. Khanam, R.; Hussain, M. What Is YOLOv5: A Deep Look into the Internal Features of the Popular Object Detector. *arXiv* **2024**, arXiv:2407.20892. [CrossRef]
24. Qian, L.; Zheng, Y.; Cao, J.; Ma, Y.; Zhang, Y.; Liu, X. Lightweight Ship Target Detection Algorithm Based on Improved YOLOv5s. *J. -Real-Time Image Process.* **2024**, *21*, 3. [CrossRef]
25. Guo, L.; Wang, Y.; Guo, M.; Zhou, X. YOLO-IRS: Infrared Ship Detection Algorithm Based on Self-Attention Mechanism and KAN in Complex Marine Background. *Remote Sens.* **2024**, *17*, 20. [CrossRef]
26. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7029–7038. [CrossRef]
27. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [CrossRef]
28. Zhou, J.; Jiang, P.; Zou, A.; Chen, X.; Hu, W. Ship Target Detection Algorithm Based on Improved YOLOv5. *J. Mar. Sci. Eng.* **2021**, *9*, 908. [CrossRef]
29. Giulietti, N.; Tombesi, S.; Bedodi, M.; Sergenti, C.; Carnevale, M.; Giberti, H. Hazelnut Yield Estimation: A Vision-Based Approach for Automated Counting of Hazelnut Female Flowers. *Sensors* **2025**, *25*, 3212. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.