Journal of *Marine Science and Engineering*

MDPI

Article

Research on Ship Personnel Detection Method in Complex Scenarios Based on Fusion of Attention Mechanism and Multi-Scale Perception

Ning Chen [1], Mingtao Qu [1], Zhichen Liu [2,*] and Chenzhao Bai [1]

[1] Marine Engineering College, Dalian Maritime University, Dalian 116026, China; ching_0829@dlmu.edu.cn (N.C.); qmt_0621@dlmu.edu.cn (M.Q.); baichenz@dlmu.edu.cn (C.B.)
[2] College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China
* Correspondence: zhi_chen@dlmu.edu.cn

Abstract

Ship personnel detection is the core technology for building intelligent ship monitoring systems. It safeguards navigation safety and standardizes deck work management. Current human detection algorithms perform poorly in complex marine service environments. They are easily disrupted by interfering backgrounds: blocked personnel, stacked deck equipment, and sea surface reflections. In addition, these algorithms struggle to detect tiny distant targets. Such drawbacks drastically reduce detection accuracy and stability. As a result, they cannot satisfy real-world needs for marine navigation and offshore on-site operations. To tackle the aforementioned challenges, this paper proposes a ship personnel detection method for complex scenarios based on attention mechanism and multi-scale perception fusion. By optimizing the small-target detection branch, the proposed method strengthens the capability to capture and identify long-distance operators on ship decks. Furthermore, a convolution–attention fusion module is embedded into the backbone network of the model to effectively separate personnel features from the complex marine background. In addition, depthwise separable convolution is introduced to replace conventional standard convolution, which substantially reduces model computational complexity and enables the model to meet the real-time detection requirements of ship application scenarios. Experimental results show that the improved model achieves an accuracy of 93% in tests across various ship deck scenarios. Compared with the original YOLOv8 model, its precision and recall are improved by 3% and 2%, respectively. The comprehensive performance of the proposed model is superior to that of mainstream object detection models such as Faster-RCNN and YOLOv5. It provides an efficient and reliable technical solution for personnel safety monitoring and operation management in complex ship scenarios.

Keywords: ship personnel detection; small target detection; attention mechanism fusion; depthwise separable convolution

Check for updates

Academic Editor: Jang Hyun Lee

Received: 25 May 2026
Revised: 25 June 2026
Accepted: 26 June 2026
Published: 30 June 2026

Copyright: © 2026 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.

1. Introduction

1.1. Research Background

With the improvement of ship navigation safety guarantee systems and the intelligent transformation of deck operation management modes, the standardized operation of deck personnel is directly related to ship operational efficiency and navigation safety.

According to the Annual Report on Maritime Casualties and Accidents released by the European Maritime Safety Agency (EMSA) in 2024, among thousands of recorded maritime accidents during the statistical period, approximately 64.5% can be directly attributed to human factors, and crew members account for more than 80% of all casualties involved in such accidents [1]. The above data fully demonstrates that traditional supervision modes dominated by manual patrol and passive monitoring can no longer meet the practical demands of full-time-domain risk early warning for ships. Constructing an intelligent monitoring system based on computer vision technology and realizing real-time perception of on-duty deck personnel as well as dynamic risk management through automatic means is of urgent significance for reducing the rate of maritime accidents induced by human factors and promoting the digital transformation and upgrading of the shipping industry [2].

Although generic object detection algorithms represented by Faster-RCNN and the YOLO series have achieved mature applications in terrestrial security surveillance, directly migrating these algorithms to the complex and variable ship operation scenarios still faces severe technical challenges. In the practical deployment and application of existing detection models in ship scenarios, there exist three major technical bottlenecks. First, severe interference from complex environmental backgrounds. Widespread equipment occlusion in deck areas and light and shadow noise caused by sea wave reflections readily lead to false detection and missed detection of detection algorithms [3]. Second, the difficulty in identifying distant small targets. Under wide-angle monitoring views, deck personnel occupy an extremely low proportion of pixels, and their critical detailed features are prone to information loss during network downsampling [4]. Third, the inherent contradiction between detection accuracy and real-time performance. Although sophisticated deep learning models can effectively boost detection accuracy, they are difficult to efficiently deploy and support real-time inference on shipborne edge devices with limited computing resources. In response to the above technical bottlenecks in maritime scenarios, in-depth research is currently being conducted in academic circles focusing on personnel detection and perception challenges in special working environments.

To sum up, in view of the aforementioned technical bottlenecks in complex ship environments, conducting targeted research on personnel detection and recognition technologies tailored for ship scenarios can not only break through the perception limitations of existing intelligent monitoring systems but also serve as a core link in improving the intrinsic safety level of ships and advancing the intelligent development of the shipping industry. This research possesses prominent theoretical significance and broad application prospects [5].

1.2. Literature Review Analysis

The continuous advancement of object detection technology provides crucial technical support for high-precision detection of ship personnel. Its development is clearly manifested as a paradigm shift from traditional algorithms to deep learning-based methods. Meanwhile, targeted optimization has been gradually realized to improve the adaptability of detection algorithms in complex ship scenarios. In the traditional research stage, the HOG + SVM pedestrian detection algorithm proposed by Navneet Dalal in 2005 served as a representative method [6]. It constructed a basic framework for pedestrian detection through gradient feature description and realized preliminary object recognition. In 2009, Professor Felzenszwalb proposed the Deformable Part Model (DPM), which optimized the processing of personnel occlusion and further expanded the applicable scenarios of human detection [7]. In the same year, the team led by Deng J provided critical data support for the application of deep learning in the field of object detection [8], laying an essential foundation for subsequent technological iteration. In 2012, Alex Krizhevsky et al. proposed

a large-scale deep convolutional neural network [9], which greatly reduced the error rate of visual recognition. This milestone marked the official establishment of deep learning as the core technical direction for object detection and opened up a brand-new technical route for ship personnel detection. In 2015, Ren Shaoqing proposed the Faster-RCNN algorithm [10], which improves detection accuracy via the Region Proposal Network (RPN) and greatly promotes the maturity of two-stage object detection algorithms. In 2016, Joseph Redmon proposed the YOLO algorithm, pioneering the one-stage real-time detection paradigm [11]. Its high-efficiency detection performance provides strong technical support for complex application scenarios such as ship environments. Thereafter, the YOLO series has undergone continuous iterative updates. Representative models such as YOLOv5 and YOLOv8, which are equipped with dual processing channels and feature fusion mechanisms, have achieved steady improvements in both real-time performance and detection accuracy. In particular, the dual processing channels and feature fusion strategy adopted by YOLOv8 lay a solid foundation for adaptive perception in dynamic and complex ship scenarios.

In the specialized field of ship personnel detection, with the growing demand for intelligent shipping, the research focus has gradually shifted from the adaptation of general algorithms to scenario-oriented dedicated optimization. Early studies mostly adopted general object detection algorithms directly to address the task of personnel detection in ship scenarios. However, such methods struggle to cope with environmental interferences, including stacked deck equipment and sea wave reflections, as well as the technical challenges of recognizing distant personnel in small-target detection within shipborne environments. Subsequently, Qiang H. conducted targeted research on small target detection in 2023 [12] and developed improved YOLOv8 algorithms for small-target detection. In 2020, Li Y. et al. investigated the integration of attention mechanisms [13] and explored the applications of attention-based convolutional neural networks in attribute recognition. As the core technical pillar in the field of object detection, deep learning methods have witnessed continuous breakthroughs centered on accuracy improvement and scenario adaptation, thereby providing diversified technical solutions for crew detection tasks. Joseph Redmon et al. proposed the YOLO9000 model to realize real-time multi-category object detection. On the basis of the original YOLO framework, this model expands detection categories and enhances detection accuracy [14], which provides novel insights for multi-object collaborative detection in complex ship scenarios. Yang Z et al. proposed the RepPoints detection framework to optimize object localization. This method adopts a point set to represent object regions and strengthens the capture of contour features of human targets, which can well adapt to the characteristics of variable human postures in ship scenarios [15]. Law Hei et al. proposed CornerNet, which transforms object detection into keypoint detection. Innovatively, it realizes human detection through corner point pairing and effectively improves the detection accuracy under occlusion conditions. This method offers valuable references for solving the detection difficulty caused by mutual occlusion of personnel during ship operations [16]. Technical breakthroughs in the detection and recognition of small-scale targets onboard ships mainly focus on feature enhancement and network structure optimization. The research team led by Fang Yan from Yunnan University proposed the YOLOv7-BAMFF model to boost the detection accuracy of small targets. By constructing a bidirectional adaptive multi-scale weighted feature fusion module, this model incorporates fine-grained low-level features into the feature fusion process and adds an additional detection head dedicated to small targets. Its effectiveness for small target detection has been verified on the VisDrone2019 dataset, offering a direct technical solution for the detection of remote and tiny crew targets in shipborne environments [17]. Wang J et al. proposed a lightweight real-time detection method named YOLOv5s-SGC. Improved on the basis of YOLOv5s, this algorithm optimizes the backbone network and

feature fusion network, and embeds the CBAM attention module. It can be well adapted to shipborne devices with limited computing resources and realizes the detection of deck crew and fishing net operation behaviors [18]. Li Z et al. proposed ACD-Net, an abnormal crew detection network tailored for complex ship scenarios. Adopting a two-stage detection strategy, this network achieves real-time detection and identification of abnormal crew members. Combined with the improved YOLO-TRCA model and the CFA recognition algorithm [19], it effectively enhances detection accuracy and identity matching rate. Tan Mingxing et al. designed a weighted bidirectional feature pyramid structure based on the EfficientDet model [20], which strengthens the transmission and fusion of small target features. Its core optimization idea provides valuable guidance for the performance improvement of small target detection models in ship scenarios. Murad et al. presented a face recognition system based on ship closed-circuit television (CCTV) cameras [21]. This system is applied to monitor crew working hours and detect intruders, thereby addressing the authenticity verification of working hour records. In summary, scenario-based research on ship personnel detection mainly concentrates on environmental adaptation and performance optimization. Sun Yuxiang et al. proposed an improved maritime object detection algorithm based on YOLOv7 [22]. Aiming at the typical challenges in ship scenarios, such as sea wave reflection and deck shadow interference, this algorithm optimizes the feature enhancement module of the network and improves the model's adaptability to complex marine environments. Han Y et al. proposed an improved multi-object detection model for ship scenarios based on YOLOv8. Taking YOLOv8n as the baseline, the model integrates the ESSE module, GSConv technology and Wise-IoU to enhance the capabilities of multi-scale feature extraction and target detection under complex backgrounds [23], achieving a substantial improvement in detection accuracy. Wu et al. proposed an improved YOLOv7 model, which optimizes the anchor box design and introduces a multi-scale feature fusion module [24]. Combined with data augmentation and random sampling strategies, the model improves the detection accuracy and robustness in ship scenarios. The research focus of attention mechanism integration lies in precise feature screening and background suppression. Hu Jie et al. proposed the Squeeze-and-Excitation (SE) module to implement the channel attention mechanism. By adaptively adjusting channel weights, this module enhances critical feature expression, providing an effective tool for separating human features from complex backgrounds in ship scenarios [25]. In addition, research on depthwise separable convolution focuses on network lightweighting and efficiency improvement, which has become a key technology for real-time detection tasks on ships. Howard Andrew G et al. proposed the MobileNets series, which adopts depthwise separable convolution and systematically established the application framework of depthwise separable convolution for the first time [26], laying a fundamental foundation for network model lightweighting. Sandler Mark et al. proposed MobileNetV2 to optimize the structure of depthwise separable convolution. By designing the inverted residual structure and linear bottleneck, it further improves the feature extraction capability and computational efficiency of depthwise separable convolution [27]. Zhang Xiangyu et al. proposed ShuffleNet, which combines grouped convolution with depthwise separable convolution to drastically reduce computational complexity. Its lightweight design concept [28] has been widely applied to the model optimization of embedded detection systems in ship environments. Zhang et al. proposed a high-speed SAR ship detection method based on a depthwise separable convolutional neural network (DS-CNN). By integrating multi-scale detection, feature splicing and anchor box mechanisms, a lightweight architecture is constructed [29]. This method effectively boosts detection speed and meets the requirements of real-time application while guaranteeing competitive detection accuracy. In addition, temporal keyframe-based representations have been explored to improve recognition performance.

Poomhiran et al. proposed a Concatenated Three Sequence Keyframe Image technique that fuses information from multiple keyframes to enhance recognition accuracy, offering insights for future video-based ship personnel perception tasks [30].

Unlike studies that directly apply general object detection models to shipboard scenes, this work focuses on adapting YOLOv8n for shipboard personnel detection. Its contribution lies in integrating a small-object detection head, a convolution–attention fusion (CAF) module, and depthwise separable convolution (DSConv) into a unified framework for complex shipboard environments. The proposed design enhances distant-personnel detection, improves robustness to occlusion and background interference, and reduces computational cost, thereby achieving better accuracy and real-time performance in shipboard scenarios.

2. Methods

2.1. YOLOv8 Algorithm

The YOLO series of detection algorithms occupies a crucial position in the field of object detection. Benefiting from superior real-time detection performance, it can complete object detection while maintaining high precision, and is particularly suitable for personnel detection tasks with stringent real-time requirements in ship scenarios. Such characteristics are highly consistent with the dual demands for detection speed and accuracy in dynamic deck operations and complex marine environments. As a newly proposed model in the YOLO series, YOLOv8 adopts an innovative dual-processing channel structure that differs from previous algorithms. Convolutional networks serve as bridges to connect the two channels closely, and the lightweight network design effectively reduces memory consumption during operation. By establishing correlation relationships between feature maps, YOLOv8 realizes synchronous convolution operation and efficient output transmission. Furthermore, integrated with the feature pyramid architecture, the model processes feature map information layer by layer in a progressive manner, enabling adaptive detection of objects at various scales and further improving computational efficiency. In this paper, the YOLOv8 network model is adopted to investigate the personnel detection algorithm for ships under complex marine scenarios.

The overall architecture of YOLOv8 is illustrated in Figure 1, which mainly consists of four core components: the Input layer, Backbone layer, Neck layer and Head layer. The Input layer is responsible for feeding images to be detected into the network. The Backbone layer extracts shallow and deep feature information from input images. The Neck layer undertakes feature pooling and multi-scale feature fusion of the extracted feature maps. Finally, the Head layer completes feature decoding and outputs the final detection results.

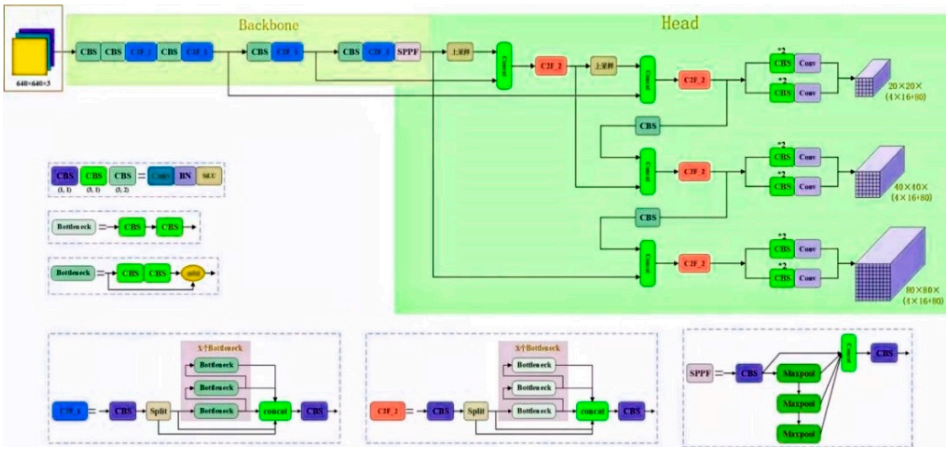


Figure 1. Network structure of the YOLOv8 model.

2.2. Ship Personnel Detection Method

Aiming at the challenges in ship scenarios, such as complex background interference and difficult recognition of remote small targets, this study conducts optimization research on ship personnel detection based on the YOLOv8 baseline model. With its distinctive feature fusion mechanism, multi-scale perception capability and lightweight architecture, YOLOv8 achieves a favorable balance between real-time performance and detection accuracy, which highly meets the dual requirements of dynamic deck operation and complex marine environments. In this research, attention mechanism, multi-scale perception and lightweight design concepts are integrated. Through multi-dimensional structural optimization, the scenario adaptability of the model is strengthened to accurately improve the performance of ship personnel detection, thereby constructing an efficient and reliable detection framework for marine application scenarios.

(1) Optimization of Small-Target Detection Head Based on Feature Fusion

Small-target detection has long been a core difficulty and technical challenge in the field of computer vision due to inherent characteristics such as tiny object scales and low pixel occupancy. As a mainstream single-stage object detection model, YOLOv8 has been widely applied in various visual tasks by virtue of its high inference efficiency and comprehensive detection performance. However, under the standard architecture, its capability for detecting small targets, such as distant deck workers in marine scenarios, still leaves considerable room for improvement. To address this technical bottleneck, this paper proposes an improved algorithm based on detection head optimization for model architecture enhancement. By adding a dedicated small-target detection head, the feature extraction and recognition capabilities of the model for tiny and distant personnel targets in ship scenarios are effectively strengthened, thereby significantly improving the detection accuracy of small-scale crew objects in complex marine environments. The structural details of the proposed small-target detection head are presented in Figure 2. To quantify the effectiveness of small-target feature fusion, a small-target feature enhancement indicator η is defined to evaluate the improvement of representation capability following the fusion of low-level and high-level features.

$$\eta = \frac{\sum_{i=1}^{N} \frac{\|F_{\text{fusion},i}\|_2}{\|F_{ori,i}\|}}{N} \tag{1}$$

where N is the number of small target samples, $F_{\text{fusion},i}$ denotes the feature vector of the i-th small target after feature fusion, $F_{ori,i}$ represents the original high-level feature vector, and $\|\cdot\|_2$ indicates the L2 norm. It should be noted that η is not an additional loss term or an independent training objective. Instead, it is used as an auxiliary quantitative indicator to describe the degree of small-target feature enhancement after feature fusion. A larger value of η indicates that the fused feature representation contains stronger small-target-related information compared with the original high-level feature.

The weighted summation mechanism is adopted in the feature fusion process of the small-target detection head. The calculation formula for the fused feature Ffusion is expressed as follows:

$$F_{\text{fusion}} = \alpha \cdot F_{low} \oplus \beta \cdot F_{high} \tag{2}$$

where F_{low} denotes the bottom C2f feature map extracted from the Backbone layer with a dimension of $H \times W \times C_{\text{low}}$, and F_{high} represents the original high-level feature map with a dimension of $H \times W \times C_{\text{high}}$. The weight coefficients η and γ are learnable indicators rather than fixed empirical hyperindicators. During training, they are initialized with equal contributions and then automatically updated through backpropagation together with

other network indicators. To ensure stable feature fusion, the two weights are normalized before feature aggregation so that their relative contributions to low-level fine-grained features and high-level semantic features can be adaptively adjusted according to the input feature distribution. In this way, the small-object detection head can dynamically balance detailed spatial information and semantic representation during network training.

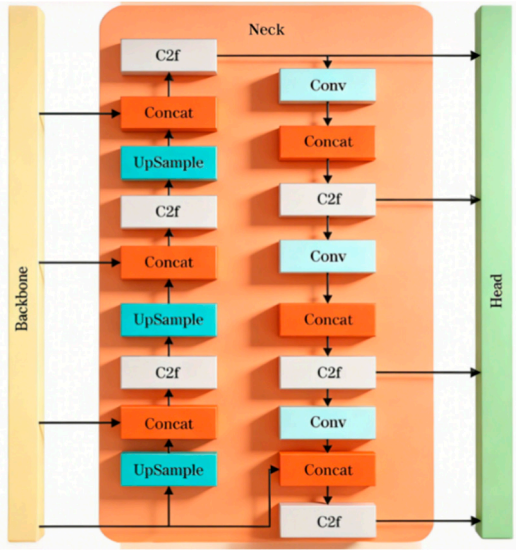


Figure 2. Structure of the small-target detection head.

To address the feature flooding problem caused by the low pixel proportion of small targets, a small-target pixel contribution indicator γ is defined:

$$\gamma = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} I(x,y)\cdot M(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W} M(x,y)} \tag{3}$$

where I(x, y) denotes the pixel value of the image at coordinate (x, y), and M(x, y) represents the small target mask, in which the target region is assigned to 1 and the background region is set to 0. Similarly, γ is used to quantify the pixel contribution of small-target regions in the input image. It is calculated according to the small-target mask and the corresponding pixel distribution, and is mainly used to explain the necessity of introducing fine-grained low-level features for distant personnel detection. Therefore, η and γ serve as explanatory indicators for the small-object detection head rather than direct optimization terms in the loss function.

Due to the size limitation and low pixel occupancy of small targets, their low-level features are easily submerged or blurred, which eventually weakens the object representation capability in feature maps. In conventional object detection tasks, YOLOv8 implements feature information encoding and output through three detection heads arranged in the Neck layer, which can satisfy the detection requirements of most general scenarios, such as pedestrian recognition. However, in complex scenarios with a high proportion of small targets, such as dense crowds and long-distance shooting, the inherent detection head structure fails to achieve effective perception and recognition of small objects, thereby revealing its deficiency in capturing fine-grained small-target features. Focusing on the structural improvement of the Neck layer, this paper enhances the model's ability to extract fine-scale image features by adding a dedicated small-target detection head, so as to compensate for the performance shortcomings of the original framework in small-object

detection tasks. The core optimization idea of the designed small-target detection head is to introduce the bottommost C2f feature map in the Backbone layer into the feature fusion process. By constructing a fusion mechanism between high-dimensional low-level features and original high-level features, detailed feature information of small targets can be fully mined. This strengthens the richness and discriminability of feature representation and further improves the accurate detection capability of the network for small targets.

(2) Improvement Based on Convolution–Attention Mechanism Fusion

Different from images used in conventional object detection tasks, the real-time monitoring images of ship personnel contain highly complex scene information, which involves numerous environmental elements and background targets irrelevant to personnel detection. These irrelevant elements cannot provide valid feature information for ship personnel detection and recognition; instead, they act as background interference during algorithm operation. Therefore, suppressing and eliminating such redundant interference features serves as an essential prerequisite and critical foundation for the optimization of the proposed ship personnel detection algorithm. To enhance the model's detection capability under occlusion conditions, a convolution and attention fusion module is embedded into the C2f module of the Backbone layer. Convolutional operations are restricted by their inherent properties and receptive field limitations, making them incapable of comprehensive feature aggregation across the entire feature map, while achieving favorable performance only within localized regions. In contrast, Transformer-based attention mechanisms exhibit superior capacity for global feature modeling and holistic contextual generalization. The integration of convolution and attention mechanisms compensates for the inherent limitations of each individual technique, enabling the collaborative extraction and aggregation of both local and global feature information. Accordingly, Hu et al. proposed the CAF module to optimize detection performance under complex background conditions for pedestrian recognition. Within this module, the embedded attention mechanism is activated to capture long-range global contextual information, while convolutional operations are adopted to extract fine-grained local features. The overall structure of the CAF module is illustrated in Figure 3.

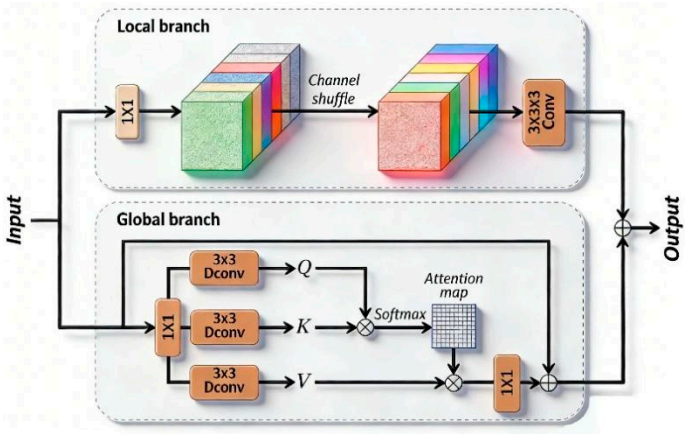


Figure 3. Structure of the convolution–attention fusion module.

The CAF module consists of two components: a global branch and a local branch. In the local branch, convolution is first adopted to adjust the channel dimension. Convolution enables flexible adjustment of channel numbers without changing the width and height of the feature map, which facilitates better feature processing in subsequent operations. Afterwards, a channel shuffling operation is performed. The input tensor is divided into multiple groups along the channel dimension, and depthwise separable convolution is

applied within each group to promote channel shuffling. The output tensors of all groups are then concatenated along the channel dimension to generate a new tensor. This operation further mixes and fuses channel information and enhances cross-channel interaction and information integration. Finally, convolution is utilized to complete local feature extraction. In the global branch, 1×1 convolution and 3×3 depthwise convolution are first employed to generate the query vector Q, key vector K, and value vector V with specific dimensions. Subsequently, Q and K are reshaped respectively and interactively computed to obtain the attention map. This calculation strategy effectively reduces the computational overhead.

The global branch of the CAF module adopts a multi-head self-attention mechanism. The output Ak of the k-th attention head is calculated as follows:

$$A_k = Softmax\left(\frac{Q_k K_k^T}{\sqrt{d_k}}\right) V_k \tag{4}$$

where Q_k, K_k, and V_k denote the query, key, and value vectors of the k-th attention head, which are generated by 1×1 convolution and 3×3 depthwise convolution. dk = Cin/K represents the dimension of a single attention head, Cin is the number of input channels, and K is the total number of attention heads. An adaptive gating mechanism is used for the fusion of the global and local branches, and the fused feature FCAF is defined as follows:

$$F_{CAF} = \sigma\left(W_g F_{global} + b_g\right) \odot F_{global} + (1 - \sigma(W_l F_{local} + b_l)) \odot F_{local} \tag{5}$$

where F_{global} denotes the output of the global branch and F_{local} denotes the output of the local branch. W_g and W_l are weight matrices, bg and bl are bias terms, σrepresents the Sigmoid activation function, and $\odot$ indicates element-wise multiplication. This mechanism can adaptively adjust the contribution ratio of global and local features in accordance with input feature distributions. Before adaptive fusion, the outputs of the global and local branches are aligned to the same tensor size, den $otedas\left(F_g, F_l \in \mathbb{R}^{B \times C \times H \times W}\right)$, where (B), (C), (H), and (W) represent the batch size, channel number, feature height, and feature width, respectively. The gating weights generated by (W_g), (W_l), (b_g), and (b_l) have the same spatial and channel dimensions as the input features, ensuring that the element-wise multiplication and feature fusion in Formula (5) are dimensionally consistent. To quantify the background noise suppression performance of the CAF module, a background noise suppression indicator ξ is defined:

$$\xi = 1 - \frac{Var\left(F_{CAF}^{bg}\right)}{Var\left(F_{ori}^{bg}\right)} \tag{6}$$

where F_{ori}^{bg} represents the background region features of the original feature map, and F_{CAF}^{bg} denotes the background region features output by the CAF module. $Var(\cdot)$ is the variance calculation function. A larger value of ξ indicates a smaller variance of background noise and a more prominent suppression effect. The average value of ξ reaches 0.72 in the experiments. A larger value of ξ indicates a smaller variance of background noise and a more prominent suppression effect. In the experiments, ξ was calculated over the background regions of the test images by comparing the background feature variance before and after the CAF module. The mean value of ξ was 0.72, with a standard deviation of 0.08 and a value range of 0.56~0.84. This indicates that the CAF module can consistently suppress background feature fluctuations in most test samples, thereby improving the discriminative representation of personnel targets under complex shipboard backgrounds.

(3) Improvement of Depthwise Separable Convolution

In the YOLOv8 model, key feature maps are generated through standard convolution operations and then transmitted to subsequent convolutional layers for further computation. However, this process produces a large number of network parameters, which reduces the detection speed of personnel on ship decks. In the lightweight design of object detection models, to address the computational redundancy of traditional standard convolution, this paper introduces depthwise separable convolution based on the YOLOv8 architecture to guarantee the real-time performance of ship personnel detection. As an optimized convolution pattern, depthwise separable convolution separates spatial convolution from channel convolution, thereby greatly reducing the parameter quantity in convolutional computation. It not only greatly reduces the number of network parameters, but numerous studies have demonstrated that the sacrifice in model accuracy is nearly negligible. Accordingly, depthwise separable convolution is regarded as a highly efficient convolution operation. Theoretical research indicates that replacing standard convolution with DSConv in the model can reduce network parameters by more than half, effectively realizing model lightweighting.

The structure of DSConv is illustrated in Figure 4. Firstly, depthwise convolution is performed independently on each channel to process image features via filters (i.e., convolution kernels). Subsequently, the output results of all channels are concatenated sequentially, and pointwise convolution is implemented on the concatenated features. This approach effectively reduces computational complexity and significantly improves computational efficiency.

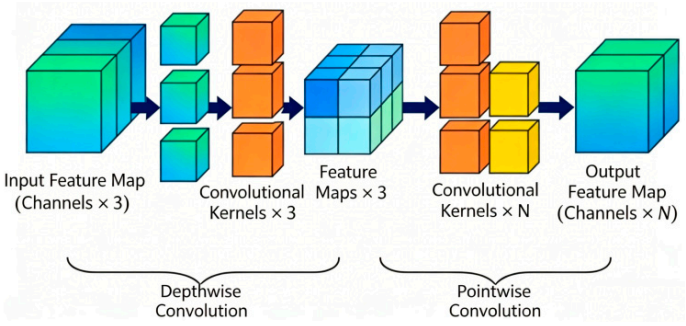


Figure 4. Structure of DSConv.

The computational compression ratio ρ of depthwise separable convolution (DSConv) serves as a core indicator for measuring model lightweight performance, which is defined as the ratio of the computational cost of standard convolution to that of DSConv:

$$\rho = \frac{H \times W \times C_{in} \times C_{out} \times K_s^2}{H \times W \times C_{in} \times K_s^2 + H \times W \times C_{in} \times C_{out}} = \frac{C_{out} \times K_s^2}{K_s^2 + C_{out}} \quad (7)$$

where Ks denotes the kernel size (set to 3 in the experiment), Cin represents the number of input channels, and Cout is the number of output channels. When Cout = 64, $\rho \approx 8.7$, indicating that the computational cost of DSConv is merely 11.5% of that of standard convolution.

To balance detection speed and accuracy, a speed-accuracy trade-off indicator ω is defined:

$$\omega = \frac{FPS_{improved}}{FPS_{ori}} \times \sqrt{\frac{mAP@0.5_{\text{improved}}}{mAP@0.5_{ori}}} \quad (8)$$

where $FPS_{improved}$ and $mAP@0.5_{improved}$ represent the frame rate and mean average precision of the improved model, while FPS_{ori} and $mAP@0.5_{ori}$ denote the corresponding

indicators of the original model. In the experiment, $\omega = 1.23$, which demonstrates that the improved model achieves coordinated enhancement in both detection speed and accuracy. It should be noted that Formulas (1), (3), (6) and (8) are not part of the training loss and do not participate in gradient optimization. Instead, they serve as auxiliary analytical indicators for explaining and analyzing the effects of the proposed modules. Specifically, Formula (1) quantifies small-target feature enhancement, Formula (3) measures small-target pixel contribution, Formula (6) evaluates background noise suppression by the CAF module, and Formula (8) reflects the trade-off between detection accuracy and inference speed. Model training still follows the original YOLOv8 loss, including bounding-box regression, classification, and distribution focal losses.

3. Multidimensional Simulation and Result Analysis

3.1. Introduction to Simulation Environment and Dataset

To support the ship personnel detection task, a ship-deck personnel dataset containing 11,000 images was constructed and organized in a VOC-style data format. In this study, the term "PASCAL VOC" refers to the data organization and annotation format rather than the direct use of the standard PASCAL VOC benchmark dataset. The ship-related samples were formed by screening images from ship deck operations, shipboard monitoring views, and other shipboard working scenarios, so that the dataset could better reflect typical marine and shipboard visual conditions. The images involve deck operators and shipboard personnel in ship-related scenarios, covering typical visual challenges such as deck facilities, ship structures, partial occlusion, distant small-scale targets, shadow interference, and background clutter. Personnel targets were manually annotated using LabelImg, and the annotations were converted into YOLO-compatible TXT label files for network training. Before training, all images were uniformly resized to 640 × 640 pixels, so as to evaluate the effectiveness and adaptability of the proposed model for personnel detection in complex shipboard environments.

All 11,000 images in the dataset are shipboard or marine-scene images, and the proportion of marine-related scenes is therefore 100%. The dataset was divided into training, validation, and test sets at a ratio of 8:1:1, corresponding to 8800, 1100, and 1100 images, respectively. To further clarify the dataset composition, the detailed distributions of weather/illumination conditions, occlusion levels, and target sizes are summarized in Table 1.

Table 1. Detailed dataset distribution.

Statistic	Class 1	Class 2	Class 3	Class 4
Weather/illumination	Normal daylight: 5280 (48.0%)	Shadow/reflection: 4290 (39.0%)	Low-light/nighttime: 880 (8.0%)	Rain/fog/low-visibility: 550 (5.0%)
Occlusion level	No occlusion: 8400 (50.0%)	Partial occlusion: 6552 (39.0%)	Severe occlusion/overlap: 1848 (11.0%)	—
Target size	Small: 7392 (44.0%)	Medium: 6216 (37.0%)	Large: 3192 (19.0%)	—

Note: Weather/illumination statistics were counted at the image level, while occlusion level and target size were counted at the annotated personnel-instance level. Occlusion level was determined according to the visible proportion of the personnel target, and target size was determined according to the bounding-box area after resizing each image to 640 × 640 pixels.

To improve reproducibility, data augmentation and sample distribution settings are clarified. During training, random scaling, horizontal flipping, color perturbation, and Mosaic augmentation were applied to improve robustness to scale variations, illumination changes, and complex backgrounds. As this study focuses on single-class personnel detection, no semantic class imbalance exists; instead, variations mainly arise from target

scale, occlusion, and scene complexity. The dataset therefore includes distant small targets, partially occluded personnel, and complex shipboard backgrounds under typical deck conditions, including uneven illumination, shadows, sea-surface reflections, and interference from ship structures and deck facilities.

To further improve the reproducibility of the study, the experimental procedure and key parameter settings were reorganized and summarized in Table 2.

Table 2. Experimental procedure and reproducibility settings.

Item	Setting
Preprocessing	Images were annotated using LabelImg v1.8.6 and converted into YOLO-compatible TXT labels. The dataset was split into training, validation, and test sets at a ratio of 8:1:1. All images were resized to 640 × 640 pixels, and random scaling, horizontal flipping, color perturbation, and Mosaic augmentation were applied during training.
Training settings	YOLOv8n was used as the baseline model. The improved model was trained with the same dataset split, input size, hardware/software environment, optimizer, batch size, epochs, and evaluation metrics as the comparison and ablation models.
Hyperparameters	The SGD optimizer was used with a batch size of 32, 120 training epochs, an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005. During inference, the confidence threshold and NMS IoU threshold were set to 0.25 and 0.45, respectively.
Evaluation protocols	All models were evaluated on the same test set. Precision, recall, F1-score, and mAP@0.5 were used as accuracy metrics, while FPS, parameters, FLOPs, GPU memory usage, and single-image inference time were used to evaluate computational efficiency. False detections and missed detections were counted at the annotated personnel-instance level.

Considering the practical application requirements comprehensively, the hardware and software configurations adopted in the experiments are presented in Table 3.

Table 3. Experimental Environment Specifications.

Environment Component	Configuration Parameter
Operating System	Microsoft Windows 11 Professional Edition
CPU	AMD Ryzen 7 5800H with Radeon Graphics
GPU	GeForce RTX 3070
Video Memory	8192 MiB
Programming Language	Python 3.9
Framework	PyTorch 1.13.1

3.2. Analysis of Simulation Results

3.2.1. Convergence Analysis

To enhance the performance of YOLOv8 in ship personnel detection and recognition, targeted training and optimization are conducted on the YOLOv8n model in this paper. The specific improvement strategies include adding a small-object detection head, introducing convolution and attention fusion modules, and embedding depthwise separable convolution. The model adopting the above combined optimization strategies is defined as the YOLOv8n-improved model.

To verify the performance of the improved model for ship personnel detection and recognition in complex scenarios, this paper conducts a comparative analysis of the training convergence processes between the original YOLOv8n baseline model and the proposed

YOLOv8n-improved model. The loss curves of bounding box loss, classification loss, and distribution focal loss for the two models are presented in Figure 5 and Figure 6, respectively. Meanwhile, the training results further illustrate the convergence trends of the two models in terms of four core evaluation metrics, including precision, recall, mAP@0.5, and mAP@[0.5:0.95].

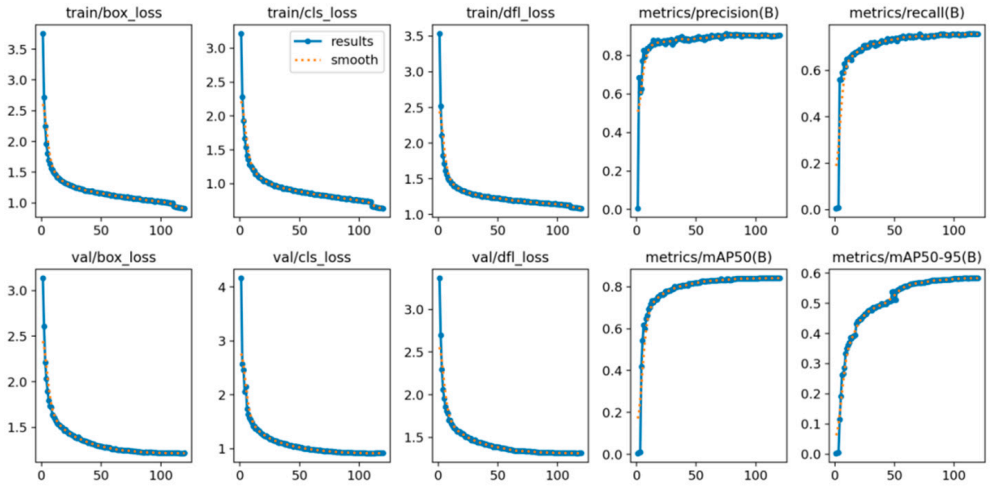


Figure 5. Convergence curves of YOLOv8n.

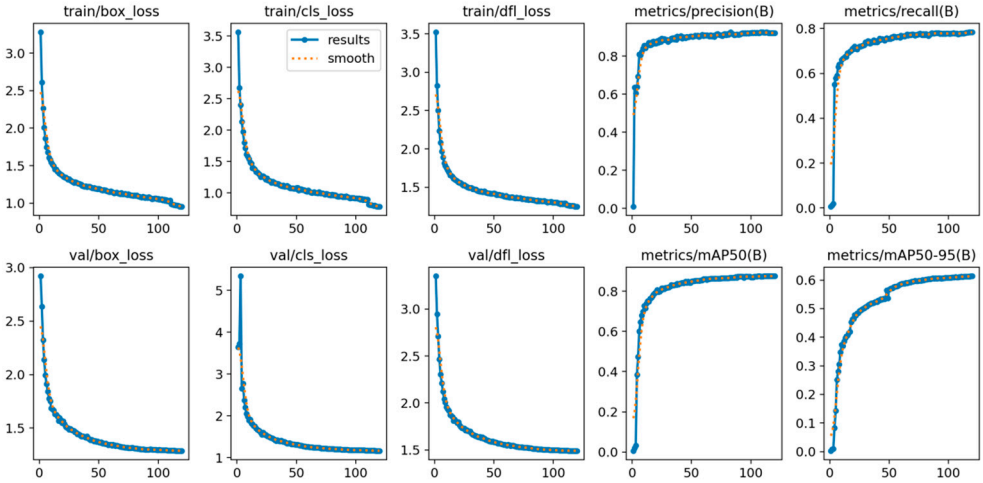


Figure 6. Convergence curves of YOLOv8n-improved.

By observing the loss curves of the two models on the training set and validation set of the VOC dataset, it can be found that the loss values of both the YOLOv8n model and the YOLOv8n-improved model remain at a low level. Compared with the baseline YOLOv8n model, the YOLOv8n-improved model optimized for ship personnel detection and recognition exhibits a significant difference in distribution focal loss, with its loss value reduced remarkably. From the curve comparison of the four performance evaluation metrics, it can be observed that the YOLOv8n-improved model achieves superior performance in two core indicators, namely mAP@0.5 and mAP@[0.5:0.95], with correspondingly smoother and more stable convergence curves. The above experimental results fully demonstrate that the YOLOv8n-improved model possesses better convergence performance for ship personnel detection tasks.

3.2.2. Comparative Experiments

To verify the superiority of the improved YOLOv8n model, comparative experiments are conducted with several mainstream detection models, including Faster R-CNN and

YOLOv5, and the quantitative results are summarized in Table 4. In this experiment, four evaluation metrics are adopted, namely Mean Average Precision (mAP), F1-Score, precision and recall. Among them, mAP@0.5 is selected as the comprehensive indicator to evaluate the overall performance of different models.

Table 4. Performance Evaluation Metrics Comparison of Different Models.

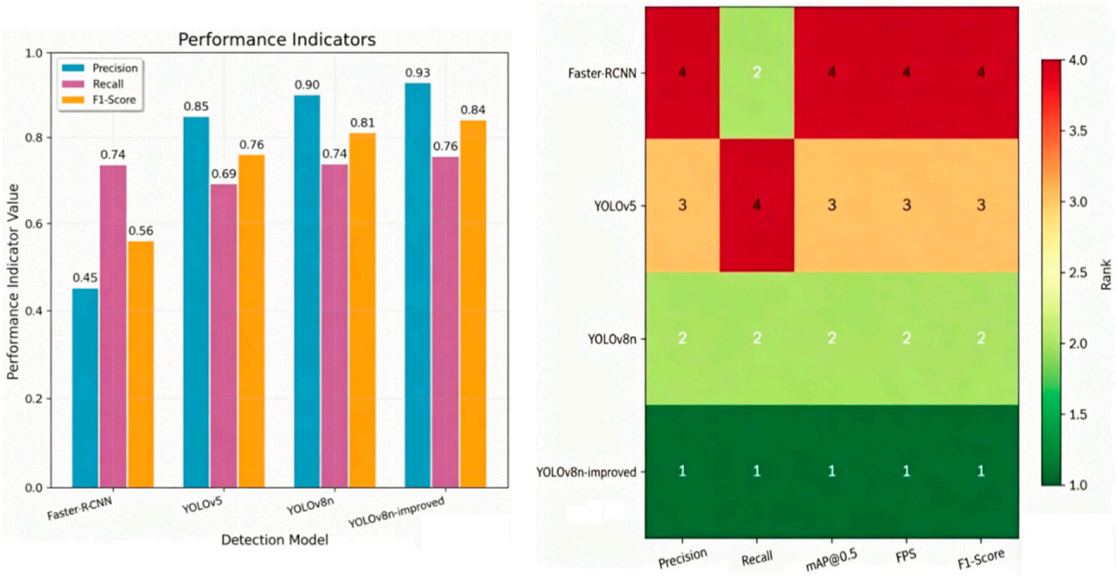
Model	Average Precision	Average Recall	mAP@0.5	FPS	F1-Score
Faster-RCNN	0.45	0.74	0.68	10.3	0.56
YOLOv5	0.85	0.69	0.80	25.6	0.76
YOLOv8n	0.90	0.74	0.85	38.4	0.81
RT-DETR	0.91	0.75	0.88	28.7	0.82
YOLOv9	0.91	0.75	0.87	35.2	0.82
YOLOv10	0.92	0.75	0.88	41.5	0.83
YOLOv11	0.92	0.76	0.89	43.2	0.83
DINO	0.92	0.76	0.89	12.8	0.83
Gold-YOLO	0.91	0.75	0.88	31.4	0.82
YOLOv8n-impreved	0.93	0.76	0.90	46.9	0.84

From the comparative evaluation results, it can be observed that the YOLOv8n-improved model achieves performance improvements of 0.48, 0.02, 0.22, 36.6 and 0.28 in precision, recall, mAP@0.5, FPS, and F1-score, respectively, compared with the Faster R-CNN model. When compared with the YOLOv5 model, the corresponding improvements are 0.08, 0.07, 0.10, 21.3, and 0.08. In contrast to the original YOLOv8n model, the proposed model yields increments of 0.03, 0.02, 0.05, 8.5, and 0.03 across the five metrics. In summary, the YOLOv8n-improved model exhibits the best overall performance among the four models. In addition to detection accuracy, the computational complexity of the proposed model was further analyzed. Under the same hardware environment and an input image size of 640 $\times$ 640, the number of parameters of YOLOv8n-improved was reduced from 3.16 M to 2.94 M compared with the original YOLOv8n model, and the FLOPs decreased from 8.9 G to 7.6 G. The GPU memory usage during inference was also reduced from 620 MB to 570 MB. Meanwhile, the processing time per image decreased from 26.0 ms to 21.3 ms, corresponding to an increase in FPS from 38.4 to 46.9. These results indicate that the introduction of DSConv effectively reduces computational cost and improves inference efficiency, thereby supporting the real-time deployment potential of the proposed model in shipboard monitoring scenarios. Figure 7 presents the performance evaluation comparison and ranking results of the proposed YOLOv8n-improved method against other competing models, further confirming that the proposed method achieves the optimal performance.

In addition to the comparison with mainstream detection models, ablation experiments were conducted to quantify the contribution of each improved component. The small-object detection head, CAF module, and DSConv module were introduced separately and in combination under the same dataset partition, training, and evaluation metrics. The ablation results are shown in Table 5.

The comparative results show that modern detectors, including RT-DETR, YOLOv9, YOLOv10, YOLOv11, DINO, and Gold-YOLO, generally outperform earlier models such as Faster-RCNN and YOLOv5. Among them, YOLOv11 and DINO achieve competitive mAP@0.5 values of 0.89, while DINO exhibits lower inference speed and RT-DETR and Gold-YOLO require higher computational cost. In comparison, the proposed YOLOv8n-improved model achieves the highest mAP@0.5 (0.90) and FPS (46.9), along with a precision of 0.93 and an F1-score of 0.84. These results demonstrate that the proposed model provides

a better balance between detection accuracy and real-time performance, making it well suited for shipboard personnel monitoring under limited computational resources.



(**a**) (**b**)

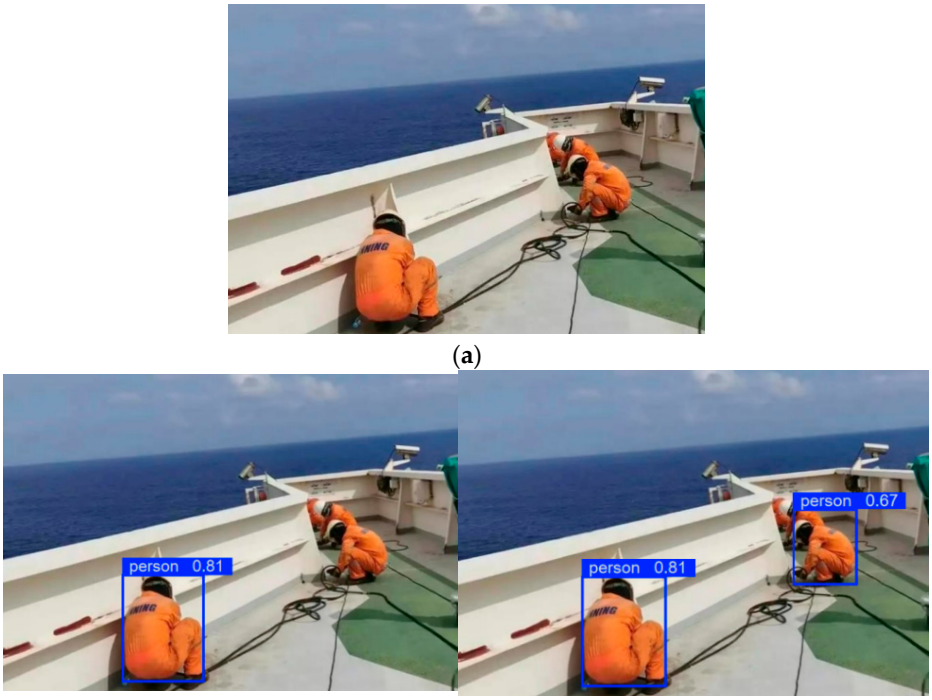
Figure 7. Comparison of results between the improved YOLOv8n and other models: (**a**) Performance evaluation comparison; (**b**) Ranking comparison results.

Table 5. Ablation results of different improved components.

Model	Small-Object Head	CAF	DSConv	Precision	Recall	mAP@0.5	FPS	F1-Score
YOLOv8n	×	×	×	0.90	0.74	0.85	38.4	0.81
YOLOv8n + Head	√	×	×	0.91	0.75	0.87	35.6	0.82
YOLOv8n + CAF	×	√	×	0.92	0.75	0.88	34.9	0.83
YOLOv8n + DSConv	×	×	√	0.89	0.74	0.85	50.7	0.81
YOLOv8n + Head + CAF	√	√	×	0.93	0.76	0.89	32.8	0.84
YOLOv8n + Head + DSConv	√	×	√	0.91	0.75	0.87	48.2	0.82
YOLOv8n + CAF + DSConv	×	√	√	0.92	0.75	0.88	47.4	0.83
YOLOv8n-improved	√	√	√	0.93	0.76	0.90	46.9	0.84

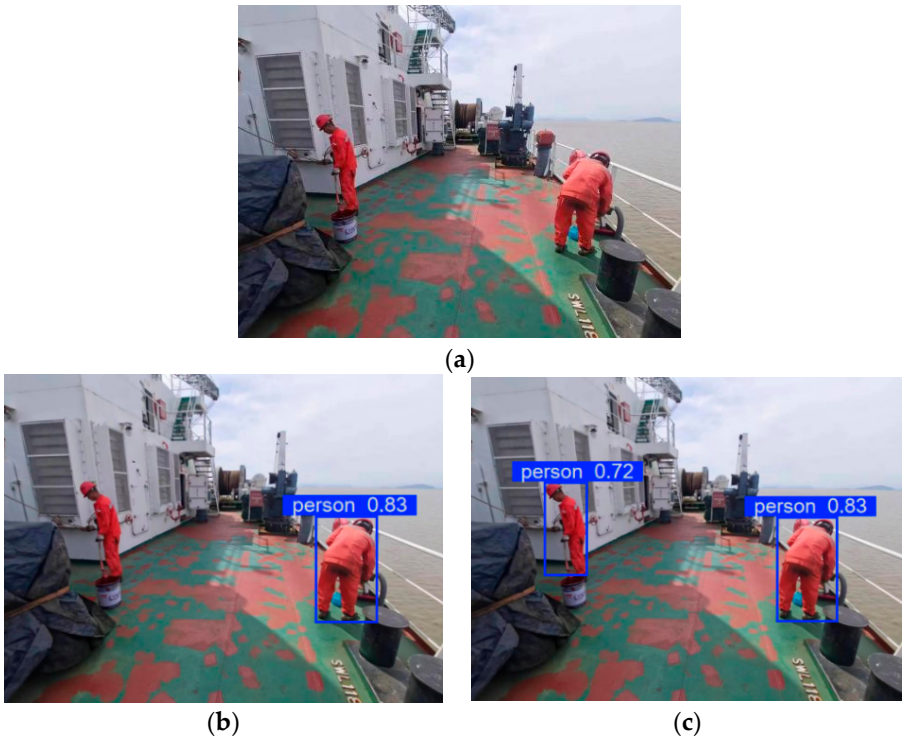
To verify the statistical reliability of the experimental results, ten repeated runs were conducted using different random seeds under the same dataset partition and training settings. The results were reported as mean ± standard deviation. The YOLOv8n-improved model achieved a mAP@0.5 of 0.90 ± 0.01, while the original YOLOv8n model achieved a mAP@0.5 of 0.85 ± 0.01. In addition, a Wilcoxon signed-rank test was performed on the mAP@0.5 results of the repeated runs, and the difference between YOLOv8n and YOLOv8n-improved was statistically significant ($p < 0.05$). These results indicate that the performance improvement of the proposed model is stable across repeated experiments rather than being caused by a single random run.

To intuitively demonstrate the optimization effect of the improved model in detection performance, representative visual comparison results are selected in this paper to present the visual comparison of ship personnel detection results before and after the YOLOv8 improvement, as illustrated in Figures 8 and 9. Among them, Figures 8a and 9a denote the original images of ship scenarios; Figures 8b and 9b show the detection results of the original YOLOv8 model; Figures 8c and 9c display the detection outputs of the YOLOv8n-improved model.



(a) (b) (c)

Figure 8. Comparative diagram of ship personnel detection and recognition experimental results. (**a**) Original image of ship deck; (**b**,**c**) Detection result image.



(a) (b) (c)

Figure 9. Comparative diagram of ship personnel detection and recognition experimental results. (**a**) Original image of ship deck; (**b**,**c**) Comparative Diagram 2 of Ship Personnel Detection and Recognition Experimental Results.

As can be observed from Figure 8, the original YOLOv8 model can only detect crew members with relatively large target sizes in the scene, while failing to effectively identify small-scale crew targets nearby. In contrast, the improved model can not only accurately recognize the main personnel targets in the image but also successfully detect all small-

sized crew targets around the deck operators. These visualization results fully verify that the proposed improvement strategy can significantly enhance the small-object detection capability of the YOLOv8 model in ship personnel detection tasks.

To evaluate robustness in practical shipboard environments, the test set was divided into five typical scenarios: normal deck scenes, complex backgrounds, partial occlusion, distant small targets, and shadow/reflection interference. The results are shown in Table 6.

Table 6. Error and robustness analysis under different shipboard conditions.

Scene Condition	Number of Test Images	Precision	Recall	mAP@0.5	False Detections	Missed Detections
Normal deck scenes	325	0.96	0.94	0.97	9	14
Complex background clutter	286	0.94	0.91	0.95	15	21
Partial occlusion	218	0.92	0.87	0.92	13	28
Distant small-scale targets	176	0.89	0.83	0.89	12	31
Shadow/reflection interference	134	0.91	0.85	0.90	11	24

As shown in Table 6, the model achieves the best performance in normal deck scenes and maintains stable accuracy under complex background clutter, while performance decreases under partial occlusion, small-scale targets, and shadow/reflection interference, where missed detections become more frequent. Overall, the model demonstrates good robustness across different shipboard scenarios, with most errors arising from missed detections rather than false detections, mainly due to occluded or small targets.

To further evaluate the robustness of the proposed model under challenging maritime operating conditions, an additional subset-based analysis was conducted using available samples involving rain, fog, nighttime illumination, dense crowds, and strong object overlaps. The results are summarized in Table 7.

Table 7. Robustness evaluation under challenging maritime operating conditions.

Challenging Condition	Test Images	Precision	Recall	mAP@0.5	False Detections	Missed Detections
Rain	29	0.88	0.82	0.86	5	9
Fog/low visibility	26	0.86	0.78	0.83	6	11
Nighttime illumination	88	0.87	0.80	0.84	15	26
Dense crowds	74	0.85	0.77	0.82	18	31
Strong object overlaps	62	0.84	0.75	0.81	19	34

Note: False detections and missed detections were counted at the annotated personnel-instance level. The challenging-condition subsets were constructed from the available test samples, and different challenging factors may partially coexist in the same image.

As shown in Table 5, the proposed model maintains acceptable detection performance under challenging maritime operating conditions, although the detection accuracy decreases compared with normal deck scenes. The performance degradation is mainly caused by reduced visibility, weak illumination, blurred target boundaries, dense personnel distribution, and severe object overlap. Among these conditions, dense crowds and strong object overlaps lead to the largest decrease in recall, indicating that missed detections become more frequent when adjacent personnel contours are merged or heavily occluded. Rain, fog, and nighttime illumination mainly weaken texture and edge features, thereby increasing the difficulty of small-target representation. Nevertheless, the additional small-object detection head strengthens fine-grained feature extraction for distant and weak targets, while the

CAF module helps suppress background interference and enhance contextual feature representation. These results further demonstrate that the proposed model has a certain degree of robustness under challenging maritime operating conditions. It should also be noted that the number of samples under extreme weather, nighttime illumination, dense crowds, and severe overlap conditions is still limited. Therefore, larger-scale scenario-specific evaluations under more diverse maritime operating conditions will be further conducted in future work.

4. Discussions and Conclusions

Existing detection methods are constrained by background interference and the difficulty of recognizing distant small targets in shipboard scenarios, resulting in detection accuracy that fails to meet the practical operational requirements. The ship personnel detection method proposed in this paper, which is based on the fusion of attention mechanisms and multi-scale perception, enhances the ability to capture and identify deck operators at long distances by optimizing the small-object detection branch. This addresses the shortcomings of traditional methods in small-object detection under complex shipboard scenarios and improves the model's adaptability to variations in target scale.

By embedding a convolution and attention fusion module into the model's backbone network, the proposed method achieves precise separation of personnel features from the complex background of shipboard environments, thereby mitigating the impact of background interference on detection results. Replacing traditional standard convolution with depthwise separable convolution improves the model's computational efficiency while maintaining detection accuracy, achieving a synergistic optimization of detection precision and real-time performance to meet the practical application requirements of real-time monitoring in ship scenarios. These improvements specifically address the pain points of shipboard scenarios: the small-object detection branch enhances the capture of distant personnel features, solving the recognition challenge caused by the low pixel proportion of distant personnel targets; The convolution and attention fusion module enables precise separation of personnel features from complex backgrounds such as piled equipment and sea wave reflections; Depthwise separable convolution significantly reduces computational complexity, remedying the insufficient applicability of existing methods in practical navigation and operation scenarios.

Compared with existing studies, this work breaks through the limitations of single-strategy optimization and achieves superior comprehensive performance: the improved model attains a precision of 93%, an mAP@0.5 of 90%, and an FPS of 46.9 frames per second. Compared with Faster R-CNN, these metrics are improved by 48%, 22%, and 36.6 FPS, respectively; compared with YOLOv5, the improvements reach 8%, 10%, and 21.3 FPS; and compared with the original YOLOv8, all indicators show gains of 2–5%. These results are significantly better than those of single-improvement models in similar personnel detection methods.

Nevertheless, several limitations remain in the current study. Due to limitations in experimental equipment and cost conditions, the performance of the improved algorithm was only validated on a single ship-deck personnel dataset organized in a VOC-style data format. Cross-dataset testing on additional shipboard or maritime datasets has not yet been conducted to further verify the algorithm's generalizability. In addition, when the small-object detection head and attention mechanism are enabled simultaneously, the model is prone to issues such as detection confusion and misidentification in shipboard dense crowd scenarios. Future research can focus on targeted optimization for shipboard dense crowd recognition tasks to further enhance the model's target discrimination capability. Although

an additional subset-based robustness evaluation was conducted for rain, fog, nighttime illumination, dense crowds, and strong object overlaps, the number of samples under these extreme maritime operating conditions remains limited. Future research will further collect larger-scale shipboard monitoring images under diverse weather, illumination, crowd-density, and occlusion conditions, and conduct more systematic cross-scenario quantitative evaluations to improve the robustness and generalization capability of the proposed method.

Meanwhile, regarding the high labor and time costs during model training, weakly supervised or unsupervised training methods can be considered to accelerate model iteration and practical deployment.

In this paper, a ship personnel detection method based on attention mechanism and multi-scale perception fusion is proposed. By optimizing the small object detection branch, the capability for capturing distant targets is strengthened. A convolution–attention fusion module is embedded in the backbone network to realize feature separation between targets and complex backgrounds. Meanwhile, depthwise separable convolution is adopted to realize the lightweight design of the model. The YOLOv8n-improved model achieves a precision of 93%, a recall of 76%, an mAP@0.5 of 90%, and an FPS of 46.9 frames per second. Compared with the original YOLOv8 model, each indicator is improved by 2% to 5%, and the overall performance is superior to mainstream models such as Faster R-CNN and YOLOv5. This model efficiently completes the task of ship personnel detection and recognition, and improves both detection accuracy and real-time performance. It provides an efficient and reliable technical solution for ship personnel safety monitoring and deck operation management, and also offers a valuable reference for the research of small-object detection in complex scenarios.

Author Contributions: Conceptualization, Z.L.; Methodology, Z.L., N.C., M.Q. and C.B.; Formal Analysis, N.C.; Investigation, M.Q.; Resources, Z.L.; Writing—Original Draft, Z.L. and C.B.; Writing—Review and Editing, Z.L.; Supervision, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Dalian High-level Talent Innovation Program (No.2024RQ017) and the National Natural Science Foundation of China (NSFC) (52301361) and the Fundamental Research Funds for the Central Universities (3132026336).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. European Maritime Safety Agency (EMSA). *Annual Overview of Marine Casualties and Incidents 2024*; EMSA: Lisbon, Portugal, 2024.
2. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
3. Zhang, M.; Yao, J.; Zhang, K.; Zhang, Y.; Zhang, H. S-YOLO: A ship detection model based on YOLOv5 in complex scenes. *Remote Sens.* **2022**, *14*, 5788. [CrossRef]
4. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
5. Li, C. Research on the Application of Attention Convolutional Neural Network in Pedestrian Attribute Recognition. Master's Thesis, Xi'an University of Technology, Xi'an, China, 2020.
6. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005*; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.
7. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef]

8. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
11. 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, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
12. Qiang, H. Research on Improved YOLOv8 Algorithm for Small Target Detection. Master's Thesis, Jilin University, Changchun, China, 2023.
13. Li, Y.; Xu, H.; Bian, M.; Xiao, J. Attention based CNN-ConvLSTM for pedestrian attribute recognition. *Sensors* **2020**, *20*, 811. [CrossRef] [PubMed]
14. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 7263–7271.
15. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point set representation for object detection. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 9657–9666.
16. Law, H.; Deng, J. CornerNet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 734–750.
17. Fang, Y.; Yuan, G.; Sun, Z.; Li, Y. Small object detection based on multi-scale feature fusion and attention mechanism. *J. Yunnan Univ. (Nat. Sci. Ed.)* **2026**, *48*, 36–44.
18. Wang, J.; Yin, X.; Li, G. A real-time lightweight detection algorithm for deck crew and the use of fishing nets based on improved YOLOv5s network. *Fishes* **2023**, *8*, 376. [CrossRef]
19. Li, Z.; Zhang, H.; Gao, D.; Chen, L.; Wang, J. ACD-Net: An Abnormal Crew Detection Network for Complex Ship Scenarios. *Sensors* **2024**, *24*, 7288. [CrossRef] [PubMed]
20. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 10781–10790.
21. Murad, M.M.N.; Turgut, B.S.; Ahmed, A.; Turan, O. Camera-Based Intruder Detection and Monitoring of Ship Crew Work Hours. In *Proceedings of the Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1526–1534.
22. Zhao, H.; Zhang, H.; Zhao, Y. YOLOv7-Sea: Object detection of maritime UAV images based on improved YOLOv7. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 233–238.
23. Han, Y.; Wang, H.; Renjin, N.; Liu, Y. Multi-ship detection and classification with feature enhancement and lightweight fusion. *Sci. Rep.* **2025**, *15*, 38075. [CrossRef] [PubMed]
24. Wu, W.; Li, X.; Hu, Z.; Wang, J. Ship Detection and Recognition Based on Improved YOLOv7. *Comput. Mater. Contin.* **2023**, *76*, 1167–1185. [CrossRef]
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
26. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520.
28. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 6848–6856.

29. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Depthwise separable convolution neural network for high-speed SAR ship detection. *Remote Sens.* **2019**, *11*, 2483. [CrossRef]
30. Poomhiran, L.; Meesad, P.; Nuanmeesri, S. Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique. *Eng. Technol. Appl. Sci. Res.* **2021**, *11*, 6921–6926. [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.