automation

Article

A Lightweight Attention-Guided and Geometry-Aware Framework for Robust Maritime Ship Detection in Complex Electro-Optical Environments

Zhe Zhang [1], Chang Lin [2,*] and Bing Fang [1]

[1] School of Mechanical and Electrical Engineering, Fujian Agriculture and Forestry University, Fuzhou 350100, China; 52312040060@fafu.edu.cn (Z.Z.); fangbing008@163.com (B.F.)
[2] College of Computer and Data Science, Putian University, Putian 351100, China
* Correspondence: linchanpt@163.com

Abstract

Reliable ship detection in complex maritime optical imagery is a fundamental requirement for intelligent maritime monitoring and maritime automation systems. However, severe image degradation, large-scale variations, and background clutter often lead to feature ambiguity and unstable detection performance in real-world maritime environments. To address these challenges, this paper proposes a lightweight one-stage ship detection framework designed for robust real-time perception under degraded maritime sensing conditions. The proposed method incorporates an Adaptive Expert Selection Attention (AESA) mechanism to perform adaptive feature selection and background suppression under visually degraded conditions, together with a Geometry-Aware MultiScale Fusion (GAMF) module that enables orientation-aware aggregation of contextual information for elongated ship targets near complex sea–sky boundaries. In addition, a geometry-aware bounding box regression refinement is introduced to improve localization consistency in image space. Extensive experiments conducted on a unified real-world maritime benchmark demonstrate that the proposed framework consistently outperforms the baseline YOLO11n model by approximately 2–5 percentage points in terms of mAP@0.5 and mAP@0.5:0.95, while maintaining moderate computational complexity and real-time inference capability. These results indicate that the proposed method provides a practical and deployment-oriented perception solution for maritime automation applications, including onboard electro-optical sensing and coastal surveillance.

Keywords: maritime perception; ship detection; maritime automation; attention mechanism; multiscale fusion; You Only Look Once (YOLO)

Academic Editor: Yi-Xin Su

Received: 22 December 2025
Revised: 22 February 2026
Accepted: 10 March 2026
Published: 12 March 2026

1. Introduction

With the rapid advancement of intelligent maritime monitoring and autonomous navigation, reliable ship detection in maritime optical imagery has become a fundamental prerequisite for automated maritime perception and real-time situational awareness. Accurate ship detection plays a critical role in navigation assistance, maritime traffic supervision, collision avoidance, and automated port management. In practical maritime automation pipelines, ship detection serves as a core perception module that directly supports downstream decision-making and control processes, including autonomous navigation and safety assurance. Recent breakthroughs in computer vision and deep learning have significantly propelled the development of object detection techniques, enabling real-time visual perception in increasingly complex and dynamic environments.

Despite this progress, most conventional object detectors are primarily optimized for terrestrial scenes and often fail to account for the unique visual characteristics and operational constraints of maritime environments. Maritime optical imagery is frequently compromised by severe image degradation caused by haze, dynamic illumination variations, atmospheric scattering, wave reflections, and complex sea–sky boundaries. These factors typically result in low contrast, background clutter, and feature ambiguity, thereby degrading the detection stability and localization reliability in practical maritime automation systems. Recent surveys on the YOLO (You Only Look Once) family have highlighted its dominance in real-time detection tasks [1]. However, achieving robust detection performance under such visually degraded maritime conditions while simultaneously maintaining real-time efficiency on resource-constrained and safety-critical automation platforms remains a formidable challenge.

Early research in maritime visual perception therefore prioritized detection efficiency by incorporating lightweight backbone networks into real-time detection frameworks to facilitate its deployment on embedded and edge platforms [2]. Such lightweight designs have been widely recognized as a fundamental requirement for practical maritime automation systems. Beyond backbone simplification, compact convolutional architectures such as LightSDNet have also demonstrated potential for maritime perception by leveraging efficient yet expressive feature representations [3]. However, these studies collectively indicate that efficiency-oriented optimization alone is insufficient to ensure robust perception under severe maritime visual degradation.

More recently, specialized detection frameworks have been proposed to address the adverse imaging conditions inherent in maritime environments. Representative studies have explored enhanced feature representation strategies to improve detection performance under low visibility, complex sea–sky backgrounds, and large-scale variations [4,5]. In addition, comprehensive reviews have consistently identified small object detection, scale imbalance, and strong background interference as the primary bottlenecks limiting reliable maritime perception [6]. These findings suggest that adaptive and context-aware feature representation is critical for achieving robust performance in heterogeneous maritime scenes.

Beyond feature representation, the geometric characteristics of maritime vessels introduce additional challenges for reliable detection. Many ships exhibit elongated and slender structures, leading to highly imbalanced aspect ratios in bounding boxes. Under such conditions, conventional Intersection over Union (IoU)-based regression losses often suffer from unstable optimization behavior, resulting in inconsistent localization accuracy. Although improved formulations such as Generalized IoU (GIoU) [7] partially alleviate these issues, regression stability remains limited, particularly in visually degraded maritime environments.

From the perspective of maritime automation systems, achieving reliable real-time perception requires not only robust feature representation but also stable geometric localization under complex operating conditions. To address these challenges, this paper proposes an enhanced YOLO-based ship detection framework tailored for complex maritime automation. The proposed method incorporates three task-oriented architectural improvements for maritime automation:

1. First, an Adaptive Expert Selection Attention (AESA) module is introduced to refine multiscale channel and spatial features by dynamically isolating discriminative information under environmental interference.

2. Second, a Geometry-Aware MultiScale Fusion (GAMF) module is integrated into the SPPF structure, forming an SPPFGAMF block that enhances orientation-aware multiscale aggregation.
3. Finally, a refined bounding box regression strategy combining WiseIoU (WIoU) [8] with image-space aspect ratio regularization is employed to improve geometric consistency and localization robustness.

Extensive experiments conducted on a unified maritime benchmark constructed from multiple publicly available data sources demonstrate that the proposed framework consistently outperforms baseline models in terms of accuracy and robustness while maintaining moderate computational complexity. These results confirm the practical applicability of the proposed approach as a reliable perception component for autonomous maritime systems operating under complex maritime visual conditions.

2. Related Work

To clarify the limitations of the existing approaches and the technical motivations behind the proposed framework, this section discusses the representative studies on maritime ship detection under complex visual conditions and organizes them into three main research directions: attention-guided feature enhancement, multiscale feature fusion for complex maritime scenes, and loss function optimization for improving geometric consistency.

2.1. Attention-Guided YOLO-Based Ship Detection

Maritime optical imagery is frequently affected by low contrast, background clutter, haze, and dynamic illumination variations, which significantly degrade feature discriminability and increase false detections in practical maritime automation scenarios. To mitigate these challenges, numerous studies have incorporated attention mechanisms into deep learning-based detection frameworks for maritime perception to enhance feature saliency and suppress background interference.

Mainstream channel attention mechanisms, such as Squeeze-and-Excitation (SE) networks [9] and Efficient Channel Attention (ECA) [10], recalibrate channel responses to emphasize informative features with limited computational overhead. These approaches have demonstrated effectiveness in maritime detection tasks, particularly in suppressing background noise near sea–sky boundaries. In addition, attention-guided feature refinement has shown promising results in degraded sensing modalities, such as infrared maritime perception, where semantic, detail, and edge information are jointly exploited to improve weak target detection performance [11]. Furthermore, multiscale attention strategies that integrate spatial and channel priors have been reported to enhance detection robustness and perceptual consistency under dynamically changing maritime environments [12]. Guo et al. [13] proposed a two-stage attention-enhanced network for ship detection in optical remote sensing imagery, achieving an improved performance for small and elongated vessels under cluttered sea backgrounds.

Despite these advances, most existing attention-based approaches rely on static attention structures with fixed weighting schemes. Such designs limit adaptability under dynamically varying atmospheric and illumination conditions, thereby constraining perceptual robustness in real-world maritime automation platforms. These observations suggest the necessity of a more adaptive feature selection and spatial modulation mechanisms that can dynamically respond to image content and environmental variations in maritime scenes.

In addition to channel and spatial attention mechanisms, recent studies have explored cross-attention strategies within YOLO-based detection frameworks to enhance inter-region feature interaction and global context modeling. For instance, CAYOLO integrates cross-attention modules into the YOLO architecture to strengthen long-range dependency modeling, demonstrating an improved localization performance under structurally complex scenarios [14].

These works reflect a growing trend of incorporating more expressive attention designs into YOLO-based detectors to improve their contextual reasoning ability. While such global interaction mechanisms improve the feature's expressiveness, they often introduce additional computational overhead. This further highlights the importance of designing lightweight yet adaptive attention mechanisms tailored for real-time maritime perception systems.

2.2. MultiScale Feature Fusion for Small and Slender Ships

Ship targets in maritime scenes often exhibit extreme scale variations and elongated geometric profiles, which necessitate effective multiscale feature representations for robust maritime perception [15]. Standard real-time detection pipelines typically employ hierarchical architectures such as the Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) to fuse semantic and spatial information across multiple feature scales [16,17]. In parallel, receptive field enhancement modules such as Spatial Pyramid Pooling (SPP) [18] and its lightweight variants are widely adopted to expand contextual awareness while maintaining computational efficiency, which is crucial for real-time maritime automation applications.

Beyond generic detection frameworks, several maritime-specific multiscale fusion strategies have been developed to better capture the visual characteristics of ship targets under complex sea–sky conditions. For example, YOLOPFA [19] improves feature alignment through adaptive fusion mechanisms, while ShipFPN [20] enhances dense small-ship detection in high-resolution maritime imagery by introducing task-oriented pyramid designs. These studies highlight the importance of task-specific multiscale fusion strategies for improving detection robustness in maritime environments.

However, most existing fusion modules still rely on static concatenation or fixed weighting strategies, which are insufficient for modeling the strong orientation bias introduced by sea–sky boundaries and the elongated aspect ratios of ship targets [21,22]. Under such conditions, rigid fusion strategies may dilute salient structural cues and degrade feature consistency. Recent studies have indicated that incorporating directional information and adaptive weighting into multiscale fusion can better accommodate the anisotropic spatial distributions commonly observed in maritime scenes [23]. Consequently, orientation-aware and adaptive multiscale fusion strategies are increasingly regarded as a key requirement for reliable ship detection in complex maritime environments.

In recent years, several advanced ship detection and feature fusion frameworks have been proposed to enhance representation capability under complex sensing conditions. Recent studies have emphasized structured knowledge integration, imbalance-aware learning, and hierarchical feature modeling to improve robustness in challenging maritime environments. For instance, KFIANet introduces knowledge fusion and imbalance-aware mechanisms to address category distribution bias in SAR ship detection [24]. Similarly, incremental recognition strategies and multigranularity feature fusion networks have demonstrated improved adaptability and cross-scale feature interaction in SAR target recognition scenarios [25,26].

These developments reflect a broader research trend toward structured feature aggregation and task-aware representation design, further underscoring the importance of lightweight yet adaptive fusion strategies for complex maritime perception tasks.

2.3. IoU-Based Loss Functions for Ship Bounding Box Regression

IoU-based loss functions constitute the foundation of bounding box regression in modern object detection systems. Conventional IoU loss suffers from score discontinuity when predicted and ground truth bounding boxes do not overlap, which often leads to unstable optimization behavior and inconsistent localization results. To address these limitations, several improved formulations have been proposed. Generalized IoU (GIoU) [7] introduces penalties for non-overlapping regions, Distance IoU (DIoU) [27] incorporates center-distance constraints, and Scaled IoU (SIoU) [28] further models orientation and geometric priors to improve regression efficiency for elongated objects.

Despite these improvements, existing IoU-based loss functions still lack effective mechanisms to suppress low-quality gradients, which frequently arise in visually degraded maritime environments affected by haze, wave reflections, and illumination interference. Such unreliable gradients can dominate the optimization process, resulting in unstable regression behavior and reduced localization reliability in practical maritime automation scenarios. Moreover, maritime ship targets typically exhibit highly imbalanced aspect ratio distributions in image space, and neglecting this dataset-specific geometric prior can further aggravate regression instability and localization inconsistency. Orientation-aware detection frameworks such as R^2-CNN [29] have also highlighted the importance of incorporating geometric constraints when detecting slender vessels.

These observations indicate that bounding box regression in maritime ship detection can benefit from loss formulations that jointly consider prediction quality and dataset-specific geometric characteristics, thereby improving localization stability under complex maritime sensing conditions.

3. YOLO11 Object Detection Model

In this work, we adopt a YOLO11-style lightweight detector implemented within the Ultralytics framework as the baseline detection architecture. YOLO11n follows the general design philosophy of recent YOLO families, featuring a lightweight backbone, a decoupled detection head, and a multiscale feature pyramid structure. By integrating the Feature Pyramid Network (FPN) [16] and the Path Aggregation Network (PAN) [17], YOLO11n enables effective top-down and bottom-up information flow across multiple feature scales, achieving a favorable balance between detection accuracy and computational efficiency. These characteristics make YOLO11n well suited for real-time deployment on resource-constrained edge devices.

Despite these advantages, when directly applied to maritime optical imagery, the baseline YOLO11n architecture still exhibits limited robustness under severe image degradation, strong background clutter near sea–sky boundaries, and the elongated geometric profiles of ship targets. As discussed in Section 2, existing detection frameworks often suffer from insufficient adaptability in feature selection, limited orientation awareness in multiscale fusion, and unstable localization under imbalanced aspect ratio distributions.

Motivated by these challenges, this study adopts YOLO11n as the baseline framework and introduces a series of task-oriented architectural refinements tailored for maritime ship detection. Specifically, the proposed model enhances feature representation through adaptive attention mechanisms, improves contextual aggregation via orientation-aware multiscale fusion, and strengthens localization robustness using geometry-aware regression constraints. Through these targeted improvements, the proposed framework aims to achieve robust real-time ship detection under complex maritime sensing conditions. The overall architecture of the proposed model is illustrated in Figure 1.

https://doi.org/10.3390/automation7020048

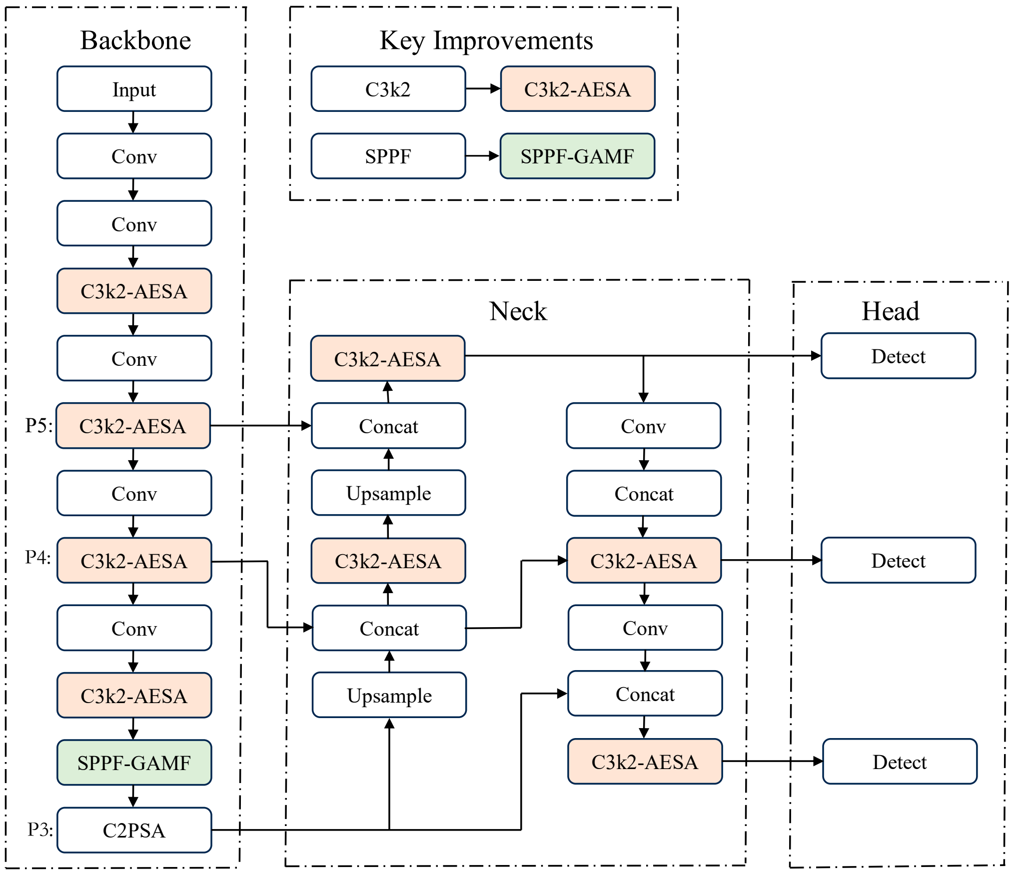


Figure 1. Architecture of the improved YOLO11 model.

3.1. *Adaptive Expert Selection Attention (AESA) Mechanism*

Motivated by the limited adaptability of static channel and spatial attention mechanisms in maritime electro-optical imagery (Section 2.1), we designed an Adaptive Expert Selection Attention (AESA) module that formulates feature enhancement as a structured selection process rather than uniform reweighting. Specifically, AESA integrates multiscale channel modeling, competitive channel selection, and shared spatial modulation into a unified attention pipeline.

As illustrated in Figure 2, the AESA mechanism consists of three sequential stages: (1) multiscale channel dependency modeling, (2) intragroup competitive routing for dynamic channel selection, and (3) shared spatial filtering for consistent background suppression.

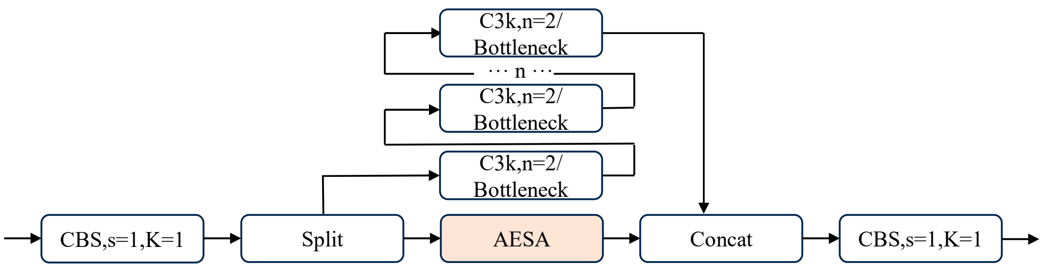


Figure 2. Structure of the C3k2-AESA module.

(1) Multiscale Channel Modeling: instead of relying on a single receptive field scale, AESA formulates channel dependency modeling as a scale-adaptive selection problem over a predefined set of receptive fields. Specifically, given an input feature map $X \in \mathbb{R}^{\{C\times H\times W\}}$, a set of parallel channel modeling branches $\{F_s \mid s \in S\}$ is constructed,

where $S = \{s_1, s_2, s_M\}$ denotes a collection of kernel sizes corresponding to different effective receptive field scales.

Each branch $F_s(\cdot)$ represents a scale-specific channel modeling operator. For each scale $s \in \mathcal{S}$, the corresponding scaleaware channel response e_s is obtained by applying the Efficient Channel Attention (ECA) operation with kernel size s to the input feature map X, i.e., $e_s = F_s(X)$.

Each branch thus captures channel dependencies under a distinct effective receptive field scale, yielding a set of scale-aware channel response functions $\{e_s\}$. Rather than aggregating these responses through fixed summation or concatenation, AESA introduces a Softmax-based competitive fusion mechanism across the scale dimension, enabling implicit scale selection without requiring explicit gating networks.

$$\begin{cases} m_c = \sum_{s \in S} w_s \cdot e_{s,c} \\ w_s = \frac{\exp(\alpha_s)}{\sum_{t \in S} \exp(\alpha_t)} \end{cases}, \tag{1}$$

where α_s denotes the learned importance score of scales. This competitive normalization encourages the network to adaptively emphasize the most informative receptive field scale for each channel, alleviating the optimization conflicts commonly observed in static multi-branch aggregation. In practice, the scale set S is instantiated as $\{3,\ 5,\ 7\}$ to balance representational diversity and computational efficiency.

(2) Intragroup Routing Competition: to further enhance feature adaptiveness, AESA introduces a lightweight routing-inspired competition mechanism at the channel group level. Specifically, each channel is expanded into K expert groups via a 1×1 expert (grouped) convolution, followed by a temperature-controlled Softmax operation across the group dimension. This design allows each channel to dynamically select the most suitable expert response in a context-aware manner, which differs from the uniform channel recalibration used in conventional attention mechanisms. By enforcing competition among expert groups, AESA reduces mutual interference and mitigates feature over-smoothing under low contrast or visually degraded conditions.

(3) Shared Spatial Filtering: before the final aggregation, all expert outputs pass through a shared spatial attention module designed to emphasize salient ship regions through a unified spatial modulation mechanism. Unlike CBAM, which applies spatial attention independently in each branch, AESA uses a unified spatial filter across all expert groups. This design prevents redundant computation and ensures consistent spatial modeling across the entire attention pipeline. The internal workflow of the AESA mechanism is illustrated in Figure 3.

In this work, the scale set S is instantiated as {3, 5, 7}, corresponding to three parallel ECA branches with different effective receptive fields. Under this specific configuration, the general formulation in Equation (1) can be explicitly rewritten as Equation (2):

$$\begin{cases} m_c = \sum_{i=1}^{3} w_i \cdot e_i \\ w_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{3} \exp(\alpha_j)} \end{cases},\ i \in \{1,\ 2,\ 3\}, \tag{2}$$

Subsequently, a 1×1 grouped convolution is employed to expand each channel into K expert groups, followed by Softmax-based routing competition across the group dimension, as formulated in Equation (3), where the routing probabilities are normalized across the expert dimension.

$$p_{i,k} = \frac{\exp(s_{i,k}/\tau)}{\sum_{j=1}^{K} \exp(s_{i,j}/\tau)},\ k = 1,\ K, \tag{3}$$

where τ denotes the temperature parameter controlling the sharpness of the routing distribution. This dynamic routing strategy enables each channel to autonomously select the most appropriate expert branch and effectively alleviates feature over-smoothing caused by naive multi-branch averaging.

Finally, a shared spatial attention module is applied to each expert feature to suppress background interference and highlight salient ship regions. The outputs of all expert branches are aggregated to generate the final enhanced feature representation, as defined in Equation (4):

$$y = \sum_{k=1}^{K} \mathrm{SA}(x_k), \tag{4}$$

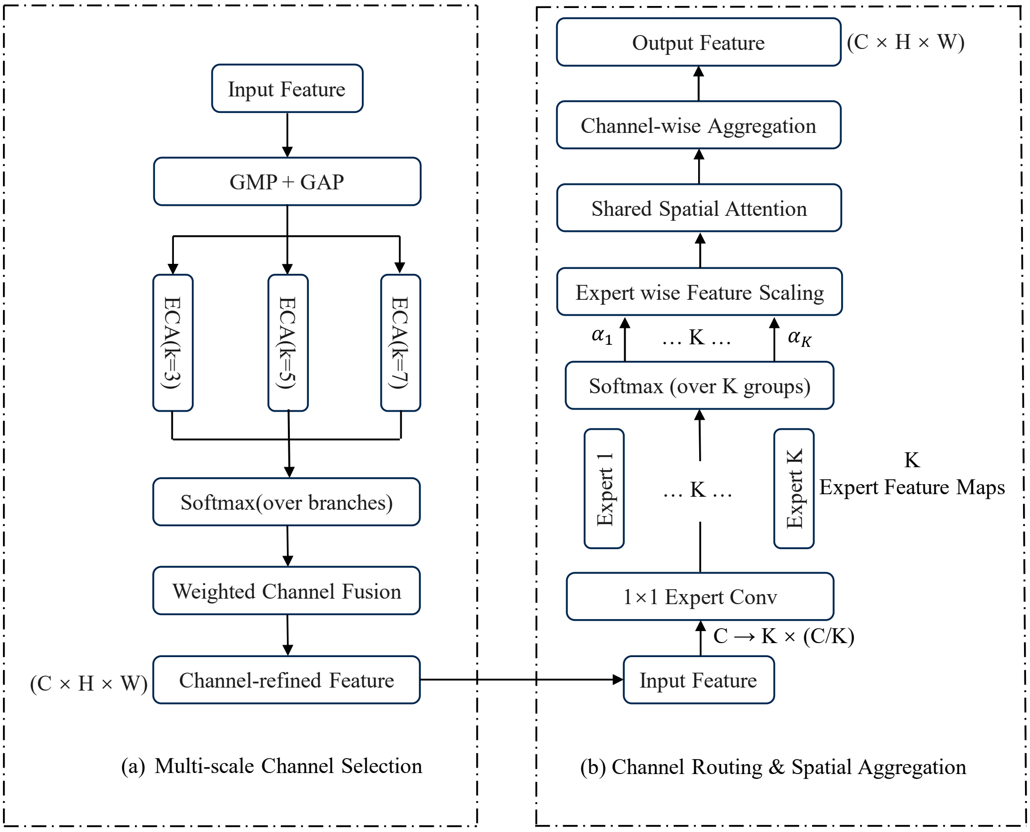


Figure 3. Internal structure of the AESA module.

Overall, AESA introduces a hierarchically organized attention mechanism that integrates multiscale channel modeling, groupwise competitive routing, and lightweight shared spatial modulation within a unified feature selection process. Unlike conventional attention operators that apply channel or spatial reweighting in isolation, AESA performs structured routing at the channel group level to selectively emphasize discriminative responses across multiple receptive scales, while preserving a single-stream feature representation.

Compared with conventional MoE-style routing mechanisms commonly used in dynamic expert models, which require explicit gating networks or auxiliary losses, AESA embeds lightweight expert routing directly into the channel aggregation process. This design enables efficient optimization without introducing additional training overhead. Designed for the characteristics of maritime electro-optical sensing, the shared spatial modulation further reinforces spatial consistency under severe scale variation, background clutter, and sensor-induced degradation, thereby producing more robust and discriminative feature representations for ship detection.

3.2. Geometry-Aware MultiScale Fusion (GAMF)

The Spatial Pyramid Pooling Fast (SPPF) module is widely adopted in the YOLO family to aggregate multiscale contextual information with minimal computational overhead.

However, existing adaptive fusion designs still rely on fixed weighting schemes or hand-crafted fusion rules, which limits their ability to model orientation bias and heterogeneous spatial distributions in maritime scenes [22,23].

To address this limitation, we propose a Geometry-Aware MultiScale Fusion (GAMF) module that enhances the SPPF block with direction-aware and sample-adaptive branch weighting. Instead of treating all multiscale branches equally, GAMF leverages coordinate attention to extract horizontal and vertical directional statistics from each branch feature and uses them to compute normalized fusion weights, enabling adaptive aggregation that better matches the current image content.

As illustrated in Figure 4, GAMF preserves the output resolution of the original SPPF to maintain full compatibility with the YOLO11 backbone. Following channel compression, multiple parallel feature representations with different receptive field scales are generated in accordance with the SPPF design. These multiscale branch features are then fed into the GAMF module for adaptive fusion.

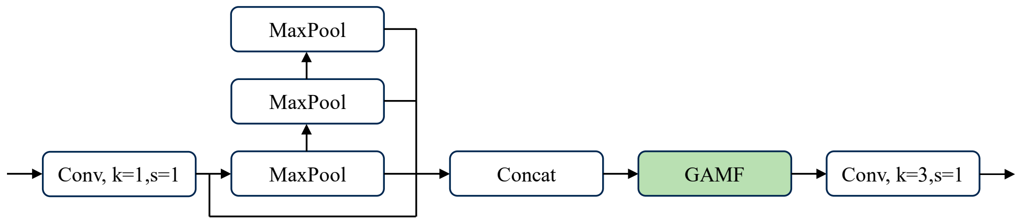


Figure 4. Architecture of the SPPFGAMF module.

Although three dilation rates {1, 3, 5} are adopted in this work to balance contextual coverage and computational efficiency, the GAMF formulation itself is not restricted to specific dilation settings. In principle, GAMF can be generalized to arbitrary multi-branch configurations with different receptive field designs, making the proposed fusion strategy adaptable to a broader range of backbone architectures and scale modeling schemes.

Inside GAMF (Figure 5), coordinate attention is applied to derive horizontal and vertical directional descriptors from each branch feature map $F_d \in \mathbb{R}^{C\times H\times W}$.

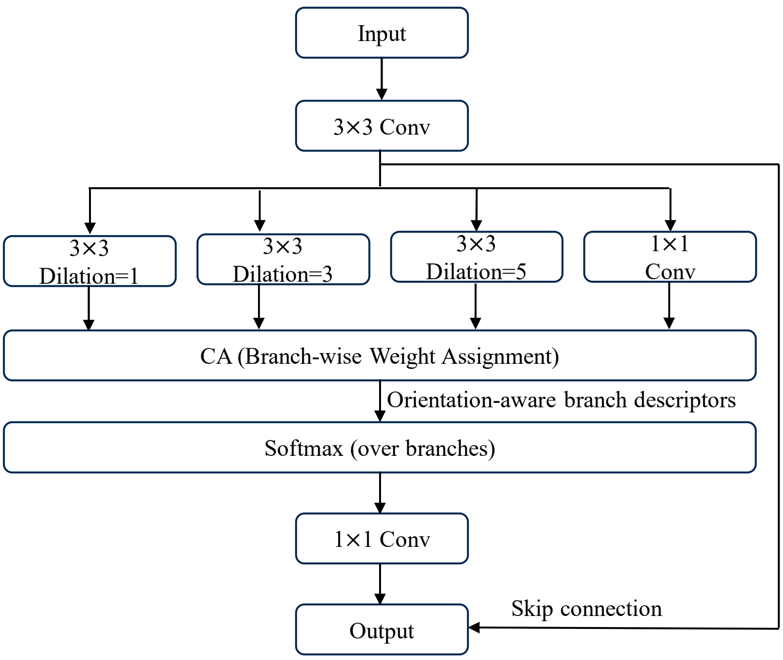


Figure 5. Structure of the GAMF module.

From a modeling perspective, GAMF formulates multiscale feature fusion as a direction-aware branch selection problem, where the relative importance of each receptive field branch is adaptively inferred from horizontal and vertical contextual statistics rather than predefined heuristics. The corresponding directional statistics are computed as follows:

$$\begin{cases} F_h(i) = \frac{1}{W}\sum_{j=1}^{W} F_d(i,j),\ i = 1,\ \ H \\ F_w(j) = \frac{1}{H}\sum_{i=1}^{H} F_d(i,j), j = 1,\ \ W \end{cases}, \tag{5}$$

where H and W denote the spatial height and width, respectively. Based on these descriptors, GAMF estimates sample-level branch scores $\{S_k\}$ and converts them into normalized weights $\{\alpha_k\}$ via a Softmax operation:

$$\begin{cases} \alpha_k = \frac{\exp(S_k)}{\sum_{t=1}^{K}\exp(S_t)},\ k = 1,\ \ K \\ \quad Y = \sum_{k=1}^{K} \alpha_k \odot B_k \end{cases}, \tag{6}$$

where B_k denotes the k-th branch feature and $\odot$ represents per-channel scaling with spatial broadcasting. Unlike linear fusion or static weighting, the Softmax normalization enforces inter-branch competition, ensuring that dominant directional cues are selectively amplified while redundant responses are suppressed.

The final fused representation is obtained through the weighted aggregation of branch features and further refined by a 1×1 convolution with a residual connection to preserve feature integrity. Overall, GAMF introduces a direction-aware and sample-adaptive fusion mechanism into the SPPF structure, enabling a more discriminative multiscale feature aggregation for elongated ship targets under complex maritime imaging conditions.

3.3. Improved Loss Function

As discussed in Section 2.3, existing IoU-based loss functions still suffer from unstable optimization and insufficient geometric constraints when applied to elongated ship targets in degraded maritime imagery.

To address this limitation in a unified manner, this work adopts WiseIoU (WIoU) [8] as the baseline regression loss due to its quality-aware weighting strategy. By dynamically adjusting gradient contributions according to prediction quality, WIoU effectively suppresses noisy gradients caused by ambiguous or low-confidence samples, resulting in more stable and reliable regression optimization. For fair comparison, the classification loss and Distribution Focal Loss (DFL) terms are kept unchanged.

Beyond quality-aware weighting, maritime ship targets typically exhibit elongated geometric structures and a highly imbalanced aspect ratio distribution that is not explicitly modeled by existing IoU-based losses. Therefore, the proposed regression loss is not a direct application of WIoU. An additional aspect ratio regularization term is introduced, derived from the empirical distribution of ship hull geometries in the training dataset. Unlike CIoU, which applies a fixed geometric penalty, the proposed formulation imposes a distribution-aware and softly bounded constraint, stabilizing gradient updates near extreme aspect ratio conditions and reducing oscillatory behavior during training.

For the ith training sample with predicted bounding box width w_i and height h_i, the aspect ratio r_i is defined as:

$$r_i = \frac{w_i}{h_i}, \tag{7}$$

Based on the aspect ratio statistics computed on the training set only, the lower and upper bounds are determined using robust percentile estimation. Specifically, r_{min} = 0.40 and r_{max} = 3.20 are selected according to the 95th and 99th percentiles, respectively. This percentilebased strategy provides stable geometric bounds under the long-tailed aspect ratio distribution of maritime ship targets while reducing the influence of extreme outliers.

A Softplus-based regularization function with temperature parameter τ = 0.2 is then constructed as shown in Equation (8), where τ controls the smoothness of the penalty near the boundary. A small τ ensures gradual gradient transition near the boundary, preventing abrupt optimization shocks during training.

$$L_{ratio} = \frac{1}{N}\sum_{i=1}^{N}\left[softplus\left(\frac{r_i - r_{max}}{\tau}\right) + softplus\left(\frac{r_{min} - r_i}{\tau}\right)\right], \quad (8)$$

When r_i lies within the valid range $[r_{min}, r_{max}]$, the regularization term remains close to zero, and increases smoothly when the predicted aspect ratio exceeds the predefined bounds. This design effectively mitigates abrupt boundary gradients and improves regression stability during training.

From a geometric prior perspective, the proposed aspect ratio regularization does not enforce a fixed shape constraint but rather constrains predictions within a statistically valid range derived from real maritime ship geometries. This design allows the network to preserve shape diversity while suppressing extreme and unreliable aspect ratio predictions.

The overall detection loss adopts a composite multiterm formulation, in which the regression component integrates quality-aware weighting and distribution-aware geometric regularization. The complete loss function is expressed in Equation (9), combining WIoU-based localization loss, aspect ratio regularization, classification loss, and distribution focal loss.

$$L_{dct} = L_{WIoU} + \lambda_{ratio}L_{ratio} + L_{cls} + L_{dfl}, \quad (9)$$

In all experiments, the coefficient λ_{ratio} is fixed to 0.1 and treated as a constant hyperparameter throughout training and evaluation. Under this setting, the proposed WIoU+AR formulation provides effective geometric constraints for elongated ship targets without introducing noticeable optimization instability, thereby improving localization robustness and mAP consistency in complex maritime environments.

4. Experimental Design and Result Analysis

4.1. Dataset

To ensure that the experimental evaluation is consistent with the practical electro-optical (EO) maritime perception scenario addressed in this study, a unified maritime ship detection dataset was constructed by integrating multiple publicly available EO benchmarks. Unlike SAR or infrared datasets, which differ significantly in imaging mechanisms and feature characteristics, the selected datasets focus exclusively on real-world visible light maritime imagery, thereby maintaining consistency with the intended application domain.

Specifically, the dataset integrates SeaShips [30], the Singapore Maritime Dataset (SMD) [31], and McShips [32], all of which are widely adopted EO maritime benchmarks captured under diverse viewpoints and environmental conditions. In addition, a subset of open-source ship images was collected from a publicly available ship dataset hosted on Roboflow Universe [33], released under the CC BY 4.0 license. This subset was incorporated to further enhance diversity in vessel categories, scales, and viewing conditions.

All images originate from real-world maritime electro-optical sensing scenarios captured under natural illumination. The dataset covers challenging environmental conditions, including haze, backlighting, wave-induced reflections, occlusions, and complex sea–sky boundaries. No synthetic or simulated imagery is included, ensuring that the evaluation reflects realistic maritime operational environments.

To ensure annotation reliability and dataset consistency, all images and corresponding labels were manually inspected to remove redundant samples, low-quality frames, and inconsistent bounding boxes. Potentially duplicated or highly similar images—particularly those extracted from continuous video sequences—were sparsely sampled to reduce redundancy and mitigate temporal bias. All annotations were first standardized in the Pascal VOC format and subsequently converted to the YOLO format for network training and evaluation. Category definitions across different source datasets were unified through manual mapping to ensure semantic consistency among heterogeneous annotations.

The final curated dataset contains 5000 images. Although moderate in scale, the dataset is constructed from heterogeneous maritime sources and strictly filtered to avoid excessive redundancy, which helps preserve effective diversity while preventing bias introduced by highly correlated video frames. The dataset is divided into training, validation, and testing subsets with a ratio of 7:2:1. Images originating from the same video sequence or highly similar scenes were assigned to the same subset to prevent potential data leakage and ensure fair evaluation.

The dataset covers seven representative maritime target categories, namely buoy, naval vessels, sailboat, ferry, vessel ship, speedboat, and boat, consistent with the class definitions shown in Figure 6. The composition of data sources is summarized in Table 1, while the statistical characteristics of class distribution, bounding box locations, and width–height relationships are illustrated in Figure 6. Notably, the dataset exhibits a highly imbalanced class and aspect ratio distribution, particularly for elongated ship targets such as the vessel ship and naval vessels. This characteristic directly motivates the geometric constraints introduced in the proposed loss function, establishing a clear connection between the dataset properties and model design.

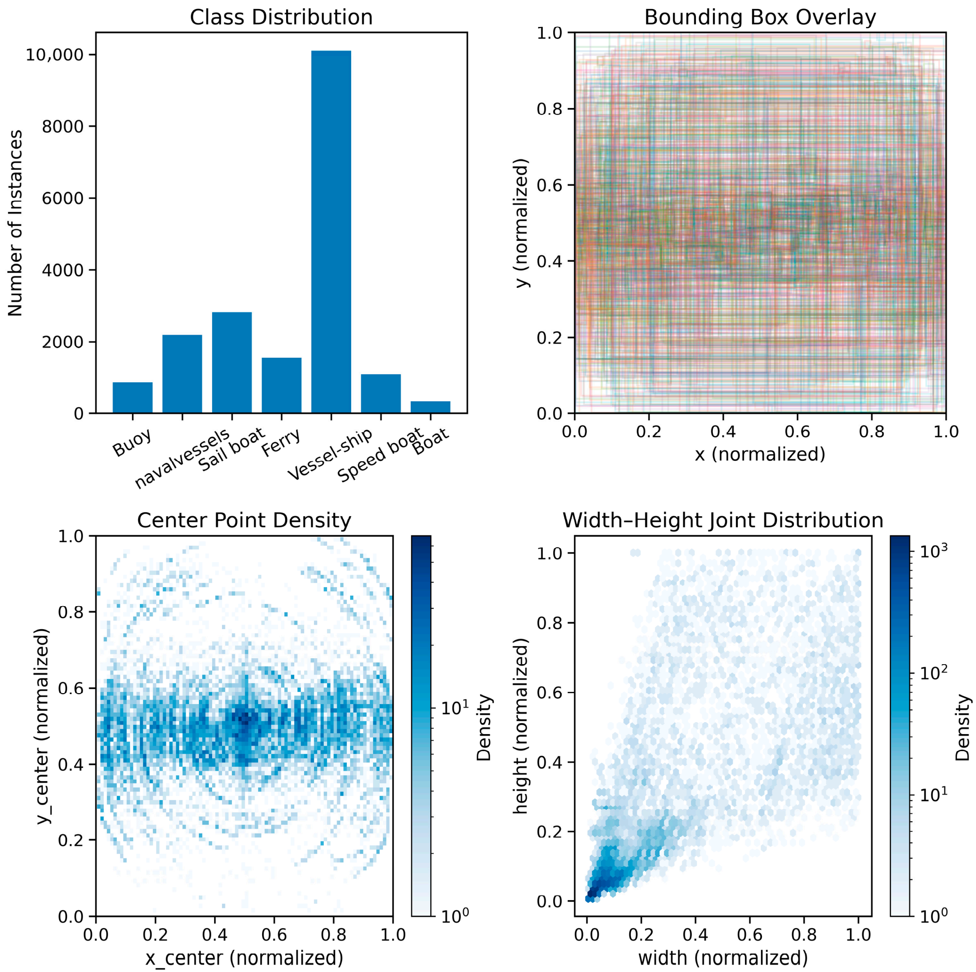


Figure 6. Dataset class distribution and annotation statistics.

Table 1. Composition of the constructed maritime ship detection dataset.

Source Dataset	Platform	Selected Images
SeaShips	Public	1500
SMD	Public	1500
McShips	Public	1500
Ship Computer Vision Dataset	Roboflow Universe	500

4.2. Experimental Environment and Parameter Settings

To evaluate the detection performance of the proposed method on maritime ship targets, all experiments were implemented in Python using the PyTorch deep learning framework. The evaluation was conducted on real-world maritime image datasets with controlled data augmentation, following a standardized offline experimental protocol commonly adopted in maritime vision research.

All experiments were performed on a workstation running Windows 10, equipped with an Intel Core i5-12600KF CPU and an NVIDIA GeForce RTX 4060 GPU. The software environment includes Python 3.10, PyTorch 2.1.0, and CUDA 11.8. The backbone network was initialized with pretrained weights to accelerate convergence, and a cosine learning rate decay strategy was applied during training.

Unless otherwise specified, all models were trained and evaluated with an input resolution of 640×640. For real-time performance evaluation, FPS was measured with an inference batch size of 1 and includes both forward inference and non-maximum suppression (NMS) under FP32 precision on a single GPU. All compared models were evaluated under identical input resolution, confidence threshold, and NMS settings to ensure a fair and consistent comparison.

The primary training hyperparameters and evaluation configurations are summarized in Table 2. In addition to detection accuracy, model complexity and real-time performance were jointly assessed using the number of parameters (Params), GFLOPs, and FPS, in order to evaluate the practical suitability of different detection models for resource-constrained maritime sensing and automation platforms.

Table 2. Hyperparameter settings and evaluation configuration.

Category	Parameter	Value
Training Hyperparameters	Optimizer	SGD
	Weight decay	0.0005
	Initial learning rate	0.01
	Epochs	100
	Batch size	16
	Momentum	0.937
	NMS IoU threshold	0.50
Evaluation Configuration	Input size	640×640
	Inference batch size	1
	FPS includes	Forward + NMS
	Hardware	NVIDIA RTX 4060 (PyTorch FP32)

4.3. Result Analysis

4.3.1. Overall Quantitative Comparison

In this study, precision, recall, and mean average precision (mAP) are adopted as the primary evaluation metrics. Precision reflects the accuracy of the predicted results, recall measures the completeness of target detection, and mAP provides an overall assessment of the detector's performance. To ensure a comprehensive and fair evaluation, mAP is computed under two Intersection over Union (IoU) settings. Specifically, mAP@0.5 fixes the IoU threshold at 0.5 and emphasizes localization accuracy, while mAP@0.5:0.95 averages the mAP values over IoU thresholds from 0.5 to 0.95 with a step size of 0.05, providing a stricter and more robust evaluation under different target scales and complex backgrounds. These metrics are widely regarded as standard benchmarks for object detection tasks.

Under these evaluation criteria, the proposed method is systematically compared with several mainstream YOLO-series detectors, including YOLOv5n, YOLOv6n, YOLOv7-tiny, YOLOv8n, and YOLO11n. In addition to detection accuracy, computational complexity and real-time performance are also considered, including GFLOPs, the number of parameters (Params), and inference speed (FPS). The comprehensive quantitative results are reported in Table 3.

Table 3. Quantitative comparison results of different detection models.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	GFLOPs	Params (M)	FPS
YOLOv5n	0.79795	0.70429	0.74256	0.50229	7.840	2.655	183.40
YOLOv6n	0.80241	0.66583	0.69338	0.47589	12.691	5.058	104.04
YOLOv7-tiny	0.8274	0.7079	0.78610	0.49780	13.860	6.227	237.55
YOLOv8n	0.82799	0.71016	0.77255	0.53943	8.858	3.157	198.11
YOLO11n	0.77149	0.70708	0.75223	0.52740	6.457	2.591	142.00
Ours	0.87374	0.7279	0.78686	0.55570	9.710	5.904	103.57

As shown in Table 3, the proposed method achieves the best overall detection performance among all compared models. The Precision reaches 0.8737, which is approximately 4.6 percentage points higher than that of YOLOv8n, demonstrating a strong capability in suppressing false detections and improving prediction reliability. The recall is improved to 0.7279, exceeding YOLOv8n by about 2.5 percentage points, indicating enhanced detection completeness in dense and cluttered maritime scenes. In terms of localization accuracy, the proposed method obtains an mAP@0.5 of 0.7869 and an mAP@0.5:0.95 of 0.5557, outperforming all comparative models. These consistent improvements indicate that the proposed architectural enhancements effectively strengthen feature discrimination and localization robustness under sensor-impaired maritime conditions.

From the perspective of computational efficiency, Table 3 reports both detection accuracy and model complexity under a unified evaluation configuration (Table 2). The proposed model requires 9.71 GFLOPs, which is higher than YOLO11n but remains within the lightweight complexity range and is lower than heavier nanoscale variants such as YOLOv6n and YOLOv7-tiny. The resulting model contains 5.904 M parameters, reflecting a moderate increase in model capacity introduced by the AESA and GAMF modules while preserving a compact architecture suitable for maritime sensing applications. With an inference speed of 103.57 FPS on an RTX 4060 GPU under the evaluation configuration summarized in Table 2, the proposed detector maintains real-time capability. Although the inference speed is lower than extremely lightweight variants such as YOLOv7-tiny and YOLOv8n, it remains sufficient for real-time maritime electro-optical applications.

Overall, the proposed method strikes a favorable balance between detection accuracy, model complexity, and real-time performance. Its lightweight yet expressive architecture and robust adaptability to complex maritime environments make it well suited for deployment in resource-constrained maritime perception platforms, contributing to more reliable and efficient ship detection in practical monitoring scenarios.

In addition to the unified evaluation setting, further verification on standard maritime benchmarks is necessary to assess generalization capability and ensure comparability with prior studies. To address this concern, we additionally conducted cross-dataset evaluations on the official test splits of SeaShips, SMD, and McShips. For fair comparison, the same training protocol was adopted, and single-class ship detection results are reported due to taxonomy differences across datasets. The quantitative results are summarized in Table 4.

Table 4. Cross-dataset evaluation results on official benchmark test splits.

Dataset	YOLO11n		Ours	
	mAP@0.5	mAP@0.5:0.95	mAP@0.5	mAP@0.5:0.95
SeaShips	0.89047	0.58107	0.91388	0.62312
SMD	0.80278	0.48741	0.81682	0.49046
McShips	0.88863	0.64864	0.90606	0.65959

As shown in Table 4, the proposed method consistently outperforms the baseline YOLO11n across all three benchmark datasets. On SeaShips, the proposed detector achieves an improvement of 2.34% in mAP@0.5 and 4.20% in mAP@0.5:0.95. On McShips, gains of 1.74% and 1.10% are observed under the two metrics, respectively. Although the improvement on SMD is relatively modest, the proposed model still maintains a stable performance advantage. These results indicate that the observed improvements can generalize effectively to official benchmark test splits, thereby enhancing the experimental robustness and reproducibility of the study.

4.3.2. Category-Level Performance Analysis

To gain deeper insight into category-specific behavior, a fine-grained evaluation is performed across major ship types. The corresponding results are presented in Table 5.

Table 5. Comparison of mAP@0.5 of various detection models across different categories.

Model	Buoy	Naval Vessels	Sailboat	Ferry	Vessel Ship	Speed	Boat
YOLOv5n	0.768	0.650	0.892	0.860	0.789	0.408	0.829
YOLOv8n	0.789	0.693	0.936	0.925	0.830	0.380	0.858
YOLO11n	0.764	0.648	0.875	0.876	0.789	0.411	0.836
Ours	0.801	0.683	0.944	0.927	0.839	0.445	0.894

Table 5 shows that different detection models exhibit significant performance differences across various ship targets. For target categories with larger scales and relatively stable structures (such as ferry, sailboat, and boat), all YOLO baseline models achieved high detection accuracy, indicating that the basic recognition capabilities of mainstream target detection frameworks for these targets are relatively mature.

In contrast, for target categories such as speedboats, which exhibit high-speed motion, are small in scale, and are susceptible to wave reflection and motion blur, the detection performance of the baseline model significantly declines, with AP values generally below 0.45. This reflects that under complex ocean electro-optic imaging conditions, small-scale high-speed targets remain a significant challenge limiting detection performance.

Benefiting from the adaptive enhancement of discriminative features by the Adaptive Expert Selection Attention (AESA) module and the effective modeling of directional contextual information by the Geometry-Aware MultiScale Fusion (GAMF) module, the proposed method achieves significant performance improvements in more challenging categories such as speedboats, vessel ships, and naval vessels. The improvement in mAP@0.5 is particularly noticeable in the speedboat category, demonstrating the method's significant advantages in suppressing complex background interference and enhancing the localization stability of slender, highspeed targets. It should be noted that in some static or large-scale target categories (such as boats), the detection accuracy of the proposed method fluctuates slightly compared to the baseline model; however, its significant improvements in overall mAP and key challenging categories indicate that the proposed method achieves a more reasonable trade-off between detection accuracy, robustness, and engineering practicality.

4.3.3. Qualitative Visualization Analysis

The effectiveness of the proposed method in handling sea–air background interference under low signal-to-noise ratio conditions is further examined using GradCAM visualization (Figure 7).

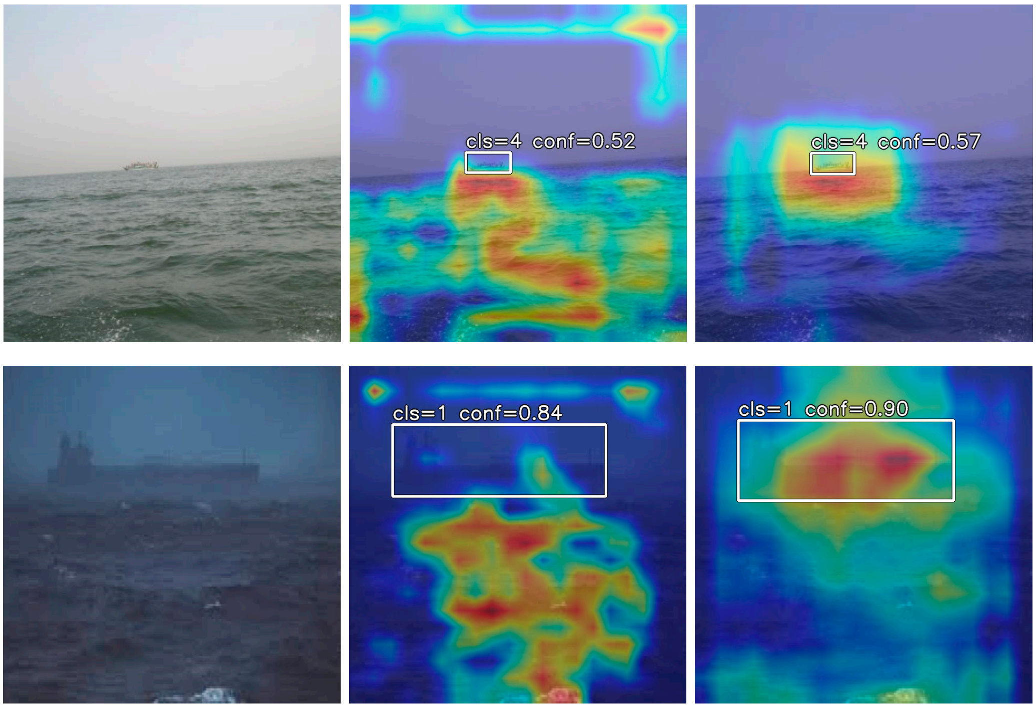


Figure 7. GradCAM visualization comparison between baseline YOLO11n and the proposed method under low-SNR maritime scenes. (**a**) Original image; (**b**) YOLO11; (**c**) ours.

Unlike purely metric-based evaluation, the visualization results reveal spatial attention redistribution toward foreground ship contours. In the first scenario (upper row), which contains small-scale distant targets under relatively low contrast conditions, the baseline YOLO11n model exhibits dispersed activation responses across the sea surface and horizon region. Strong responses are observed in large background areas with wave reflections, indicating that the detector is partially distracted by sea clutter. In contrast, the proposed method produces more concentrated activation around the actual ship target, while significantly reducing irrelevant responses in the surrounding sea–sky background. This demonstrates that the enhanced feature selection and contextual modeling mechanisms help the network better distinguish foreground objects from background interference.

In the second scenario (lower row), captured under heavy haze and complex illumination conditions, the baseline model shows extensive high-intensity activation across non-target regions, particularly over wave patterns and high-frequency textures. Such responses suggest sensitivity to background noise. However, the improved model exhibits a notably more compact activation region tightly aligned with the ship contour, while suppressing spurious responses over the sea surface. The confidence score in the representative example increases from 0.84 to 0.90.

These visualization results confirm that the proposed Adaptive Expert Selection Attention (AESA) and Geometry-Aware MultiScale Fusion (GAMF) modules effectively enhance discriminative feature representation and reduce the influence of complex maritime backgrounds. Therefore, the proposed method demonstrates stronger robustness in low-contrast and background-interfered maritime electro-optical scenes.

4.3.4. Quantitative Robustness Evaluation Under Adverse Weather Conditions

To examine model behavior under degraded maritime imaging conditions, additional quantitative evaluations were conducted on dense fog and low-contrast subsets extracted from the test set. The results are shown in Table 6.

Table 6. Quantitative robustness evaluation under adverse maritime imaging conditions.

Condition	YOLO11n		Ours	
	mAP@0.5	mAP@0.5:0.95	mAP@0.5	mAP@0.5:0.95
Complex background	0.731	0.512	0.7831	0.537
Dense fog	0.710	0.489	0.766	0.533
Low contrast	0.728	0.508	0.775	0.542

As shown in Table 6, the proposed method consistently outperforms the baseline YOLO11n under all three adverse conditions. The improvement is particularly notable under dense fog, where mAP@0.5 increases by 5.6% and mAP@0.5:0.95 by 4.4%, indicating enhanced localization stability under severe visibility degradation. Similar performance gains are observed under complex background and low-contrast conditions, demonstrating that the proposed attention-guided multiscale feature modeling effectively suppresses background interference and improves feature discriminability in visually degraded maritime scenes.

4.4. Ablation Experiment Analysis

4.4.1. Ablation Experiment Analysis of Each Component of the AESA Module

To quantify the contribution of each key design component within the AESA module, a series of structured ablation experiments are performed. Specifically, four core components—spatial attention, branch Softmax fusion, intragroup competition, and multiscale ECA—are selectively removed while keeping all other settings unchanged, enabling a clear assessment of their functional roles. The quantitative results are summarized in Table 7.

Table 7. Ablation results of key components in the proposed AESA module.

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Without Spatial Attention	0.82508	0.71355	0.77204	0.54594
Without Branch Softmax Fusion	0.83348	0.71901	0.77787	0.54764
Without Intragroup Competition	0.82855	0.71769	0.78023	0.54939
Without Multiscale ECA	0.82952	0.70535	0.77366	0.53642
Complete Improved Model	0.87374	0.7279	0.78686	0.5557

As shown in Table 7, removing any submodule leads to a consistent degradation in detection performance across precision, recall, and mAP metrics, confirming that each component contributes positively to the overall effectiveness of the AESA mechanism. Among all components, disabling the multiscale ECA branch and the spatial attention block results in the most pronounced performance drops. When the multiscale ECA branch is removed, precision and recall decrease from 0.8737 and 0.7279 to 0.8295 and 0.7054, respectively, while mAP@0.5:0.95 drops by 1.93 percentage points. Similarly, removing the spatial attention block causes precision and recall to decline to 0.8251 and 0.7136, accompanied by an approximately 1.0 percentage point reduction in mAP@0.5:0.95. These results indicate that multiscale channel modeling and spatial attention play critical roles in suppressing background interference and enhancing discriminative feature representation in complex maritime environments.

In contrast, removing the branch Softmax fusion or the intra-group competition mechanism results in relatively smaller, yet still noticeable, performance degradation. Under these settings, mAP@0.5:0.95 decreases by 0.81 and 0.63 percentage points, respectively, suggesting that these components provide auxiliary but meaningful contributions by facilitating adaptive channel selection and coordinated multi-branch feature integration.

Overall, the complete model that integrates all AESA components achieves the best performance across all evaluation metrics. These ablation results demonstrate that the proposed AESA design is not a simple aggregation of attention operations, but a complementary and synergistic combination of multiscale channel modeling, competitive routing, and shared spatial modulation, which together enhance robustness and detection accuracy under complex maritime sensing conditions.

4.4.2. Sensitivity Analysis of Expert Group Number K

To verify the rationality of the number of expert groups K in the proposed AESA module, this paper conducted a comparative experiment on $K \in \{1,2,4,8\}$ while keeping the experimental configurations, including the training dataset, number of epochs, optimizer settings, and input resolution (imgsz = 640) completely consistent. The results are shown in Table 8.

Table 8. Detection performance and computational complexity under different expert group numbers K.

K	mAP@0.5	GFLOPs	Params (M)	FPS
1	0.7650	9.685	5.901	126.63
2	0.7704	9.694	5.901	117.43
4	0.7869	9.710	5.904	103.57
8	0.7694	9.746	5.905	80.59

From the perspective of detection accuracy, as the number of expert groups increases from 1 to 4, mAP@0.5 improves from 0.7650 to 0.7869. This trend indicates that moderate expert partitioning enhances feature diversity and strengthens intragroup competitive routing, thereby improving discriminative capability in complex maritime scenarios. However, when K further increases to 8, mAP@0.5 decreases to 0.7694, suggesting that excessive expert partitioning may lead to fragmented feature aggregation and reduced optimization stability under the current dataset scale.

From the perspective of computational complexity and inference efficiency, GFLOPs exhibit a gradual growth trend (9.685 $\rightarrow$ 9.746) and the number of parameters increases slightly (5.901M $\rightarrow$ 5.905M), which aligns with the linear expansion characteristic of the grouped expert projection in the AESA structure. Meanwhile, the end-to-end detection speed (including preprocessing and NMS postprocessing) decreases monotonically with K (126.63 FPS $\rightarrow$ 80.59 FPS), indicating that larger K values inevitably introduce additional computational overhead.

Considering both detection performance and efficiency, the model achieves the highest mAP@0.5 (0.7869) at K = 4 while maintaining acceptable real-time inference speed (103.57 FPS). Therefore, $K = 4$ represents a balanced configuration between representational flexibility and computational cost and is selected as the default setting in subsequent experiments.

4.4.3. Ablation Experiment Analysis of Main Modules

An additional ablation study is conducted based on the YOLO11n baseline to evaluate the respective contributions of the three main components: the AESA module, the GAMF module, and the WIoU with aspect ratio regularization (WIoU + AR). The quantitative results are reported in Table 9.

Table 9. Ablation study of the main components in the proposed detector.

Model	AESA	GAMF	WIoU + AR	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLO11n (Baseline)				0.77149	0.70708	0.75223	0.52740
YOLO11n + AESA	√			0.85577	0.68986	0.76582	0.54002
YOLO11n + GAMF		√		0.82695	0.71858	0.77178	0.53693
YOLO11n + AESA + GAMF	√	√		0.83793	0.73234	0.77672	0.53008
Ours (AESA + GAMF + WIoU + AR)	√	√	√	0.87374	0.7279	0.78686	0.55570

As shown in Table 9, introducing the AESA module alone leads to a substantial improvement in precision, increasing from 0.77149 to 0.85577, which demonstrates its strong capability in suppressing background interference and reducing false detections. Meanwhile, mAP@0.5:0.95 improves to 0.54002. However, a slight decrease in recall is observed, indicating that the enhanced feature selectivity introduced by AESA may also suppress certain weak or ambiguous ship targets.

In contrast, introducing the GAMF module alone yields a more balanced performance gain. Precision increases from 0.77149 to 0.82695, while recall is improved from 0.70708 to 0.71858. Meanwhile, mAP@0.5 and mAP@0.5:0.95 increase to 0.77178 and 0.53693, respectively. These results suggest that GAMF effectively enhances multiscale feature aggregation and improves detection completeness, particularly for ship targets with large scale variations and elongated structures, without introducing excessive false positives.

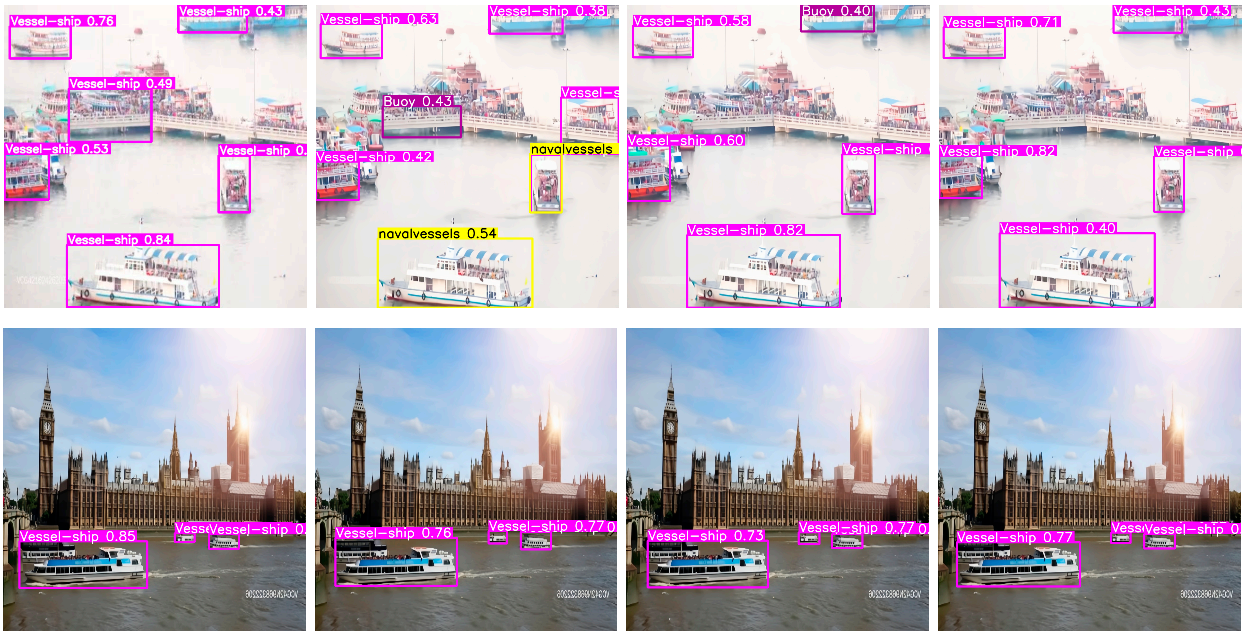
When AESA and GAMF are jointly integrated, the model achieves further improvements, particularly in recall, confirming their complementary roles in feature enhancement and multiscale fusion. Finally, when all three components—AESA, GAMF, and WIoU+AR—are incorporated, the proposed full model achieves the best overall performance, reaching 0.87374 in precision, 0.72790 in recall, 0.78686 in mAP@0.5, and 0.55570 in mAP@0.5:0.95.

These results indicate that AESA and GAMF serve as the primary contributors to performance improvement through adaptive feature selection and orientation-aware multiscale fusion, while WIoU+AR provides an effective complementary refinement by enhancing bounding box localization consistency. By jointly addressing feature noise, multiscale inconsistency, and geometric instability, the integrated framework achieves a well-balanced improvement across precision, recall, and multi-IoU evaluation metrics under complex maritime environments.

4.5. Visual Comparison and Analysis

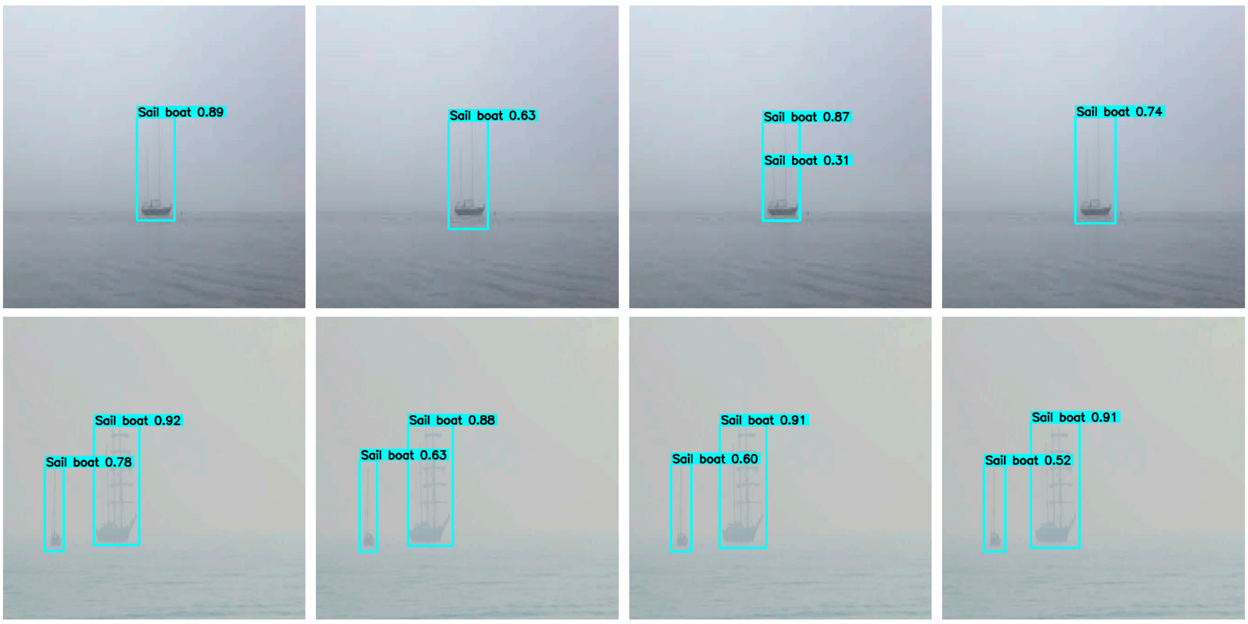
To further examine the robustness of the proposed method under representative maritime degradation conditions, qualitative comparisons are conducted on three typical adverse scenarios derived from the test set, including complex background interference, dense fog obscuration, and low-contrast environments. These scenarios correspond to the quantitative robustness subsets presented in Table 6 and reflect distinct degradation mechanisms encountered in real-world maritime electro-optic imaging.

For each scenario, detection results of the proposed model and several comparative methods are visually analyzed under identical inference settings. The visualization results are presented in Figures 8–10.



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

Figure 8. Qualitative comparison of detection results for complex background interference. The upper and lower rows present two representative comparison cases. (**a**) Ours; (**b**) YOLO11; (**c**) YOLOv8; (**d**) YOLOv6. All results are generated using the same input resolution, confidence threshold, and non-maximum suppression settings.



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

Figure 9. Qualitative comparison of target detection results in dense fog weather. The upper and lower rows present two representative comparison cases. (**a**) Ours; (**b**) YOLO11; (**c**) YOLOv8; (**d**) YOLOv6. All results are generated using the same input resolution, confidence threshold, and non-maximum suppression settings.

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

Figure 10. Qualitative comparison of target detection results in low-contrast images. The upper and lower rows present two representative comparison cases. (**a**) Ours; (**b**) YOLO11; (**c**) YOLOv8; (**d**) YOLOv6. All results are generated using the same input resolution, confidence threshold, and non-maximum suppression settings.

Figure 8 presents the qualitative comparison results under complex maritime background conditions. In this scenario, the presence of background textures, water reflections, and visually similar structures increases the risk of false positives and class confusion. It can be observed that baseline models such as YOLOv8 and YOLOv6 tend to produce false detections in background regions with structural similarity to ship targets or exhibit unstable confidence scores for partially occluded vessels. Although YOLO11n improves detection robustness to some extent, it still shows inconsistencies in suppressing background interference and occasionally misclassifies small-scale vessels in cluttered areas.

In contrast, the proposed method demonstrates stronger background suppression capability and more stable confidence predictions. Adjacent vessels are more accurately localized, and the background structures demonstrated are less likely to trigger false responses. This indicates that the introduced adaptive attention and multiscale feature fusion mechanisms effectively enhance foreground–background discrimination under complex maritime scenes.

Figure 9 illustrates the qualitative comparison results under dense fog conditions. In this scenario, fog scattering significantly reduces imaging contrast, blurs the ship target outline, and weakens the grayscale difference between the background and the target. The baseline models used in the comparison generally exhibit problems such as reduced detection confidence and unstable localization under these conditions, with some samples showing a risk of false detection. In contrast, the proposed method can more stably activate discriminative features related to the ship target under dense fog conditions, maintain a high and consistent confidence output for the target, and effectively suppress background noise response, demonstrating superior robustness in low-contrast and strong fog-occlusion environments.

Figure 10 illustrates the qualitative contrast results under low-contrast sea conditions. In this scenario, the overall grayscale difference between the vessel target and the surrounding sea and air background is significantly compressed, increasing the risk of missed detections and unstable confidence levels. It can be seen that the baseline model exhibits incomplete target localization or low confidence levels under these conditions. In some cases, small vessels partially blend into the background texture, resulting in insufficient fine-grained boundary separation capabilities. The proposed method can more effectively suppress responses from non-target areas and maintain stable and accurate detection results for ship targets, demonstrating stronger generalization ability under complex background conditions.

In contrast, the proposed method generates more compact bounding boxes and more stable confidence levels for both near and far vessels, effectively reducing missed detections in visually compressed scenes. These results demonstrate that the adaptive attention mechanism and geometrically aware multiscale fusion contribute to improved feature amplification under low-contrast sea conditions.

Overall, the qualitative comparison results are highly consistent with the aforementioned quantitative experimental conclusions, indicating that the proposed method not only achieves a stable improvement in detection accuracy but also significantly enhances the visual reliability and temporal consistency of the detection results under complex maritime electro-optic imaging conditions. These characteristics have significant engineering application value for real-time marine monitoring and electro-optic sensing systems operating in degraded environments.

5. Conclusions

In this study, an enhanced ship detection framework is developed to support reliable real-time perception in complex maritime automation environments. By integrating the Adaptive Expert Selection Attention (AESA) module and the Geometry-Aware MultiScale Fusion ((GAMF) module, the proposed approach explicitly addresses feature instability, background interference, and scale inconsistency commonly encountered in degraded maritime electro-optical imagery. In addition, a refined regression loss combining WiseIoU with aspect ratio regularization further strengthens geometric consistency and localization stability for ships with diverse scales and elongated structures.

From an imaging and perception perspective, the proposed framework enhances feature stability and geometric consistency under degraded electro-optical imaging conditions, thereby enabling more reliable object perception in complex maritime scenes. Extensive experiments conducted on a unified real-world maritime benchmark integrating multiple publicly available datasets demonstrate that the proposed method consistently outperforms the baseline detector and representative YOLO-series models in terms of precision, recall, and mean average precision.

Importantly, these performance gains are achieved while maintaining a lightweight architecture and real-time inference capability. This makes the proposed method well suited for deployment as a perception component in onboard electro-optical cameras, shore-based surveillance systems, and embedded maritime automation platforms where computational resources are limited.

Overall, the proposed framework achieves a practical balance between detection accuracy, computational efficiency, and sensing robustness. However, certain limitations remain to be addressed. Specifically, the detection of extremely small-scale ship targets (e.g., those occupying fewer than 16×16 pixels) remains a performance bottleneck; despite the multiscale capabilities of the AESA module, critical structural features of such targets are often lost during successive down-sampling in the backbone, leading to localization instability in far-field scenarios. Furthermore, while this study focuses on electro-optical (EO) imagery, the potential for multi-sensor fusion remains a significant area for enhancement. Integrating the GAMF module's spatial-aware fusion logic with cross-modal data—such as EOInfrared (IR) or EORadar—could leverage thermal signatures and range–azimuth priors to overcome the inherent physical limitations of optical sensors in visually degraded or low-light conditions.

Future work will therefore focus on overcoming the small target bottleneck through high-resolution feature compensation or generative adversarial refinement. Additionally, we will investigate advanced multi-sensor fusion architectures to achieve cross-modal complementarity, extending the framework toward robust, all-weather, and day–night maritime perception to further support real-time deployment in fully autonomous maritime platforms.

Author Contributions: Writing—original draft, software, methodology, investigation, formal analysis, visualization, and data curation: Z.Z.; Supervision, conceptualization, writing—review and editing, project administration, and funding acquisition: C.L.; Resources, methodology, and writing—review and editing: B.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fujian Provincial Natural Science Foundation (No. 2022J05257), the Fujian Provincial Guiding Science and Technology Project (No. 2024H0022), the 2024 Central Government Guiding Local Science and Technology Development Project (No. 2024L3030), and the Scientific Research Start-up Project of Putian University (No. 2026018).

Data Availability Statement: The dataset used in this study was constructed by integrating multiple publicly available maritime electro-optical datasets (e.g., SeaShips, SMD, and McShips), together with additional open-source samples. Due to the heterogeneous license terms and redistribution restrictions associated with some data sources, the unified dataset and revised annotations cannot be released as a single public package. However, all original datasets remain publicly accessible through their respective official repositories. To facilitate reproducibility, detailed descriptions of data sources, category mapping rules, data cleaning and filtering procedures as well as the train/validation/test split protocol are provided in this paper.

Acknowledgments: The authors used ChatGPT (GPT-5.3) for language editing and grammar checking to improve the readability of the manuscript. The authors take full responsibility for the content of this article.

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

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
AESA	Adaptive Expert Selection Attention
GAMF	Geometry-Aware MultiScale Fusion
SPPF	Spatial Pyramid Pooling—Fast
FPN	Feature Pyramid Network
PAN	Path Aggregation Network
ECA	Efficient Channel Attention
CA	Coordinate Attention
IoU	Intersection over Union
mAP	mean Average Precision
WIoU	Wise Intersection over Union
DIoU	Distance Intersection over Union
CIoU	Complete Intersection over Union
SIoU	Scaled Intersection over Union
DFL	Distribution Focal Loss
GAP	Global Average Pooling
GMP	Global Max Pooling
SGD	Stochastic Gradient Descent
NMS	Non-Maximum Suppression

References

1. Terven, J.; CórdovaEsparza, D.M.; RomeroGonzález, J. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLONAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
2. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [CrossRef]
3. Zhang, M.; Rong, X.; Yu, X. LightSDNet: A Lightweight CNN Architecture for Ship Detection. *IEEE Access* **2022**, *10*, 10862–10873. [CrossRef]
4. Liang, S.; Liu, X.; Yang, Z.; Liu, M.; Yin, Y. Offshore Ship Detection in Foggy Weather Based on Improved YOLOv8. *J. Mar. Sci. Eng.* **2024**, *12*, 1641. [CrossRef]
5. Jiang, Z.; Su, L.; Sun, Y. YOLOv7-Ship: A Lightweight Algorithm for Ship Object Detection in Complex Marine Environments. *J. Mar. Sci. Eng.* **2024**, *12*, 190. [CrossRef]
6. Zhao, T.; Wang, Y.; Li, Z.; Gao, Y.; Chen, C.; Feng, H.; Zhao, Z. Ship Detection with Deep Learning in Optical Remote Sensing Images: A Survey of Challenges and Advances. *Remote Sens.* **2024**, *16*, 1145. [CrossRef]
7. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
8. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. WiseIoU: Bounding Box Regression with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
9. Hu, J.; Shen, L.; Sun, G. SqueezeandExcitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
10. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECANet: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
11. Yao, J.; Xiao, S.; Deng, Q.; Wen, G.; Tao, H.; Du, J. An Infrared Maritime Small Target Detection Algorithm Based on Semantic, Detail, and Edge Multidimensional Information Fusion. *Remote Sens.* **2023**, *15*, 4909. [CrossRef]
12. Zheng, J.; Liu, Y. A Study on Small-Scale Ship Detection Based on Attention Mechanism. *IEEE Access* **2022**, *10*, 77940–77949. [CrossRef]
13. Guo, Y.; Wang, Z.; Chen, Z. A TwoStage AttentionEnhanced Network for Ship Detection in Optical Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2487. [CrossRef]
14. Zhang, Z.; Zhao, Q.; Li, X.; Wang, C.; Zhu, G.; Zhang, Y.; Huo, Y.; Yu, H.; Zhang, Y. CAYOLO: Cross Attention Empowered YOLO for Biomimetic Localization. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 3598306. [CrossRef]
15. Ren, Y.; Zhu, C.; Xiao, S. Small Object Detection in Remote Sensing Images Based on Improved YOLOv5. *Sensors* **2022**, *22*, 3319. [CrossRef]
16. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2017; pp. 2117–2125. [CrossRef]
17. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. [CrossRef]
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 346–361. [CrossRef]
19. Liu, S.; Liu, P.; Zhang, Z.; Sun, M.; He, P. YOLOPFA: Advanced MultiScale Feature Fusion and Dynamic Alignment for SAR Ship Detection. *J. Mar. Sci. Eng.* **2025**, *13*, 1936. [CrossRef]
20. Mbietieu, M.A.; Tapamo, K.H.M.; Kouamou, G.E.; Tsopze, N. ShipFPN: New Feature Pyramid Network Architecture for Object Detection in HighResolution Satellite Images: Application to Ship Detection. *ISPRS J. Photogramm. Remote Sens.* **2025**, *12*, 100270. [CrossRef]
21. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790. [CrossRef]
22. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for SingleShot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10593–10602. [CrossRef]
23. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722. [CrossRef]

24. Li, Y.; Zhang, H.; Xu, J.; Chen, L. KFIANet: A Knowledge Fusion and ImbalanceAware Network for MultiCategory SAR Ship Detection. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *125*, 105127. [CrossRef]
25. Wang, T.; Liu, Q.; Zhao, Y.; Sun, Z. A Ship Incremental Recognition Framework via Unknown Extraction and Joint Optimization Learning. *Remote Sens.* **2026**, *18*, 149. [CrossRef]
26. Zhou, K.; Liu, H.; Zhang, Y.; Wu, P. MultiGranularity Feature Fusion Network Integrating ASC for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 3649343. [CrossRef]
27. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. DistanceIoU Loss: Faster and Better Learning for Bounding Box Regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
28. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740. [CrossRef]
29. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position Detection and Direction Prediction for Arbitrary-Oriented Ships via Multitask Rotation Region Convolutional Neural Network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
30. Shao, Z.; Cai, J.; Wang, Z.; Zhao, H.; Zhang, Y. SeaShips: A LargeScale Precisely Annotated Dataset for Ship Detection. *IEEE Trans. Multimed.* **2018**, *20*, 2596–2607. [CrossRef]
31. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video Processing from ElectroOptical Sensors for Object Detection and Tracking in a Maritime Environment. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2006. [CrossRef]
32. Zheng, Y.; Zhang, S. McShips: A LargeScale Ship Dataset for Detection and FineGrained Categorization in the Wild. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6. [CrossRef]
33. Roboflow. Ship Dataset. Roboflow Universe. 2023. Available online: https://universe.roboflow.com/shipmjq2z/ship-9pfw0 (accessed on 21 December 2024).