

Article

Robust Ship Detection Algorithm Under Complex Occlusion Conditions

Jiahang Li , Yan Zhang ^{*}, Yu Sun  and Churuo Zhang 

School of Navigation and Shipping, Shandong Jiaotong University, Weihai 264200, China;
ljh20012024@163.com (J.L.); sunyu@sdjtu.edu.cn (Y.S.); 19510161612@163.com (C.Z.)

^{*} Correspondence: zhangyan1503@sdjtu.edu.cn

Abstract

To address the accuracy degradation of ship detection caused by occlusion from adjacent vessels, shore-based facilities and meteorological obscuration in complex maritime-surveillance scenes, this paper proposes an occlusion-robust detection model named OAR-YOLO. An adaptive dual-path downsampling module termed ADown was embedded at the three backbone levels P3, P4 and P5, in which low-frequency contextual information and high-frequency edge information were preserved separately through parallel average-pooling and max-pooling branches, alleviating the information loss caused by conventional strided-convolution downsampling. An attention-driven intra-scale feature interaction module termed AIFI was embedded at the top level P5 to establish semantic associations between spatially separated visible regions through global self-attention, compensating for the insufficient cross-region connectivity caused by the locality of convolution. The two modules formed a dual compensation mechanism of information conservation and semantic connectivity. On a self-built ship dataset, OAR-YOLO achieved a Precision of 81.0%, an mAP@0.5 of 74.7% and an mAP@0.5–0.95 of 47.1%, with gains of 2.7, 2.6 and 1.4 percentage points over the YOLO11n baseline. The model has only 2.89 M parameters and 5.7 GFLOPs, with an inference time of 0.8 ms per frame, meeting the real-time deployment requirements of complex maritime applications.

Keywords: ship detection; OAR-YOLO; adaptive downsampling; feature interaction

1. Introduction

1.1. The Occlusion Problem in Ship Detection

With the continuous expansion of global shipping throughput and the increasingly refined requirements of maritime regulation, automatic ship detection based on optical imagery has become an indispensable perception front-end in smart-port operations [1], maritime traffic management and emergency response. However, the visual composition of maritime imagery is far more complex than that of ordinary terrestrial scenes: ships lie adjacent to one another in dense anchorages, are blocked by cranes and quay structures inside harbour basins, and are frequently covered by clouds, mist and spray on open-sea routes. These occlusion phenomena are not occasional noise but structural regularities of maritime perception scenes. When occlusion occurs, the morphological integrity of a target is broken, the textural continuity is interrupted, the discriminative information available to the network is markedly reduced, and detection accuracy decreases accordingly.

Recent studies have quantified the magnitude of this degradation. On the widely used CityPersons pedestrian-detection benchmark, where the Reasonable subset covers



Academic Editor: Eugen Rusu

Received: 20 June 2026

Revised: 24 July 2026

Accepted: 27 July 2026

Published: 1 August 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

pedestrians whose visibility ratio exceeds 65% and who are thus at most lightly occluded, and the Heavy Occlusion subset covers those whose visibility ratio ranges from 20% to 65%, mainstream two-stage detectors have been reported to yield log-average miss rates in the range of 13–16% on the Reasonable subset but of 51–57% on the Heavy Occlusion subset [2]—a three-to-fourfold increase in miss rate attributable directly to occlusion, and one that persists across subsequent state-of-the-art single-stage and two-stage detectors. Under adverse-weather corruption benchmarks, YOLO-family detectors have further been shown to lose approximately 10–20% of their clean-image mAP@0.5 when tested on real-world foggy, rainy and snowy scenes [3], with performance dropping more sharply under additional noise augmentation. Comparable stratified benchmarks are not yet publicly available for maritime imagery, but recent studies on ship detection in complex scenes have consistently identified near-shore mutual occlusion, shore-based structure occlusion and atmospheric visibility degradation as the dominant sources of missed detections and false alarms for YOLO-family baselines [4]. These findings collectively indicate that the accuracy drop under occlusion is not incidental noise but a systematic bottleneck of current detection pipelines, motivating occlusion-oriented architectural design.

It should be noted that, in current publicly available ship-detection datasets, the proportion of occluded samples has often been underestimated, while the enrichment of occluded samples is one of the important means of improving model performance under occlusion. Even when researchers have recognised this difficulty, the mainstream responses have remained confined to two categories. The first category is data-driven mitigation, in which strategies such as Mosaic [5], CutMix [6] and Copy-Paste [7] have been used to artificially expand the diversity of occluded samples. The second category is architecture-driven compensation, in which attention mechanisms [8], deformable convolutions [9] and multi-scale alignment modules [10] have been adopted to broaden the network's ability to discriminate partially visible targets. Both lines of work have made progress, but neither has examined a more fundamental question: in mainstream single-stage detection networks, at which stages of feature extraction—and by what mechanisms—the discriminative information of an occluded target is progressively attenuated.

1.2. Status of Detection Network Research

The methodological evolution of ship-target detection has roughly passed through three stages. The first generation was represented by the constant false-alarm rate (CFAR) detector and its variants [11], which relied on the scattering difference between target and background; such methods worked reasonably well over homogeneous backgrounds but degraded markedly in clutter-dense regions such as harbours. The second generation was represented by two-stage deep detectors such as Faster R-CNN [12], in which the region proposal network (RPN) and convolutional feature learning substantially raised accuracy, although inference speed could hardly meet the requirement of real-time maritime monitoring. The third generation has been dominated by single-stage detectors of the YOLO family; early representatives included SSD [13] and YOLOv1 [14], and subsequent updates—YOLOv5 [15], YOLOv7-tiny [16], YOLOv8 [17], YOLOv10 [18] and YOLO11 [19]—have each refined the balance between accuracy and speed. Recent maritime-specific extensions of YOLO11, such as MSC-YOLO [20], have further introduced scene-adaptive attention and class-prototype guidance to address background clutter and fine-grained category confusion in surveillance imagery. The principal driver behind this development has not been merely “deeper networks” or “more parameters”; rather, it has been the progressive refinement of information processing. From early sliding-window classification, through anchor-box regression, to anchor-free key-point prediction [21], the network's way of extracting and judging target features has been continuously im-

proved. Nevertheless, even in newer frameworks such as YOLO11, the improvements have mainly occurred along the semantic-aggregation axis, while the spatial-compression axis has largely inherited the early strided-convolution paradigm.

Mainstream single-stage detection networks such as YOLOv5 [15], YOLOv8 [17] and YOLO11 [19] possess an internal data flow that can be decomposed into two relatively independent axes: a spatial-compression axis and a semantic-aggregation axis. The former downsamples a high-resolution input level by level and builds a multi-scale feature pyramid [22] with a controllable computational cost, while the latter aggregates multi-layer features [23] into joint class and location predictions. In conventional architectures these two axes have evolved independently: the downsampling module has focused on how features are refined, whereas the semantic-aggregation layer has focused on how multi-scale information is fused, with no targeted cooperative design between them under the specific information constraint of occlusion.

It is precisely this lack of cooperation that gives rise to a twofold limitation under occlusion. The first limitation stems from a geometric defect in downsampling: a classical stride-2 convolution compresses a 2×2 neighbourhood into a single pixel along a single path, which forces the network to trade off low-frequency mean response against high-frequency edge response [24]. For unoccluded targets this trade-off can be compensated for by subsequent semantic-aggregation layers; for occluded targets, however, whose visible parts are usually high-frequency edge components, the trade-off causes the discriminative information to be systematically diluted before it reaches the higher layers. The second limitation stems from the locality of the convolution operator: even when high-level feature maps retain the representations of visible regions on either side of an occluded target, the limited receptive field of convolution keeps these visible regions spatially separated, so that mutual identity verification across regions is difficult to achieve. In summary, information is hard to preserve at low levels and semantics are hard to connect at high levels; this is the underlying reason why accuracy drops under occlusion.

1.3. Motivation for OAR-YOLO

Against the problems above, this paper argues that occlusion-oriented architectural improvements should satisfy two core objectives simultaneously—information conservation and semantic connectivity—rather than emphasise only one of them. On this basis, two mechanisms are introduced into the multi-scale processing path of YOLO11n, the Nano-scale variant of YOLO11 chosen to prioritise lightweight deployment in real-time maritime surveillance scenarios where edge-computing resources are constrained. ADown is used to realise information-geometric conservation at low levels; its dual-branch design preserves low-frequency context and high-frequency edge response separately during spatial compression, and the convolutional parameters of the two branches are learned adaptively so that the retention ratio between the two components is dynamically balanced. AIFI is used to realise semantic connectivity at high levels; it applies global self-attention at the top level of the feature pyramid and establishes direct semantic associations between the spatially separated visible regions of occluded targets.

The two mechanisms are not a simple concatenation but constitute a mutually reinforcing pair. ADown preserves the high-frequency information of visible regions before they enter the top layer and thus provides effective input features for AIFI's global reasoning, while AIFI completes cross-position integration at the stage where top-level semantic information is richest and converges the separated visible regions retained by ADown into a complete identity judgement. This dual relationship is the central claim of the present paper and is verified by the ablation experiments reported below.

1.4. Main Contributions

The main contributions of this work are as follows:

- An occlusion-oriented ship-detection model, OAR-YOLO (Occlusion-Aware Robust YOLO), is proposed, with the adaptive dual-path downsampling module ADown and the attention-driven intra-scale feature interaction module AIFI at its core; the two modules jointly form a dual compensation mechanism of information conservation and semantic connectivity, which effectively alleviates the detection-performance drop caused by occlusion.
- A cooperatively designed YOLO11n detection framework, OAR-YOLO, is constructed; spatial-information loss is reduced through dual-path downsampling, and semantic associations between occluded regions are established through intra-scale self-attention, so that information conservation and semantic connectivity together form a dual compensation mechanism.
- A detection dataset covering seven ship categories and various occlusion situations was built for real maritime regulation scenarios, and the proposed method was validated on this dataset. Experimental results indicate that OAR-YOLO improves mAP@0.5 by 2.6 percentage points over the YOLO11n baseline and reaches a class-leading Precision of 0.810, with only 2.89 M parameters and 5.7 GFLOPs of computation; the proposed method therefore outperforms representative methods such as YOLOv5n, YOLOv6n, YOLOv8n and YOLOv10n in the joint trade-off between accuracy and lightweight deployment.

2. Related Work

2.1. Evolution of Downsampling Mechanisms

The downsampling module is an easily overlooked yet important component of detection networks. A classical stride-2 convolution compresses a 2×2 neighbourhood into a single pixel, which is mathematically equivalent to a learnable weighted average; while computational efficiency is improved, high-frequency detail is inevitably lost. MaxPool [25] preserves edge response to some extent by taking the maximum value over a neighbourhood, but it discards low-frequency contextual information. SPDConv [26], proposed in recent years, achieves lossless downsampling through spatial and channel rearrangement, but at the cost of a relatively high parameter count and a heavy computational cost. ADown [27] adopts a more concise compromise: it performs average-pooling and max-pooling in parallel along two paths, concatenates the results along the channel dimension and then projects them, so that low-frequency context and high-frequency edge information are both preserved with only a small additional computational cost. Along a similar dual-path line, EMBSLaw [28] introduces adaptive weighting across multiple branches but at a higher computational cost than ADown's concatenation-based design.

Most evaluations of downsampling designs have, however, remained at the engineering level of feature refinement; few studies have quantitatively analysed the information loss caused by each downsampling step, and even fewer have examined the role of downsampling under the particular constraint of occlusion. This paper holds that, under occlusion, downsampling is no longer a neutral feature-refinement operation but tends to attenuate the discriminative information of partially visible targets. This shift in perspective provides a mechanistic—rather than empirical—justification for adopting ADown.

2.2. Development of Attention Mechanisms

Attention mechanisms have likewise been introduced into object detection in paradigmatically different ways. Channel and spatial attentions such as SE [8], CBAM [29] and CA [30] are usually embedded in intermediate layers of the backbone, aiming to guide the

network to focus on salient regions at an early stage of feature extraction; window-based self-attentions such as Swin Transformer [31] replace convolution throughout the backbone but incur a significant increase in parameter count and computation. AIFI (Attention-based Intra-scale Feature Interaction), which originates from RT-DETR [32], differs noticeably from the above paradigms: a single global self-attention operation [33,34] is applied only at the top level of the feature pyramid, neither replacing convolution nor being embedded in the shallow or middle stages of the backbone [35].

The choice of this location is not arbitrary: the top-level feature map carries the richest semantic information and the lowest spatial resolution, so that applying global attention at this level enables any two spatial positions to interact directly while keeping the $O(HW \cdot HW)$ computational cost of self-attention within an acceptable range. For occluded scenes, the visible regions of an occluded hull may lie at non-adjacent positions in the feature map, and AIFI is therefore able to establish direct semantic associations between these spatially separated visible regions and thus realise mutual identity verification.

3. Methodology

3.1. Problem Formalisation

The failure of ship detection under occlusion was formalised as an information-bottleneck problem. Under the unoccluded condition, the input image was assumed to carry the complete representation of the target; under occlusion, only a subset of the target pixels was assumed to remain observable, while the remaining pixels were obscured. The class and bounding box of the complete target were therefore required to be recovered by the detection network from this visible subset alone. The feasibility of this recovery task can be characterised by two criteria that describe how well the detection pipeline preserves and integrates the discriminative information carried by the visible portion of the target.

These criteria are formulated as conceptual design targets rather than directly measurable quantities; their empirical validation is realised indirectly through the ablation study in Section 4.4, in which the marginal effects of ADown and AIFI on Recall and Precision, respectively, serve as proxies for improvements along the two criteria.

Downsampling information fidelity F_d was defined as the proportion of discriminative information carried by the visible portion of an occluded target that was preserved after each level of pyramid downsampling. The higher F_d was, the less information of these visible regions was lost before the high-level layers were reached.

Cross-region semantic coherence C_s was defined as the strength of the semantic association between the visible regions on either side of an occluded area in the high-level feature map. The higher C_s was, the stronger the mutual-verification ability between visible regions became.

The classical YOLO detection pipeline was found to suppress both F_d and C_s along the multi-stage P3–P5 processing path: the former was caused by the single-path compression of strided convolution, and the latter by the local receptive field of convolution. The methodological design of this paper was therefore developed around the joint elevation of these two indicators.

3.2. Overall Architecture

The proposed OAR-YOLO framework was built on the skeleton of YOLO11n, the Nano-scale variant of the YOLO11 family. All subsequent baselines are likewise evaluated at n-scale to ensure fairness of comparison. ADown was embedded at the three downsampling nodes P3, P4 and P5 in the backbone, and AIFI was inserted at the top of the backbone, immediately before the SPPF module; ADown was likewise used on the P3→P4 and P4→P5 bottom-up paths of the neck, so that downsampling consistency was maintained. The

whole network was composed of 264 layers, with 2,894,045 parameters and 5.7 GFLOPs of computation. This scale was kept very close to that of the YOLO11n baseline, so that the performance gains observed in the subsequent comparison experiments could be attributed to the proposed modules rather than to differences in model size.

Figure 1 illustrates the overall architecture of OAR-YOLO. The three-stage design of YOLO11—backbone, neck and head—was retained, and on top of the original data-flow framework the strided-convolution downsampling at the three backbone levels P3, P4 and P5 was replaced by ADown modules, an AIFI module was embedded at the top of the backbone, immediately before SPPF, and ADown was likewise adopted on the two bottom-up paths of the neck. After bi-directional fusion by PANet, the three-scale features were sent to the detection heads corresponding to P3, P4 and P5, by which the detection results of small, medium and large targets were respectively output.

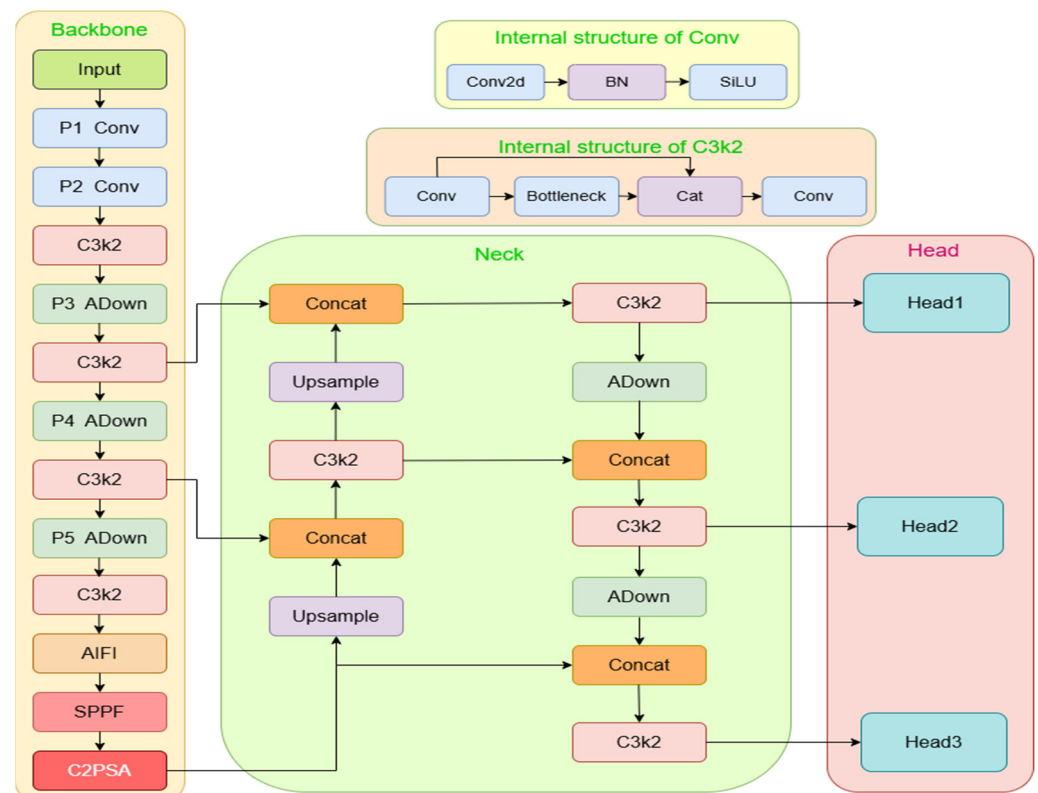


Figure 1. Overall architecture of OAR-YOLO.

The first two layers of the network were kept as a standard Conv- 3×3 stride-2 combination that completed the initial downsampling at P1 and P2. From layer P3 onwards, all downsampling operations were replaced by ADown. This choice was based on the following consideration: after P3 the resolution of the feature maps had already been markedly reduced, and any subsequent loss of information fidelity was amplified layer by layer, so that the P3–P5 segment was precisely the critical region in which ADown was expected to exert its information-conserving role. AIFI was inserted only between P5 and SPPF, where the feature map had the lowest resolution and the highest semantic density, so that the computational cost of global self-attention was controlled and the signal-to-noise ratio of the attention map was kept high.

3.3. ADown: Information Conservation in Downsampling

The classical stride-2 convolutional downsampling operation was formulated as follows:

$$F_{out}(i, j) = \sum_{m, n} W(m, n) \cdot F_{in}(2i + m, 2j + n) \quad (1)$$

in which W denotes a learnable convolution kernel. In essence, a 2×2 neighbourhood was compressed into a single pixel along a single weighted path, which forced the network to make a trade-off between low-frequency mean response and high-frequency edge response at every downsampling step. For unoccluded targets this trade-off could be remedied by subsequent layers; for occluded targets, however, whose visible parts were usually high-frequency edge components, the trade-off implied that discriminative information was systematically attenuated before the higher layers were reached.

The design idea of ADown was to separate low-frequency and high-frequency components explicitly. The input feature map was first smoothed by an average-pooling layer that preserved the spatial resolution, and was then split along the channel dimension into two halves that were routed to two parallel branches: one in which the spatial compression was carried out by a strided convolution so that low-frequency contextual information was retained, and another in which the compression was carried out by max-pooling followed by a point-wise projection so that high-frequency edge response was retained. The two branch outputs were finally concatenated along the channel dimension. The operation was expressed as:

$$F_1, F_2 = \text{Split} \left(\text{AvgPool}_{(2 \times 2, s=1)}(F_{in}) \right) \quad (2)$$

$$F_{out} = \text{Conv}_{(3 \times 3, s=2)}(F_1) \parallel \text{Conv}_{(1 \times 1)} \left(\text{MaxPool}_{(3 \times 3, s=2)}(F_2) \right) \quad (3)$$

where Split denotes an even division along the channel dimension, $\parallel$ denotes channel-wise concatenation, and the subscripts of Conv, AvgPool and MaxPool give the kernel size and stride of each operator. Each of the two branches therefore processed only half of the input channels, and both frequency components were retained during downsampling instead of being traded off against each other.

The implication for occluded scenes was that the visible regions of an occluded ship usually exhibited two kinds of signal characteristics: high-frequency edge components such as the gunwale, mast and funnel; and low-frequency texture components such as the hull paint and deck layout. The former was diluted by the mean response under classical stride-2 convolution, whereas the latter was suppressed by the edge response under pure MaxPool. The dual-path parallel design of ADown was therefore matched to the signal characteristics of occluded scenes, so that both kinds of signals were preserved across the downsampling step and carried into the high-level semantic-aggregation stage. Figure 2 depicts the processing flow of ADown.

Regarding computational cost, ADown reduced both the parameter count and the GFLOPs at each downsampling node compared with the strided convolution originally used at the same location in YOLO11n. In YOLO11n, each downsampling node was realised by a single Conv module with kernel size 3, stride 2 and padding 1, mapping the full input channels to the full output channels through one dense convolution. In ADown, the input feature map was first passed through an average-pooling smoothing layer with kernel size 2 and stride 1 that preserved the spatial resolution, and then split along the channel dimension into two equal halves. One half was processed by a 3×3 stride-2 convolution that performs the spatial downsampling directly, while the other half was first downsampled by a max-pooling layer with kernel size 3 and stride 2 and then projected by a 1×1 convolution; the two branch outputs were concatenated along the channel dimension.

Because each of the two convolutional branches operated on only half of the input channels, the total number of multiply–accumulate operations was markedly lower than that of the original full-channel 3×3 stride-2 convolution at the same spatial resolution. Table 1 compares the two configurations at the three backbone downsampling nodes P3, P4 and P5. The same substitution was applied at the two additional downsampling nodes on the neck’s bottom-up paths, from which analogous savings were obtained.

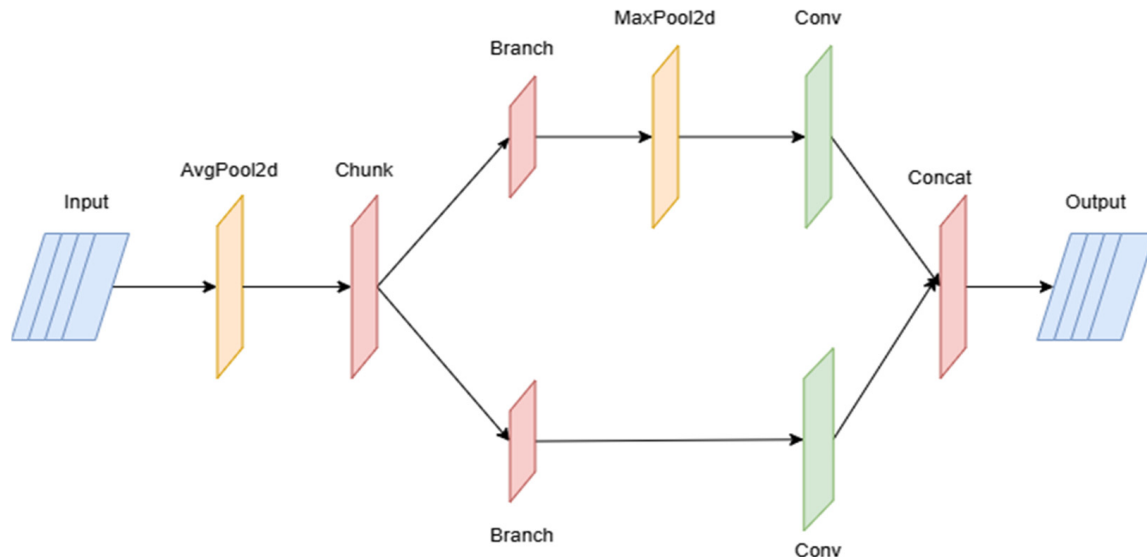


Figure 2. Processing flow of ADown.

Table 1. Params and GFLOPs: strided convolution vs. ADown.

Downsampling Node	Module	In-ch	Out-ch	Params	GFLOPs
P3	YOLO11n Conv ($k = 3, s = 2$)	64	64	36,992	0.236
P3	ADown	64	64	10,368	0.066
P4	YOLO11n Conv ($k = 3, s = 2$)	128	128	147,712	0.236
P4	ADown	128	128	41,216	0.066
P5	YOLO11n Conv ($k = 3, s = 2$)	128	256	295,424	0.118
P5	ADown	128	256	82,432	0.033

Across the three backbone downsampling nodes, ADown collectively reduced the parameter count from 480,128 to 134,016 and the GFLOPs from 0.590 to 0.165, corresponding to a 72% reduction in both metrics.

3.4. AIFI: Intra-Scale Feature Interaction

Even when the discriminative information of visible regions had been preserved by ADown, this information was found to face a second limitation upon reaching the high-level feature map—namely, the locality of the convolution operator. The expansion of a convolutional receptive field depended on the linear stacking of network depth, and even at P5, although the semantics of a single feature position were the richest, cross-position semantic associations were still limited by the effective receptive field. When the visible regions of an occluded hull lay at non-adjacent positions in the feature map, convolution alone was therefore insufficient for mutual identity verification between them.

The design idea of AIFI was that a complete multi-head self-attention operation was applied once at the P5 level, by which any two spatial positions were allowed to interact directly. The operation was expressed mathematically as:

$$AIFI(F) = F + MHSA(F + PE) \quad (4)$$

where *MHSA* denotes multi-head self-attention and *PE* denotes positional encoding, and the residual connection was used so that the original features were not diluted by the attention operation.

Keeping the computational cost of global reasoning within an acceptable range was a key consideration of the AIFI design. The computational complexity of self-attention is $O(HW \cdot HW)$, which grows rapidly on high-resolution feature maps; however, after several downsampling steps the HW of the P5 layer had been reduced to roughly 1/1024 of the original input. Global attention applied at this point kept the computational cost on a par with that of convolution, while the semantic signal-to-noise ratio of the attention operation was high—because each P5 feature position already carried higher-order class semantics, the association between any two positions carried strong discriminative significance. The processing flow of AIFI is shown in Figure 3.

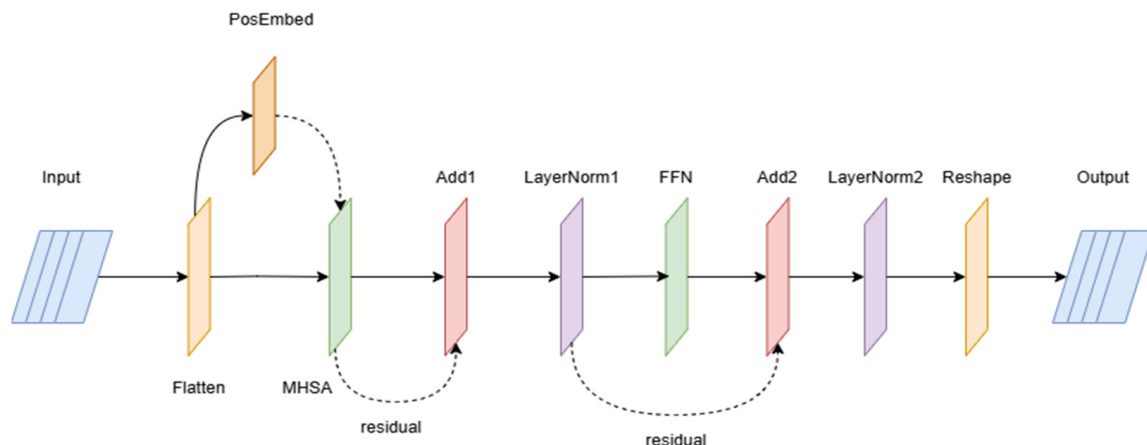


Figure 3. Processing flow of AIFI.

The implication for occluded scenes was that direct semantic associations were established by AIFI between the spatially separated visible regions of an occluded hull, so that the identities of the regions could be mutually verified. For instance, when the bow of a container ship was occluded by a crane at an adjacent berth, the stern and bow visible regions were spatially separated within a convolutional field of view; in the attention map of AIFI, however, they were nevertheless linked by the shared “container ship” semantic feature, by which a complete target judgement was supported.

3.5. Synergistic Coupling Mechanism

The central claim of this paper was that the two modules constituted complementary operators for solving the occlusion problem.

When ADown was used alone, the high-frequency information of low layers was preserved, but at the high-level stage this information still existed in the form of spatially separated visible regions, and the local receptive field of convolution was unable to connect them into a complete identity. This is reflected in the subsequent “+ADown” configuration of the ablation experiments, where the mAP@0.5 was raised to 0.744 yet the Precision reached only 0.796: information was preserved but not fully exploited.

When AIFI was used alone, global reasoning was applied at the top level, but if the low-level downsampling had already attenuated the discriminative information of the visible regions, the top-level attention operation lacked effective input. The “+AIFI” configuration of the ablation experiments, in which the Precision reached 0.806 but the Recall reached only 0.681, is exactly an empirical reflection of this phenomenon: global reasoning was introduced, but the information available for reasoning was insufficient.

When the two modules were used together, the high-frequency information of the visible regions preserved by ADown was supplied as the input to the global attention operation of AIFI, and the cross-position integration of AIFI converged the spatially separated visible regions retained by ADown into a complete identity judgement. This synergistic effect was manifested as a non-additive gain: the Precision of the cooperative model was raised to 0.810, which exceeded the simple linear sum of the gains of the two individual modules.

In summary, ADown and AIFI were not two parallel engineering refinements but were, respectively, the dual operators that responded to the two core objectives of “information conservation” and “semantic connectivity”; their cooperation was found to be a necessary condition for occlusion-robust ship detection.

4. Experiments

4.1. Dataset Construction

To support empirical validation of the proposed occlusion-robust ship-detection method, a self-built optical-imagery dataset covering seven common maritime ship categories was constructed in this work. The dataset contained 4803 images in total, which were split into 3869 training, 510 validation and 424 test images. The seven categories were liner, container ship, bulk carrier, sailboat, other ship, fishing boat and tug, and they were chosen to cover diverse occlusion situations: navigation channels contributed a large number of mutually adjacent ships, harbour operation areas contributed samples in which shore-based cranes, gangways and buildings occluded the hull, and offshore navigation segments contributed samples in which clouds, mist and spray covered the vessel. Figure 4 shows several sample images from the dataset.

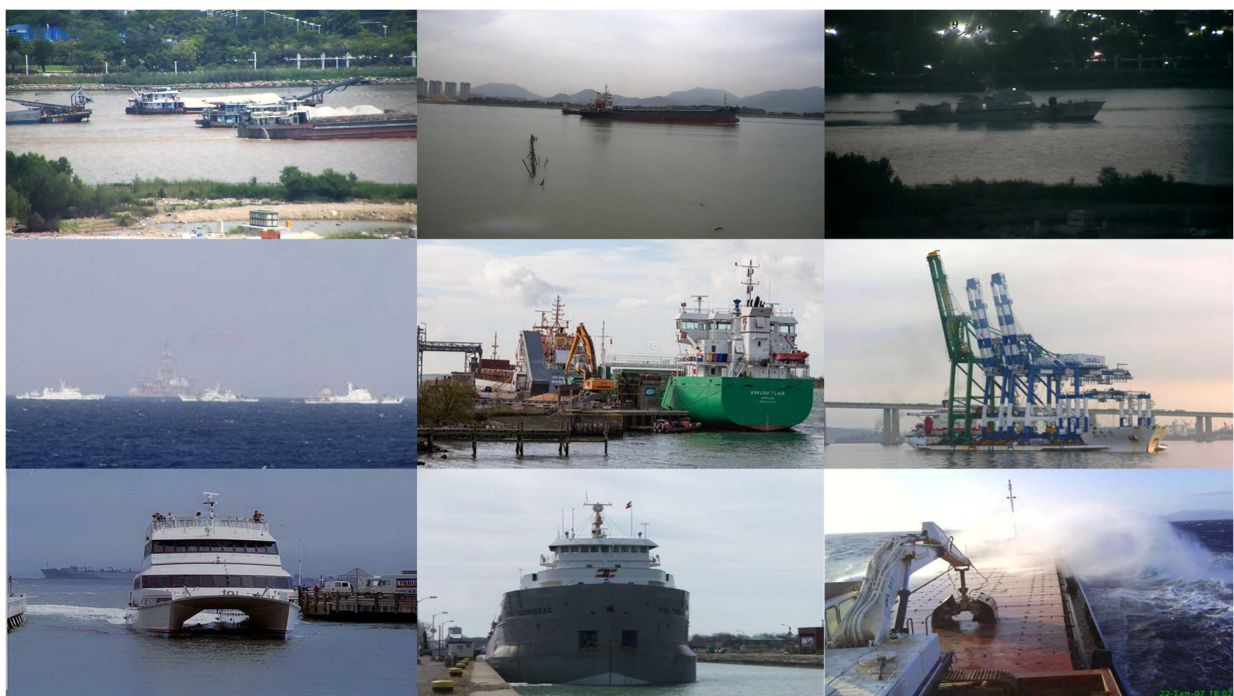


Figure 4. Dataset examples.

This coverage strategy does not pursue a simple increase in sample volume; rather, it ensures that the constructed dataset statistically reflects the structural regularities of maritime perception scenes and thereby provides a sound data basis for the subsequent validation of the proposed method. The per-class distribution of images and instances across the three subsets is summarised in Table 2.

Table 2. Dataset split distribution.

Category	Train Images	Train Instances	Val Images	Val Instances	Test Images	Test Instances
liner	919	975	88	92	148	191
container ship	406	421	49	49	55	55
bulk carrier	1668	1891	176	185	144	153
sailboat	271	450	22	26	49	94
other ship	506	807	124	150	125	231
fishing boat	387	709	68	131	29	43
tug	218	298	37	46	15	30
Total	3869	5551	510	679	424	797

As shown in Table 2, all seven categories are present in every subset, so that per-class performance can be evaluated consistently. The class distribution is moderately imbalanced: bulk carrier is the most represented category, accounting for roughly one third of all instances, while tug is the least represented. This imbalance is not a defect of the dataset but rather reflects the actual frequency spectrum of maritime shipping. To further characterise the occlusion prevalence of the dataset, an instance-level occlusion estimate was computed based on the geometric overlap of ground-truth bounding boxes. Table 3 reports the resulting per-class figures.

Table 3. Per-class inter-ship occlusion.

Category	Total Instances	Occluded Instances	Occluded %
liner	1258	72	5.7%
container ship	525	23	4.4%
bulk carrier	2229	279	12.5%
sailboat	570	37	6.5%
other ship	1188	112	9.4%
fishing boat	883	175	19.8%
tug	374	32	8.6%

Because the dataset does not carry explicit occlusion or visibility annotations, this estimate was computed by a bounding-box IoU heuristic: within the same image, any pair of ground-truth bounding boxes whose IoU exceeds 0.10 was regarded as mutually occluding, and both participating instances were counted as occluded. The percentages in Table 3 therefore quantify inter-ship occlusion only—the case in which one ship is partially blocked by another ship in the same image. Other occlusion types present in the dataset, such as shore-based structural occlusion by cranes, gangways or dock installations, and atmospheric obscuration by fog, mist or spray, are not captured by this heuristic and are not reflected in the reported percentages. The overall occlusion prevalence of the dataset is therefore higher than the ship-to-ship figures shown in Table 3; the diversity of the remaining occlusion types was ensured at the source-selection stage of dataset construction, as described earlier in this section.

4.2. Experimental Setup and Evaluation Metrics

All experiments were carried out on a unified hardware platform: an NVIDIA RTX 5090 GPU with 32 GB of memory was used as CUDA device 0, and the framework was Ultralytics 8.3.9 with PyTorch 2.7.0+cu128. Training was performed with the SGD optimiser, the initial learning rate was set to 0.01, the momentum to 0.937 and the weight decay to 0.0005, and the batch size and number of training epochs were kept identical to the officially recommended YOLO11 configuration so that the fairness of the comparison was preserved.

To prevent overfitting on the moderately sized self-built training set, several standard measures were adopted. First, the default Ultralytics data-augmentation pipeline was applied throughout training, including Mosaic with probability 1.0, HSV colour jittering with hgain 0.015, sgain 0.7 and vgain 0.4, random horizontal flip with probability 0.5, random scaling of ± 0.5 and random translation of ± 0.1 . Second, both models were trained from scratch without pre-trained weights, so that no external representational capacity was transferred into training and every parameter was learned from the target distribution. Third, the weight decay of 0.0005 configured for the SGD optimiser above provided implicit L2 regularisation on the parameter norm.

The evaluation metrics cover four standard dimensions: Precision, Recall, mean Average Precision at an IoU threshold of 0.5 (mAP@0.5), and mean Average Precision over IoU thresholds from 0.5 to 0.95 (mAP@0.5–0.95). Precision measures the proportion of true targets among all samples predicted to be positive, and Recall measures the proportion of true targets that are correctly detected; the two are computed as follows:

$$Precision = TP / (TP + FP) \quad (5)$$

$$Recall = TP / (TP + FN) \quad (6)$$

where TP (true positive) denotes the number of correctly detected targets, FP (false positive) denotes the number of incorrect detections, and FN (false negative) denotes the number of missed true targets. Whether a prediction is judged to be TP or FP depends on whether the intersection-over-union between the predicted box and the ground-truth box exceeds the preset threshold.

At a given IoU threshold, the precision–recall curve can be drawn with Recall on the horizontal axis and Precision on the vertical axis, and the area under the curve is the average precision of a single class:

$$AP = \int_0^1 P(R) dR \quad (7)$$

where P denotes the precision corresponding to recall R . Taking the arithmetic mean of the AP over all N classes yields the mean average precision (mAP@0.5):

$$mAP@0.5 = \frac{1}{N} \sum_{i=1}^N AP_i \quad (8)$$

in which N is the total number of classes (here $N = 7$) and AP_i is the average precision of the i -th class.

mAP@0.5 refers to the mAP computed at a fixed IoU threshold of 0.5, which emphasises a model's ability to detect and coarsely localise targets, whereas mAP@0.5–0.95 is

computed by averaging the mAP values obtained at ten IoU thresholds from 0.5 to 0.95 with a step of 0.05, and imposes a stricter requirement on localisation accuracy:

$$mAP@0.5-0.95 = \frac{1}{10} \sum_{t \in \{0.50, 0.55, \dots, 0.95\}} mAP_{IoU=t} \quad (9)$$

Compared with mAP@0.5, which is evaluated at a single threshold, mAP@0.5–0.95 provides an integrated view of model performance under different degrees of localisation strictness.

As supplementary indicators, the parameter count and GFLOPs are used to quantify the computational cost, while the inference speed reflects the practical feasibility of deployment.

To further verify that overfitting did not occur under this training regime, the training and validation losses of both models were monitored throughout training and are shown in Figure 5. Averaged over the last ten epochs, the validation loss exceeded the training loss by 20.3–27.3% for YOLO11n and by 20.1–31.4% for OAR-YOLO across the three loss components. These stable gaps, together with the absence of any upward divergence of the validation loss at late epochs and the smooth plateau of the mAP curves, indicate that both models converged to a well-generalising solution rather than memorising the training distribution. Notably, the train–val gap of OAR-YOLO is comparable to that of YOLO11n despite its slightly larger parameter count, demonstrating that the ADown and AIFI modules introduce no additional overfitting risk over the baseline.

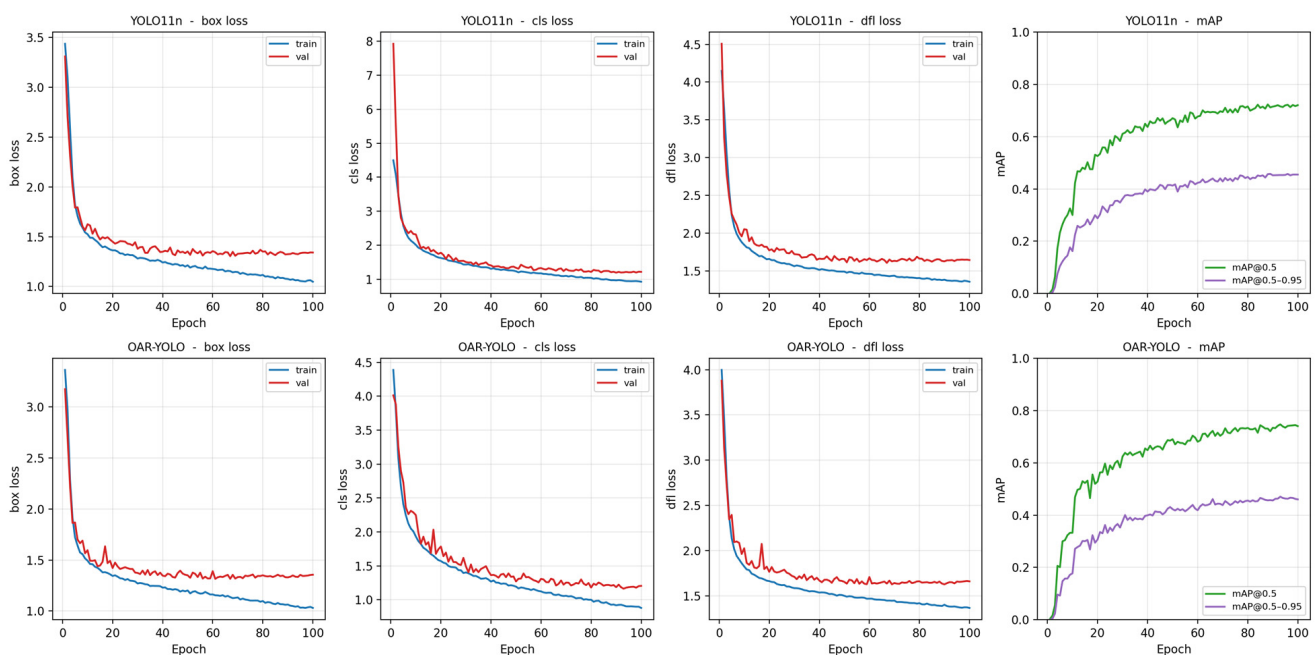


Figure 5. Training and validation curves for YOLO11n and OAR-YOLO.

4.3. Comparison with Representative Methods

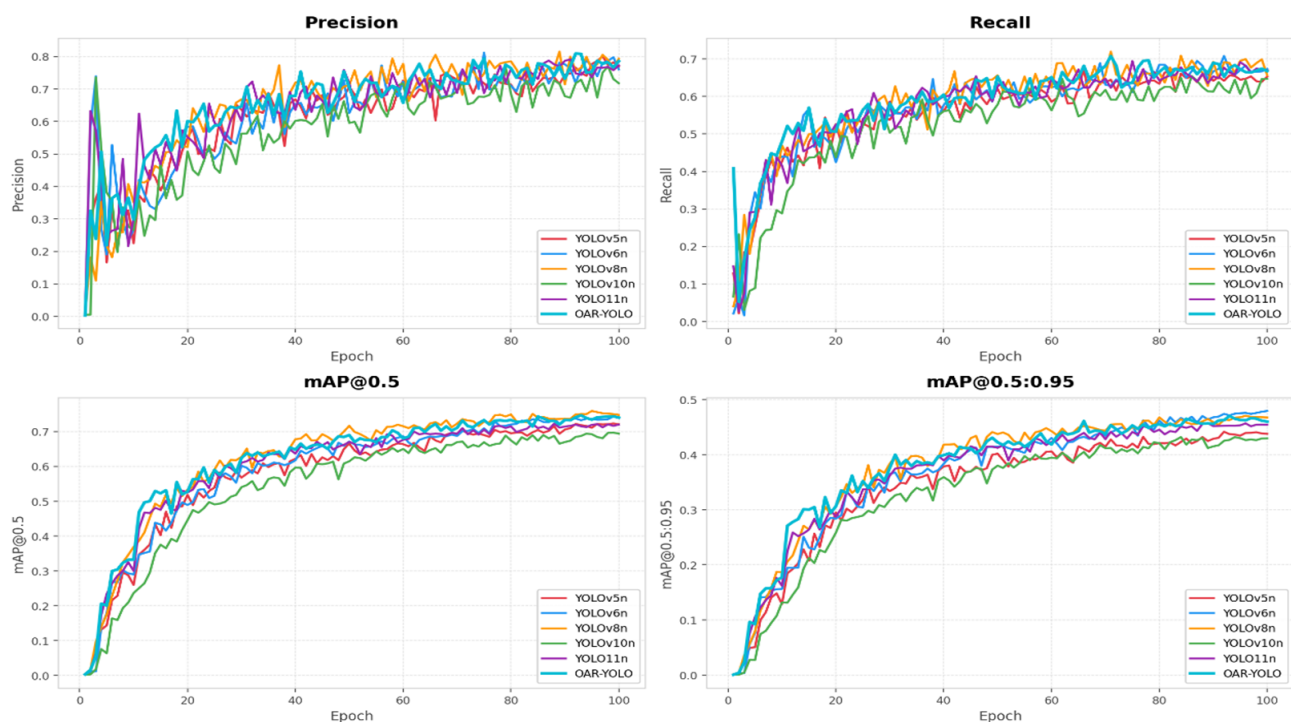
To position the proposed method within the mainstream lineage of single-stage detectors, five representative YOLO variants—YOLOv5n, YOLOv6n, YOLOv8n, YOLOv10n and YOLO11n—were selected as comparison baselines, and all methods were trained from scratch on the same dataset and on the same hardware so that cross-method comparability was guaranteed. The comparative results are listed in Table 4.

Table 4. Comparative experiments.

Method	GFLOPs	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv5n	5.8	0.749	0.656	0.722	0.442
YOLOv6n	11.5	0.760	0.675	0.739	0.479
YOLOv8n	6.8	0.776	0.699	0.756	0.471
YOLOv10n	6.5	0.691	0.645	0.694	0.432
YOLO11n	6.3	0.783	0.672	0.721	0.457
OAR-YOLO	5.7	0.810	0.675	0.747	0.471

Three key signals are revealed by Table 4. First, the proposed method reaches a Precision of 0.810, which is 2.7 percentage points higher than that of the YOLO11n baseline and the highest among all compared methods; this indicator is particularly meaningful under occlusion. Because occlusion is the situation most likely to trigger false detections, the high Precision provides direct evidence of the contribution of AIFI's global reasoning to false-detection suppression. Second, the mAP@0.5 of the proposed method exceeds that of the YOLO11n baseline by 2.6 percentage points and that of the weakest baseline, YOLOv10n, by 5.3 percentage points; although it is slightly lower than the 0.756 obtained by YOLOv8n, this must be weighed against the computational cost—YOLOv8n requires about 6.8 GFLOPs, whereas the proposed method requires only 5.7 GFLOPs, so that a better balance between accuracy and efficiency is achieved. Third, the proposed method reaches an mAP@0.5–0.95 of 0.471, which is on a par with YOLOv8n and clearly higher than YOLO11n and YOLOv10n, indicating that the proposed architectural improvements enhance not only the discrimination ability under a loose IoU threshold but also the localisation accuracy.

To present the convergence behaviour and stability of the various models more intuitively, Figure 6 plots the curves of Precision, Recall, mAP@0.5 and mAP@0.5–0.95 over 100 training epochs for the six methods.

YOLO Model Training Comparison**Figure 6.** Training-process metric comparison of the various models.

The curves show that the Precision of the proposed method is consistently high throughout training and that its late-stage fluctuations are smaller than those of baselines such as YOLOv5n and YOLOv10n, which reflects a more stable decision behaviour. On both mAP@0.5 and mAP@0.5–0.95 the proposed method converges faster than most baselines in the early phase, and the final accuracy is close to that of YOLOv8n while exceeding the YOLO11n baseline—consistent with the numerical conclusions of Table 4.

4.4. Ablation Study

To identify the marginal contribution of each of the two mechanisms—ADown and AIFI—and the non-additive gain of their cooperation, an ablation study was designed with the following configurations: YOLO11n, +AIFI, +ADown and OAR-YOLO. All configurations were trained independently on the same dataset under the same training protocol. The ablation results are summarised in Table 5.

Table 5. Ablation study of ADown and AIFI on the self-built dataset.

Model	ADown	AIFI	GFLOPs	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLO11n			6.3	0.783	0.672	0.721	0.457
+ADown	✓		5.3	0.796	0.691	0.744	0.465
+AIFI		✓	6.6	0.806	0.681	0.745	0.473
OAR-YOLO	✓	✓	5.7	0.810	0.675	0.747	0.471

The data in Table 5 provide empirical support for the mechanistic analysis in this paper at three levels.

First, the marginal contribution of AIFI is concentrated on the Precision dimension. The introduction of AIFI alone raises Precision from 0.783 to 0.806 and mAP@0.5 from 0.721 to 0.745. This result is consistent with the mechanistic analysis in Section 3.4: the core function of global self-attention is to enable cross-position mutual verification of the visible regions of occluded targets, whose direct effect is the suppression of false detections. With AIFI, the detection results are built on a “globally negotiated” basis, and the rise in Precision is therefore mechanistically coherent.

Second, the marginal contribution of ADown is concentrated on the Recall dimension. The introduction of ADown alone raises Recall from 0.672 to 0.691, which is the highest among the four configurations. This result is likewise consistent with the analysis in Section 3.3: the dual-path parallel downsampling preserves the high-frequency edge components, so that occluded targets that would otherwise have been missed because of downsampling-induced information loss are retained for the high-level decision. Information is preserved and recall therefore rises; this causal relationship is directly supported by the data.

Third, the cooperative effect manifests itself as a non-additive gain on Precision and mAP@0.5. AIFI raises Precision to 0.806 and ADown raises it to 0.796, whereas the joint configuration reaches 0.810, the highest among all four configurations. Although the joint gain is not strictly super-additive, the fact that combining the two modules yields a higher Precision than either alone indicates that their contributions are complementary rather than redundant. In addition, mAP@0.5 reaches its highest value of 0.747 under the joint configuration, which indicates that the two mechanisms—information conservation and semantic connectivity—reinforce each other when they work together. The high-frequency information preserved by ADown serves as the input of AIFI’s global reasoning, and the cross-position integration of AIFI converges the spatially separated visible regions retained by ADown into a complete identity judgement.

Read in aggregate, these three observations serve as the indirect empirical validation of the two design criteria introduced in Section 3.1. The Recall gain contributed by ADown reflects an improvement of downsampling information fidelity F_d , because the additional targets recovered under the “+ADown” configuration are precisely those whose visible regions would otherwise have been attenuated by single-path strided-convolution downsampling before reaching the high-level layers. The Precision gain contributed by AIFI reflects an improvement of cross-region semantic coherence C_s , because the false detections suppressed under the “+AIFI” configuration are precisely those that a purely local receptive field would fail to reject through cross-region mutual verification. The non-additive Precision gain of the joint configuration, in turn, reflects the mutual reinforcement between F_d and C_s —high-frequency edge information retained by ADown provides the substrate on which AIFI’s cross-region reasoning becomes effective. In this sense, Table 5 constitutes the empirical counterpart of the analytical framework laid out in Section 3.1, without requiring F_d and C_s to be directly quantified as scalar measurements.

It should be noted that the Recall of the joint configuration is slightly lower than the 0.691 of the ADown-only configuration. This phenomenon does not indicate a failure of the mechanism; rather, it represents a trade-off in the joint optimisation of Precision and Recall. While AIFI suppresses false detections through global reasoning, it also imposes a stricter decision condition on low-confidence visible regions, so that some marginal regions are judged to be non-targets. The cost of this trade-off is outweighed by its benefit, and the overall effect remains positive.

4.5. Per-Class Performance Analysis

To clarify how the proposed method performs differentially across the seven ship classes, the per-class detection performance of the proposed model on the validation set was recorded; the results are listed in Table 6.

Table 6. Detection results on the seven ship classes.

Class	Images	Instances	Precision	Recall	mAP@0.5	mAP@0.5:0.95
liner	88	92	0.863	0.887	0.913	0.629
container ship	49	49	0.932	0.755	0.880	0.648
bulk carrier	176	185	0.745	0.800	0.766	0.422
sailboat	22	26	0.871	0.692	0.761	0.541
other ship	124	150	0.670	0.433	0.572	0.359
fishing boat	68	131	0.754	0.611	0.686	0.386
tug	37	46	0.834	0.544	0.652	0.309
all	510	679	0.810	0.675	0.747	0.471

Table 6 reveals a phenomenon worth analysing: the mAP@0.5 of the seven classes shows a clear hierarchy—liner and container ship are at the top, bulk carrier, sailboat and fishing boat sit in the middle, and other ship and tug fall at the bottom. This gap is not solely determined by sample size: sailboat has only 26 instances yet still reaches 0.761, while other ship has 150 instances yet reaches only 0.572.

This phenomenon can be explained by intra-class representation sparsity. Liner and container ship possess strongly distinguishable morphology, such as a slender hull or a regular container array, whose features are effectively encoded in the early convolutional stages, so that the cooperation of ADown and AIFI is able to take full effect. By contrast, other ship is a mixed category of multiple ship types with substantial intra-class variation, and even under a better architecture the network is still unable to learn stable discriminative representations for it. This result does not undermine the overall contribution of the proposed method and suggests that future work may further refine the intra-class granularity.

For a direct visual comparison of the detection effect before and after improvement, Figure 7 shows the original images, the detection results of the YOLO11n baseline and those of the proposed method under occlusion. It can be observed that, in situations involving mutual occlusion among ships, partial occlusion by shore-based facilities, and small targets close to the sea–sky line, the YOLO11n baseline exhibits missed detections and low confidence scores, whereas the proposed method produces more complete bounding boxes and higher confidence scores for occluded container ships, tugs and fishing boats, and also suppresses false detections for confusing categories such as other ship and sailboat.

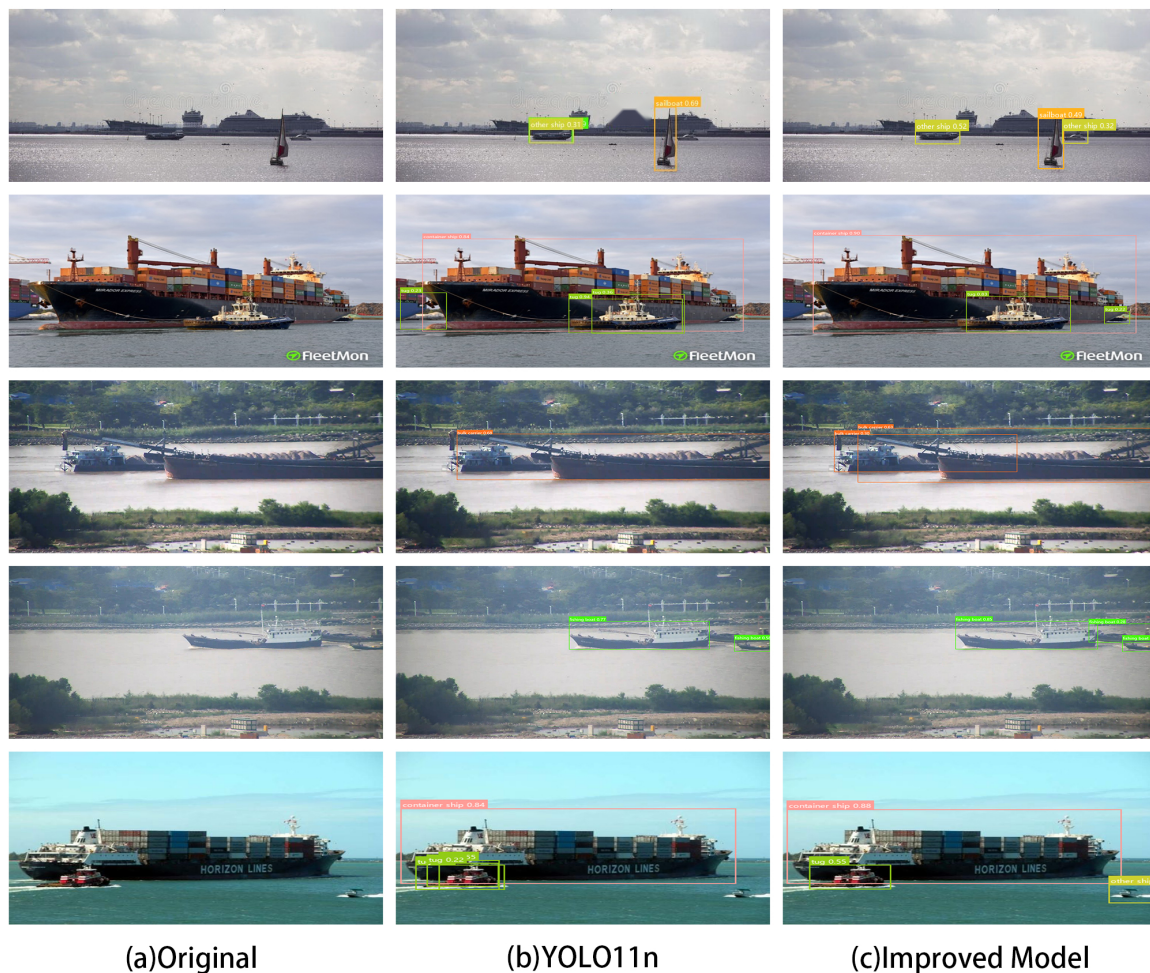


Figure 7. Detection-result comparison of different methods.

To further reveal the inner source of the performance gain, heat maps are used in Figure 8 to compare the attention distributions of YOLO11n and the proposed method. In the baseline model, high-response regions often spread to background areas such as shore-based buildings and water reflections, which is prone to inducing false detections; in the heat map of the proposed method, the response is more concentrated on the hull itself, and a coherent response is also formed over the visible parts of occluded targets. This is consistent with the mechanistic analysis whereby AIFI achieves cross-occlusion semantic mutual verification through global self-attention while ADown preserves high-frequency edge information through dual-path downsampling, and it provides qualitative visual support for the central claim of this paper.

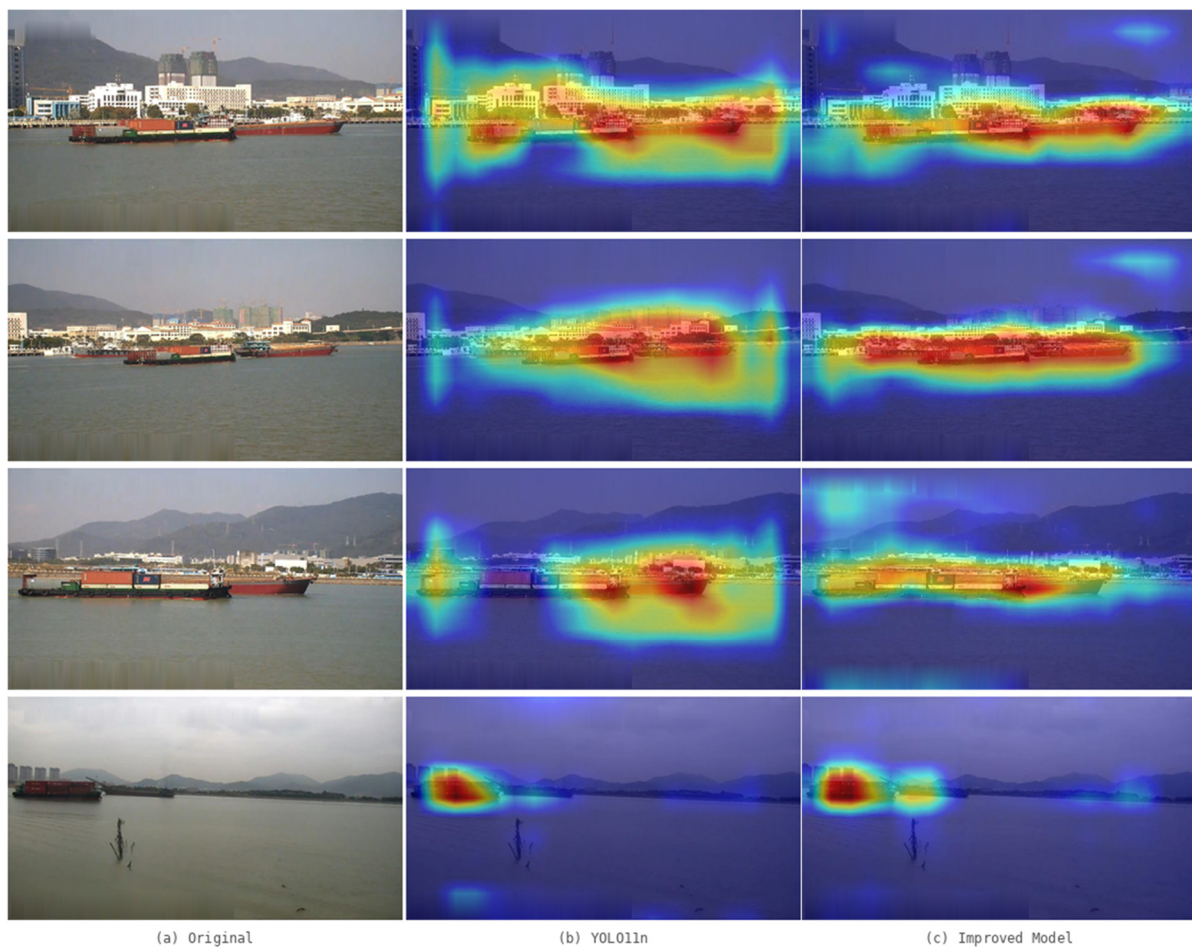


Figure 8. Heat-map comparison of different methods.

4.6. Computational Efficiency and Deployment Feasibility

The final model produced by the proposed method consisted of 264 layers, 2,894,045 parameters and 5.7 GFLOPs. On the RTX 5090 platform, the single-image inference latency was measured at 0.8 ms, with 0.0 ms for preprocessing, 0.8 ms for inference and 0.4 ms for post-processing—1.2 ms in total—which translates into a theoretical throughput of more than 800 FPS. Even when the model was migrated to an embedded platform, the actual inference rate was estimated to remain above 50 FPS under empirical performance-scaling factors, which satisfies the deployment requirement of real-time maritime monitoring.

This level of computational efficiency is not achieved at the cost of accuracy: the mAP@0.5–0.95 of the proposed method equals that of the computationally heavier YOLOv8n while using 16% fewer GFLOPs, and its Precision exceeds that of YOLOv8n by 3.4 percentage points. The reason why lightweight design and high accuracy can be realised simultaneously is that the positions of ADown and AIFI within the network are deliberately chosen: ADown is applied only at the P3–P5 downsampling nodes rather than throughout the entire network, and AIFI is applied only at the top P5 layer rather than throughout the backbone. This minimal-intervention design, which concentrates computation at critical nodes, accounts for the favourable trade-off achieved by the proposed method between accuracy and efficiency.

4.7. Evaluation on SeaShips: Baseline Performance and Occlusion Robustness

To address two complementary concerns—whether the improvements reported on the self-built dataset generalise to a widely used public benchmark, and whether OAR-YOLO delivers a quantifiable robustness advantage under controlled occlusion—a joint evaluation

was conducted on the SeaShips dataset [36]. SeaShips contains 7000 optical images of six ship categories captured under a variety of coastal surveillance scenarios and is publicly released with YOLO-compatible annotations.

Both YOLO11n and OAR-YOLO were trained from scratch on the SeaShips training split under exactly the same hyper-parameter configuration as in the main experiment, comprising 100 epochs at input size 640×640 , batch size 16 and the SGD optimiser with initial learning rate 0.01, momentum 0.937 and weight decay 0.0005. No pre-trained weights were used, so that the observed performance can be fully attributed to the model architecture rather than to transferred external capacity.

Evaluation was performed on the SeaShips test split under four occlusion levels of $r = 0.0, 0.1, 0.3$ and 0.5 , where r denotes the fraction of each ground-truth bounding box area covered by a rectangular occlusion mask. The mask side length was set to $\sqrt{r}$ times the corresponding bounding box side, so that its area equalled r times the box area. For every ground-truth bounding box in every test image, the mask anchor was seeded deterministically by a hash of the image filename and the box index, so that the mask position remained consistent across occlusion levels and larger masks were geometrically nested within smaller ones; this design ensured that increasing r corresponds to progressively occluding the same visible region rather than sampling a fresh random region. Each mask was filled with a background colour estimated as the median colour of the four image corners, providing a naturalistic approximation of maritime background. Ground-truth annotations were kept unchanged, so that the evaluation measured the model's ability to recover complete detections from partially visible targets. The setting $r = 0.0$ corresponds to the clean SeaShips test set and simultaneously serves as the public-dataset generalisation benchmark.

Table 7 reports the mAP@0.5 of both models at each occlusion level, together with the retention rate defined as the ratio between the mAP@0.5 at the given r and the mAP@0.5 at $r = 0.0$. Figure 9 visualises the corresponding degradation curves.

Table 7. Detection performance on the SeaShips test set under progressive artificial occlusion.

Occlusion r	YOLO11n mAP@0.5	OAR-YOLO mAP@0.5	Δ	YOLO11n Retention	OAR-YOLO Retention
0.0	0.9815	0.9856	+0.004	100.0%	100.0%
0.1	0.9272	0.9422	+0.015	94.5%	95.6%
0.3	0.6009	0.6276	+0.027	61.2%	63.7%
0.5	0.3086	0.3782	+0.070	31.4%	38.4%

On the clean SeaShips test set at $r = 0.0$, YOLO11n and OAR-YOLO attain closely matched mAP@0.5 of 0.9815 and 0.9856 respectively, with OAR-YOLO also exhibiting a marginally higher mAP@0.5–0.95 of 0.776 against 0.768 for YOLO11n. This equivalence on a public benchmark that differs from the self-built dataset in class taxonomy, imaging viewpoint and scene composition demonstrates that the improvements introduced by the ADown and AIFI modules generalise beyond the training distribution used in the main experiment, and confirms that they do not compromise baseline detection quality on clean imagery.

Under artificial occlusion, however, the two models diverge in a monotonically widening manner. At $r = 0.1$ the mAP@0.5 gap is 1.5 percentage points; it grows to 2.7 points at $r = 0.3$, and reaches 7.0 points at $r = 0.5$, where YOLO11n retains only 31.4% of its clean performance whereas OAR-YOLO retains 38.4%. This expanding advantage under increasing occlusion severity is the intended empirical signature of the ADown and AIFI modules: the high-frequency edge information preserved by ADown and the cross-region

semantic association enabled by AIFI become increasingly load-bearing as the visible signal is progressively suppressed. The observation that the two models are indistinguishable on clean data yet increasingly separated under occlusion further corroborates the analytical framework introduced in Section 3.1, in which Fd and Cs were postulated as design criteria specifically targeted at the occluded regime rather than at overall detection capacity.

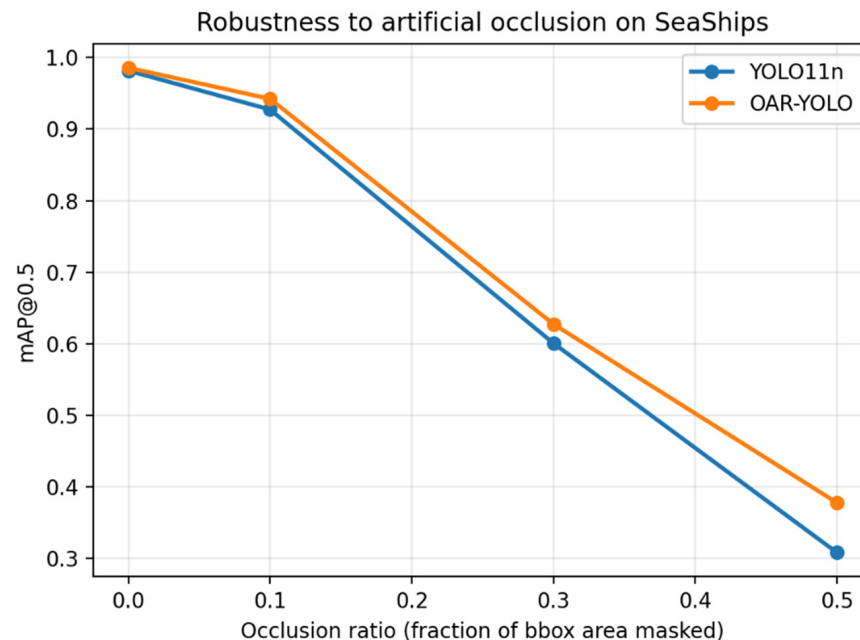


Figure 9. Robustness curves under artificial occlusion on SeaShips.

5. Conclusions and Discussion

5.1. Conclusions

In this paper, we identified and analysed a twofold mechanistic deficiency in mainstream YOLO detection pipelines under complex occlusion, namely the low-level downsampling information loss and the high-level convolutional semantic limitation. Rather than treating it from the traditional perspective of sample scarcity, we re-interpreted the occlusion problem as a structural deficiency in the way the detection network itself processes information, and on this basis we constructed the cooperatively designed YOLO11n detection framework OAR-YOLO: we used the dual-path parallel downsampling of ADown to reduce spatial-information loss, and we used the intra-scale self-attention of AIFI to establish semantic associations between occluded regions, thereby building a dual compensation mechanism of information conservation and semantic connectivity. On the self-built dataset, the proposed method raised mAP@0.5 by 2.6 percentage points over the YOLO11n baseline, attained a class-leading Precision of 0.810, clearly outperformed representative single-stage detectors such as YOLOv5n, YOLOv6n and YOLOv10n, and matched the detection accuracy of the more expensive YOLOv8n with only 5.7 GFLOPs of computation. The ablation study further revealed the marginal contribution of AIFI to false-detection suppression on the Precision dimension, the marginal contribution of ADown to missed-detection compensation on the Recall dimension, and the non-additive gain produced by their cooperation.

OAR-YOLO therefore provides a feasible path that balances accuracy and efficiency for maritime perception under complex occlusion. The proposed framework demonstrates that information conservation and semantic connectivity act as dual operators rather than independent engineering refinements, and that their cooperative design constitutes a necessary condition for occlusion-robust ship detection. We believe that the perspective

and design principles presented here generalise to other detection tasks in which occlusion is a structural rather than incidental property, and that they offer a useful starting point for future research on robust detection under information-bottleneck conditions.

5.2. Discussion

Performance boundaries under different occlusion types. Physical occlusion by adjacent vessels or shore-based structures produces sharp local discontinuities while leaving the visible parts semantically rich, so AIFI's cross-region reasoning is expected to dominate in this regime. Meteorological occlusion by clouds, mist and spray instead degrades the visible signal itself in a low-contrast, spatially diffuse manner, favouring ADown's high-frequency edge preservation. A scenario-stratified evaluation was not conducted here and is left for future work.

Category-level limitations. The "other ship" class remained substantially below the overall average, reaching an mAP@0.5 of 0.572 against 0.747 for the dataset as a whole, because its catch-all labelling merges multiple sub-types with large intra-class morphological variation. This ceiling is imposed by the label taxonomy rather than the detection architecture, and cannot be resolved by architectural improvements alone.

Future directions. Three directions are promising: refining the intra-class taxonomy of "other ship" into finer subcategories so that consistent discriminative features become learnable; incorporating temporal cues from surveillance video streams to recover occluded targets that become visible in adjacent frames; and exploring occlusion-aware training losses such as visibility-weighted regression to complement the architectural modifications proposed here.

Author Contributions: Conceptualization, J.L. and Y.Z.; methodology, J.L., Y.S. and C.Z.; software, J.L.; validation, J.L. and Y.Z.; formal analysis, J.L., Y.S. and C.Z.; investigation, J.L., Y.S. and C.Z.; resources, Y.S. and C.Z.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, Y.Z., Y.S. and C.Z.; visualisation, J.L.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) under Grant No. 2024CXGC010804.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study, including a related master's thesis and follow-up research that have not yet been completed. Requests to access the datasets should be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

OAR-YOLO	Occlusion-Aware Robust YOLO
ADown	Adaptive Downsampling
AIFI	Attention-based Intra-scale Feature Interaction
AP	Average Precision
RT-DETR	Real-Time Detection Transformer
FN	False Negative
FP	False Positive
PE	Positional Encoding

References

1. Yin, Y.; Cheng, X.; Shi, F.; Liu, X.; Huo, H.; Chen, S. High-order spatial interactions enhanced lightweight model for optical remote sensing image-based small ship detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4201416. [[CrossRef](#)]
2. Pang, Y.; Xie, J.; Khan, M.H.; Anwer, R.M.; Khan, F.S.; Shao, L. Mask-Guided Attention Network for Occluded Pedestrian Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4967–4975.
3. Gholinavaz, S.; Saeedi, N.; Gharehveran, S.S. Robustness Analysis of YOLO and Faster R-CNN for Object Detection in Realistic Weather Scenarios with Noise Augmentation. *Sci. Rep.* **2025**, *15*, 44888. [[CrossRef](#)] [[PubMed](#)]
4. Guo, Y.; Liu, R.W.; Qu, J.; Lu, Y.; Zhu, F.; Lv, Y. Asynchronous Trajectory Matching-Based Multimodal Maritime Data Fusion for Vessel Traffic Surveillance in Inland Waterways. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 12779–12792. [[CrossRef](#)]
5. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
6. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
7. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 2918–2928.
8. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
9. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
10. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
11. El-Darymli, K.; McGuire, P.; Power, D.; Moloney, C. Target detection in synthetic aperture radar imagery: A state-of-the-art survey. *J. Appl. Remote Sens.* **2013**, *7*, 071598. [[CrossRef](#)]
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Kim, J.H.; Kim, N.; Park, Y.W.; Won, C.S. Object detection and classification based on YOLO-V5 with improved maritime dataset. *J. Mar. Sci. Eng.* **2022**, *10*, 377. [[CrossRef](#)]
16. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
17. Gong, Y.; Chen, Z.; Deng, W.; Tan, J.; Li, Y. Real-time long-distance ship detection architecture based on YOLOv8. *IEEE Access* **2024**, *12*, 116086–116104. [[CrossRef](#)]
18. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): Vancouver, BC, Canada, 2024; Volume 37, pp. 107984–108011. [[CrossRef](#)]
19. Khanam, R.; Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
20. Lu, B.; Liu, L.; Wang, C.; Wang, D.; Xu, H.; Cao, J. MSC-YOLO: An Accurate and Effective Maritime Ship Detection Model Based on Improved YOLOv11n. *J. Mar. Sci. Eng.* **2026**, *14*, 1066. [[CrossRef](#)]
21. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
22. Liu, S.; Qi, L.; Qin, H.; Li, Z.; Sun, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
23. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Scherer, D.; Müller, A.; Behnke, S. Evaluation of pooling operations in convolutional architectures for object recognition. In *International Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 92–101.

26. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer Nature: Cham, Switzerland, 2022; pp. 443–459.
27. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. YOLOv9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 1–21.
28. Yuan, J.; Wan, L. YOLO-ELR: A High-Precision Lightweight Object Detection Model in Marine Environment. *J. Mar. Sci. Eng.* **2026**, *14*, 998. [[CrossRef](#)]
29. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 3–19.
30. 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*, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
32. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
35. Zheng, M.; Gao, P.; Zhang, R.; Li, K.; Wang, X.; Li, H.; Dong, H. End-to-end object detection with adaptive clustering transformer. *arXiv* **2020**, arXiv:2011.09315.
36. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [[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.