Journal of Engineering, Project, and Production Management
2026, 16(3), 2026-226

AFB-YOLO: An Improved Infrared Ship Detection Method for Maritime Applications Based on YOLOv7

Jia Li
Undergraduate Student, Queen Mary Hainan College, Beijing University of Posts and Telecommunications Lingshui, Hainan, 572400, China, E-mail: jp2023213791@qmul.ac.uk

Project Management
Received February 10, 2026; revised April 23, 2026; accepted April 26, 2026
Available online May 4, 2026

Abstract: Infrared ship images suffer from issues such as blurred target features, complex backgrounds, and a high proportion of small targets, which render existing detection methods prone to missed detections and false alarms. This paper proposes the Attention and Feature Balanced You Only Look Once (AFB-YOLO) method. The model adopts YOLOv7 (You Only Look Once version 7) as its baseline architecture. In the backbone network, an ELAN-O module is introduced to enhance feature extraction capability, ODConv dynamic convolution is employed to adaptively perceive multi-scale features, and a Similarity-based Attention Module (SimAM) is incorporated to intensify focus on ship regions. Within the neck structure, a weighted fusion mechanism is constructed to improve the effect of path aggregation. The detection head utilizes an Adaptively Spatial Feature Fusion (ASFF) module to mitigate spatial feature conflicts and elevate multi-scale perception. With respect to the loss function, the Efficient Intersection over Union (EIOU) loss is introduced to optimize the bounding box regression process and improve localization accuracy. Experimental evaluations on a public infrared ship dataset demonstrate that AFB-YOLO achieves an mAP@0.5 (mAP@0.5 stands for mean Average Precision at Intersection over Union (IoU) threshold 0.5) of 90.7%, markedly outperforming the original YOLOv7 and surpassing YOLOv12 by 12.3 percentage points. These results confirm that the proposed method effectively addresses the inherent limitations of infrared imaging and substantially enhances ship detection performance.

Keywords: Ship detection; feature extraction; dynamic convolution; attention mechanism; path Aggregation.

DOI 10.32738/JEPPM-2026-226

1. Introduction

All-weather and high-precision marine vessel detection is a key technology for ensuring marine economic activities, achieving effective sea area control, and safeguarding national security. Thermal imaging technology does not rely on visible light and has strong anti-interference ability, making it a core means for all-weather maritime surveillance. In complex marine environments, the physical principles of infrared imaging face severe challenges. When the ship-sea temperature gap narrows, or thermal crosstalk arises, target-background contrast drops to a very low level. Background elements such as waves, clouds, and islands have inherent thermal radiation characteristics and often generate strong background noise. Some of their features are very similar to those of actual ships. These problems lead to a low signal-to-noise ratio in infrared images, blurred and unstable target features, and when general target detection algorithms are applied to such scenarios, false positives and false negatives are prone to occur (Mo and Pei, 2022). The research and development of ship target detection algorithms that can overcome these challenges and achieve robust and accurate detection in complex infrared environments have significant theoretical value and practical significance.

In object detection research, many related studies and theories have been put forward. The existing methods are divided into two mainstream technologies: anchor-based (Kong et al., 2020) and anchor-free (Gao et al., 2023). Anchor-based methods constitute a highly classical line of approach. The fundamental procedure entails pre-placing a large number of candidate boxes with varying scales and aspect ratios across an image, after which the model determines whether an object is present within each box and subsequently refines its position and dimensions. Representative algorithms include the two-stage Faster Region-based Convolutional Neural Network (Faster R-CNN) (Ren et al., 2015) and the single-stage YOLO series (Redmon et al., 2016; Khanam and Hussain, 2024; Tian et al., 2025). Although this direction has reached considerable maturity, hyperparameters such as box size and aspect ratio exert a pronounced influence on final performance, rendering tuning laborious, imposing non-negligible computational overhead, and predisposing the process to an imbalance between positive and negative samples. To circumvent these limitations, a number of methods that forgo reliance on preset anchor boxes have been introduced. These approaches directly output object locations, for instance, by predicting center

points (CenterNet (Duan et al., 2019)) or distances to bounding box edges Fully Convolutional One-Stage (FCOS) object detection (Tian et al., 2019), thereby streamlining the overall detection pipeline. Nevertheless, in scenarios involving crowded scenes or severe occlusion, such methods frequently encounter ambiguities in spatial assignment. A more recent direction frames object detection as a set prediction problem, exemplified by Detection Transformer (DETR) (Zhu et al., 2020), which incorporates the Transformer architecture and furnishes a novel paradigm for detection tasks.

Most general detection models have been validated on large-scale visible light datasets such as Microsoft Common Objects in Context (MS COCO) or Pattern Analysis, Statistical Modeling and Computational Learning Visual Object Classes (PASCAL VOC), which typically have rich colors, clear textures, and distinct contours. In sharp contrast, the infrared ship images studied in this paper have fundamental differences and unique detection challenges. Due to the lack of color and fine texture information in infrared imaging, detection can only rely on objects.

In response to these problems, this paper proposes the infrared ship target detection model AFB-YOLO, aiming to achieve high detection accuracy in infrared ship images when object features are blurred, and small targets exist.

This work makes the following primary contributions:

1) In the backbone network, to enhance the network's feature extraction ability for weak targets, the original Efficient Layer Aggregation Network (ELAN) module is improved by introducing ODConv dynamic convolution to enhance the network's feature extraction performance. In addition, integrate the SimAM attention mechanism to enhance the network's ability to focus on key features.
2) For the neck structure, a weighted fusion mechanism is designed to optimize the path aggregation of multi-scale features and effectively improve the fusion efficiency of features at different levels.
3) In the head detection section, the Adaptively Spatial Feature (ASFF) module is integrated to address spatial feature conflicts, significantly enhancing the network's multi-scale feature perception capability. On this basis, the Efficient Intersection over Union (EIOU) loss function is introduced to optimize the bounding box regression process and improve the accuracy of the model in target positioning.
4) Compared with the six mainstream object detection methods, the method proposed in this paper has better detection performance.

The remainder of this paper is organized as follows. Section 2 surveys related research in object detection, emphasizing infrared imagery. Section 3 details the proposed architecture and its core components. Section 4 reports experiments, including ablations and comparisons with alternative detectors. Section 5 concludes and outlines future directions.

2. Literature Review

2.1. General Object Detection Algorithms

Mainstream general-purpose detectors fall into three groups: anchor-based, anchor-free, and transformer-driven approaches. Anchor-based detectors are a classic example of object detection. The core method is to pre-define a set of prior boxes (anchors) of different sizes and aspect ratios on the feature map. This transforms the detection problem into classifying these anchors and performing position regression. Representative algorithms under this paradigm include the high-precision two-stage detector Faster Region-based Convolutional Neural Network (Faster R-CNN) (Ren et al., 2015) and single-stage detectors such as the YOLO series (Redmon et al., 2016; Khanam and Hussain, 2024; Tian et al., 2025) and Single Shot Detector (SSD) (Liu et al., 2016). Single-stage detectors integrate neck structures such as Feature Pyramid Networks (FPN) (Ghiasi et al., 2019) or Path Aggregation Networks (PANet) (Piao et al., 2021) after the backbone network thereby achieving effective detection of multi-scale targets and laying the foundation for real-time target detection. This paradigm is mature and powerful, but its performance heavily relies on the hyperparameter adjustment of anchor boxes, usually requiring cluster analysis based on specific datasets, which limits the generalization ability of the model. In addition, the dense anchor boxes lead to severe imbalance of positive and negative samples and a large amount of computational overhead. To ease information bottlenecks in deep models and stabilize gradient propagation, YOLOv9 (Wang et al., 2024) introduces Programmable Gradient Information (PGI) and the General Efficient Layer Aggregation Network (GELAN), marking a recent advance in information-flow optimization.

Anchor-free detector: To eliminate the reliance on anchor boxes, anchless detectors emerged. These methods aim to locate objects more directly and mainly follow two technical paths. The first type is keypoint-based detection, such as CornerNet (Duan et al., 2019) and CenterNet (Duan et al., 2019), which regard object detection as predicting the corner points or center points of the object. The second type is detection based on dense prediction. Taking Fully Convolutional One-Stage (FCOS) (Tian et al., 2019) as an example, it directly regains the distance from each pixel position to the four edges of the target bounding box. Anchor-free approaches indeed streamline the detection pipeline and obviate the need for extensive preset hyperparameters. However, in their early stages, these methods were prone to localization errors when confronted with severe object overlap or extreme variations in object scale, thereby compromising overall performance. In response, researchers have turned to dynamic label assignment strategies for improvement. Methods such as SimOTA (Ge et al., 2021) and Task Alignment Learning (TAL) (Feng et al., 2021) eschew fixed assignment rules, instead framing label assignment as an optimal transport or task alignment problem. This enables the flexible selection of more appropriate positive samples tailored to the specific characteristics of individual objects. Such refinements have yielded substantial gains for anchor-free detectors and have propelled the anchor-free paradigm to become the prevailing choice in contemporary practice. Models including YOLOv10(Wang et al., 2024) now concurrently adopt anchor-free architectures and dynamic label assignment strategies, with the explicit aim of further reconciling detection speed with accuracy.

Transformer-based detectors represent another significant direction that has emerged in recent years. Exemplified by DETR (Detection Transformer) (Zhu et al., 2020), this line of work reframes object detection as a set prediction task, leveraging the self-attention mechanism of the transformer (Han et al., 2021) to establish global contextual associations and employing bipartite matching for label assignment. This approach dispenses entirely with the need for anchor design and obviates Non-Maximum Suppression (NMS) in post-processing, thereby realizing true end-to-end detection from input to output for the first time. While conceptually compelling, early instantiations exhibited pronounced limitations, including notably slow training convergence and suboptimal performance on small-scale objects (Zhu et al., 2020; Dai et al., 2021). In response to these shortcomings, numerous subsequent works have introduced targeted optimizations. Deformable DETR (Zhu et al., 2020) proposed a deformable attention module that restricts the model's focus to a sparse set of key sampling points across the image, thereby reducing computational burden and accelerating convergence. DINO (Zhang et al., 2022) further introduced a contrastive denoising training strategy, which enhances both training efficiency and overall detection accuracy. Of particular note is RT-DETR (Huang et al., 2019), which incorporates a more efficient hybrid encoder and an IoU-aware query selection mechanism, enabling transformer-based detectors to achieve real-time performance that rivals, for the first time, CNN-based models, such as the YOLO series. Moreover, this end-to-end paradigm has exerted influence beyond transformer-centric architectures. For instance, YOLOv10 (Wang et al., 2024) adopts a "one-to-many" matching strategy in its post-processing stage, thereby successfully eliminating NMS from the traditionally real-time-oriented YOLO framework and achieving end-to-end real-time detection as well.

2.2. Infrared Ship Target Detection

In the specific domain of infrared ship target detection, researchers have predominantly pursued targeted enhancements built upon established frameworks, with modifications to the YOLO series being the most prevalent. To address the challenges posed by the complex and variable marine environment, investigators have explored a range of distinct technical avenues. In terms of feature enhancement, attention modules constitute a common choice: Huang et al. (Huang et al., 2019) introduced channel attention to recalibrate the response weights of different feature channels, while the CBAM module proposed by Woo et al. (2018) which has been substantiated to confer discernible benefits for object recognition in maritime scenes. With respect to multi-scale fusion features, various architectural adjustments have been undertaken. For example, the BiFPN structure advanced by Tan et al. (2020) facilitates more efficient information exchange across hierarchical feature levels through weighted bidirectional fusion. Zhao et al. (2019) focused on refining cross-scale connectivity to render the feature pyramid more adaptive to scale variations. As for the persistent difficulty of missed detections and imprecise localization concerning distant small targets, current efforts center primarily on innovations in loss function design. One approach entails improving localization metrics: Yang et al. (2022) substituted the conventional IoU measure with the Normalized Gaussian Wasserstein Distance (NWD), which substantially enhances localization accuracy for small targets. Chen et al. (2020) proposed the Spatial Constraint Intersection over Union (SC-IoU) loss, which incorporates spatial constraints into the loss computation to stabilize bounding box regression and better align predicted boxes with actual object contours. Second, data augmentation techniques are utilized: the SRGAN developed by Ledig et al. (2017) provides an effective solution for the super-resolution reconstruction of infrared images, and Wang et al. proposed a data generation method to alleviate the challenges of small sample learning. In the field of model lightweighting, ShuffleNet by Zhang et al. (2018) and MobileNet by Howard et al. (2017) provide efficient infrastructure for real-time detection systems. In addition, the emerging Transformer architecture has been introduced into object detection. Carion et al. (2020) proposed DETR, which initiated a new paradigm for end-to-end detection. Liu et al. (2018) continuously optimized the performance of the detection network through structural re-parameterization technology.

The core objective of this study is to break away from this "single-point" and "static" improvement approach. The underlying rationale involves the incorporation of a number of dynamic adaptive modules, thereby enabling the detector, when processing infrared imagery, to respond flexibly to image content in both the extraction of fine-grained details and the fusion of multi-scale features. This approach permits a more holistic handling of the diverse complexities inherent in infrared ship detection.

3. Methodology

At the model level, we designed and implemented AFB-YOLOv7. As illustrated in Fig. 1, the proposed improvements span the three core modules of the model: the backbone, the neck, and the detection head. In the backbone, we incorporate the ELAN-O module, which integrates ODConv dynamic convolution (Li et al., 2022) and the SimAM attention mechanism (Sun et al., 2025). The neck component adopts the weighted fusion mechanism of BiFPN (Doherty et al., 2025) to construct the MP-B module and optimize the feature pyramid. The Detection Head component utilizes ASFFDetect (Qiu et al., 2022) to enhance the adaptive fusion capability of spatial features. The following section will provide detailed explanations of each optimized component.

3.1. Backbone Network Improvements

In infrared images, there are significant differences in the shape, texture, illumination, and even direction of ships. The traditional static convolution kernel is a fixed feature extraction template and is difficult to effectively adapt to these dynamic changes. We replaced the original ELAN module with ElAN-O (Fig. 2 (b)) and introduced OD convolution in ELAN to enhance the model's ability to adapt to dynamic features.

The core concept of ODConv (Fig. 2 (a)) is to dynamically synthesize a dedicated convolution kernel in real time for each input sample. This process is decomposed into continuous multiplicative modulation of the fundamental kernel in four mutually orthogonal dimensions.

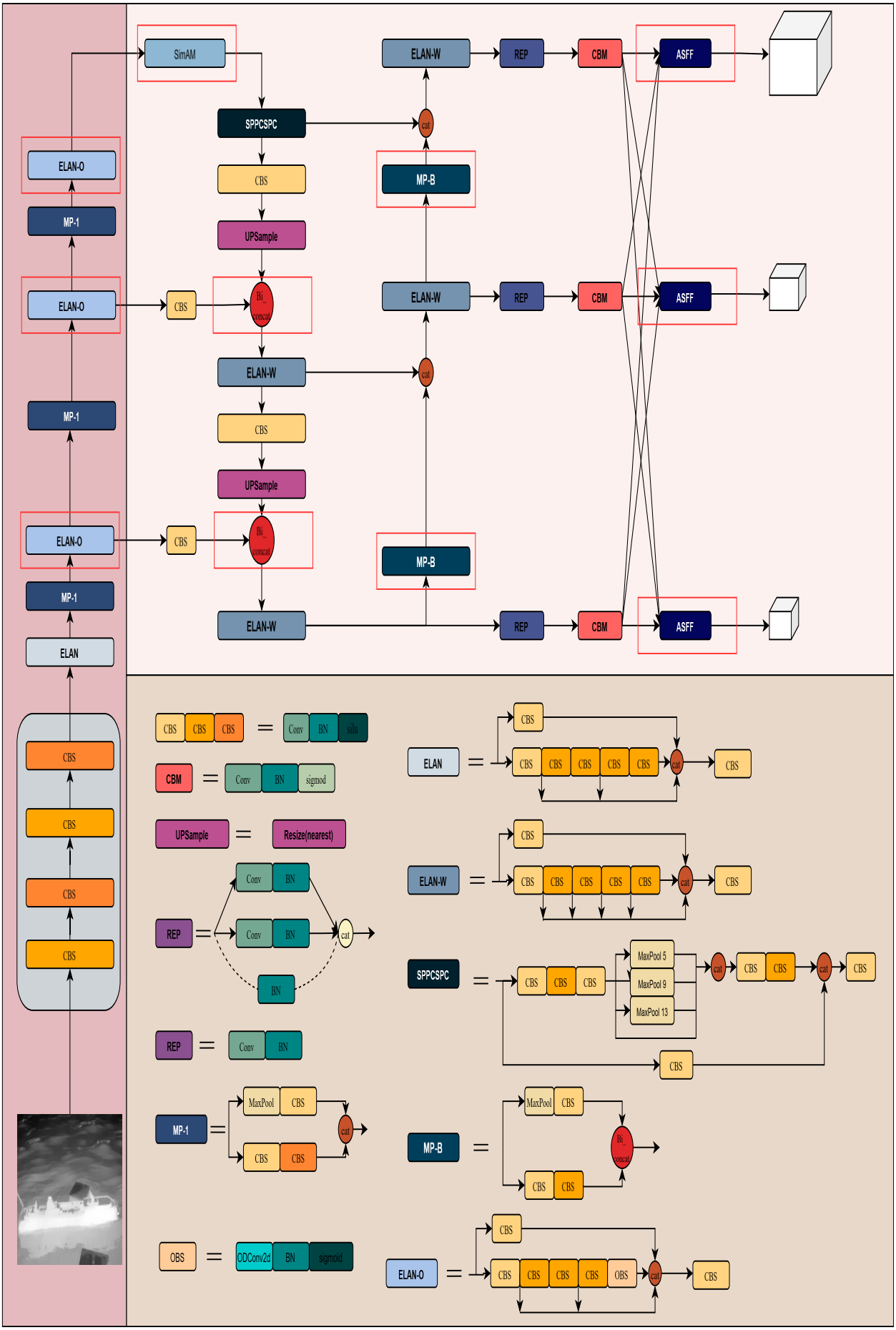
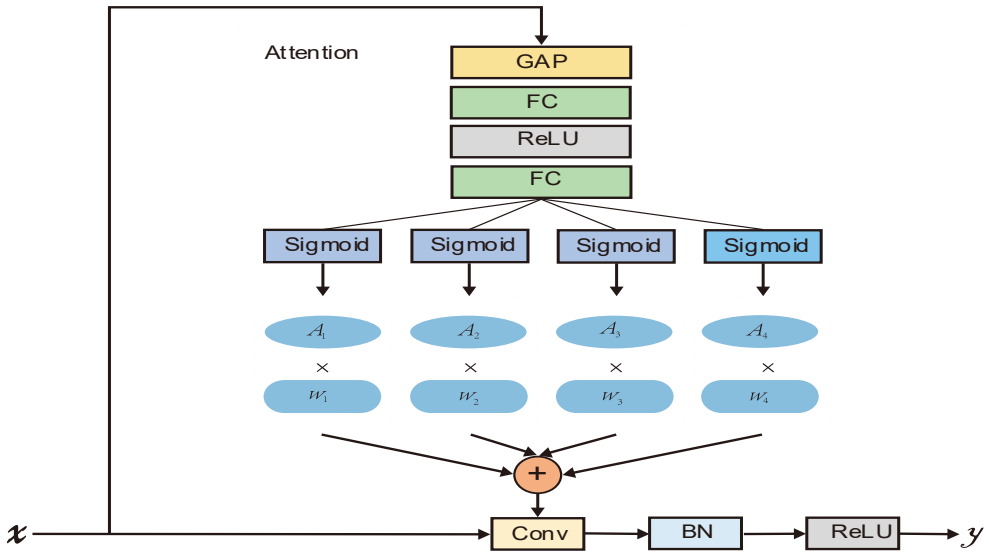
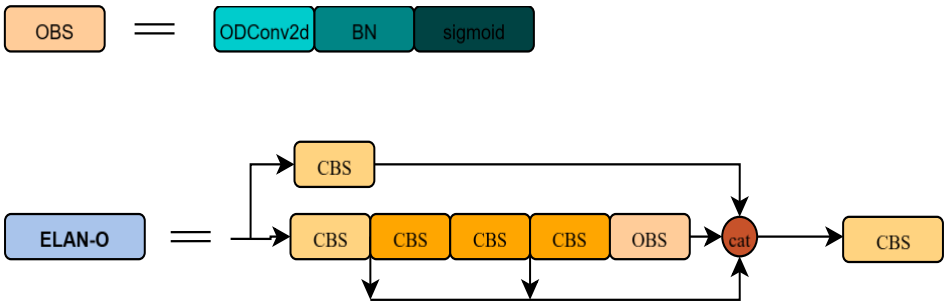


Fig. 1. Overall model architecture diagram (afb-yolov7 model structure)



(a) ODConv dynamic convolution details



(b) Improved weighted fusion architecture

Fig. 2. ODConv and ELAN-O dynamic convolution modules

Convolution kernel attention (αw): Perform a linear combination of k predefined basic kernels based on the input features.

Output Channel Attention (αo): Independently scales each output channel filter within the aggregated convolutional kernel.

Input Channel Attention (αi): Adjusts the weights corresponding to each input channel within the filter.

Spatial Attention (αs): Re-weights the spatial positions of the convolution kernel (e.g., a 3x3 grid).

This process is driven by a lightweight attention network that compresses and refines the input feature map x, ultimately generating the aforementioned four sets of attention weights. Consequently, the final dynamic convolution kernel W′ generated for sample x is the product of the base convolution kernel W undergoing a series of content-related modulations. As shown in Eq. (1).

$$W'(x) = (\alpha w(x) \otimes W) \odot \alpha s(x) \odot \alpha i(x) \odot \alpha o(x) \tag{1}$$

Here, ⊗ denotes weighted summation along the number of dimensions of the convolution kernel, while ⊙ represents element-wise multiplication. This makes each convolution a highly customized feature extraction operation.

SimAM Parameter-Free Attention Module

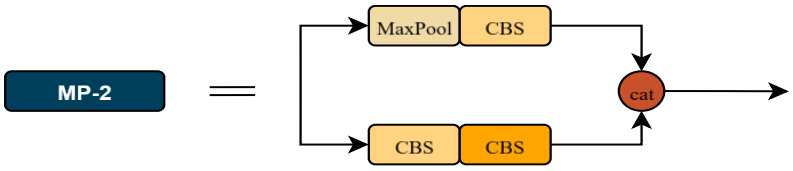
To enhance the model's focusing ability on ship targets, a SimAM parameter-free attention mechanism is introduced between the final output of the backbone network and the SPPCSPC module of the neck network (Sun et al., 2025). In infrared images with complex backgrounds, the real target signal is often masked by a large number of redundant background features (such as sea clutter). The core strength of SimAM lies in its neuroscience-inspired approach, which infers importance by evaluating the uniqueness of each neuron (feature) relative to adjacent neurons without injecting additional parameters. The network can independently identify the neurons with the highest discriminative ability in the feature space, and these neurons correspond to the true ship targets.

Specifically, SimAM defines an energy function et for each neuron t to measure its linear separability from other neurons in the channel, as shown in Eq. (2).

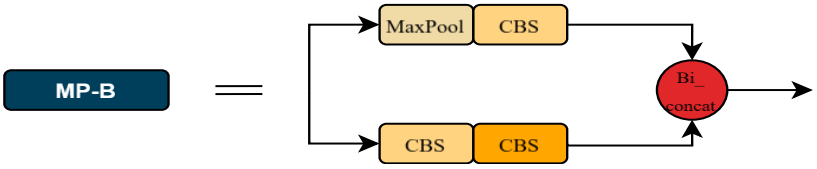
$$e_t = \frac{1}{M-1}\Sigma_{\{i=1\}}^{\{M-1\}(-1-(w_t x_i + b_t))^2} + \left(1 - (w_t t + b_t)\right)^2 + \lambda w_t^2 \quad (2)$$

The lower the energy e_t, the greater the pattern difference between that neuron and its neighbors, indicating higher importance. Through this mechanism, SimAM assigns higher attention weights to more distinctive features representing ship targets while suppressing background noise features. Ultimately, by converting energy into attention weights and applying them to the original feature maps, the network's focus on critical targets is effectively enhanced, achieving precise targeting of vessel objects at the feature level.

3.2. Neck Structure Improvements



(a) Neck structure before improvement



(b) Neck structure after improvement

Fig. 3. Comparison of standard and weighted attribute integration in the Yolov7 network

The original Yolov7 network (Fig. 3(a)) was used to control the fragments of the standard puzzle (concat) to integrate various attributes. The limit of an action is to treat all input attributes as equivalent internal mechanisms. Useful information is based on the different characteristics of multi-functional targets (such as ships). Treating these factors equally will weaken the wording and stifle the accuracy of the test.

To address this issue, we developed a weighted integration strategy to optimize the attribute distortion of a given path (see Fig. 3(b)), thereby improving the network history. Without providing a simple alternative, the various attributes integrated in ns-netzwerk were reprogrammed, and the existing standard interrupts were replaced by weighted merging.

The main idea is to introduce the learned scores for each integrity scheme, allowing the network to evaluate and adjust the meanings of different parameters. To integrate the two attributes P_1^{in} and P_2^{in}, the combined attribute is shown in Eq. (3).

$$P^{out} = Conv\left(\frac{w_1 \cdot P_1^{in} + w_2 \cdot P_2^{in}}{w_1 + w_2 + \varepsilon}\right) \quad (3)$$

Here, w_1 and w_2 are weights learned through backpropagation, while ε is a small value (e.g., 0.0001) to ensure numerical stability. The weights are passed through the ReLU function to guarantee non-negativity. This approach enables the network to dynamically assign higher weights to more important feature maps based on the data, achieving an optimal weighted combination. Consequently, it enhances the feature representation capability for multi-scale targets, particularly small ones.

3.3. Detection Head Improvement

After the neck network outputs multi-scale feature maps, feature maps at different scales may contain inconsistent or even conflicting information for the same object. To address this issue, we replace the original IDetect detection head with the ASFFDetect detection head (Wang et al., 2020). The core idea of ASFF (Adaptive Spatial Feature Fusion) is to achieve dynamic, per-pixel feature selection and fusion.

For feature maps from different levels $l \in \{1, 2, 3\}$, ASFF first spatially unifies them to the same resolution. Subsequently, a lightweight convolutional network generates three spatial attention weight maps α^l, β^l, γ^l for each scale. For the output feature map at level l, the feature vector at position (i, j) is computed as shown in Eq. (4).

$$y_{\{ij\}}^l = \alpha_{\{ij\}}^l \cdot x_{\{ij\}}^{\{1\to l\}} + \beta_{\{ij\}}^l \cdot x_{\{ij\}}^{\{2\to l\}} + \gamma_{\{ij\}}^l \cdot x_{\{ij\}}^{\{3\to l\}} \quad (4)$$

Here, $x_{\{ij\}}^{\{n\to l\}}$ denotes the result of adjusting the feature vector at position (i, j) of the feature map at level n to the resolution of level l. The weights α^l, β^l, and γ^l are generated by the softmax function, ensuring that $\alpha_{\{ij\}}^l + \beta_{\{ij\}}^l + \gamma_{\{ij\}}^l = 1$ at each spatial location. Through end-to-end learning, the network intelligently determines which scale of information is most reliable at each position in the image and assigns it the maximum weight, thereby effectively resolving feature

conflicts.

3.4. Loss Function Improvement

YOLOv7's loss function consists of three supporting components: restore box boundaries (L_{box}), sense of loss of trust (L_{obj}), and loss category (L_{cls}). The optimization in this article focuses on the loss of the border box (L_{box}), which is essential for the precision of the model's location.

The original Model uses the Complete Intersection over Union (CIoU) loss (Howard et al., 2017) as the loss function for regression to the border frame. The restriction imposed by the iou is that the punishment for exterior relationships focuses only on expected "exterior relationships" and actual exterior relationships, while ignoring "completely different exterior relationships." This optimization does not immediately slow the rate of model convergence and reduce the optimization accuracy.

To solve this problem and correct model precision, we use EIOU losses (Zhang et al., 2018). The core advantage of EIOU Loss is that it directly decomposes the aspect ratio penalty into individual components, which are used to measure the differences between the predicted box width w and height h and the actual box width w^{gt} and height h^{gt} on the ground. This design enables EIOU to directly minimize the discrepancies in width and height, rendering the optimization direction more explicit and the gradients more effective. Consequently, convergence is accelerated, and localization accuracy is enhanced. Its complete definition is as shown in Eq. (5).

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp} = (1 - IOU) + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{C_w^2} + \frac{\rho^2(h, h^{gt})}{C_h^2} \quad (5)$$

Where L_{IOU} denotes the conventional Intersection over Union loss, L_{dis} accounts for the offset between center points, and the critical term L_{asp} directly constrains the discrepancies in width and height between the predicted bounding box and the ground truth.

4. Experiments and Results Analysis

4.1. Experimental Environment and Parameter Settings

Set the training epoch to 100 and the batch size to 16. The input image size is uniformly adjusted to 384x288. The optimizer adopts SGD, with an initial learning rate of 0.001, and combines cosine annealing to achieve learning rate attenuation. All experiments were conducted on the same high-performance server, configured with an Intel® Xeon® Gold 6248R CPU, 1.0 TB of system memory, and an NVIDIA GeForce RTX 3090 (24GB) GPU.

4.2. Dataset and Evaluation Metrics

This study uses the publicly available infrared ship dataset (http://openai.raytrontek.com/apply/E_Sea_shipping.html/) for model training and evaluation. This dataset contains a variety of scenarios, covering infrared images of ships in real Marine environments such as coastal areas and the high seas. The dataset involves ships of different types and sizes, including various lighting conditions such as day and night, and can comprehensively test the robustness of the algorithm in complex infrared scenarios. In the experiment, the dataset was strictly divided into a training set, a validation set and a test set in an 8:1:1 ratio.

Evaluation index

To comprehensively evaluate the model's performance, widely recognized standard metrics in the field of object detection are adopted: accuracy, recall rate, and average accuracy (mAP).

4.3. Ablation Studies

To validate the effectiveness of each proposed enhancement module and their combined impact, we designed comprehensive ablation experiments. Using YOLOv7 as the baseline, we progressively or concurrently integrated the proposed enhancements and evaluated their performance on the same test dataset. The experimental results are presented in Table 1.

Table 1 shows that compared to the original reference model, the YOLOv7 model has some limitations in this dataset: the mAP@0.5 was only 37.3%. Very low verification rates (35.2%) indicate that the model produces a large amount of false positives, directly reflecting the difficulty of detecting many types in the complex infrared ocean scene.

The key role of the neck network is simply to optimize the neck network and head positioning, and achieve a jump in quality in model performance. mAP@0.5 increased from 37.3% to 88.4%. Optimizing connectors and resolving feature inconsistencies are important for improving the performance of models.

The synergy effect of backbone optimization: When ODConv or ODConv+SimAM is introduced alone on the baseline model, the mAP@0.5 of the model even slightly decreases. Only enhancing the feature extraction capability of the backbone network without optimizing the feature fusion path may lead to unsatisfactory performance. When these backbone modules are combined with the robust Neck Network (BiFPN+ASFF), the performance will steadily improve.

The advantages of multi-module collaboration: The complete AFB-YOLOv7 model integrating all four modules achieved the best performance in all metrics, with an accuracy rate of 90.3%, a recall rate of 86.0%, and a mAP@0.5 rate of 90.7%. This fully validates the effectiveness of the systematic and multi-module collaborative improvement strategy

proposed in this paper.

Table 1. Ablation study results

Model	P (%)	R (%)	mAP@0.5 (%)	Parameter Quantity (M)	GFLOPS
Baseline (YOLOv7)	35.2	68.0	37.3	37.23	105.2
Baseline + ODConv	35.0	69.1	35.8	37.26	101.0
Baseline + ODConv + SimAM	36.5	67.2	36.6	38.18	102.7
Baseline + BiFPN + ASFF	85.9	85.4	88.4	58.89	138.7
Baseline + BiFPN + ASFF + ODConv	87.4	87.3	90.2	58.92	134.5
AFB-YOLOv7 (Model in This Paper)	90.3	86.0	90.7	59.84	136.1

4.4. Comparative Analysis

Table 2 summarizes the performance comparison between AFB-YOLOv7 and several representative YOLO detectors under identical testing conditions. Five example images, (a) to (e), are shown in Fig. 4.

Table 2. Performance comparison of different models

Model	P (%)	R (%)	mAP@0.5 (%)
YOLOv7 (Baseline) (Wang et al., 2023)	35.2	68.0	37.3
YOLOv5 (Jocher et al., 2020)	85.2	73.7	79.8
YOLOv8	88.8	73.7	80.6
YOLOv9m (Wang et al., 2024)	89.4	80.4	86.3
YOLOv10b (Wang et al., 2024)	90.4	78.2	85.9
YOLOv11 (Khanam and Hussain, 2024)	83.6	73.1	78.7
YOLOv12 (Tian et al., 2025)	84.7	72.8	78.4
AFB-YOLOv7 (Model in This Paper)	90.3	86.0	90.7

Comparative experimental results are demonstrated as folows:

1) Performance of AFB-YOLOv7: In terms of mAP@0.5, AFB-YOLOv7 attained a value of 90.7%, ranking first among all models included in the comparison. For reference, YOLOv9m and YOLOv10b achieved scores of 86.3% and 85.9%, respectively. The magnitude of this margin alone substantiates the high efficacy of the modifications tailored to the characteristics of infrared imagery and elevates the detection performance to the current state-of-the-art level.

2) Balance between Precision and Recall: In comparison with other models, AFB-YOLOv7 simultaneously achieves the highest recall (86.0%) while maintaining precision at a high level of 90.3%, thereby exhibiting no evident trade-off between the two metrics. The balance between high precision and high recall rate indicates that the model can accurately identify targets and minimize missed detections, which is a key capability for important applications

such as maritime safety.

3) The significant improvement potential of the YOLOv7 baseline: Initially, YOLOv7 performed poorly on this infrared dataset (mAP@0.5 only 37.3%), and our target system enhancement achieved a performance leap of over 53 percentage points. Selecting an appropriate baseline model and applying in-depth, domain-specific optimizations is an effective way to improve the performance of specific detection tasks.

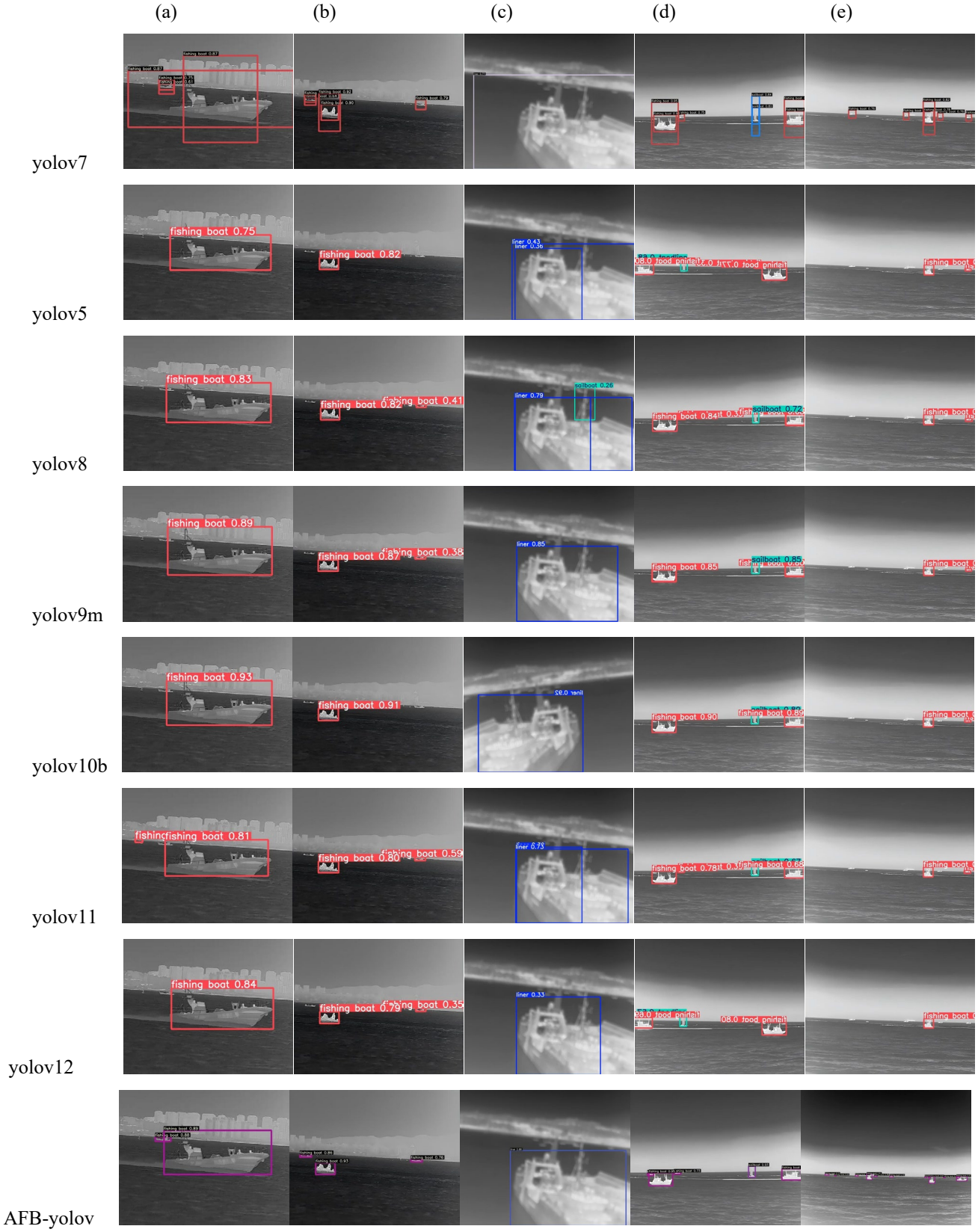


Fig. 4. Detection of different models

5. Conclusion

This paper proposes an infrared vessel detection model named AFB-YOLO to address the specific challenges of vessel detection in infrared images. The model is based on YOLOv7, with improvements made to the backbone network, neck structure and detection head, respectively. Experimental results demonstrate that, compared to the mainstream YOLOv11 and YOLOv12 models, the proposed method achieves improvements of 6.7% and 5.6% in accuracy and recall, 12.9% and 13.2% in mAP@0.5, and 12.0% and 12.3%, respectively, indicating that it achieves high detection accuracy. Future work will explore model compression and quantization techniques to further enhance the model's inference speed without sacrificing too much accuracy, making it more suitable for edge computing devices. For port security or maritime managers, these results can directly influence equipment selection and personnel scheduling. For instance, during nighttime or foggy conditions, if the accuracy of infrared surveillance algorithms is sufficiently high, it can reduce the pressure on on-duty personnel to monitor screens continuously, whilst also minimizing the number of unnecessary emergency call-outs caused by false alarms. Managers can redirect the human resources saved to other tasks requiring greater judgment, rather than expending them on repeatedly verifying false alarms. Furthermore, there is a related issue worth exploring: most current smart port systems rely too heavily on visible-light cameras, which become unreliable in low-light conditions. The improved accuracy of AFB-YOLO in infrared applications suggests that, in the future, it may be possible to address the shortcomings of night-time perception using lower-cost infrared solutions, rather than necessarily pursuing the expensive route of radar upgrades. If, in the next phase, the model can be made smaller and faster so that it can run on standard edge devices, then an all-weather, comprehensive vessel monitoring network will be much more likely to be implemented. This would provide practical assistance in shifting maritime regulation from 'reactive response' to 'proactive warning'.

Author Contributions

The sole author of this manuscript is responsible for the entire research process.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used Gemini solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. *In European Conference on Computer Vision*, 213-229. Springer.

Chen, Z., Chen, K., Lin, W., See, J., Yu, H., Ke, Y., and Yang, C. (2020). PIOU loss: Towards accurate oriented object detection in complex environments. *In European Conference on Computer Vision,* 195-211. Springer.

Dai, Z., Cai, B., Lin, Y., and Chen, J. (2021). UP-DETR: Unsupervised pre-training for object detection with transformers. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 1601-1610.

Doherty, J., Gardiner, B., Kerr, E., and Siddique, N. (2025). BiFPN-YOLO: One-stage object detection integrating bi-directional feature pyramid networks. *Pattern Recognition*, 160, 111209.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). CenterNet: Keypoint triplets for object detection. In Proceedings of the *IEEE/CVF International Conference on Computer Vision, 6569*-6578.

Feng, C., Zhong, Y., Gao, Y., Scott, M. R., and Huang, W. (2021). Tood: Task-aligned one-stage object detection. In Proceedings of the *IEEE/CVF International Conference on Computer Vision*, 3490-3499.

Gao, F., Cai, Y., Deng, F., Yu, C., and Chen, J. (2023). Feature alignment in anchor-free object detection. *IEEE Transactions on Circuits and Systems for Video Technology,* 33(8), 3799-3810.

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430.

Ghiasi, G., Lin, T. Y., and Le, Q. V. (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7036-7045.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, 34, 15908-15919.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W. (2019). CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the *IEEE/CVF International Conference on Computer Vision,* 603-612.

Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., Ingham, F., Poznanski, J., Fang, J., Yu, L., and Wang, M. (2020). ultralytics/yolov5: v3. 1-bug fixes and performance improvements. *Zenodo*.

Khanam, R., and Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.

Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., and Shi, J. (2020). FoveaBox: Beyond anchor-based object detection. *IEEE Transactions on Image Processing,* 29, 7389-7398.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the *IEEE conference on computer vision and pattern recognition* 4681-4690.

Li, C., Zhou, A., and Yao, A. (2022). Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition,* 8759-8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision,* 21-37. Springer.

Mo, W., and Pei, J. (2022). Nighttime infrared ship target detection based on two-channel image separation combined with saliency mapping of local grayscale dynamic range. *Infrared Physics & Technology*, 127, 104416.

Piao, Y., Jiang, Y., Zhang, M., Wang, J., and Lu, H. (2021). PANet: Patch-aware network for light field salient object detection. *IEEE Transactions on Cybernetics,* 53(1), 379-391.

Qiu, M., Huang, L., and Tang, B. H. (2022). ASFF-YOLOv5: Multielement detection method for road traffic in UAV images based on multiscale feature fusion. *Remote Sensing,* 14(14), 3498.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, 779-788.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems,* 28.

Sun, S., Zheng, S., Xu, X., and He, Z. (2025). GD-YOLO: A lightweight model for household waste image detection. *Expert Systems with Applications,* 279, 127525.

Tan, M., Pang, R., and Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781-10790.

Tian, Y., Ye, Q., and Doermann, D. (2025). YOLOv12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524.

Tian, Z., Shen, C., Chen, H., and He, T. (2019). FCOS: Fully convolutional one-stage object detection. In Proceedings of the *IEEE/CVF International Conference on Computer Vision*, 9627-9636.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., and Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems,* 37, 107984-108011.

Wang, C. Y., Bochkovskiy, A., and Liao, H. Y. M. (2023). 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*, 7464-7475.

Wang, C. Y., Yeh, I. H., and Liao, H. Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision,* 1-21.

Wang, Y., Zhang, J., Kan, M., Shan, S., and Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12275-12284.

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the *European Conference on Computer Vision,* 3-19.

Yang, C., Huang, Z., and Wang, N. (2022). QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 13668-13677.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H. Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605.

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition, 6848*-6856.

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., and Ling, H. (2019). M2Det: A single-shot object detector based on multi-level feature pyramid network. In Proceedings of the *AAAI Conference on Artificial Intelligence,* 33(1), 9259-9266.

Zhao, Z., Chen, S., Ge, Y., Yang, P., Wang, Y., and Song, Y. (2024). RT-DETR-Tomato: Tomato target detection algorithm based on improved RT-DETR for agricultural safety production. *Applied Sciences*, 14(14), 6287.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Li Jia is enrolled in the Information and Computing Science program at Queen Mary Hainan College, Beijing University of Posts and Telecommunications. During her studies, she won the Third Prize in Hainan Province at the Chinese National College Students Mathematics Competition and has passed the College English Test Band 4 (CET-4). Her current curriculum focuses on data analysis, algorithm design, and information security, with proficiency in programming and in computational tools such as Python. Her research interests include machine learning, data mining, and information security.