# MSA-YOLO: An Optimized UAV Object Detection Algorithm for Low-Visibility Maritime

Longcheng Huang , Mengguang Liao * , Shaoning Li , Chuanguang Zhu , Sichun Long

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2492.v1

Keywords: You Only Look Once (YOLO) algorithm; Unmanned Aerial Vehicle (UAV) imagery; object detection; attention mechanism

 doi:10.20944/preprints202603.2492.v1



*Article*

# MSA-YOLO: An Optimized UAV Object Detection Algorithm for Low-Visibility Maritime

**Longcheng Huang [1,2], Mengguang Liao [1,2,*], Shaoning Li [1,2], Chuanguang Zhu [1,2] and Sichun Long [2]**

[1] Sanya Institute of Hunan University of Science and Technology, Sanya 572024, China
[2] School of Earth Sciences and Spatial Information Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
* Correspondence: 1100114@hnust.edu.cn

## Highlights

**What are the main findings?**

- A lightweight detection model, termed MSA-YOLO, is proposed for UAV-based maritime object detection in low-visibility environments.
- The proposed model achieves systematic optimization of the backbone, neck, and detection head by jointly leveraging multi-scale convolution and attention mechanisms.

**What are the implications of the main findings?**

- Experimental results on AFO, Zhoushan Island, and Shandong Province datasets demonstrate that the proposed model achieves a favorable balance between detection accuracy and computational efficiency, making it suitable for low-visibility maritime detection.
- Experimental results demonstrate that systematic optimization of the model architecture effectively improves detection performance under low-visibility conditions, providing a reference for the design of related models.

## Abstract

Maritime search and rescue is an important component of emergency response frameworks and primarily relies on UAVs for maritime object detection. However, maritime accidents frequently occur in low-visibility environments, such as foggy or low-light conditions, which lead to low contrast, blurred object boundaries, and degraded texture representations. Most existing maritime object detection algorithms are developed for natural light scenes, and their performance deteriorates markedly when deployed directly in low-visibility environments, primarily due to reduced image quality that hinders feature extraction and semantic information aggregation. Although several studies incorporate image enhancement techniques prior to detection to improve image quality, these approaches often introduce significant additional computational overhead, limiting their practical deployment on UAV platforms. To tackle these challenges, this paper proposes a lightweight model built upon a recent YOLO framework, termed Multi-Scale Adaptive YOLO (MSA-YOLO), for maritime detection using UAVs in low-visibility environments. The proposed model systematically optimizes the backbone, neck, and detection head networks. Specifically, an improved StarNet backbone is designed by integrating ECA mechanisms and multi-scale convolutional kernels, which strengthen feature extraction capability while maintaining low computational overhead. In the neck network, a high-frequency enhanced residual block branch is inserted into the C3k2 module to capture richer detailed information, while depthwise separable convolution is utilized to further reduce computational cost. Moreover, a non-parametric attention module is incorporated into the detection head to adaptively optimize features in the classification and regression branches. Finally, a joint loss function that combines bounding box regression, classification, and distribution focal losses is utilized to improve detection accuracy and training stability. Experimental results on the constructed AFO, Zhoushan Island, and Shandong Province datasets demonstrate that, relative to

 doi:10.20944/preprints202603.2492.v1



YOLOv11-s, MSA-YOLO reduces model parameters and FLOPs by 52.07% and 41.36%, respectively, while achieving improvements of 1.11% and 1.33% in mAP@0.5:0.95 and mAP@0.5. These results indicate that the proposed method effectively balances computational efficiency and detection accuracy, rendering it suitable for for practical maritime search and rescue applications in low-visibility environments.

**Keywords:** You Only Look Once (YOLO) algorithm; Unmanned Aerial Vehicle (UAV) imagery; object detection; attention mechanism

## 1. Introduction

As maritime traffic density continues to increase, maritime accidents are becoming more frequent [1]. Traditional search and rescue methods have limitations in response speed and coverage area, necessitating more efficient technological solutions. An Unmanned Aerial Vehicle (UAV), due to its ability to reach remote or hard-to-access areas and its low cost, has become an effective and convenient tool for maritime rescue [2]. Their flexible operational capability in complex marine environments gives them significant potential for object detection tasks under harsh conditions. However, in low-visibility environments, such as foggy or low-light conditions, the contrast between targets and the background decreases [3], texture and edge features are weakened [4], and large scale variations occur [5], making maritime object detection in low-visibility environments still challenging.

Object detection methods based on deep learning can typically be categorized into two types—two-stage and one-stage detectors—and have been widely applied across diverse practical tasks. Two-stage detectors adopt a two-stage pipeline, which initially generates candidate regions and then classifies and regresses the bounding boxes for these regions. Representative methods in this category, including the original R-CNN [6], the improved Fast R-CNN [7], and Faster R-CNN [8], the latter of which features a region proposal network (RPN), demonstrate high detection accuracy in most conventional object detection tasks. However, due to their multi-stage processing, the inference speed is relatively slow, which makes them unsuitable for latency-sensitive tasks. One-stage detectors dispense with explicit candidate regions, directly predict class labels and object coordinates for each location on the feature map. Compared to two-stage detectors, one-stage detectors offer lower inference latency, facilitating the advancement of deep learning-based maritime object detection research. Zhou et al. [9] improved YOLO by redesigning the anchor box clustering strategy and optimizing the loss function, thereby enhancing detection performance for vessels. Sun et al. [10] proposed HRNet, which employs hierarchical multi-scale fusion to preserve high-resolution representations, performing well in capturing ship boundary details. Kim et al. [11] proposed LiM-YOLO, which adjusts pyramid level outputs to better capture small and narrow vessel scales and incorporates group-normalized convolutional blocks to stabilize training on high-resolution imagery. Wang et al. [12] proposed YOLO11s-APFAN, which enhances multi-scale feature representation by incorporating an adaptive pyramid focus and diffusion network, while adopting Wise-IoU for improved bounding box regression, enabling the model to better focus on ships amid wake interference. Xie et al. [13] proposed MiSSNet, which mitigates background-class semantic distribution discrepancies via class-specific regularization and local semantic distillation. In response to the practical needs of unmanned surface vehicles for surface object detection, Cheng et al. [14] improved the YOLO backbone by applying network pruning and integrating focal loss with blank label training strategies.

However, the aforementioned algorithms are mainly designed for natural light conditions. When directly applied to low-visibility environments, their performance tends to degrade, primarily due to the reduction in image quality, which makes it difficult to extract discriminative features and aggregate the related semantic information. Specifically, contrast degradation in low-visibility environments weakens the discriminative features of small-scale objects, making them more prone

 doi:10.20944/preprints202603.2492.v1



to missed detections. Meanwhile, the reduction in signal-to-noise ratio caused by low-visibility imaging enhances pseudo-structures in background regions, such as noise-induced false edges, thereby degrading object localization accuracy. Figure 1 provides a visual comparison of YOLOv8 detection results under natural-light and foggy conditions, clearly demonstrating the severe performance degradation caused by low visibility, where red boxes indicate false positives and blue boxes denote false negatives.

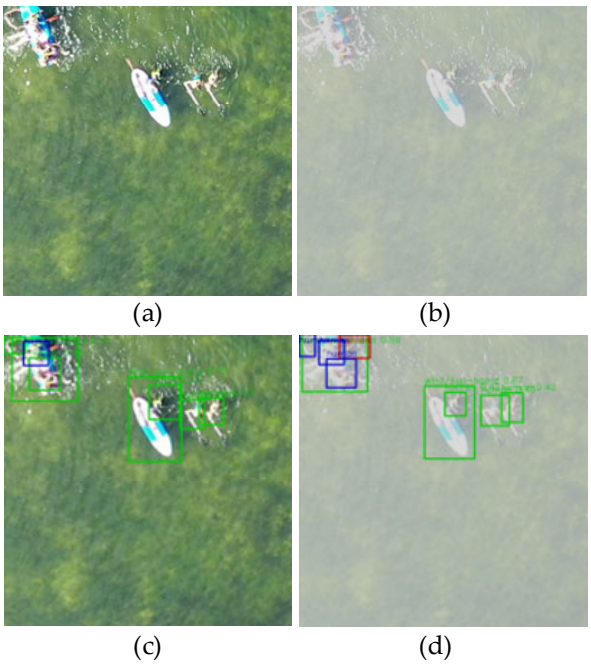



**Figure 1.** Visual comparison of YOLOv8 detection performance under natural-light and foggy conditions: (a) natural-light image; (b) foggy image; (c) YOLOv8 detection result under natural-light conditions; and (d) YOLOv8 detection result under foggy conditions.

In response, several solutions have been proposed. Zhang et al. [15] introduced a multi-task learning strategy that enables end-to-end optimization of image restoration and object detection, effectively reducing false negatives and false positives caused by environmental interference. Ma et al. [16] designed a YOLOX-based network that enhances the detection of multiple objects in difficult weather using a joint framework of domain adaptation and image restoration. To address the issue of blurred features under foggy conditions, Liu et al. [17] integrated an adaptive dehazing module with an improved YOLOv5 detection network, enhancing vessel recognizability. The aforementioned methods require image processing to enhance image quality before object detection, which incurs substantial computational overhead, thereby making them difficult to deploy effectively on UAVs. Therefore, achieving efficient maritime object detection in low-visibility environments requires detection models to effectively balance environmental robustness, detection accuracy, and computational efficiency, which remains a challenging research problem.

To tackle the aforementioned challenges, we propose a lightweight model, termed Multi-Scale Adaptive YOLO (MSA-YOLO), aimed at maritime detection using UAVs in low-visibility environments. Instead of relying on computationally expensive image enhancement preprocessing, the proposed model focuses on improving feature representation and information extraction within the detection network through lightweight and task-oriented module design. The main contributions of this work are outlined below:

1. On the basis of the C3k2 module, a high-frequency enhanced residual block (HFERB) is introduced to capture richer detailed information. Additionally, a Simple, Parameter-Free Attention Module (SimAM) is incorporated within the detection head to adaptively optimize

 doi:10.20944/preprints202603.2492.v1



features for both classification and regression branches, improving the model's detection performance and stability.

2. An improved StarNet network is adopted as the backbone, integrating efficient channel attention (ECA) mechanisms and multi-scale convolutional kernels, which enable effective feature representation while maintaining low computational complexity.
3. Extensive evaluations on several datasets demonstrate that the proposed model delivers superior performance with respect to detection accuracy and computational efficiency, enabling effective application for UAV-based maritime object detection in low-visibility environments.

## 2. Related Work

### *2.1. YOLO-Driven Detection Approaches*

The YOLO series of algorithms has made a significant contribution to object detection by reformulating the detection workflow as a unified regression framework, thereby simplifying the traditional multi-stage detection pipeline. Through continuous architectural refinement and training strategy optimization across successive versions, the YOLO family has continuously incorporated advanced techniques and optimization strategies, leading to substantial improvements in detection performance. YOLOv2 introduces multi-scale training strategies and anchor boxes, which enhance its capability to detect objects across varying scales. YOLOv3 employs a deeper backbone, Darknet-53, along with multi-scale feature prediction, which enhances detection accuracy while maintaining real-time inference speed. Subsequent versions, including YOLOX, YOLOv7, and YOLOv9, have further improved robustness in complex scenarios and detection accuracy through innovations such as anchor-free detection, Bag-of-Freebies training strategies, and programmable gradient information mechanisms. Owing to its real-time inference capability and strong multi-scale representation, YOLO has attracted increasing attention in maritime object detection, where targets often exhibit large scale variations and complex background interference. As a result, a growing body of research has explored the application of YOLO-based models to maritime detection tasks. Huang et al. [18] applied guided filtering and grayscale enhancement for image preprocessing and simplified certain convolutional operations in YOLOv3 to mitigate feature redundancy, thereby achieving higher ship detection performance. Xue et al. [19] optimized the YOLOv8 backbone by refining the activation functions and incorporating additional pooling layers, and employed a sliding loss function to address class imbalance in ship detection, thereby improving detection accuracy across multiple ship categories. To improve performance on small maritime objects, Wang et al. [20] introduced the ConvNeXt module into the neck feature fusion stage of YOLOv11 and employed WIoU as the bounding box loss to alleviate the adverse effects caused by sparse pixels in small objects. Building upon these research advances, we utilize YOLOv11 as the baseline detection model, as it integrates the strengths of previous YOLO variants while further enhancing detection accuracy and robustness. Moreover, its flexible architecture facilitate targeted optimization in feature fusion and scale sensitivity.

### *2.2. Attention-Driven Feature Refinement*

The attention mechanism (AM) draws inspiration from the human visual system and aims to selectively emphasize informative regions while highlighting features that are most relevant to the target [21]. By adaptively reweighting feature representations, attention mechanisms can effectively suppress background interference and enhance discriminative target-related features. For object detection tasks, integrating attention mechanisms allows the model to dynamically attend to key target regions across different feature levels, thus enhancing detection performance. Sun et al. [22] introduced a hybrid attention module within the feature fusion stage, facilitating interaction between features at different scales, thereby enhancing detection accuracy for objects with significant size variations. Haruna et al. [23] designed a SaRPFF module, which integrates multi-head self-attention with a register-based pyramid mechanism to improve multi-scale object representation. Bu et al. [24]

 doi:10.20944/preprints202603.2492.v1



proposed OD-YOLO, which incorporates a hybrid attention transformer and a dynamic head, aimed at improving robustness to geometric variations. Tian et al. [25] designed a hybrid cross-feature interaction attention module to strengthen channel and spatial interactions within feature maps. Shen et al. [26] incorporated the CBAM into the backbone of YOLOX to adaptively recalibrate feature responses along both spatial and channel dimensions, enabling the network to more effectively highlight salient target-related features. Xiong et al. [27] designed a channel-shuffling spatial attention module based on subspace attention, which captures subspace weight discrepancies and promotes inter-channel information interaction through channel shuffling, effectively improving small-object recognition while suppressing background noise effects. Motivated by these studies, we introduce attention mechanisms at multiple key stages of the network, aiming to jointly enhance feature discriminability and suppress background noise, thereby improving the detection of small and scale-varying targets under low-visibility conditions.

## 3. Methods

### *3.1. The Structure of MSA-YOLO*

To better satisfy the practical demands of maritime search and rescue under low-visibility conditions, MSA-YOLO is designed in a problem-driven manner. In real-world maritime environments, object detection models typically face three major challenges: significant variations in object scale, degradation of discriminative features caused by low-visibility conditions, and strict constraints on computational efficiency for real-time deployment. To address scale variation, the backbone adopts StarNet with multi-scale convolutions, which introduces convolutional kernels of different sizes at various feature extraction stages, thereby enhancing feature representations of objects at different scales. Meanwhile, an efficient channel attention module is introduced to enhance the representation of target-related features by adaptively recalibrating channel responses, thereby mitigating discriminative feature degradation in low-visibility conditions. In the neck, considering the requirement for lightweight computation, depthwise separable convolutions replace standard convolutions to reduce computational cost. Meanwhile, to compensate for detail loss under degraded imaging conditions, high-frequency enhancement residual blocks are combined with the C3k2 module to strengthen detailed information extraction. In the detection head, a SimAM module is embedded into both the classification and regression branches to achieve adaptive feature optimization, which further refines detection performance. Figure 2 illustrates the overall architecture of MSA-YOLO.

### *3.2. Multi-Scale Kernel-Based StarNet Backbone Network*

YOLOv11's backbone employs a cross-stage partial connection structure, in which the feature map is partitioned and processed along two distinct paths, enhancing feature representation. This structure also includes several downsampling layers, enabling improved performance in detecting large-scale objects. While this design improves performance, it unavoidably increases the computational complexity. Moreover, standard convolution has limited capability for nonlinear combination across channels, making it difficult to effectively extract high dimensional semantic information from feature maps in complex scenarios. To tackle the issues, we develop a lightweight network based on StarNet [28]. The core of StarNet is built upon the star-shaped computation, which interactively maps features in a low-dimensional space to implicitly expand feature dimensions, thereby enhancing feature representation without additional computational cost, while multi-path interactions further improve information fusion across channels. The star-shaped computation is formulated as:

$$w_1^T x \otimes w_2^T x = \sum_{j=1}^{d+1} \sum_{i=1}^{d+1} w_1^i w_2^j x^i x^j \quad (1)$$

where $\otimes$ denotes element-wise multiplication, w denotes weight matrix.

 doi:10.20944/preprints202603.2492.v1



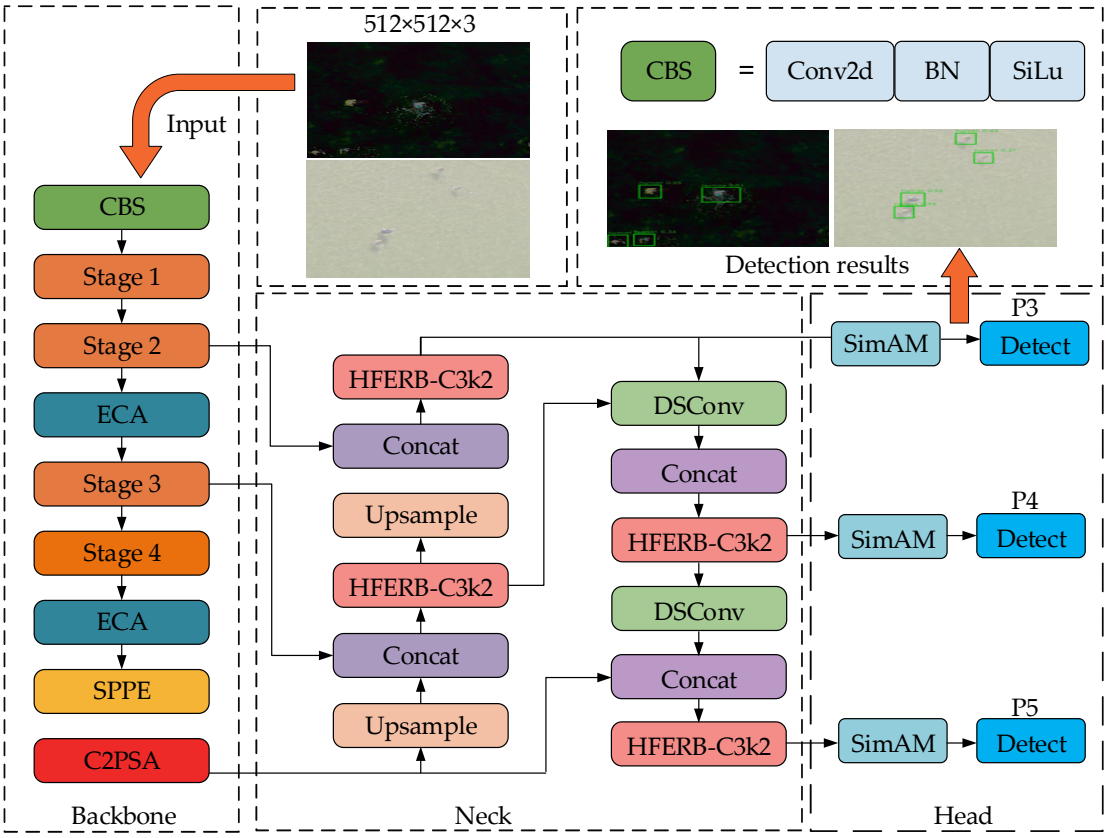



**Figure 2.** The overall structure of Multi-Scale Adaptive YOLO (MSA-YOLO); ECA represents Efficient Channel Attention; SPPE indicates Spatial Pyramid Pooling Enhancement; C2PSA denotes a Cross-Stage Partial Spatial Attention module; and DSConv denotes Depthwise Separable Convolution.

Furthermore, to address the issue of weak object feature responses in low-visibility environments, we introduced an efficient channel attention (ECA) module [29] after Stage 2 and Stage 4 of StarNet. ECA computes the inter-channel dependencies and automatically assigns a dynamic weight to each channel, thereby enhancing object features while suppressing background interference. The structure of ECA is shown in Figure 3.

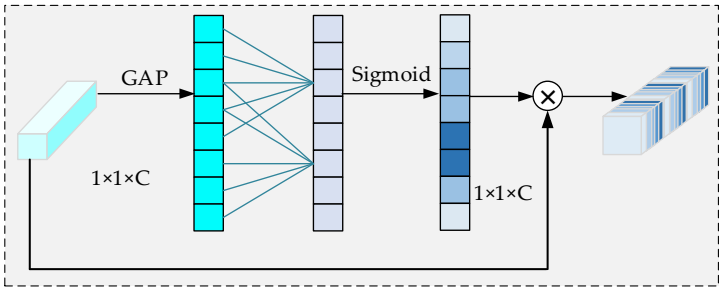



**Figure 3.** The overall structure of efficient channel attention.

Specifically, given an input feature $F_1$, it undergoes global average pooling and one-dimensional convolution of size 3, after which the sigmoid function is applied to produce the attention weights. These weights are multiplied with initial feature $F_1$, resulting in the channel-enhanced feature representation $F_1^{'}$. Mathematically,

$$F_1^{'} = \sigma(f^{1\times1}(GAP(F_1))) \otimes F_1 \quad (2)$$

where $\mathrm{GAP}(\cdot)$ denotes the global average pooling, and $\sigma(\cdot)$ denotes the sigmoid function.

 doi:10.20944/preprints202603.2492.v1



On the basis of the receptive field theory, it can be observed that the use of smaller convolutional kernels in shallow layers facilitates local feature extraction, while larger kernels in deeper layers effectively broaden the receptive field, enabling the extraction of more comprehensive global features. Therefore, we optimized the convolutional kernel design of StarNet. Differing from the original StarNet, which utilizes only a 7×7 kernel, we utilize 3×3 and 5×5 small-scale kernels in the lower stages (Stage 1 and Stage 2), while utilizing 7×7 and 9×9 large-scale kernels in the higher stages (Stage 3 and Stage 4). Ultimately, we constructed a lightweight StarNet with multi-scale convolutional kernels, as illustrated in Figure 4.

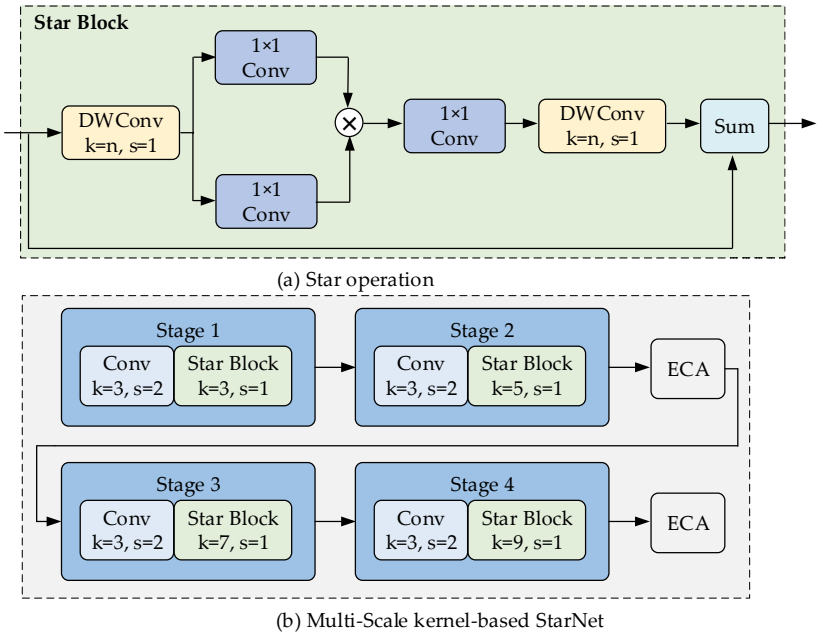



**Figure 4.** The overall structure of multi-scale kernel-based StarNet.

### *3.3. HFERB-C3k2-Based Lightweight Neck Network*

In low-visibility environments, the edges and textures of small-scale objects are easily blurred, which hinders the detection of small-scale objects. Meanwhile, standard convolution layers in the neck structures incur high computational complexity, limiting the efficiency of model deployment. To address these issues, we refined the traditional neck network, with the main improvements focusing on two key aspects. On one hand, depthwise separable convolutions are utilized in place of standard convolutions with the aim of reducing computational cost. On the other hand, we add a new branch based on the C3k2 module, in which the high-frequency enhanced residual block (HFERB) [30] is adopted to strengthen the extraction of detailed information, as illustrated in Figure 5.

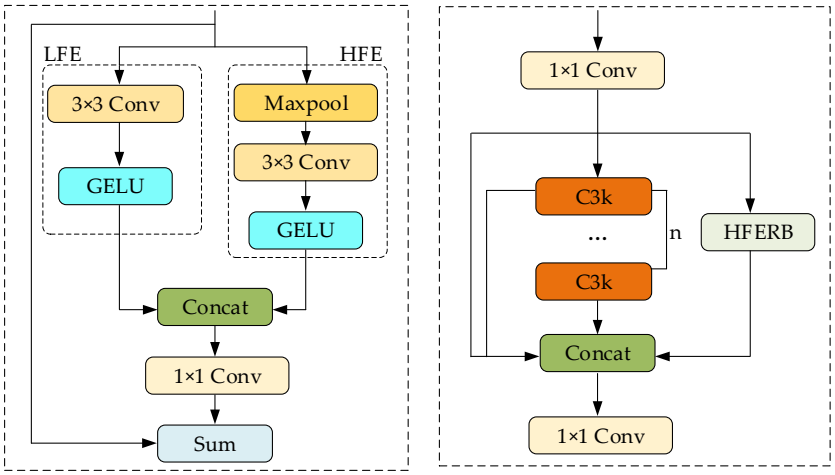
The HFERB consists of two branches, comprising a local feature extraction branch (LFE) and a high-frequency enhancement branch (HFE). Provided an initial feature $F_2$, it is partitioned into two sub-features $F_2^{(t)}$, where $t = 1, 2$, which are fed into the HFE and LFE branches for processing, respectively. In the HFE, $F_2^{(1)}$ is sequentially processed by a max-pooling operation, a 1×1 convolution, and a GELU activation to generate $F_2^{\mathrm{HFE}}$ with enhanced high-frequency responses. Mathematically,

$$F_2^{\mathrm{HFE}} = GELU(f^{1\times1}(Max(F_2^{(1)}))) \tag{3}$$

where $Max(\cdot)$ denotes the max pooling operation, $f^{1\times1}(\cdot)$ denotes the operation using a 1×1 convolution kernel, and GELU(·) denotes the GELU activation.

 doi:10.20944/preprints202603.2492.v1






(a) HFERB (b) HFERB-C3k2

**Figure 5.** Structure of HFERB-C3k2 Module.

In the LFE, $F_2^{(2)}$ is sequentially processed by a 3×3 convolution and a GELU activation to generate $F_2^{\mathrm{LFE}}$. This branch is mainly concerned with relatively smooth regions of the image. Mathematically,

$$F_2^{\mathrm{LFE}} = GELU(f^{3\times3}(F_2^{(2)})) \tag{4}$$

where $f^{3\times3}(\cdot)$ denotes the operation performed with a 3×3 convolution kernel.

Finally, $F_2^{\mathrm{HFE}}$ and $F_2^{\mathrm{LFE}}$ undergo concatenation followed by a 1×1 convolution, subsequent to which the fused result is added to $F_2$. This output is concatenated with the outputs of the C3k2 branches, followed by 1×1 convolution layer for further fusion, producing the output $F_2^{\mathrm{Out}}$. Mathematically,

$$F_2^{\mathrm{Out}} = f^{1\times1}(C(f^{1\times1}(C(F^{\mathrm{HFE}}, F^{\mathrm{LFE}})) + F_2,\ \mathrm{C3k2}(F_2))) \tag{5}$$

where C(·) denotes the concatenation function.

### *3.4. Head Network Based on SimAM*

Traditional detection heads typically adopt a decoupled structure that models classification and regression separately to reduce computational complexity. However, the current structures overlook the semantic and spatial differences between the two tasks, causing the shared features to potentially interfere with each other in each task, thereby reducing detection stability. To tackle this issue, we introduce a Simple, Parameter-Free Attention Module (SimAM) [31] in the classification and regression branches. The SimAM is a non-parametric attention mechanism leveraging the minimization of an energy function, which measures the contribution of each feature through the local statistical contrast between each position in the feature map and its neighborhood, thereby adaptively enhancing informative features and suppressing less relevant ones, effectively alleviating conflicts between features from different tasks. Figure 6 illustrates the structure of SimAM.

The energy function of SimAM is defined as the variance-normalized squared deviation between the local feature and the global mean, which can be mathematically expressed as follows:

$$E_{i,j} = \frac{(X_{i,j} - u)^2}{4(v^2 + \lambda)} + 0.5 \tag{6}$$

where $u$ denotes the spatial mean, $\lambda$ denotes a smoothing term for numerical stability, and $v$ denotes the variance.

 doi:10.20944/preprints202603.2492.v1



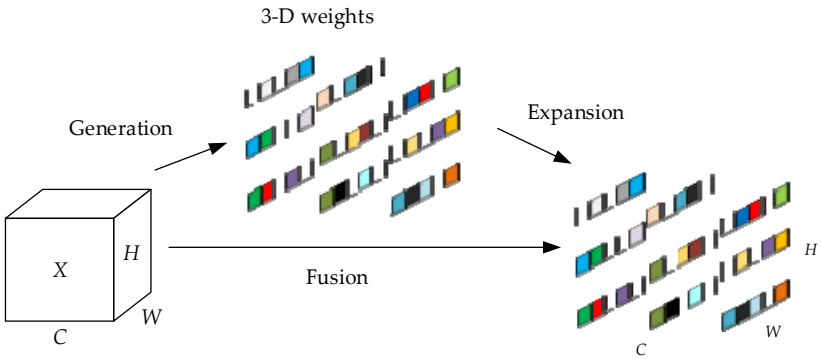



**Figure 6.** Overview of the SimAM architecture.

Following the calculation of the energy function, a sigmoid function is applied to generate attention weights, which are subsequently multiplied by the initial features. Mathematically,

$$X'_{i,j} = X_{i,j} \otimes \sigma\left(\frac{(X_{i,j}-u)^2}{4(v^2+\lambda)}+0.5\right) \tag{7}$$

We place the SimAM after the 3×3 convolution because the preceding convolution has preliminarily modeled the classification and regression features, endowing them with basic task-level discriminability, upon which SimAM is further introduced to reweight the task-relevant features for enhanced representation, as presented in Figure 7.

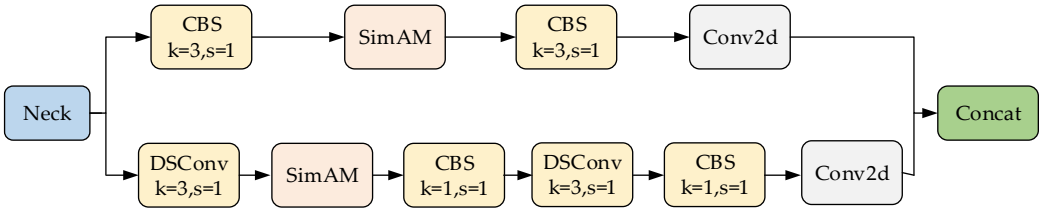



**Figure 7.** The overall structure of the MSA-YOLO detection head.

### *3.5. Joint Loss Function*

In object detection, the loss function is composed of bounding box and classification losses. For the classification loss, we employ Binary Cross-Entropy (BCE) loss to guide model to correctly classify each object category. The BCE loss is formulated as follows:

$$L_{\text{BCE}} = -\sum_{i=0}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \tag{8}$$

where N denotes the total number of samples, $\hat{y}_i$ denotes the predicted value, and $y_i$ denotes the ground truth value.

For the bounding box loss, we adopt a joint loss function that simultaneously considers multiple aspects of object localization, thereby improving the accuracy of predicted bounding boxes. Intersection over Union (IoU) can be expressed as the ratio of the intersection area between the predicted and ground truth bounding boxes to their union area, and is mathematically expressed as IoU=$|a \cap a^{gt}|/|a \cup a^{gt}|$. Considering the variation in shapes and positions of objects in UAV images, we introduce the Complete Intersection over Union (CIoU) loss [32], which incorporates the aspect ratio, the distance between bounding box centers, and the IoU, aiming to achieve improved geometric optimization. The CIoU is formulated as:

$$L_{\text{IoU}} = 1 - \text{IoU} \tag{9}$$

$$L_{\text{CIoU}} = L_{\text{IoU}} + \frac{\rho^2(b,b^{gt})}{(c_w)^2+(c_h)^2} + \frac{4}{\pi^2}\left(\tan^{-1}\frac{w^{gt}}{h^{gt}} - \tan^{-1}\frac{w}{h}\right) \tag{10}$$

 doi:10.20944/preprints202603.2492.v1



where $\rho(b, b^{gt})$ denotes L2 distance between centers of the ground truth and predicted boxes, $h$ and $w$ denote height and width of the predicted box, $w^{gt}$ and $h^{gt}$ denote width and height of the ground truth box, and $c_w$ and $c_h$ denote width and height of minimum enclosing box computed from the predicted and ground-truth boxes

However, maritime objects often exhibit an imbalanced spatial distribution, with dense objects in nearshore areas and sparse objects in offshore regions. Such imbalance may lead to imprecise annotations, which can adversely affect models trained with CIoU-based loss. To mitigate the training bias, we introduce WIoU loss [33], which assigns differentiated weights to samples, alleviating the effect of poor-quality samples during gradient updates. The WIoU is formulated as follows:

$$\beta = \frac{L_{\mathrm{IoU}}^{*}}{\overline{L_{\mathrm{IoU}}}} \tag{11}$$

$$L_{\mathrm{WIoU_{v3}}} = \frac{\beta}{\delta \cdot \alpha^{\beta-\delta}} \cdot \exp\left(\frac{(x_p - x_{gt})^2 + (y_p - y_{gt})^2}{(W_g^2 + H_g^2)}\right) \cdot L_{\mathrm{IoU}} \tag{12}$$

where $x_p$ and $y_{gt}$denote coordinates of the predicted box, $H_g$ and $W_g$ denote height and width of the minimum enclosing rectangle, $y_{gt}$ and $x_{gt}$ denote coordinates of the ground truth box, $\delta$ and $\alpha$ denote tunable parameters for different model configurations, $\overline{L_{\mathrm{IoU}}}$ denotes the average of $L_{\mathrm{IoU}}$, and $L_{\mathrm{IoU}}^{*}$ denotes the monotonic focus coefficient.

Distribution Focal Loss (DFL) [34] mitigates the ambiguity in bounding box localization by modeling multiple possible locations for each coordinate. Therefore, we incorporate DFL to further mitigate the localization errors caused by blurred object boundaries in low-visibility environments. The DFL is formulated as follows:

$$S_i = \frac{d_{i+1} - d}{d_{i+1} - d_i} \tag{13}$$

$$L_{\mathrm{DFL}} = -\left(\left(d_{i+1} - d\right) log\left(S_i\right) + \left(d - d_i\right)\right) log\left(S_{i+1}\right) \tag{14}$$

where $d$ denotes the ground truth position, $d_i$ and $d_{i+1}$ denote right and left positions of the predicted box. The total loss for model training is formulated as follows:

$$L_{\mathrm{total}} = \lambda_1 L_{\mathrm{BCE}} + \lambda_2 L_{\mathrm{CIoU}} + \lambda_3 L_{\mathrm{WIoU_{v3}}} + \lambda_4 L_{\mathrm{DFL}} \tag{15}$$

where $\lambda_1$ $\lambda_2$ $\lambda_3$ and $\lambda_4$ denote the coefficients for BCE, CIoU, WIoU$_{v3}$, and DFL losses, respectively. The overall weight configuration of the joint loss function follows the default design of YOLOv11. Specifically, for the bounding box regression loss, while keeping the total regression loss constant, we performed a grid search within a limited range to adjust the weights of CIoU and WIoU$_{v3}$. Based on performance comparisons on the validation set, the weight ratio of CIoU to WIoU$_{v3}$ was finally determined to be 6:4.

## 4. Experiment

### *4.1. Dataset Construction*

Due to the difficulty in acquiring maritime UAV imagery in low-visibility environments, the availability of real-world samples is limited. To conduct our study, we utilize the existing AFO dataset [35], which was collected using UAVs by Gasienica-Jozkovy et al. The original image resolution ranges from 1280×720 to 3840×2160, and the dataset includes six categories of targets: kayak, surfboard,human, sailboat, buoy, and boat. To meet the experimental requirements, we crop the images to a unified resolution of 512×512, resulting in 4216 samples. Based on these images, we construct low-visibility environments by applying a channel-based fog synthesis algorithm [36] and a nighttime synthesis algorithm [37], thereby generating foggy and low-light scenes for our study.

 doi:10.20944/preprints202603.2492.v1



Finally, the dataset was partitioned into training and testing subsets, with a split of 20% for testing and 80% for training. The sample images are illustrated in Figure 8.

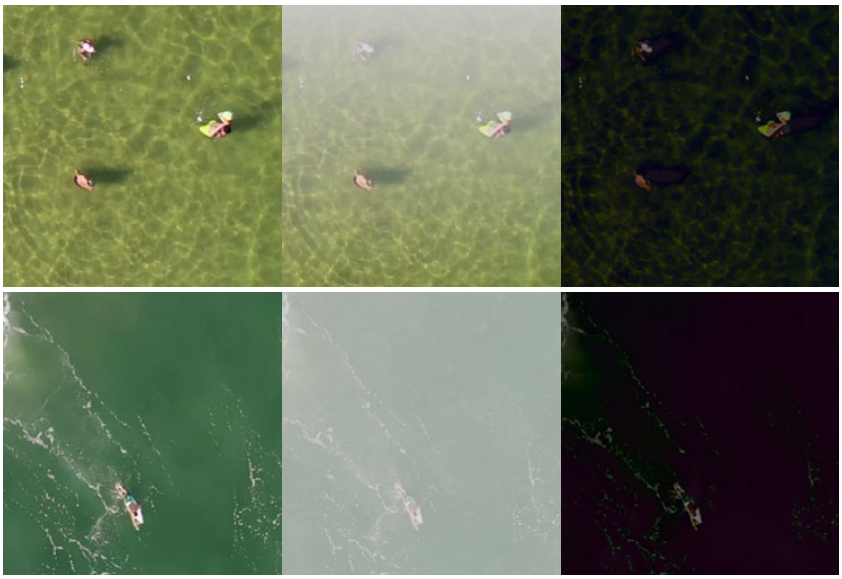



**Figure 8.** Experimental dataset.

*4.2. Metrics for Model Evaluation*

For quantitative evaluation, Recall, Precision, and mean Average Precision (mAP) are adopted as performance metrics. The definitions are formulated below:

Precision is formulated as:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

Recall is formulated as:

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

AP is formulated as:

$$AP = \int_0^1 P(R) \cdot dR \tag{18}$$

mAP is formulated as:

$$mAP = \frac{1}{K} \sum_{i=0}^{K} AP_i \tag{19}$$

where FP denotes the count of background s incorrectly detected as targets, TP denotes the count of targets correctly detected, FN denotes the count of targets incorrectly detected as background, TN denotes the count of background correctly detected, and K denotes the count of object categories.

*4.3. Model Training*

The proposed model is implemented using PyTorch and trained on an NVIDIA RTX 3060 GPU featuring 8 GB of memory. For optimization, SGD is employed with a momentum of 0.94 and a weight decay of 0.0004. The initial learning rate is set to 0.001, and a warm-up procedure is adopted during the first five epochs, in which the learning rate is gradually increased to mitigate excessively high learning rates and promote smoother convergence. Taking into account the available computational resources and the dataset size, the model is trained for 200 epochs with a batch size of 10.

 doi:10.20944/preprints202603.2492.v1



Figure 9 illustrates the progression of the model's mAP@0.5 and losses during training, from which it is evident that both the validation and training losses decline as the number of epochs increases, with a more pronounced decline during the first 50 epochs and convergence is reached around 150 epochs. Overall, the loss curves are stable, and no obvious overfitting is observed. Meanwhile, as the number of epochs increases, the mAP@0.5 metric rises rapidly and gradually stabilizes at a high level after approximately 150 epochs, which demonstrates the effectiveness of the training. Particularly, extending the training beyond the convergence point helps further stabilize performance and reduce metric fluctuations, which motivates the adoption of 200 epochs as a conservative training setting.

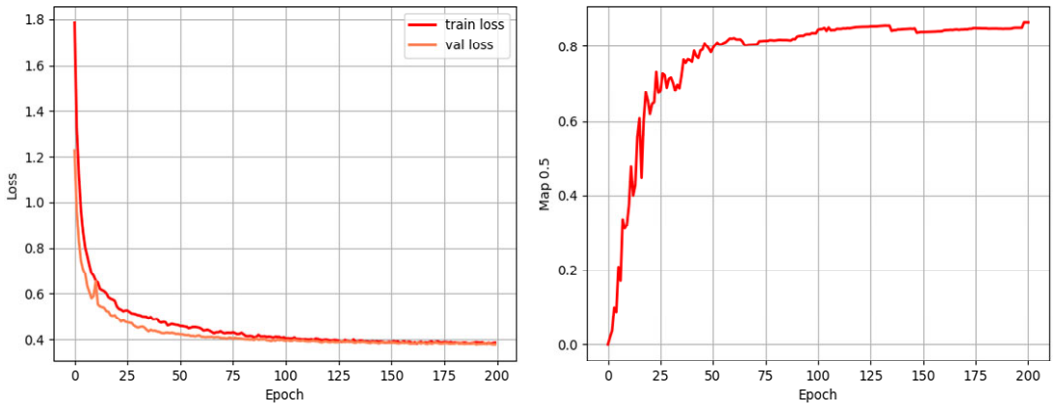



**Figure 9.** Training metrics curves: (a) validation and training loss, (b) mAP@0.5 metrics.

*4.4. Experimental Result*

To evaluate the effectiveness of the proposed MSA-YOLO, we conducted comparative experiments against thirteen representative lightweight detection models, including EfficientDet [38], NanoDet, RetinaNet [39], PicoDet, YOLOv4-tiny [40], LMSD-YOLO, G-YOLOv11 [41], YOLOv11-RGBT [42], YOLOv5-s, YOLOX-s [43], YOLOv7-tiny [44], YOLOv8-s, and YOLOv11-s. Specifically, NanoDet, EfficientDet and PicoDet are designed for deployment on mobile platforms, while LMSD-YOLO is developed for maritime object detection. Both G-YOLOv11 and YOLOv11-RGBT are advanced algorithms based on YOLOv11. For YOLOv11-RGBT, the red channel from the input is duplicated to form four channels, satisfying the input requirements of YOLOv11-RGBT. All models were benchmarked with consistent hyperparameter configurations, and the corresponding results listed in Table 1. From these results, MSA-YOLO achieves optimal performance in terms of Recall, mAP@0.5, and mAP@0.5:0.95, with values of 79.71%, 86.39%, and 52.36%, respectively. This demonstrates that MSA-YOLO detects targets more accurately in low-visibility environments compared with the other models. Additionally, while MSA-YOLO had comparatively low Precision, it achieved the highest mAP@0.5:0.95 and Recall, demonstrating that it had better stability in detecting objects in low-visibility environments. With respect to resource consumption, the number of FLOPs and parameters of MSA-YOLO are approximately 4.04 G and 4.52 M, respectively, which are higher than those of EfficientDet, NanoDet and PicoDet. Meanwhile, MSA-YOLO outperforms EfficientDet, LMSD-YOLO, RetinaNet, YOLOv5-s, YOLOv8-s, YOLOv11-RGBT and YOLOv11-s in terms of inference time, suggesting that it enables faster detection while maintaining lower computational costs. Specifically, although MSA-YOLO has slightly higher numbers of parameters and FLOPs than EfficientDet, NanoDet, and RetinaNet, it achieves faster inference speed compared with EfficientDet, indicating that it remains promising for deployment on mobile devices.

 doi:10.20944/preprints202603.2492.v1



**Table 1.** Performance comparison of various models on AFO images (%).

| Method | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | Params(M) | FLOPs(G) | Inference Time (ms) |
|---|---|---|---|---|---|---|---|
| EfficientDet | 82.78 | 77.35 | 84.38 | 50.38 | 3.88 | 2.42 | 37.43 |
| NanoDet | 82.35 | 66.63 | 83.29 | 45.57 | 3.79 | 3.19 | 20.79 |
| RetinaNet | 79.39 | 77.18 | 83.81 | 49.39 | 19.87 | 39.89 | 53.86 |
| PicoDet | 69.23 | 75.45 | 76.53 | 43.36 | 1.09 | 2.17 | 23.79 |
| LMSD-YOLO | 70.23 | 79.24 | 84.16 | 50.09 | 3.56 | 5.61 | 29.35 |
| G-YOLOv11 | 84.12 | 79.05 | 85.11 | 51.33 | 7.79 | 6.66 | 26.81 |
| YOLOv11-RGBT | 85.66 | 76.78 | 85.18 | 51.12 | 15.99 | 9.71 | 30.89 |
| YOLOv4-tiny | 84.63 | 72.36 | 82.92 | 47.02 | 5.89 | 5.18 | 18.65 |
| YOLOv5-s | 81.78 | 74.64 | 84.25 | 49.71 | 7.08 | 5.29 | 33.23 |
| YOLOX-s | 85.29 | 75.82 | 82.94 | 48.98 | 8.94 | 8.56 | 25.28 |
| YOLOv7-tiny | 82.28 | 75.79 | 83.21 | 50.19 | 6.03 | 6.61 | 26.82 |
| YOLOv8-s | 82.19 | 78.81 | 85.27 | 50.36 | 11.14 | 14.33 | 31.05 |
| YOLOv11-s | 84.79 | 78.49 | 85.06 | 51.25 | 9.43 | 6.89 | 29.02 |
| MSA-YOLO | 84.94 | 79.71 | 86.39 | 52.36 | 4.52 | 4.04 | 28.27 |

Figure 10 illustrates the detection results of four representative test images selected out of the AFO dataset, where red boxes represent false positives and blue boxes represent false negatives. The following analysis is qualitative in nature and is based on visual inspection of the detection results. The six models exhibit varying degrees of errors in their detection results. Specifically, the results of EfficientDet, YOLOv7-tiny, and YOLOv8-s contain relatively more red boxes, suggesting that these models incorrectly detect many non-target regions as target regions. The errors of EfficientDet and YOLOv7-tiny may stem from insufficient semantic representation in high-level features, while the low-level features contain background noise. During feature fusion, the background noise is difficult to suppress effectively, resulting in wave patterns on the sea surface being incorrectly detected as targets. On the other hand, the errors of YOLOv8-s stem from the limitation of the regression branch in accurately fitting the object boundaries, leading to detection errors. Meanwhile, the results of YOLOX-s contain relatively more blue boxes, indicating that this method incorrectly detects many target regions as non-target regions. The errors of YOLOX-s may stem from its limited ability to model discriminative features of small targets under low-contrast interference, which leads to missed detections. Despite some errors in our results, likely due to the limited sensitivity to minor variations in input, such as subtle pixel shifts or noise, the total count of red and blue boxes is relatively small. Additionally, through enhancing the global feature representation and extraction of detailed information, our method achieves more accurate detection results in all foggy and low-light scenes of the AFO dataset, demonstrating that it outperforms the other models on the AFO dataset.

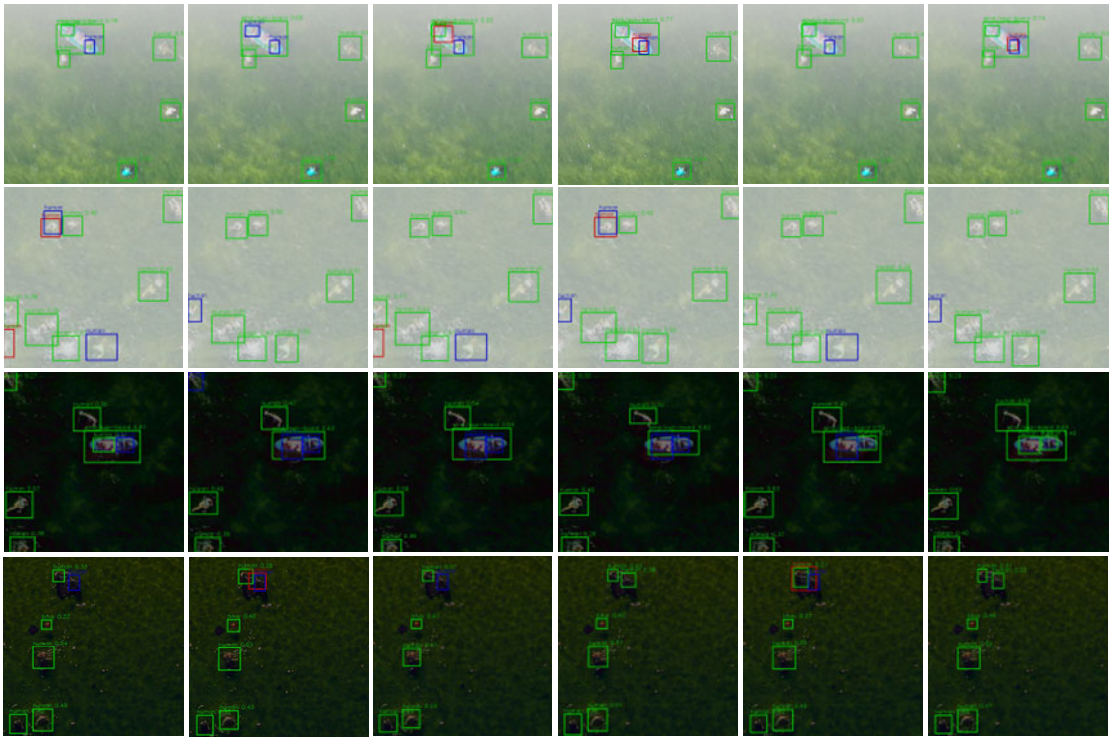
### *4.5. Ablation Experiment*

To evaluate the impact of each module on model performance, a set of ablation studies was conducted on the AFO dataset. In these studies, the modules were added sequentially, including the multi-scale kernel-based StarNet network (MSK-StarNet), the HFERB-C3k2 neck network, and the SimAM module. Table 2 presents the results of different module combinations. Model 1 adopts

 doi:10.20944/preprints202603.2492.v1



StarNet as the backbone network, which achieves performance in terms of Precision, Recall, mAP@0.5, and mAP@0.5:0.95, with values of 89.45%, 75.63%, 84.21%, and 49.33%, respectively, and in terms of FLOPs and parameters, with values of 3.98 G and 4.43 M, respectively. Model 2 utilizes the MSK-StarNet as the backbone network, from which it can be observed that the Recall, mAP@0.5:0.95, and mAP@0.5 outperform those of Model 1, with a marked improvement of 3.22% in terms of Recall. This indicates that exploiting features from different receptive fields contributes to the recognition of objects of various sizes, thereby reducing missed detections. This capability is particularly beneficial in low-visibility conditions, where small or distant objects are more likely to be obscured. Additionally, compared with Model 2, Model 3 shows improvements of 1.86%, 0.82%, and 3.09% with respect to mAP@0.5:0.95, mAP@0.5, and Precision, respectively, indicating that the introduced HFERB-C3k2 enhances the extraction of detailed information, allowing the network to better distinguish objects from complex backgrounds, which improves detection accuracy under low-visibility conditions. Furthermore, Model 4 introduces the SimAM module, achieving increases of 1.15%, 0.25%, and 0.45% with respect to Recall, mAP@0.5:0.95, and mAP@0.5, respectively, relative to Model 3, while retaining identical computational complexity, which highlights its effectiveness in optimizing feature representation.



(a) (b) (c) (d) (e) (f)

**Figure 10.** Visualization of detection results obtained by different models on the AFO images: (a) EfficientDet, (b) YOLOX-s, (c) YOLOv7-tiny, (d) YOLOv8-s, (e) YOLOv11-s, and (f) MSA-YOLO (ours).

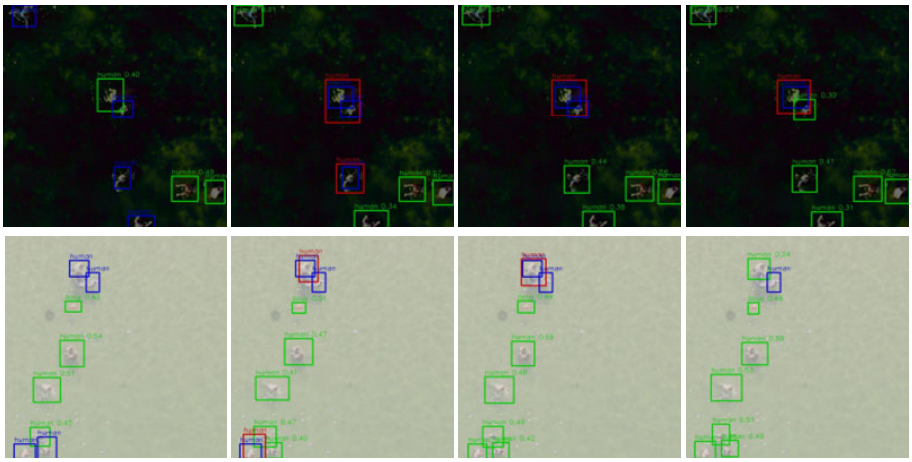
**Table 2.** Performance metrics for different combination strategies (%).

| Method | MSK-StarNet | DSC | HFERB-C3k2 | SimAM | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | Params (M) | FLOPs (G) |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | × | × | × | × | 89.45 | 75.63 | 84.21 | 49.33 | 4.43 | 3.98 |
| Model 2 | ✓ | × | × | × | 82.32 | 78.85 | 85.12 | 50.25 | 4.61 | 3.72 |
| Model 3 | ✓ | ✓ | ✓ | × | 85.41 | 78.56 | 85.94 | 52.11 | 4.52 | 4.04 |
| Model 4 | ✓ | ✓ | ✓ | ✓ | 84.94 | 79.71 | 86.39 | 52.36 | 4.52 | 4.04 |

 doi:10.20944/preprints202603.2492.v1



Detection results for different combination methods are illustrated in Figure 11. The following analysis is qualitative in nature and is based on visual inspection of the detection results. It can be noted that with sequential addition of each module, both the false negative and false positive rates declined progressively, which is reflected in the changes of the blue and red box counts. Specifically, compared with Model 1, Model 2 contains fewer blue boxes and more red boxes, indicating that the introduced MSK-StarNet effectively enhances the detection capability for targets at different scales. However, relying solely on the MSK-StarNet makes it difficult to accurately constrain the target regions while detecting more targets, causing an increase in the false positive rate. Additionally, after introducing HFERB-C3k2, the results of Model 3 contain fewer red boxes than those of Model 2, indicating that HFERB-C3k2 yields a beneficial effect, especially in lowering the false positive rate. This is because HFERB-C3k2 enhances detailed information extraction, allowing the network to better distinguish objects from complex backgrounds, such as noise or pseudo-structures in fog and low-light conditions. These results further confirm that the two modules function cooperatively rather than independently. Furthermore, after introducing SimAM into the head structure, the results of Model 4 contain fewer blue and red boxes than those of Model 3, indicating that SimAM optimizes feature representation by emphasizing salient regions and suppressing irrelevant background, which enhances robustness under low-visibility conditions. Therefore, each module is effective and contributes positively to detection performance to varying degrees.



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

**Figure 11.** Visualization of detection results for different combination strategies: (a) Model 1, (b) Model 2, (c) Model 3, (d) Model 4.

*4.6. Loss Function Analysis*

To investigate the suitability of CIoU-WIoU$_{v3}$ loss in MSA-YOLO, comparative experiments were performed in which the CIoU-WIoU$_{v3}$ loss was substituted with either EIoU [45] or WIoU$_{v3}$ loss during training. The EIoU loss builds upon the CIoU loss by removing the aspect ratio penalty, replacing it with separate constraints on the width and height differences between the ground-truth and predicted boxes, thereby more directly optimizing the geometric alignment of bounding boxes. Figure 12 illustrates the results of experiments, from which it can be observed that the CIoU-WIoU$_{v3}$ outperforms the other two loss functions with respect to mAP@0.5:0.95, mAP@0.5, and Recall. When processing objects with extreme aspect ratios, the linear-squared penalty in the EIoU may amplify the width and height differences, leading to excessive modification of the predicted box shape. Additionally, the WIoU$_{v3}$ loss reduces the influence of low-quality samples on gradients through dynamic weight adjustment, which improves training stability. However, since it does not explicitly incorporate geometric constraints such as center distance or width and height differences, its ability to optimize the geometric accuracy of bounding boxes is relatively limited. In comparison, the CIoU-WIoU$_{v3}$ loss merges the strengths of the CIoU and WIoU$_{v3}$ losses, where the CIoU enhances

 doi:10.20944/preprints202603.2492.v1



constraints on object geometry while the $WIoU_{v3}$ ensures better training stability, indicating that it is better suited for MSA-YOLO with regard to detection accuracy and stability.

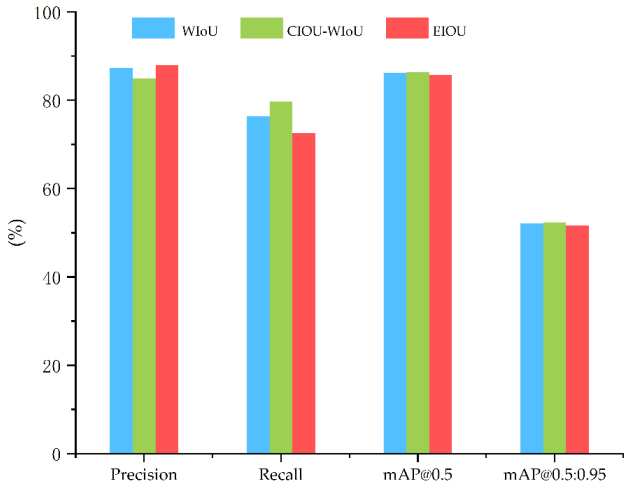



**Figure 12.** Performance metrics of different loss functions on the AFO dataset.

*4.7. Impact of Different Backbone Networks*

To evaluate whether MSK-StarNet exhibits performance advantages compared to other architectures, we compared it with two typical lightweight networks—MobileNetV2 [46] and GhostNet [47]. Table 3 presents the results, from which it can be seen that GhostNet outperforms MobileNetV2 in terms of mAP@0.5, mAP@0.5:0.95, and Precision. This may be ascribed to the fact that Ghost convolution facilitates more comprehensive inter-channel information interaction than depthwise separable convolution, thereby enhancing feature representation. In addition, MSK-StarNet outperforms GhostNet and MobileNetV2 with respect to mAP@0.5, mAP@0.5:0.95 and Recall, while having the fewest parameters and FLOPs, indicating that it can extract richer features at a low computational cost, thereby achieving more efficient detection.

**Table 3.** Comparison of results across different backbones (%).

| Method | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | Params(M) | FLOPs(G) |
|---|---|---|---|---|---|---|
| MobileNetv2 | 83.59 | 74.61 | 81.81 | 48.85 | 5.41 | 4.29 |
| GhostNet | 86.88 | 70.21 | 84.47 | 49.07 | 6.08 | 3.81 |
| Ours | 82.32 | 78.85 | 85.12 | 50.25 | 4.61 | 3.72 |

*4.8. Impact of Attention Mechanisms*

To evaluate the suitability of the ECA module in MSA-YOLO, we further introduce two representative attention mechanisms for comparison, including Split-Attention (SA) and Coordinate Attention (CA). Table 4 presents the comparison results, from which it can be seen that SA outperforms CA in terms of Precision, mAP@0.5, and mAP@0.5:0.95, which may be attributed to its branch-wise competition mechanism along the channel dimension, enabling the model to better emphasize discriminative features. In addition, compared with SA and CA, ECA achieves superior performance in terms of Precision and mAP@0.5:0.95 while maintaining the lowest parameter count, indicating that incorporating ECA into MSA-YOLO achieves a more favorable trade-off between detection accuracy and computational cost.

**Table 4.** Comparison of results across different attention mechanisms (%).

| Method | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 | Params(M) | FLOPs(G) |
|---|---|---|---|---|---|---|
| ECA | 84.94 | 79.71 | 86.39 | 52.36 | 4.52 | 4.04 |

 doi:10.20944/preprints202603.2492.v1



| | | | | | | |
|---|---|---|---|---|---|---|
| SA | 83.41 | 79.86 | 87.13 | 52.12 | 4.87 | 4.09 |
| CA | 77.03 | 81.35 | 85.68 | 51.79 | 4.54 | 4.03 |

*4.9. Impact of Joint Loss Weight Configuration*

To further optimize the performance of the joint loss function, we compared different weight ratios of CIoU and $WIoU_{v3}$ while keeping the overall loss scale unchanged, specifically 2:8, 6:4, and 9:1. Table 5 shows the comparison results, from which it can be seen that the 6:4 configuration outperforms the other two combinations in terms of Precision and Recall. Additionally, although the mAP@50 is slightly lower, the mAP@0.5:0.05 reaches the highest value, indicating that this weight configuration provides better stability across different thresholds.

**Table 5.** Comparison of results under different weight ratios (%).

| CIoU: $WIoU_{v3}$ | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 |
|---|---|---|---|---|
| 2:8 | 82.85 | 78.56 | 86.06 | 52.11 |
| 6:4 | 84.94 | 79.71 | 86.39 | 52.36 |
| 9:1 | 81.31 | 78.07 | 87.22 | 51.47 |

## 5. Discussion

*5.1. Analysis of Feature Visualization*

To visualize MSA-YOLO's attention distribution in images, we employ the Grad-CAM method to generate attention heatmaps. Specifically, Grad-CAM is applied to the last convolutional layer of the detection head, which contains rich high-level semantic information relevant to object localization. The generated heatmaps are normalized to the range [0, 1] and upsampled to the input image resolution for visualization. In these heatmaps, blue pixels represent regions of low attention, while red pixels indicate regions of high attention. Figure 13 shows the heatmaps of YOLOv11-s and MSA-YOLO on the same test samples, from which it can be observed that, compared with YOLOv11-s, the heatmaps generated by MSA-YOLO better align with the true object regions and exhibit less attention to noise. In addition, under various conditions, the heatmaps produced by MSA-YOLO consistently outperform those of YOLOv11-s, further demonstrating the superior detection performance of MSA-YOLO.

*5.2. Analysis of Model Performance Under Different Degradation Levels*

To further analyze the model's performance under different synthesis parameter settings, we employ the Peak Signal-to-Noise Ratio PSNR (PSNR) and the Structural Similarity Index Measure SSIM (SSIM) to quantitatively characterize the degradation degree of the synthesized images. Specifically, a PSNR threshold of 30 dB is used to categorize low-visibility scenes into light and severe degradation levels. Figure 14 presents a qualitative comparison among YOLOv-11s, G-YOLOv11, and MSA-YOLO. It can be observed that under low-light conditions, the texture and edge information of objects are significantly degraded, causing YOLOv11s to miss some objects, while G-YOLOv11 exhibits insufficient localization accuracy for small-scale objects. In comparison, MSA-YOLO achieves comprehensive and precise detection of objects. In addition, in dense regions under foggy conditions, MSA-YOLO demonstrates more accurate object localization, with its predicted boxes better matching the object positions compared to YOLOv11-s and G-YOLO. Overall, MSA-YOLO exhibits stronger robustness in complex degraded environments, indicating its ability to achieve stable object detection in low-visibility maritime scenarios.

 doi:10.20944/preprints202603.2492.v1



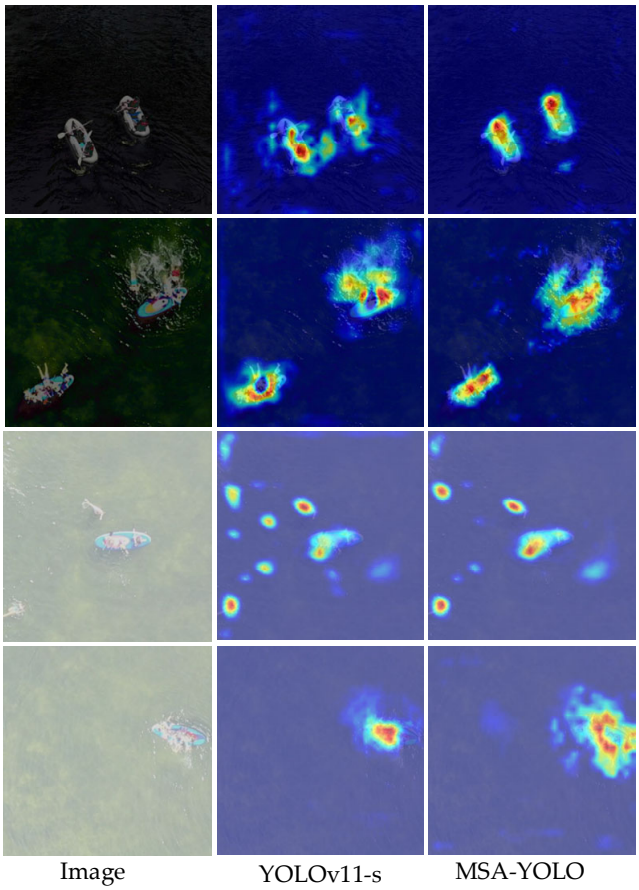



**Figure 13.** Comparison of heatmaps.

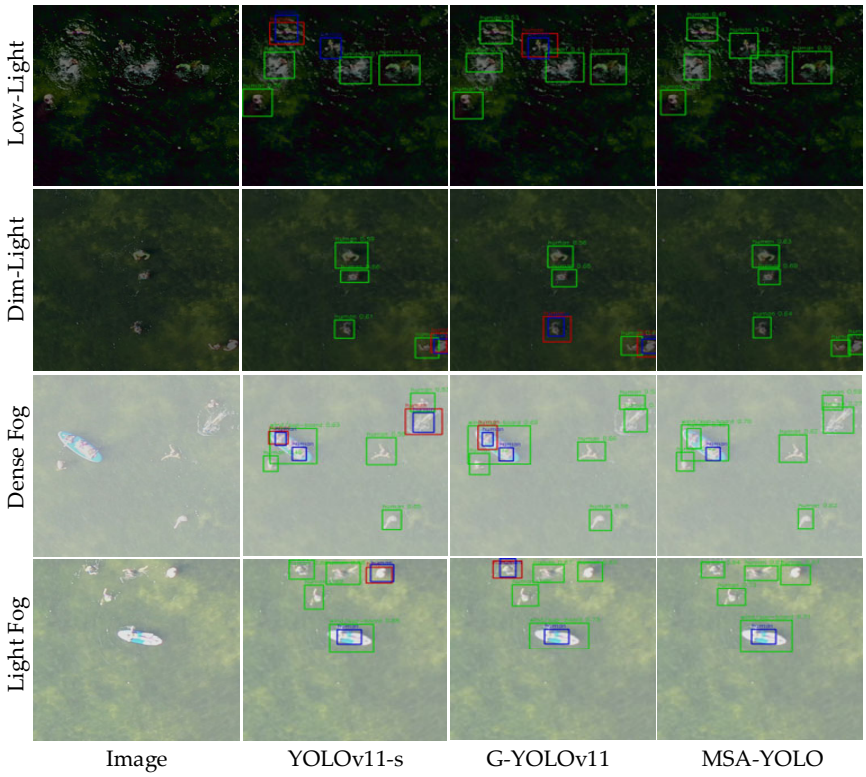



**Figure 14.** Qualitative comparison of YOLOv11-s, G-YOLOv11 and MSA-YOLO under different degradation conditions.

 doi:10.20944/preprints202603.2492.v1



*5.3. Applicability of the Method*

5.3.1. Zhoushan Island

To assess the generalization performance of MSA-YOLO, we constructed a validation dataset using GF-2 satellite imagery from Zhoushan Island area. Two main considerations motivate this choice. On the one hand, Zhoushan Island, as one of the major maritime transport corridors in China, features dense vessel traffic and complex routes, making it representative of typical maritime monitoring scenarios. On the other hand, there are certain differences between GF-2 satellite images and UAV images, which are manifested in lower spatial resolution and vulnerability to atmospheric scattering factors. These differences help us evaluate the generalization of models more comprehensively. To facilitate cross-regional knowledge transfer, we employed transfer learning [48] to transfer the knowledge acquired from the AFO to Zhoushan Island. In particular, we froze the backbone weights and fine-tuned the neck and detection head for 60 epochs with a batch size of 4, using an SGD optimizer configured with a learning rate of 0.0004, momentum of 0.8, and weight decay of 0.0002. The backbone was initialized with pretrained weights, while the neck and detection head were randomly initialized and updated during training. Subsequently, the trained models were applied to evaluate performance on the Zhoushan Island dataset.

Zhoushan Island dataset consists of 337 images from 2016 to 2019. Similarly, we generated foggy and low-light scene images using a channel-based fog synthesis algorithm and a nighttime synthesis algorithm, respectively, with example images shown in Figure 15. The synthesized images were further split into training and validation sets with a ratio of 7:3.

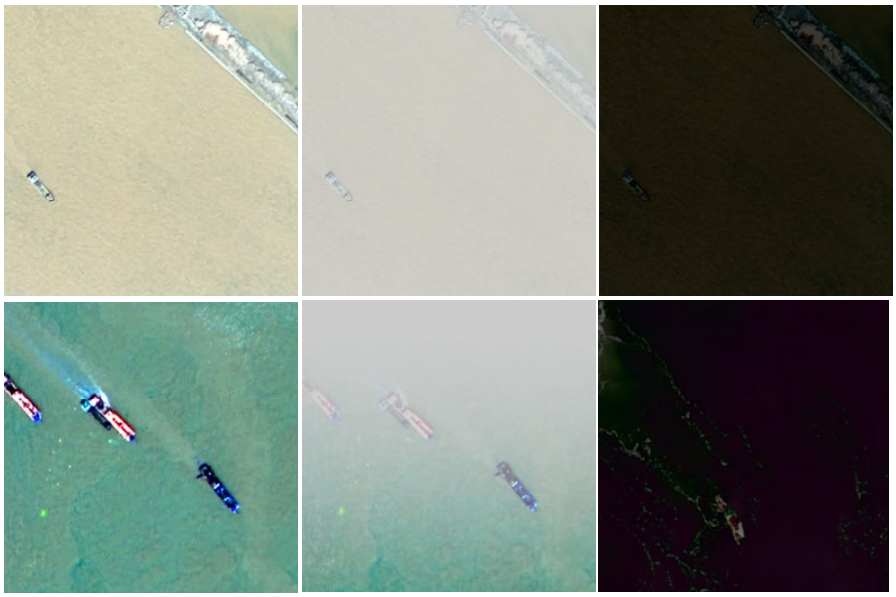



**Figure 15.** Zhoushan Island.

Figure 16 illustrates the detection results of four representative test images. The following analysis is qualitative in nature and is based on visual inspection of the detection results. It can be seen that six models show varying degrees of errors in their results. Specifically, EfficientDet and YOLOv7-tiny contain relatively more red boxes, while YOLOX-s contains relatively more blue boxes, indicating that these three models have difficulty in transferring the learned knowledge to Zhoushan Island. Additionally, although our model exhibits degraded performance when ships are arranged side by side—owing to the close spatial proximity and overlapping boundaries between adjacent ships, which reduce inter-object separability and increase boundary ambiguity, thereby impairing localization accuracy—the detection results contain fewer red and blue boxes. This indicates that our

 doi:10.20944/preprints202603.2492.v1



model can better adapt to the Zhoushan Island scenario compared with other models, demonstrating superior generalization and applicability.

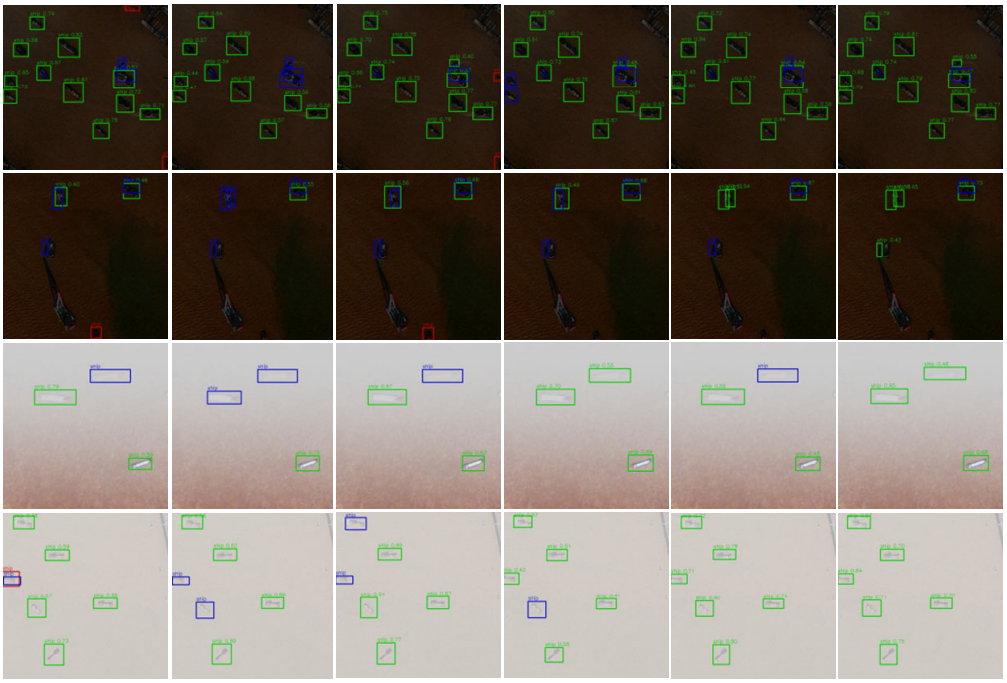



**Figure 16.** Visualization of detection results for various models on Zhoushan Island images: (a) EfficientDet, (b) YOLOX-s, (c) YOLOv7-tiny, (d) YOLOv8-s, (e) YOLOv11-s, and (f) MSA-YOLO (ours).

### 5.3.2. Shandong Province

To further evaluate the performance of MSA-YOLO in real-world scenarios, we constructed a validation dataset using UAV imagery captured under foggy and low-light conditions in Shandong Province. UAV flights were conducted at altitudes ranging from approximately 100 to 600 m, depending on the scenario, resulting in a spatial resolution of roughly 2–35 cm/pixel. The dataset comprises 256 images collected in October 2025, as illustrated in Figure 17. These images were subsequently divided into training and validation sets at a 7:3 ratio. Following the same transfer learning strategy, the trained models were applied to this dataset to assess their detection performance.


**Figure 17.** Shandong Province.

Figure 18 illustrates the detection results of four representative test images. The following analysis is qualitative in nature and is based on visual inspection of the detection results. It can be seen that six models show varying degrees of errors in their results. Specifically, YOLOv7-tiny contains relatively more red boxes, while YOLOX-s contains relatively more blue boxes, indicating that these two models have difficulty in transferring the learned knowledge to Shandong Province,

 doi:10.20944/preprints202603.2492.v1



resulting in poor applicability. Additionally, although our model still contains certain detection errors, the results contain fewer red and blue boxes than other models, demonstrating that it has more stable detection performance in real-world scenarios, which enables better applicability in low-visibility maritime environments.

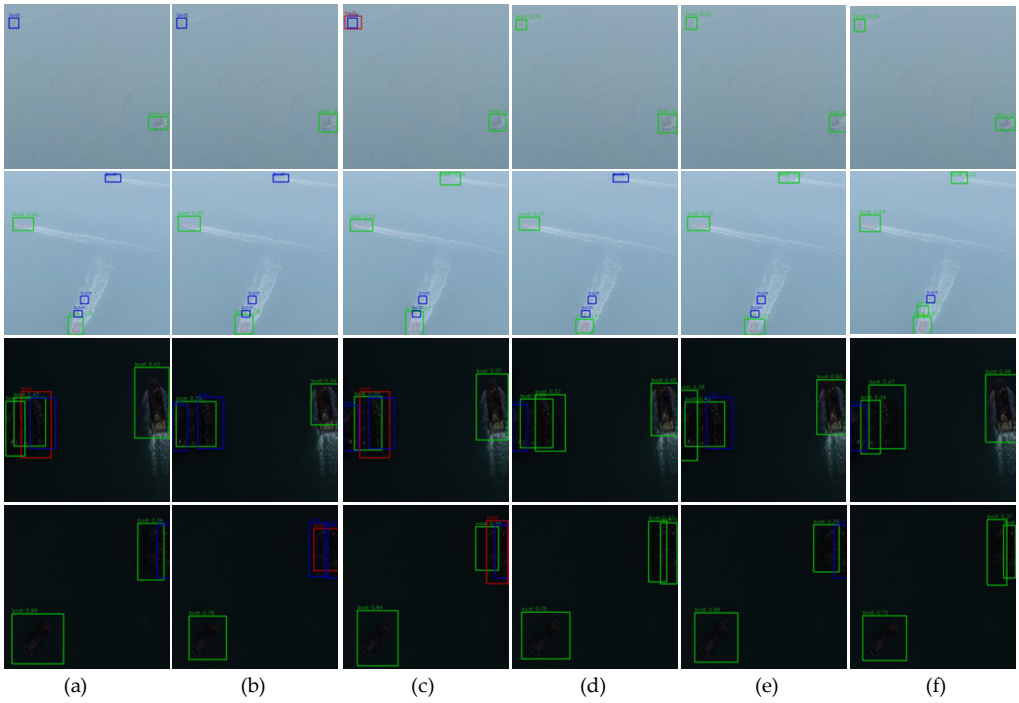



**Figure 18.** Visualization of detection results for various models on Shandong Province images: (a) EfficientDet, (b) YOLOX-s, (c) YOLOv7-tiny, (d) YOLOv8-s, (e) YOLOv11-s, and (f) MSA-YOLO (ours).

## 6. Conclusions

This study explores major challenges of using UAVs for maritime search and rescue in low-visibility environments, including blurred object edges and unclear texture features. These factors make object detection more challenging than in natural-light environments. Therefore, we analyze existing detection approaches and discuss the advantages and disadvantages of different approaches. In particular, current methods often fail to strike a balance between efficiency and accuracy. For example, detection models designed for natural-light conditions perform poorly when directly applied to low-visibility environments, while image enhancement techniques can improve detection accuracy by enhancing image quality but often suffer from real-time processing limitations. To tackle these challenges, we develop a lightweight detection model termed MSA-YOLO, which incorporates several efficient modules to achieve a better trade-off between computational cost and detection accuracy. In the backbone, the StarNet structure, combined with the ECA module and multi-scale convolutional kernels, enhances feature extraction while reducing computational cost. In the neck, high-frequency enhancement residual block branches are introduced into the C3k2 module to improve detailed information extraction, thereby alleviating the negative impact of blurred target edges in low-visibility environments. Meanwhile, standard convolutions are substituted with depthwise separable convolutions to further reduce computational complexity. In the detection head, SimAM is introduced to adaptively optimize features in both the regression and classification branches. Furthermore, a joint loss is applied to constrain the training process, achieving optimal performance. Experiments conducted on the AFO, Zhoushan, and Shandong Province datasets indicate that MSA-YOLO has 4.52 M parameters and 4.04 G FLOPs, demonstrating strong adaptability to resources. In terms of detection accuracy, MSA-YOLO outperforms existing models on all three datasets, demonstrating its effectiveness. Subsequent research will investigate cross-

 doi:10.20944/preprints202603.2492.v1



domain generalization in low-visibility object detection, with an emphasis on mitigating domain discrepancies through domain adaptation techniques, so as to ensure stable detection performance under heterogeneous low-visibility conditions.

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

**Funding:** This research was funded by project supported by the Key Program of the National Natural Science Foundation of China (Grant No. 42530710，and 42377453), Hainan Provincial Natural Science Foundation of China(Grant No. 626MS0247), and Natural Science Foundation of Hunan Province (Grant No. 2026JJ80759).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset used for this study is publicly available. The AFO can be downloaded at https://universe.roboflow.com/datasetlabel/afo-y5ti6 (accessed on 1 May 2025).

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

## References

1. Martinez-Esteso, J.P.; Castellanos, F.J.; Calvo-Zaragoza, J.; Gallego, A.J.J.C.S.R. Maritime search and rescue missions with aerial images: A survey. *Computer Science Review* **2025**, *57*, 100736. [CrossRef]
2. Shi, R.; Zhang, L.; Wang, G.; Jia, S.; Zhang, N.; Wang, C.J.R.S. GD-Det: Low-Data Object Detection in Foggy Scenarios for Unmanned Aerial Vehicle Imagery Using Re-Parameterization and Cross-Scale Gather-and-Distribute Mechanisms. *Remote Sensing* **2025**, *17*, 783. [CrossRef]
3. Zhao, C.; Liu, R.W.; Qu, J.; Gao, R.J.E.A.o.A.I. Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons. *Engineering Applications of Artificial Intelligence* **2024**, *128*, 107513. [CrossRef]
4. Singh, A.P. Fast and Lightweight UAV-based Road Image Enhancement Under Multiple Low-Visibility Conditions. 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events. **2023**. [CrossRef]
5. Duan, R.; Wu, B.; Zhou, H.; Zuo, H.; He, Z.; Xiao, C.; Fu, C. E 3-Net: Event-Guided Edge-Enhancement Network for UAV-based Crack Detection. In Proceedings of the 2024 International Conference on Advanced Robotics and Mechatronics (ICARM), 2024; pp. 272-277. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014; pp. 580-587. [CrossRef]
7. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision **2015**, 1440-1448. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; Sun, J.J.A.i.n.i.p.s. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*. [CrossRef]
9. Zhou, J.; Jiang, P.; Zou, A.; Chen, X.; Hu, W.J.J.o.M.S.; Engineering. Ship target detection algorithm based on improved YOLOv5. *Journal of Marine Science Engineering* **2021**, *9*, 908. [CrossRef]
10. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019; pp. 5693-5703. [CrossRef]
11. Kim, S.-H.; Sim, H.; Jung, Y.; Jung, O.-C.; Kim, Y.J.a.p.a. LiM-YOLO: Less is More with Pyramid Level Shift and Normalized Auxiliary Branch for Ship Detection in Optical Remote Sensing Imagery. *arXiv preprint arXiv:09700* **2025**. [CrossRef]

 doi:10.20944/preprints202603.2492.v1




12. Wang, Z.; Li, W.; Fan, Y.; Wang, L.; Yao, G.J.O.E. Ship target detection algorithm in remote sensing images based on improved YOLO11s. *Ocean Engineering* **2026**, *343*, 123204. [CrossRef]
13. Xie, J.; Pan, B.; Xu, X.; Shi, Z.J.I.T.o.G.; Sensing, R. MiSSNet: Memory-inspired semantic segmentation augmentation network for class-incremental learning in remote sensing images. *IEEE Transactions on Geoscience* **2024**, *62*, 1-13. [CrossRef]
14. Cheng, L.; Deng, B.; Yang, Y.; Lyu, J.; Zhao, J.; Zhou, K.; Yang, C.; Wang, L.; Yang, S.; He, Y.J.I.A. Water target recognition method and application for unmanned surface vessels. *IEEE Access* **2021**, *10*, 421-434. [CrossRef]
15. Zhang, K.; Yan, X.; Wang, Y.; Qi, J. Adaptive dehazing yolo for object detection. In Proceedings of the International Conference on Artificial Neural Networks, 2023; pp. 14-27. [CrossRef]
16. Ma, J.; Lin, M.; Zhou, G.; Jia, Z. Joint image restoration for domain adaptive object detection in foggy weather condition. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), 2024; pp. 542-548. [CrossRef]
17. Liu, T.; Zhang, Z.; Lei, Z.; Huo, Y.; Wang, S.; Zhao, J.; Zhang, J.; Jin, X.; Zhang, X. An approach to ship target detection based on combined optimization model of dehazing and detection. *Engineering Applications of Artificial Intelligence* **2024**, *127*, 107332. [CrossRef]
18. Huang, H.; Sun, D.; Wang, R.; Zhu, C.; Liu, B.J.M.P.i.E. Ship target detection based on improved YOLO network. *Mathematical Problems in Engineering* **2020**, *2020*, 6402149. [CrossRef]
19. Xue, Y.; Zhan, L.; Liu, Z.; Bing, X.J.R.S. SAR Ship Target Instance Segmentation Based on SISS-YOLO. *Remote Sensing* **2025**, *17*, 3118. [CrossRef]
20. Wang, Y.; Zeng, W.; Xu, H.; Jiang, Y.; Liu, M.; Xiao, C.; Zhao, K.J.P. Multi-Type Ship Target Detection in Complex Marine Background Based on YOLOv11. *Processes* **2025**, *13*, 249. [CrossRef]
21. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H.J.I.T.o.G.; Sensing, R. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Transactions on Geoscience Remote Sensing* **2021**, *60*, 1-13. [CrossRef]
22. Sun, G.; Wang, S.; Xie, J.J.E. An image object detection model based on mixed attention mechanism optimized YOLOv5. *Electronics* **2023**, *12*, 1515. [CrossRef]
23. Haruna, Y.; Qin, S.; Chukkol, A.H.A.; Bello, I.; Lawan, A.J.M.T.; Applications. SaRPFF: A self-attention with register-based pyramid feature fusion module for enhanced rice leaf disease (RLD) detection. *Multimedia Tools Applications* **2025**, 1-27. [CrossRef]
24. Bu, Y.; Ye, H.; Tie, Z.; Chen, Y.; Zhang, D.J.S. OD-YOLO: Robust small object detection model in remote sensing image with a novel multi-scale feature fusion. *Sensors (Basel)* **2024**, *24*, 3596. [CrossRef]
25. Tian, D.; Han, Y.; Liu, Y.; Li, J.; Zhang, P.; Liu, M.J.R.S. Hybrid cross-feature interaction attention module for object detection in intelligent mobile scenes. *Remote Sensing* **2023**, *15*, 4991. [CrossRef]
26. Shen, C.; Ma, C.; Gao, W.J.S. Multiple attention mechanism enhanced YOLOX for remote sensing object detection. *Sensors (Basel)* **2023**, *23*, 1261. [CrossRef]
27. Xiong, X.; He, M.; Li, T.; Zheng, G.; Xu, W.; Fan, X.; Zhang, Y.J.I.I.o.T.J. Adaptive feature fusion and improved attention mechanism-based small object detection for UAV target tracking. *IEEE Internet of Things Journal* **2024**, *11*, 21239-21249. [CrossRef]
28. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the stars. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024; pp. 5694-5703. [CrossRef]
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp. 11534-11542. [CrossRef]
30. Li, A.; Zhang, L.; Liu, Y.; Zhu, C. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023; pp. 12514-12524. [CrossRef]
31. Yang, L.; Zhang, R.-Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the International conference on machine learning, 2021; pp. 11863-11874. [CrossRef]

 doi:10.20944/preprints202603.2492.v1




32. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W.J.I.t.o.c. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics* **2021**, *52*, 8574-8586. [CrossRef]
33. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: bounding box regression loss with dynamic focusing mechanism. *arXiv preprint arXiv:10051* **2023**. [CrossRef]
34. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems* **2020**, *33*, 21002-21012. [CrossRef]
35. Gasienica-Jozkowy, J.; Knapik, M.; Cyganek, B.J.I.C.-A.E. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integrated Computer-Aided Engineering* **2021**, *28*, 221-235. [CrossRef]
36. Zhang, N.; Zhang, L.; Cheng, Z. Towards simulating foggy and hazy images and evaluating their authenticity. In Proceedings of the International conference on neural information processing, 2017; pp. 405-415. [CrossRef]
37. Tjia, M.; Kim, A.; Wijaya, E.W.; Tefara, H.; Zhu, K.J.a.p.a. Enhancing Robustness of Human Detection Algorithms in Maritime SAR through Augmented Aerial Images to Simulate Weather Conditions. *arXiv preprint arXiv:13766* **2024**. [CrossRef]
38. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp. 10781-10790. [CrossRef]
39. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision* **2017**, 2980-2988. [CrossRef]
40. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M.J.a.p.a. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:10934* **2020**. [CrossRef]
41. Ferdi, A. Lightweight g-yolov11: Advancing efficient fracture detection in pediatric wrist x-rays. *Biomedical Signal Processing Control* **2026**, *113*, 108861. [CrossRef]
42. Wan, D.; Lu, R.; Fang, Y.; Lang, X.; Shu, S.; Chen, J.; Shen, S.; Xu, T.; Ye, Z. Yolov11-rgbt: Towards a comprehensive single-stage multispectral object detection framework. *arXiv preprint arXiv:.14696* **2025**. [CrossRef]
43. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J.J.a.p.a. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:08430* **2021**. [CrossRef]
44. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023; pp. 7464-7475. [CrossRef]
45. Peng, H.; Yu, S.J.I.T.o.I.P. A systematic IOU-related method: Beyond simplified regression for better localization. *IEEE Transactions on Image Processing* **2021**, *30*, 5032-5044. [CrossRef]
46. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018; pp. 4510-4520. [CrossRef]
47. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020; pp. 1580-1589. [CrossRef]
48. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge data engineering* **2009**, *22*, 1345-1359. [CrossRef]