frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 25 June 2025
DOI 10.3389/fmars.2025.1547747

OPEN ACCESS

EDITED BY
Salvatore Antonio Biancardo,
University of Naples Federico II, Italy

REVIEWED BY
Xinqiang Chen,
Shanghai Maritime University, China
Yuanyuan Wang,
Dalian Maritime University, China
Xinjian Wang,
Dalian Maritime University, China
Qiang Luo,
Guangzhou University, China

*CORRESPONDENCE
Jianchuan Yin
yinjianchuan@gdou.edu.cn

RECEIVED 15 January 2025
ACCEPTED 30 May 2025
PUBLISHED 25 June 2025

CITATION
Xu G, Yin J, Zhang J and Wang N (2025) Maritime man-overboard search based on MOB-Detector with modulated deformable convolution and bi-directional feature fusion network.
Front. Mar. Sci. 12:1547747.
doi: 10.3389/fmars.2025.1547747

Maritime man-overboard search based on MOB-Detector with modulated deformable convolution and bi-directional feature fusion network

Guokang Xu[1], Jianchuan Yin[1,2*], Jinfeng Zhang[3] and Nini Wang[4]

[1]Naval Architecture and Shipping College, Guangdong Ocean University, Zhanjiang, China, [2]Guangdong Provincial Key Laboratory of Intelligent Equipment for South China Sea Marine Ranching, Zhanjiang, China, [3]School of Navigation, Wuhan University of Technology, Wuhan, China, [4]College of Mathematics and Computer, Guangdong Ocean University, Zhanjiang, China

Introduction: Maritime transport is vital for global trade and cultural exchange, yet it carries inherent risks, particularly accidents at sea. Drones are increasingly valuable in marine search missions. However, Unmanned Aerial Vehicles (UAV) operating at high altitudes often leave only a small portion of a person overboard visible above the water, posing challenges for traditional detection algorithms.

Methods: To tackle this issue, we present the Man-Overboard Detector (MOB-Detector), an anchor-free detector that enhances the accuracy of man-overboard detection. MOB-Detector utilizes the bi-directional feature fusion network to integrate location and semantic features effectively. Additionally, it employs modulated deformable convolution (MDConv), allowing the model to adapt to various geometric variations of individuals in distress.

Results: Experimental validation shows that the MOB-Detector outperformed its nearest competitor by 8.6% in [Metric 1 AP50] and 5.2% in [Metric 2 AP_{small}], demonstrating its effectiveness for maritime search tasks. Furthermore, we introduce the ManOverboard Benchmark to evaluate algorithms for detecting small objects in maritime environments.

Discussion: In the discussion, the challenge faced by the MOB-Detector in low-visibility environments is discussed, and two future research directions are proposed: optimizing the detector based on the Transformer architecture and developing targeted data augmentation strategies.

KEYWORDS

man-overboard, modulated deformable convolution, bi-directional feature fusion network, anchor-free detector, maritime search and rescue, small object detection

1 Introduction

With 70 percent of the Earth's surface covered by oceans, maritime transport plays a crucial role as a vital link connecting nations, facilitating global trade, fostering economic growth, and promoting cultural exchange (Chen et al., 2022). Despite its significant contributions to international development, the maritime industry faces inherent risks, with maritime accidents posing a persistent threat. In recent years, the frequency of marine incidents has been notable, with 23,814 casualties reported between 2014 and 2022, averaging 2,646 per year (European Maritime Safety Agency (EMSA), 2023). Factors such as adverse weather conditions, collisions, and grounding incidents can lead to accidents at sea, potentially resulting in individuals going overboard (Chen S. et al., 2023). The time lapse between an accident and the initiation of emergency responses may cause those overboard to drift away from the scene amid turbulent waters, emphasizing the urgency and complexity of maritime rescue operations. Prompt and efficient rescue efforts are paramount in safeguarding the lives of ship occupants, mitigating casualties, and addressing the challenges posed by maritime emergencies.

With the rapid development of drone-related technologies, drone search and rescue operations are gradually becoming a prominent player in maritime rescue missions. Drones possess the ability to swiftly locate and identify individuals in distress at sea. Leveraging their outstanding flexibility, portability, and extensive operational capabilities, drones can quickly reach targeted search areas, providing prompt assistance to those in peril, and significantly enhancing rescue efficiency and survival rates. Despite facing challenges in computing power and storage space, optimization and lightweight processing of algorithms for detecting people overboard are gradually overcoming these obstacles, bringing forth more possibilities and prospects for the development of drone search and rescue operations (Bai et al., 2022; Zhu et al., 2023; Zhang et al., 2023). In UAV sea search and rescue, the detection algorithm of people in the water is undoubtedly the key link. Currently, mainstream target detection algorithms are based on deep learning convolution neural networks, which are realized by feature learning from raw input data (Lei et al., 2022; Liu et al., 2024; Wang et al., 2024). These detection algorithms convert images or video frames into more abstract and high-dimensional feature representations. By analyzing and reasoning about these features, the algorithms can determine the target class and locate its position in the image or video frame. Traditional target detection algorithms typically perform well and accurately when dealing with surface targets. However, on the vast sea surface, the person overboard exposes only a small part of their body (head, shoulders, and arms), which makes them occupy fewer pixels in the image, thus increasing the difficulty of extracting effective feature information from the backbone network (Zhang et al., 2021; Li et al., 2021). In addition, the high altitude at which the UAVs fly and the complexity of the marine environment (including issues such as occlusion, blurring, and light reflection from the sea surface) further increase the difficulty of detection.

To address these issues, this study introduces a MOB-Detector based on anchor-free detection networks specifically designed for locating man-overboard, as illustrated in Figure 1. Within the realm of current detection networks, they are typically categorized into

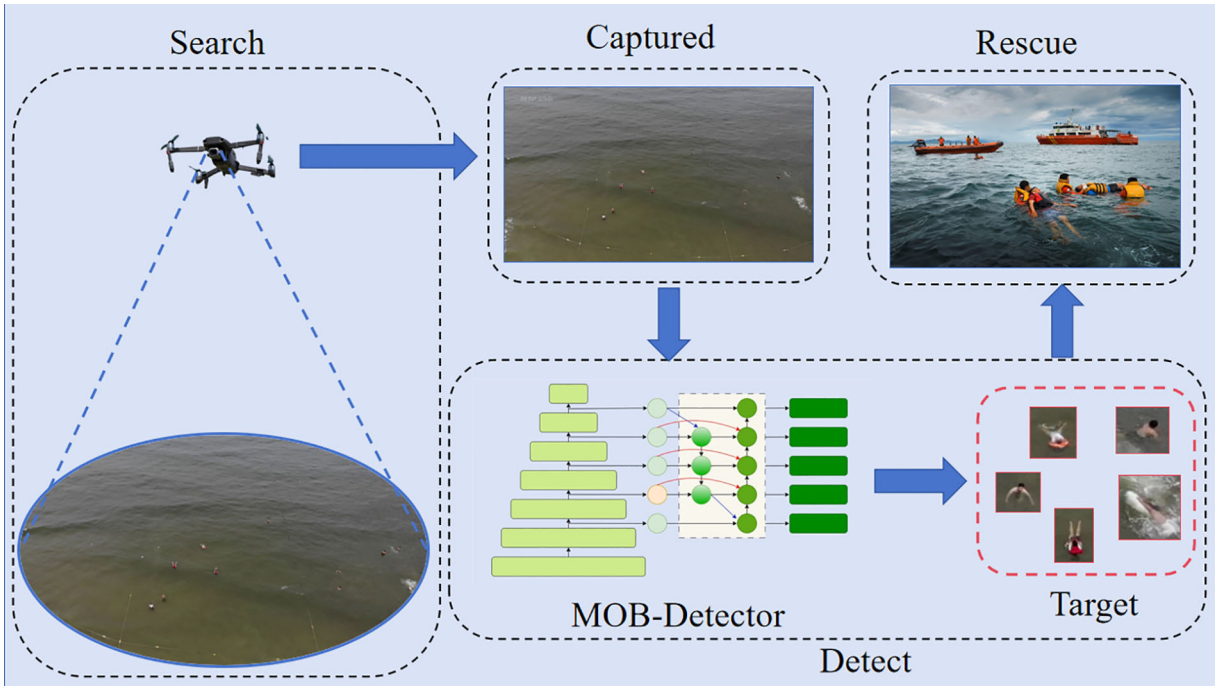

FIGURE 1
Multi-stage maritime rescue process: UAV-based man-overboard target scanning, MOB-detector detection, and dynamic search and rescue mission execution.

two groups: the anchor-frame detection network exemplified by SSD (Liu et al., 2016), Mask R-CNN (He et al., 2017), Faster R-CNN (Ren et al., 2016), etc, and the anchor-free detection network represented by approaches of CenterNet (Zhou et al., 2019), CornerNet (Law and Deng, 2018), RepPoints (Yang et al., 2019), etc. Anchor-Frame detection networks and anchor-free detection networks employ distinct strategies in target detection, differing in their approach to object boundaries, candidate region generation, and model training methodologies. Anchor-Frame detection networks rely on a predefined set of frames with specific proportions and sizes as the foundation for target detection, achieved through regressing the position and class of these anchors. Conversely, most anchor frame-based detection models necessitate manual configuration of various anchor frames of different sizes and aspect ratios, as fixed anchor frames may not adequately cater to objects of varying scales. In comparison, the anchor-free detector predicts the object's position and category directly on the feature map, circumventing intricate anchor frame computations and demonstrating heightened adaptability, particularly advantageous in detecting small and irregularly shaped objects (Tong et al., 2020).

MOB-Detector incorporates a Bi-directional Feature Pyramid Network (BiFPN) (Tan et al., 2020). This novel module enhances the typical feature fusion process (Lin et al., 2017a) by implementing a bi-directional feature transfer structure. By repeatedly integrating top-down, bottom-up, and lateral connection pathways, it enables efficient information exchange among various levels of features, facilitating the comprehensive capture of rich semantic details at higher levels and precise spatial information at lower levels (Chen X. et al., 2023). MOB-Detector employs MDConv (Zhu et al., 2019). MDConv is an improved convolution operation (Dai et al., 2017). Introducing learnable offsets enhances the model's ability to adapt to geometric variations resulting from changes in scale, pose, and viewpoint. This enables the network to more effectively process man-overboard or features of various scales, irregular shapes, and perspectives in marine rescue scenarios. However, deformable convolution may have a receptive field that extends beyond the region of interest, resulting in features influenced by the image content and background. MDConv addresses this limitation by adding additional convolutional layers with offset learning capabilities and incorporating a modulation mechanism, thereby enhancing its ability to focus on relevant regions of interest. To demonstrate the effectiveness and efficiency of the research-proposed anchor-free detector, ablation experiments and comparisons with other state-of-the-art detector algorithms will be performed. In addition, this paper introduces the ManOverboard benchmark, a novel benchmark designed for detecting and recognizing small targets at sea. Based on this benchmark, this paper conducts ablation and comparison experiments. The experimental results show that the anchor-free detector has the capability of detecting a target person overboard. The following are the main contributions of this paper:

1. A detection method called MOB-Detector is proposed, which is based on anchor-free detection networks. The MOB-Detector is designed to significantly strengthen the capability of detecting man-overboard from high-altitude UAV platforms.
2. The MOB-Detector is equipped with a BiFPN module. This design combines top-down and bottom-up pathways, allowing for effective information exchange across various feature levels. This integration enhances the ability to capture rich semantic details at higher levels and precise spatial information at lower levels, thereby facilitating comprehensive object detection.
3. The MOB-Detector utilizes the MDConv, which introduces an adaptive convolution kernel position adjustment mechanism. This allows deformable convolutions to effectively handle complex and unknown geometric transformations, thereby enhancing the model's capacity to learn complex object invariance.
4. This paper first introduces the ManOverboard benchmark. To validate the effectiveness of MOB-Detector, the experiments evaluated its performance based on the ManOverboard benchmark by performing ablation and comparison experiments. These experiments demonstrate the enough capability of MOB-Detector in the field of maritime rescue and search.

The remainder of this paper is arranged as follows: Section 2 introduces the related methods of the model, namely, BiFPN and MDConv. Section 3 presents the architecture of MOB-Detector. Section 4 is the details of the ManOverboard benchmark and experiments. Section 5 is the conclusion.

2 Methods

2.1 Bi-directional feature pyramid network

Most target detection algorithms, such as Region-CNN (R-CNN) (LeCun et al., 1998), Fast R-CNN (Girshick, 2015), etc., rely only on the final feature map output for direct prediction, which has limitations when dealing with objects of different sizes. Multi-scale feature fusion can effectively alleviate the influence of limiting the detection capability of targets at different scales in the network model due to the lack of rich semantic information and precise location information in the final generated features after the input information has gone through multiple convolution layers, pooling layers, and activation functions.

The single-stage detector (SSD) (Liu et al., 2016) is proposed to predict objects of different sizes at six different scales of feature maps. As depicted in Figure 2A, the bottom layer feature retains more spatial information, which is used to detect relatively small targets. In contrast, the top layer feature preserves rich semantic information and is responsible for detecting larger objects. SSD does

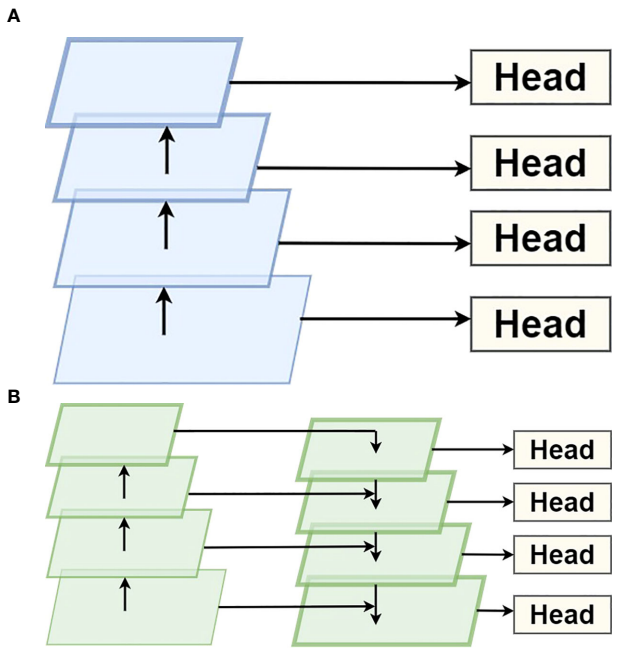

FIGURE 2
(A) Multi-scale feature fusion in SSD: utilizing hierarchical convolutional feature maps from the VGG backbone for object detection across multiple resolutions. **(B)** Multi-scale feature fusion in FPN: a top-down pathway with lateral connections for enhanced object detection across multiple scales.

not reuse the features computed at each layer. In contrast, as shown in Figure 2B, Lin et al. (2017) proposed an FPN structure: a bottom-up line, a top-down line, and lateral connections. Leveraging the distinctive traits of high-level and low-level features, they established connections between high-level features with low resolution and high semantic content and low-level features with high resolution and low semantic information through top-down and bottom-up lateral connections, ensuring that features at all scales are enriched with semantic details.

In contrast to the unidirectional information flow of FPN and path aggregation network (PANet) (Liu et al., 2018), BiFPN employs bi-directional cross-scale connections and weighted feature fusion to efficiently facilitate information exchange between features at different levels (Chen X. et al., 2023). This approach aids in capturing rich semantic details comprehensively and acquiring precise spatial information at lower levels. This involves inputting a sequence containing features from multiple scales $\vec{P}^{in} = (P_{l_1}^{in}, P_{l_2}^{in}, \ldots)$, where $P_{l_i}^{in}$ represented input features at the layer l_i, and then after aggregation transformation $f()$ obtaining a new column of multi-scale features $\vec{P}^{out} = f(\vec{P}^{in})$.

Firstly, it eliminates nodes with only one input edge. Secondly, when the original input and output nodes are at the same level, an extra edge is added for cost-effective feature fusion. Thirdly, each bi-directional layer is considered an independent feature network layer and repeats the same bi-directional layer to achieve advanced feature fusion. Finally, given that the contribution of different input features to the output features often varies across different resolutions, BiFPN proposes adding weight to each input feature. This allows the network to learn the importance of features of varying sizes. Building on this concept, the network employs Fast Normalized Fusion. The formula is shown in Equation 1:

$$O=\sum_i \frac{w_i}{\varepsilon+\sum_j w_j}\cdot I_i \tag{1}$$

where w_i is a learnable weight at level i, $w_i \geq 0$, ε=0.0001 is a small value to avoid numerical instability and I_i represent the i stage of the input feature.

For a special example, as depicted in Figure 3, we discuss the two fused features at level 5 for BiFPN. The formula is shown in

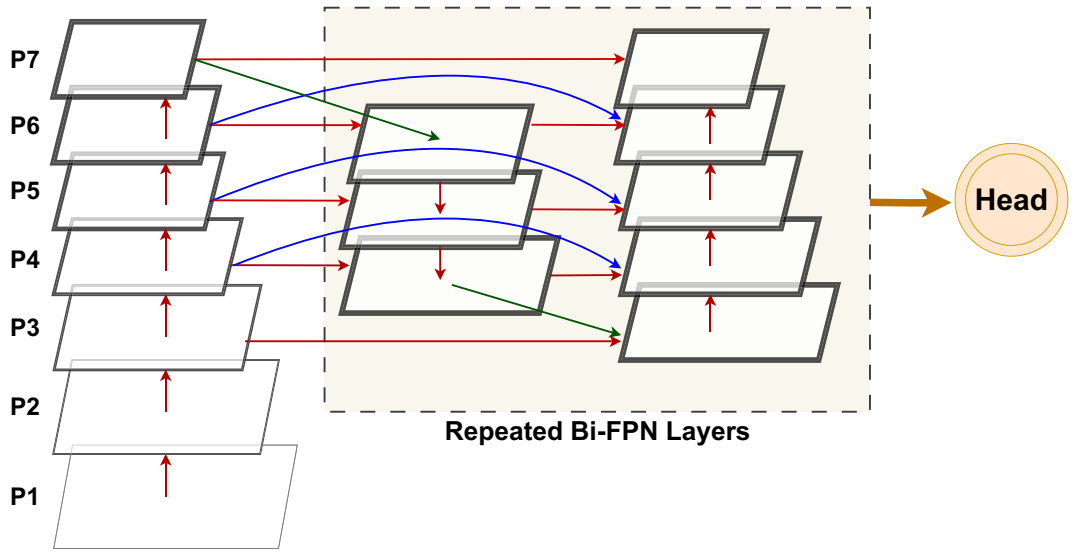

FIGURE 3
Structure and feature fusion diagram for BiFPN: bidirectional multi-scale feature fusion with cross-scale connections and weighted feature aggregation.

Equations 2 and 3:

$$P_5^{td} = Conv\left(\frac{w_1 \cdot P_5^{in} + w_2 \cdot Resize(P_6^{in})}{w_1 + w_2 + \varepsilon}\right) \quad (2)$$

$$P_5^{out} = Conv\left(\frac{w_1' \cdot P_5^{in} + w_2' \cdot P_5^{td} + w_3' \cdot Resize(P_4^{out})}{w_1' + w_2' + w_3' + \varepsilon}\right) \quad (3)$$

where *Conv* is the convolution operation, *Resize* is the operation of resolution matching, P_i^{td} is the intermediate feature at the *i* stage of the top-down pathway and P_i^{out} is the intermediate feature at the *i* stage of the down-up pathway.

2.2 Modulated deformable convolution

As shown in Figure 4, MDConv introduces two-dimensional offsets at the regular grid sampling positions of standard convolution, allowing the sampling points of the convolution kernel to move into irregular areas flexibly. Additionally, it incorporates a modulation factor that assigns a weight (ranging from [0, 1]) to each sampling point, dynamically controlling its contribution to the output features. Specifically, when the modulation factor approaches 1, the feature value at that sampling point is fully retained; conversely, when it approaches 0, the feature value is suppressed or even ignored. This mechanism effectively addresses the issue in Deformable Convolution where the spatial sampling range may extend beyond the region of interest. When the sampling area of Deformable Convolution exceeds the object region, the modulation factor can adjust weights to mitigate interference from irrelevant areas, thereby enabling a more precise focus on the object region. In maritime man-overboard search tasks, since different locations at sea may correspond to objects with different scales and attitudes, such as lifebuoys, lifeboats, or man-overboard, the MOB-Detector, which incorporates MDConv, can flexibly adjust the size of the scale or the receptive field to achieve more accurate detection. Through the mechanism of MDConv, the MOB-Detector can adaptively adjust the position and weight of the sampling points to better capture targets with different scales and poses.

Considering a convolutional kernel that includes *K* sampling points, let us define w_k and p_k as the weights and preset offsets for the *k*-th sampling position, respectively. The notation x(*p*) and y(*p*) is used to represent the feature value at position *p* in the input

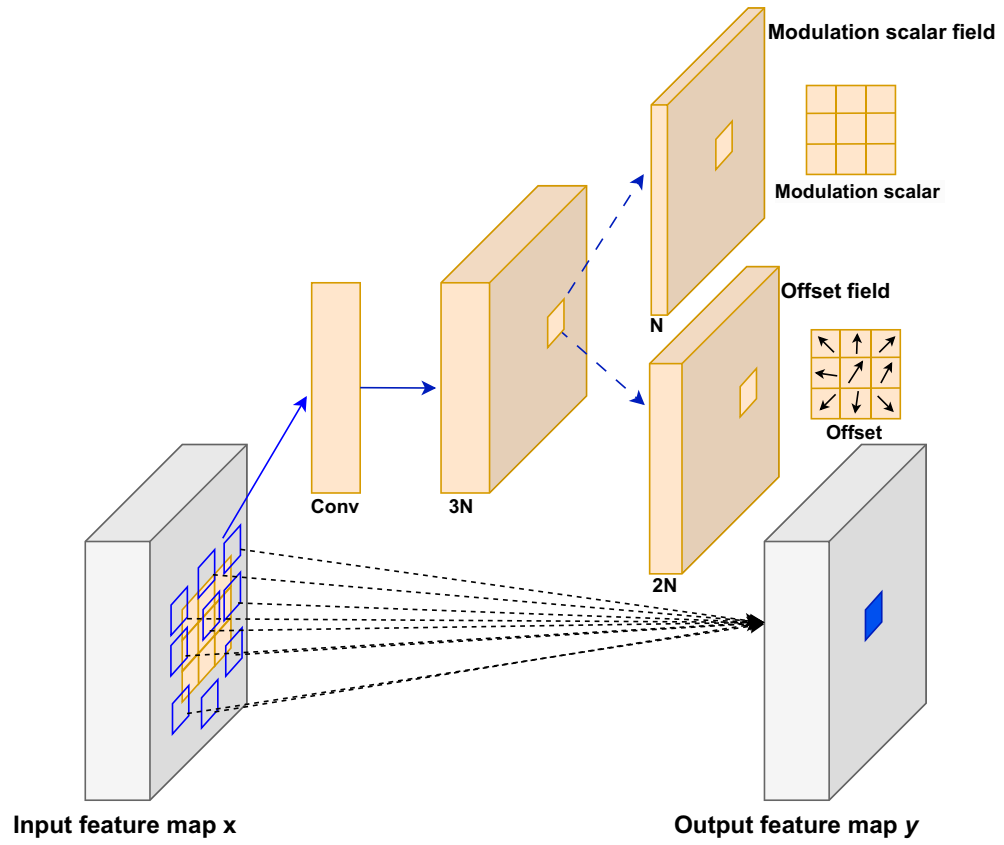

FIGURE 4
Modulated deformable convolution (3x3): adaptive receptive field for feature extraction based on modulation scalars and convolution offsets.

feature map x and the output feature map y, respectively. The MDConv can consequently be represented in Equation 4:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \quad (4)$$

where Δp_k and Δm_k represent the adjustable offset and modulation factor for the k-th position, respectively. The modulation scalar Δm_k is constrained within the interval [0, 1], whereas Δp_k is a real number that can take any value within an unbounded range. Given $p + p_k + \Delta p_k$ are small values, bilinear interpolation is utilized in the calculation of $x(p + p_k + \Delta p_k)$. Both Δp_k and Δm_k are derived by applying a distinct convolutional layer to the identical input feature map x. This particular convolutional layer mirrors the current layer in terms of spatial dimensions and the number of filters. The output consists of $3K$ channels, where the initial $2K$ channels are associated with the learned offset $\{\Delta p_k\}_{k=1}^{K}$ and the subsequent K channels are processed through a sigmoid layer to determine the modulation factors $\{\Delta m_k\}_{k=1}^{K}$.

3 MOB-Detector

3.1 Fully convolutional one-stage object detection

Tian et al. (2019) introduced a fully convolutional single-stage target detector (FCOS) that does not rely on predefined anchor frames or proposal regions, but solves the target detection problem in a pixel-by-pixel prediction manner. By eliminating the predefined set of anchor frames, FCOS completely avoids the complex computation associated with anchor frames during the training process and achieves an anchor-free and proposal-free solution. The architecture of FCOS consists of a Backbone, FPN, and Head, as depicted in Figure 5. Unlike conventional FPN architectures, feature maps P6 and P7 are generated by applying a convolutional layer on P5 and P6. In addition, the multi-level prediction of FPN is utilized to limit the range of bounding box regressions at each level, allowing objects of different sizes to be assigned to different feature layers, which greatly solves the problem of ambiguity due to overlap in ground-truth boxes.

The head consists of three branches: Classification, Center-ness, and Regression, where Classification and Center-ness share the same feature map. Center-ness is a metric that measures the distance from a point within the ground-truth box to the center of the bounding box. Score ranking is computed by multiplying the predicted centrality by the corresponding classification score, which greatly suppresses low-quality prediction borders produced by locations far from the target center. The center-ness formula is shown in Equation 5:

$$\text{centerness}^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}. \quad (5)$$

where l^*, t^*, r^*, and b^* are the distances from the position to the four sides of the enclosing box, respectively. It denotes a 4D vector $t^* = (l^*, t^*, r^*, b^*)$.

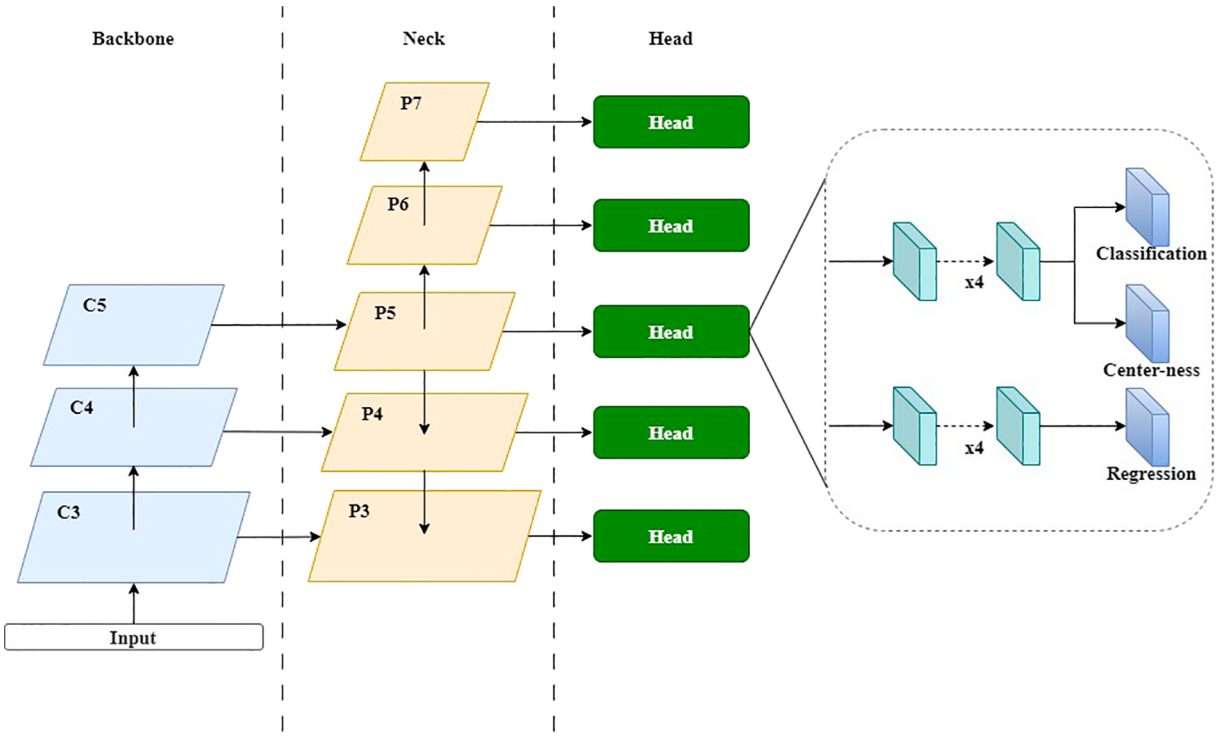

FIGURE 5
The architecture of the FCOS network: a fully convolutional one-stage network comprising backbone, neck, and head components.

3.2 The architecture of MOB-Detector

MOB-Detector is an efficient and lightweight maritime man-overboard detector. MOB-Detector is an improvement on the anchor-free network FCOS. As shown in Figure 6, this detector comprises three key components: backbone, neck, and head.

Backbone: Unlike the FCOS backbone, the MOB-Detector utilizes the ImageNet (Deng et al., 2009) pre-trained EfficientNet-B0 as its backbone network. EfficientNet-B0 is the base model of the EfficientNet series (Tan and Le, 2019), and its core design is inspired by MobileNetV2's Inverted Residual Block (Sandler et al., 2018). It introduces the Mobile Inverted Bottleneck Convolution (MBConv) module, which combines Depthwise Separable Convolution and the Squeeze-and-Excitation module, improving feature extraction while significantly reducing computation. Compared to ResNet-50 and other traditional networks, EfficientNet-B0 requires fewer FLOPs and fewer parameters for the same accuracy, giving it an advantage in situations where UAV computing resources are limited.

Neck: The neck of the MOB-Detector is composed of two repeated BiFPN layers, which extract multi-scale features from levels 3 to 7 $\{P_3, P_4, P_5, P_6, P_7\}$ of the backbone network. These layers iteratively apply top-down, bottom-up, and lateral connections to achieve weighted feature fusion, enhancing the representation of multi-scale features through bi-directional cross-scale interactions. Specifically, the BiFPN takes the features output from the P4 layer of the backbone network and processes them through a modulatable deformed convolutional layer, rather than a standard convolutional layer, for input into the neck network.

Head: In the head, the structure of FCOS is followed, including the classification branch, the regression branch, and the centerness branch.

Loss Function: A location (x, y) on the feature map falls within any ground-truth box, and its class label c^* matches the class label of that ground-truth box, it is classified as a positive sample. Otherwise, it is labeled as a negative sample, with its class label c^* set to 0 (background class). In the regression, each position is associated with a 4D vector $t^* = (l^*, t^*, r^*, b^*)$. If a location falls within multiple object bounding boxes simultaneously, the bounding box with the smallest area is selected as the regression target for that location. The loss function comprises three components: classification loss, regression loss, and center-ness loss. The classification loss is computed using Focal Loss (Lin et al., 2017b), which is specifically designed to address the challenge of class imbalance between positive and negative samples. Focal Loss introduces two key parameters, α_t and γ, to dynamically adjust the contribution of each sample to the total loss. The parameter α_t balances the importance of positive and negative samples by assigning higher weights to the minority class (typically positive samples), while γ reduces the loss contribution from well-classified samples (usually the majority class) by applying a modulating factor $(1 - p_t)^{\gamma}$, where p_t is the model's estimated probability for the ground-truth class. This mechanism ensures that the model focuses more on hard-to-classify samples, which are often underrepresented, thereby improving the overall accuracy and robustness of category prediction. By suppressing the dominance of easily classified negative samples and emphasizing the learning of challenging positive samples, Focal Loss effectively mitigates the

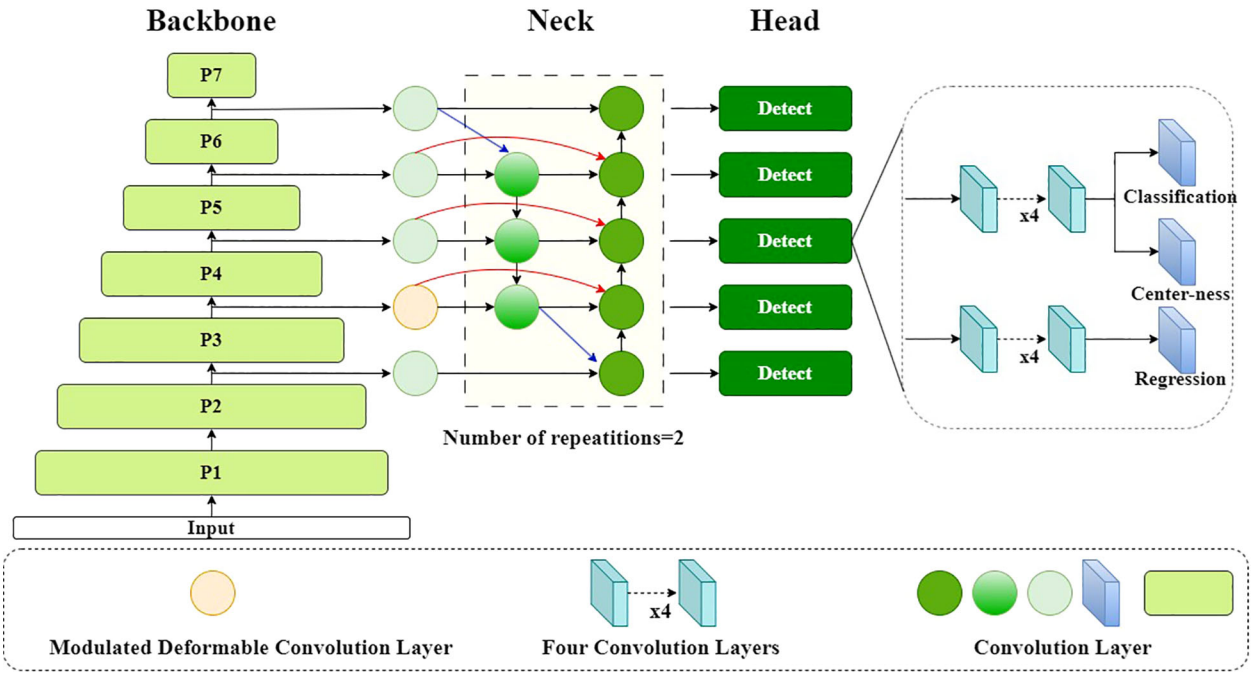

FIGURE 6
Overview of the MOB-Detector: An Integrated Framework for Maritime Man-Overboard Detection Divided into Three Parts—Backbone (EfficientNet), Neck (BiFPN), and Head (Anchor-Free Detector). This architecture leverages Multi-Scale Feature Learning and Adaptive Local Context.

class imbalance issue, leading to better performance in object detection tasks, especially for scenarios with a significant imbalance between positive and negative samples. The regression loss employs IOU Loss (Yu et al., 2016) to effectively minimize the positional discrepancies between the predicted bounding boxes and the ground truth. Lastly, the center-ness loss is calculated using Binary Cross Entropy (BCE) (Krizhevskv et al., 2017) to evaluate how well the predicted positions align with the center of the targets, thereby suppressing low-quality detections.

The training loss function is shown in Equation 6:

$$L(\{c_{x,y}\},\{t_{x,y}\}) = \frac{1}{N_{pos}}\sum_{x,y}L_{cls}(c_{x,y},c^{*}_{x,y}) + \frac{\lambda}{N_{pos}}\sum_{x,y}1_{\{c^{*}_{x,y}>0\}}L_{\mathrm{reg}}\left(t_{x,y},t^{*}_{x,y}\right) + \frac{\beta}{N_{\mathrm{pos}}}\sum_{\mathrm{x,y}}L_{centerness}\left(\mathrm{C_{x,y}},\mathrm{C^{*}_{x,y}}\right) \quad (6)$$

Where $c_{x,y}$ denotes the classification scores in position (x,y), $t_{x,y}$ denotes the regression prediction in position (x,y), $C_{x,y}$ denotes the center-ness scores in position (x,y) and * denotes the truth scores. L_{cls} refers to focal loss, L_{reg} indicates the IOU loss and $L_{centerness}$ denotes the BCE loss. N_{pos} denotes the number of positive samples and λ, β are weights for L_{reg} and $L_{centerness}$. $1_{\{c_i^{*}>0\}}$ is the indicator function, being 1 if $c_i^{*} > 0$ and 0 otherwise.

4 Experiments

4.1 ManOverboard benchmark

The ManOverboard Benchmark is a benchmark proposed by Professor Yin Jianchuan's team from the Naval Architecture and Shipping College at Guangdong Ocean University, aimed at detecting small targets at sea. The video obtained from the UAV acquisition was taken every 5 seconds and filtered to finalize 956 images, and the resolution of the images is 3840 x 2160. The team confirmed that the benchmark contained 956 images and 35119 objects with bounding boxes, as shown in Figures 7A–C. The detailed data is illustrated in Table 1. Each image has been manually annotated and categorized among COCO (Lin et al., 2014), YOLO, and VOC formats, and all annotations were reviewed by vision experts. The annotations are divided into three categories: person (person on the beach and person at sea), person

FIGURE 7
(A) Examples of various images in the ManOverboard benchmark include scenes depicting the sea, land, and a combination of both. **(B)**. Image of the ManOverboard benchmark dataset. The image scene is an open sea, and several people in the water can be seen scattered on the surface of the sea, with some of them wearing lifebuoys or lifejackets. **(C)**. Image of the ManOverboard benchmark dataset. The image scene is half beach and half sea, the image in the picture can be seen several people in the water, scattered in the sea and the beach, and in the sea part of the person wearing a life ring or wearing a life jacket.

TABLE 1 Detailed results of images and annotations of the ManOverboard benchmark.

Item	Train	Test	Sum
Images	860	96	956
Annotations	32115	3004	35119
person	28719	2665	31384
person with buoy	2181	227	2408
person with jacket	1216	112	1328

with buoy (person with a buoy), and person with jacket (person wearing a life jacket). This benchmark provides researchers with a comprehensive foundation to advance the detection and recognition of small maritime objects. The Github URL: https://github.com/YinJianchuan/ManOverboard.

4.2 Evaluation metrics and environment

To validate the performance of MOB-Detector on the ManOverboard benchmark, the evaluation metrics shown in Table 2 below were used through comparison and ablation experiments. As shown in Table 3, the experiments were conducted in an environment running the Linux operating system, with hardware consisting of an Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz, an NVIDIA HFX A800 GPU, and CUDA 12.4. The deep learning framework used was Pytorch 2.5.1.

TABLE 2 Evaluation metrics on the experiments.

Metrics	Formula	Description
Precision(P)	$P = \frac{TP}{TP + FP}$	The ratio between the number of targets correctly detected by the model and the number of all predicted as targets predicted by the model
Recall(R)	$R = \frac{TP}{TP + \mathrm{FN}}$	The ratio between the number of targets correctly detected by the model and the total number of targets
Average Precision(AP)	$AP = \int_0^1 P(r)dr$	The corresponding P and R are calculated based on different confidence thresholds in the precision-recall curve, and the area under the precision-recall curve is calculated as AP
Mean Average Precision(mAP)	$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$	Mean score of AP across all categories
AP_{50}	–	AP at IOU=.50
AP_{75}	–	AP at IOU=.75
AP_{small}	–	AP for small objects: area< 32^2
AP_{medium}	–	AP for medium objects: 32^2< area< 96^2

*TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

4.3 Ablation experiments

In the ablation study, three experiments were conducted: the first evaluated the effectiveness of the MOB-Detector's design, the second investigated the optimal number of layers in the BiFPN architecture, and the third experiment aimed to determine the best receptive field size.

The MOB-Detector enhances the feature information of small objects through the use of BiFPN and incorporates MDConv to adapt to the scale transformation of small objects. To validate the effectiveness of these designs, the first experiments were conducted, involving four experimental groups: the first group is FCOS, the second group is FCOS combined with BiFPN, the third group is FCOS combined with MDConv, and the last group is the complete MOB-Detector. In the experiments, all the groups keep the same learning rate, epoch, batch size, and other hyperparameters to ensure the effect of control variables. In Tables 4–6, bold values indicate the optimal results within each column. The symbol "√" in Table 4 indicates the application or adoption of the corresponding module.

Based on the analysis in Table 4, the MOB-Detector excels across all performance indicators. AP_{50}, AP_{75}, AP_{small}, and AP_{medium}, achieving its highest scores with values of 59.9, 23.7, 32.1, and 38.6, respectively. These scores represent improvements over the FCOS model utilizing the EfficientNet-b0 backbone network, with enhancements of 5.9, 4.8, 6.2, and 5.4, respectively. Additionally, performance gains were observed when the FCOS model was augmented with either BiFPN or MDConv. Specifically, the FCOS+BiFPN and FCOS+MDConv configurations achieved improvements of 3.5 and 4.1 in AP_{small}, respectively, compared to the original FCOS model.

Furthermore, as depicted in Figure 8, the Precision-Recall curve of the MOB-Detector, with an Average Precision of 0.62, demonstrates its competitive performance. It maintains a reasonable balance between precision and recall, indicating that the detector is capable of effectively identifying positive samples while managing false positives. This performance highlights the MOB-Detector as a promising approach in the field of object detection.

The second set of ablation experiments aimed to determine the optimal number of BiFPN layers to optimize the network architecture. The experiments were divided into five groups, with the number of BiFPN layers ranging from 1 to 5. Before this, the MDConv was fixed at layer P4, as illustrated in Figure 6.

Based on the analysis in Table 5, the performance of the MOB-Detector initially increases and then decreases as the number of BiFPN layers increases. Specifically, the MOB-Detector achieves optimal performance across all metrics when the number of BiFPN layers is set to two, with values of 59.9, 23.7, 32.1, and 38.6, respectively. This indicates that a two-layer BiFPN configuration provides the best balance for optimal performance in the MOB-Detector.

The choice of kernel size significantly impacts the performance of detecting small objects in maritime man-overboard search and rescue. Since the detection of such targets often focuses on smaller

TABLE 3 Detailed description of the hardware and software setup.

OS	CPU	GPU	Python	Pytorch	CUDA
Linux	Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz	NVIDIA HFX A800	3.12	2.5.1	12.4

TABLE 4 Results of the ablation experiments to verify the MOB-Detector's reasonableness.

Method	BiFPN	MDConv	AP_{50} (%)	AP_{75} (%)	AP_{small} (%)	AP_{medium} (%)
FCOS (pure)			54.0	18.9	25.9	33.2
FCOS	√		50.7	21.7	29.4	33.5
FCOS		√	56.2	22.5	30.0	35.8
FCOS (MOB-Detector)	√	√	**59.9**	**23.7**	**32.1**	**38.6**

All group experiments use the EfficientNet-b0 network as the backbone.
The symbol '√' indicates the application or adoption of the corresponding module (e.g., BiFPN or MDConv).
Bold values indicate the optimal results within each column.

object, and considering that a 1 ⋆1 kernel size is generally not suitable for capturing spatial contextual information, while a 7 ⋆7 kernel size would substantially increase the number of channels for offsets and modulation factors, leading to higher GPU and memory usage, we have grouped the receptive field sizes into 3 ⋆3 and 5 ⋆5. Based on the experimental results presented in Table 6, the results indicate that the 3 ⋆3 kernel size achieves superior performance across various metrics. In contrast, the 5 ⋆5 kernel size yields slightly lower performance, with an AP_{50} of 56.3, AP_{75} of 21.9, AP_{small} of 29.1 and AP_{medium} of 37.3. These findings suggest that a 3 ⋆3 kernel size is more effective for maritime man-overboard rescue and search tasks.

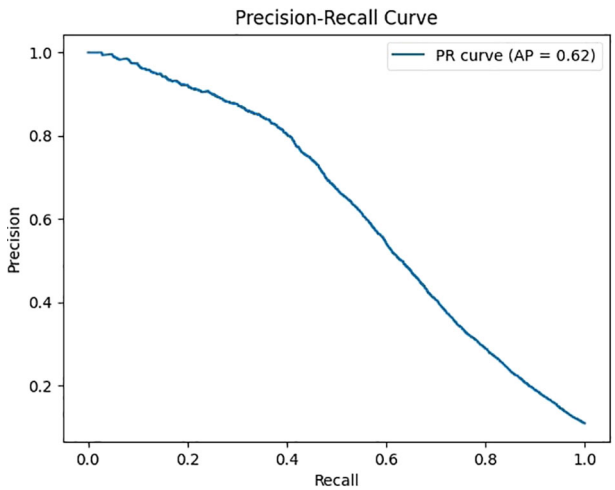

FIGURE 8
The precision-recall curve of the MOB-detector. The horizontal axis of the figure is Recall, and the vertical axis is Precision, and the curves reflect the balanced relationship between the detection precision and coverage of the model under different confidence thresholds.

4.4 Comparison experiments

In comparison experiments, the MOB-Detector proposed in this paper is compared with several state-of-the-art detectors which divided into one-stage detector [CenterNet, and RetinaNet, YOLOv5, and YOLOv8 (Xu et al., 2024)] and two-stage detector (Faster R-CNN, SSD). Compared to the aforementioned detectors, the MOB-Detector exhibits enhanced feature fusion capabilities in multi-scale feature learning. Furthermore, rather than relying on traditional local context information learning, the MOB-Detector employs modulated factors and convolution offsets to adapt to targets that vary in irregular shapes. As depicted in Figures 9A–D, the detection results of MOB-Detector, YOLOv5, and YOLOv8 are roughly consistent with the image annotations except for three models that fail to effectively differentiate between "person with buoy" and "humans" in a small area. Specifically, YOLOv5 misidentifies the buoy in the upper right corner as a human, while YOLOv8 misses the detection of the figure in white in the center of the image. In Table 7, bold values indicate the optimal results within each column. For the detailed metric results in Table 7, MOB-Detector outperforms YOLOv8, which is ranked second in all four metrics AP_{50}, AP_{75}, AP_{small}, and AP_{medium}, with leads of 8.6, 3, 4.6 and 5.2, respectively. As well as, MOB-Detector is better than the representatives two-stage detector, SSD.

4.5 Discussion

The experimental results demonstrate the effectiveness of the MOB-Detector in man-overboard detection tasks, particularly for small and medium-sized objects. As shown in Table 4, the ablation experiments reveal that the integration of BiFPN and MDConv in FCOS significantly enhances performance, with the MOB-Detector achieving the highest scores across all metrics AP_{50}: 59.9%, AP_{75}: 23.7 %, AP_{small}: 32.1%, and AP_{medium}: 38.6%. These improvements highlight the importance of adaptive local context and multi-scale feature learning in detecting small objects.

TABLE 5 Experimental results of optimal BiFPN layers within the MOB detector.

Number of Layers	AP_{50} (%)	AP_{75} (%)	AP_{small} (%)	AP_{medium} (%)
1	57.2	23.0	30.1	36.9
2	**59.9**	**23.7**	**32.1**	**38.6**
3	52.7	17.9	28.3	32.5
4	54.1	19.0	26.3	33.9
5	51.3	21.0	28.6	34.0

Bold values indicate the optimal results within each column.

TABLE 6 Experimental results of the optimal kernel size of modulated deformable convolution.

Kernel size	AP_{50} (%)	AP_{75} (%)	AP_{small} (%)	AP_{medium} (%)
3 ⋆3	**59.9**	**23.7**	**32.1**	**38.6**
5 ⋆5	56.3	21.9	29.1	37.3

Bold values indicate the optimal results within each column.

As illustrated in Tables 5, 6, further analysis of the BiFPN layers indicates that a two-layer configuration optimizes performance, suggesting a balance between feature fusion complexity and computational efficiency. Additionally, the choice of a 3×3 kernel size for the receptive field proves more effective than larger kernels, as it better captures global context without excessive resource consumption.

In Table 7, the MOB-Detector outperforms state-of-the-art algorithms like YOLOv8, which is the second in comparison, particularly in AP_{50} (59.9% vs. 51.3%) and AP_{small} (32.1% vs. 27.5%). The results collectively validate the MOB-Detector as a promising solution for object detection, especially in tasks involving small and densely packed objects.

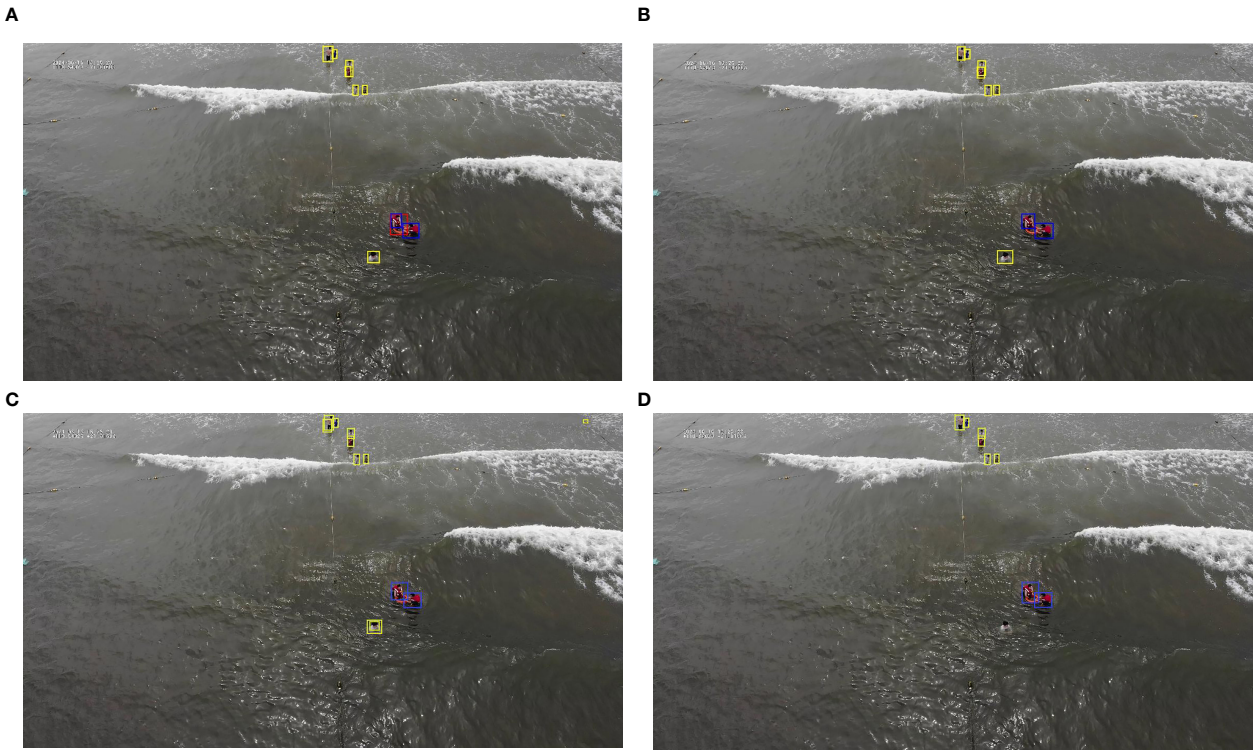

FIGURE 9
(A) The original annotations of the image. The yellow box denotes 'person', the blue box signifies 'person with buoy', and the red box indicates 'person with jacket'. **(B)** The detection results of the MOB-detector. The yellow box denotes 'person', the blue box signifies 'person with buoy', and the red box indicates 'person with jacket'. **(C)** The detection results of YOLOv5. The yellow box denotes 'person', the blue box signifies 'person with buoy', and the red box indicates 'person with jacket'. **(D)** The detection results of YOLOv8. The yellow box denotes 'person', the blue box signifies 'person with buoy', and the red box indicates 'person with jacket'.

TABLE 7 Comparison experiments using faster R-CNN, SSD, RetinaNet, YOLOv5, YOLOv8, and MOB-detector.

Model	AP_{50} (%)	AP_{75} (%)	AP_{small} (%)	AP_{medium} (%)
Faster R-CNN	35.9	9.4	15.5	18.7
SSD	37.2	13.8	17.3	23.6
RetinaNet	30.6	8.5	11.6	19.5
CenterNet	38.4	10.3	13.5	18.9
YOLOv5 (Ultralytics, 2020)	48.5	15.0	22.6	27.2
YOLOv8	51.3	20.7	27.5	33.4
MOB-Detector	**59.9**	**23.7**	**32.1**	**38.6**

Bold values indicate the optimal results within each column.

Although the AP_{50} of the MOB-Detector is close to 60%, its AP for objects with dimensions smaller than 32x32 pixels is less than satisfactory. For instance, as illustrated in Figure 10, two distant objects resembling knots are incorrectly identified as 'person' due to their small size. Furthermore, since the experimental datasets is primarily based on clear visibility, the generalization ability of the MOB-Detector needs to be validated in low visibility scenarios. Therefore, future work could focus on challenging scenarios.

5 Conclusion

In this work, this paper develops MOB-Detector, a man-overboard detection designed for UAV search and rescue missions. MOB-Detector first employs an anchor-free detection head based on the FCOS network, overcoming the limitations of traditional anchor-based methods in detecting small-scale man-overboard targets. Next, MOB-Detector introduces BiFPN to enhance the effective fusion of man-overboard features, improving detection accuracy. Finally, MDConv is incorporated, enabling the model to adapt to the geometric variations of the man-overboard in different postures on the sea surface. In addition, this paper introduces the ManOverboard Benchmark, alleviating the gap in existing datasets for small maritime objects. Using this benchmark, we conducted a comprehensive evaluation of the performance of MOB-Detector. The results from our ablation experiments indicate the effectiveness of both BiFPN and MDConv within the model. It was found that MOB-Detector achieves optimal performance when the number of BiFPN layers is

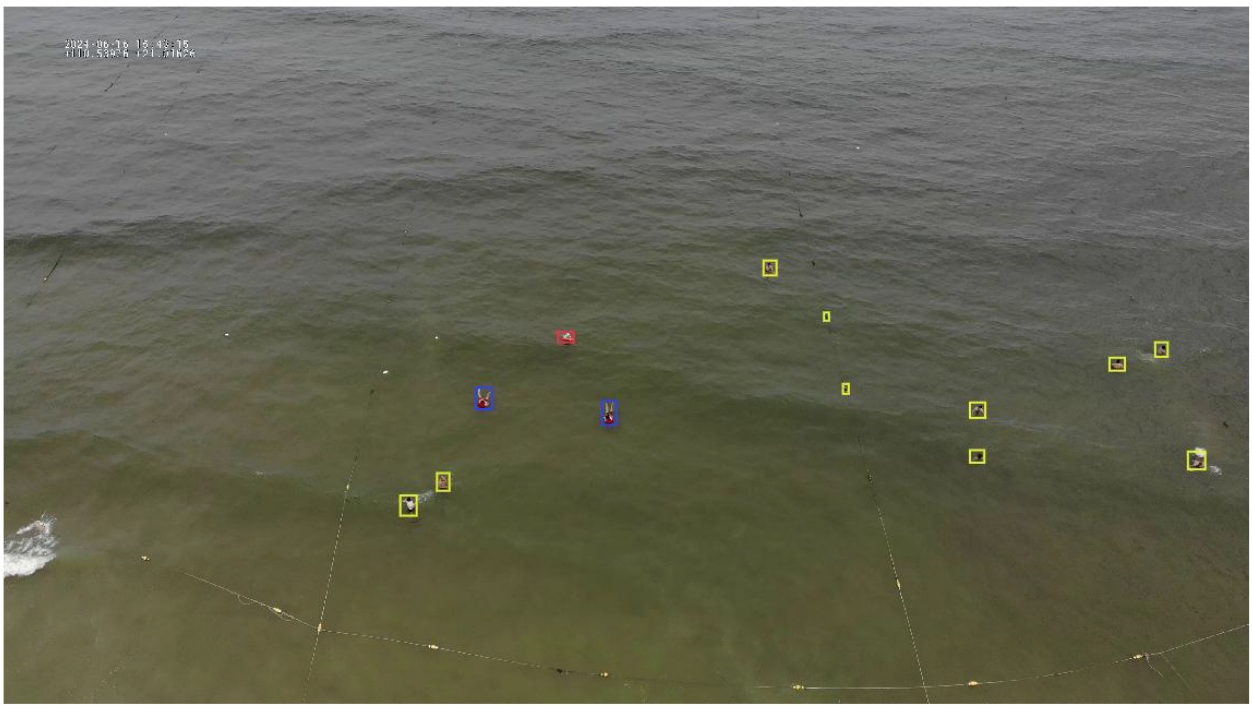

FIGURE 10
The detection results of the MOB-Detector. The yellow box denotes 'person', the blue box signifies 'person with buoy', and the red box indicates 'person with jacket'. Due to long distances, two targets were misdiagnosed as 'person' in this image.

set to 2 and the receptive field of the MDConv is configured to 3*3. In comparative experiments of one-stage algorithms and two-stage algorithms, MOB-Detector outperforms other state-of-the-art algorithms in terms of detection capability. It indicates that the MOB-Detector will serve as an effective tool on UAV for maritime search and rescue operations.

6 Future work

In future work, the MOB-Detector model designed for UAV will undergo further enhancements in accuracy to ensure that stringent performance requirements are met in resource-constrained environments with limited computational power and on low-power devices. Additionally, our future research will explore the following directions: (1) Detector Design: Detector design aims to further investigate target detection networks that leverage the Transformer architecture. The self-attention mechanism inherent in Transformers enables effective capture of global contextual information within images, thereby enhancing the accuracy and robustness of target detection in complex scenarios. Compared to traditional convolutional neural networks, Transformers offer notable advantages in modeling long-range dependencies, particularly when addressing multi-scale targets and small target detection tasks. (2) Data Augmentation: Advanced data augmentation techniques will be explored to simulate challenging environmental conditions that are difficult to collect data for, such as typhoons, heavy rain, and snowy days. This will help improve the robustness of the MOB-Detector in complex real-world scenarios, including maritime search and rescue operations.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/YinJianchuan/ManOverboard.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

GX: Writing – original draft, Writing – review & editing. JY: Writing – original draft, Writing – review & editing. JZ: Writing – original draft, Writing – review & editing. NW: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Natural Science Foundation of China under Grants 52231014 and 52271361; the Special Projects of Key Areas for Colleges and Universities in Guangdong Province under Grant 2021ZDZX1008; the Natural Science Foundation of Guangdong Province under Grant 2023A1515010684; the Technology Breakthrough Plan Project of Zhanjiang under Grant 2023B01024; the Postgraduate Education Innovation Project of Guangdong Ocean University (202546); and the Program for Scientific Research Start-Up Funds of Guangdong Ocean University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Bai, J., Dai, J., Wang, Z., and Yang, S. (2022). A detection method of the rescue targets in the marine casualty based on improved YOLOv5s. *Front. Neurorobotics*. 16, 1053124.

Chen, S., Piao, L., Zang, X., Luo, Q., Li, J., Yang, J., and Rong, J. (2023). Analyzing differences of highway lane-changing behavior using vehicle trajectory data. *Physica A: Statistical Mechanics and its Applications*, 624, 128980. doi: 10.1016/j.physa.2023.128980

Chen, X., Wei, C., Xin, Z., Zhao, J., and Xian, J. (2023). Ship detection under low-visibility weather interference via an ensemble generative adversarial network. *J. Mar. Sc. Engineer.* 11 (11), 2065. doi: 10.3390/jmse11112065

Chen, X., Wu, X., Prasad, D. K., Wu, B., Postolache, O., and Yang, Y. (2022). Pixel-wise ship identification from maritime images via a semantic segmentation model. *IEEE Sensors J.*, 22 (18), 18180–18191.

Dai, J., Qi, H., Xiong, Y., Li, Y., and Zhang, G. (2017). Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. (Piscataway, New Jersey, USA: IEEE). 764–773.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. (Piscataway, New Jersey, USA: IEEE headquarters). 2009, 248–255.

European Maritime Safety Agency (EMSA). (2023). *Annual overview of marine casualties and incidents*. Available online at: https://emsa.europa.eu/csn-menu/items.html?cid=14&id=5052 (Accessed November 1, 2024).

Girshick, R. (2015). "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*. (Piscataway, New Jersey, USA: IEEE), 1440–1448.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2961–2969.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 25. doi: 10.1145/3065386

Law, H., and Deng, J. (2018). "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, (Cham, Switzerland: Springer International Publishing). 734–750.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*. 86 (11), 2278–2324.

Lei, F., Tang, F., and Li, S. (2022). Underwater target detection algorithm based on improved YOLOv5. *J. Marine Sci. Eng.* 10, 310. doi: 10.3390/jmse10030310

Li, Y., Li, Z., Zhang, C., Luo, Z., Zhu, Y., and Ding, Z. (2021). Infrared maritime dim small target detection based on spatiotemporal cues and directional morphological filtering. *Infrared Physics & Technol.* 115, 103657.

Lin, T Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. (Piscataway, New Jersey, USA: IEEE). 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). "Focal loss for dense object detection," in *Proceedings, IEEE conference on computer vision and pattern recognition*, (Piscataway, New Jersey, USA: IEEE Institute of Electrical and Electronics Engineers), 2980–2988.

Lin, T Y., Maire, M., and Belongie, S. (2014). Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, *Proceedings, Part V 13*. (Zurich, Switzerland: Springer International Publishing), 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., et al. (2016). "Ssd: Single shot multibox detector," in *Computer vision–ECCV 2016: 14th european conference, amsterdam, the Netherlands, october 11–14, 2016, proceedings, part I 14* (Amsterdam, Netherlands: Springer International Publishing), 21–37.

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*. (Piscataway, New Jersey, USA: IEEE), 8759–8768.

Liu, X., Qiu, L., Fang, Y., Wang, K., and Rodríguez, J. (2024). Event-driven based reinforcement learning predictive controller design for three-phase NPC converters using online approximators. *IEEE Trans. Power Electron.* 40. doi: 10.1109/TPEL.2024.3510731

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Piscataway, USA: IEEE)., 4510–4520.

Tan, M., and Le, Q. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning* (Long Beach, California, USA: PMLR), 6105–6114. doi: 10.1109/ICCV.2019.00972

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* (Piscataway, New Jersey, USA: IEEE/CVF IEEE Computer Society and Computer Vision Foundation), 10781–10790.

Tian, Z., Shen, C., Chen, H., and He, T. (2019). FCOS: Fully convolutional one-stage object detection. *arxiv*. arxiv:1904.01355: 9627–9636.

Tong, K., Wu, Y., and Zhou, F. (2020). Recent advances in small object detection based on deep learning: A review. *Image Vision Computing*. 97, 103910. doi: 10.1016/j.imavis.2020.103910

Ultralytics (2020). *Ultralytics documentation*. Available online at: https://docs.ultralytics.com/models/yolov5/ (Accessed November 3, 2024).

Wang, N., Wang, Y., Feng, Y., and Wei, Y. (2024). AodeMar: Attention-aware occlusion detection of vessels for maritime autonomous surface ships. *IEEE Trans. Intelligent Transportation Syst.* 10, 25 doi: 10.1109/TITS.2024.3398733

Xu, G., Yin, J., and Zhang, Z. (2024). "Marine ship detection under fog conditions based on an improved deep-learning approach," in *International conference on neural computing for advanced applications* (Springer Nature Singapore, Singapore), 92–103.

Yang, Z., Liu, S., Hu, H., Wang, L., and Lin, S. (2019). "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*. (Piscataway, New Jersey, USA: IEEE/CVF). 9657–9666.

Yu, J., Jiang, Y., Wang, Z., Cao, Z., and Huang, T. (2016). "Unitbox: An advanced object detection network," in *Proc. ACM int. Conf. Multimedia* (New York, USA: ACM headquarters), 516–520.

Zhang, M., Dong, L., Zheng, H., and Xu, H. (2021). Infrared maritime small target detection based on edge and local intensity features. *Infrared Phys. Technol.* 119, 103940. doi: 10.1016/j.infrared.2021.103940

Zhang, Y., Tao, Q., and Yin, Y. (2023). A lightweight man-overboard detection and tracking model using aerial images for maritime search and rescue. *Remote Sens.* 16, 165. doi: 10.3390/rs16010165

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arxiv*. arxiv:1904.07850. doi: 10.48550/arXiv.1904.07850

Zhu, X., Hu, H., Lin, S., and Dai, J. (2019). "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (Piscataway, New Jersey, USA: IEEE/CVF), 9308–9316.

Zhu, Q., Ma, K., Wang, Z., and Shi, P. (2023). YOLOv7-CSAW for maritime target detection. *Front. neurorobotics* 17, 1210470. doi: 10.3389/fnbot.2023.1210470