Journal of *Marine Science and Engineering*

MDPI

Article

HPSNet: A Three-Stage Enhanced YOLOv11 Detector for Person-Overboard Detection in Maritime UAV Imagery

Yuqing Ren [1,†], Guohao Wen [1,†], Lili Zhou [1,2], Xiaoming Fan [3,*] and Yingbang Huang [2,4,*]

[1] School of Navigation and Naval Architecture, Dalian Ocean University, Dalian 116023, China; renyuqing@dlou.edu.cn (Y.R.); wgh_dlou@163.com (G.W.); zhou01212026@163.com (L.Z.)
[2] South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510300, China
[3] China Fisheries Mutual Insurance Society Dalian Branch, Dalian 116000, China
[4] College of Ocean and Meteorology, Guangdong Ocean University, Zhanjiang 524088, China
* Correspondence: fanxiaoming@cfmi.cn (X.F.); hkbone99@163.com (Y.H.)
† These authors contributed equally to this work and should be considered co-first authors.

Abstract

Person-overboard detection from maritime unmanned aerial vehicle (UAV) imagery is challenging because the targets occupy very few pixels, sea-surface clutter is severe, human appearance varies substantially, and real-time processing is required. Existing detectors therefore struggle to meet the demands of practical maritime search and rescue. This paper presents HPSNet, an accuracy- and recall-oriented detector designed for maritime UAV imagery. HPSNet uses YOLOv11 as its baseline and introduces three complementary modifications. A channel transposed attention (CTA) module is embedded in the backbone to improve the discrimination of target features from complex sea-surface interference. A Giraffe feature pyramid network (GFPN) replaces the original feature-fusion network to strengthen multiscale information exchange and preserve cues from extremely small targets. A Dynamic Head (DyHead) adapts the predictions to variations in target scale, location, and appearance. HPSNet is evaluated against 12 representative detectors on the public Person Detection in Water and AFO datasets, and ablation experiments examine the contribution of each component. On Person Detection in Water, HPSNet achieves 78.2% mAP@50, 37.5% mAP@50:95, and 67.2% recall, improving the YOLOv11 baseline by 2.7, 2.2, and 4.2 percentage points, respectively. On AFO, it achieves 88.4% mAP@50 and 59.4% mAP@50:95, with gains of 0.7 and 1.9 percentage points over the baseline. The model contains 4.18 M parameters, requires 9.6 GFLOPs, and processes an image in 13.4 ms on an RTX 4090. These results demonstrate improved detection of small and visually weak maritime targets relative to YOLOv11 while maintaining a moderate model scale, providing a foundation for future deployment and optimization on embedded maritime UAV platforms.

Keywords: person-overboard detection; object detection; YOLOv11; attention mechanism; multiscale feature fusion; maritime UAV

Check for updates

Academic Editor: Chia-Hung Yeh

Received: 11 August 2026
Revised: 6 September 2026
Accepted: 10 September 2026
Published: 12 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The continued growth of the global marine economy increases the importance of rapid maritime emergency response. According to statistics compiled over more than two decades in the China Fishery Safety Analysis Report, more than 3800 fishers died in overboard incidents, accounting for approximately 33% of all recorded fisher fatalities [1].

A person immersed in cold water may have a survival window of less than 30 min, which makes search speed a direct determinant of survival. Conventional operations rely mainly on visual searches from vessels and helicopters. Their coverage is limited, their effectiveness depends heavily on sea conditions, and they require substantial personnel and operating costs. UAVs offer high mobility, broad coverage, and rapid deployment. When equipped with optical sensors and an automatic detector, they can scan large sea areas, identify people in the water, and shorten emergency response times.

Reliable detection from maritime UAV imagery nevertheless remains difficult. A UAV often operates at a high altitude; so, a person in the water occupies only a small number of pixels and provides little discriminative information. Waves, foam, sun glint, and floating debris can resemble human targets in shape or intensity, which produces both false detections and missed targets. Changes in surface motion, body pose, immersion level, and illumination also cause large variations in target scale, contour, and texture. A fixed detection head cannot always adapt to these variations. At the same time, search-and-rescue systems must process large areas quickly, whereas accurate two-stage detectors often impose excessive computational and latency costs on resource-constrained airborne hardware. Maritime person-overboard detection therefore requires a framework that balances accuracy and efficiency while addressing weak small-target features, complex background clutter, and variable target appearance.

This paper proposes HPSNet, a detector tailored to these requirements. HPSNet improves YOLOv11 at three stages of the detection pipeline. In the backbone, CTA strengthens target-related responses and suppresses confusing sea-surface patterns. In the neck, GFPN increases cross-scale information exchange so that fine spatial detail can interact with high-level semantics. In the prediction stage, DyHead adjusts its response to changes in target scale, location, and task. These components address different but related causes of detection failure and form a complementary architecture. The contribution thus lies in the coordinated, task-oriented integration of CTA, GFPN, and DyHead, each targeting a distinct failure mode of maritime person-overboard detection.

This study also establishes a consistent evaluation of the public Person Detection in Water and AFO datasets and compares HPSNet with 12 representative detectors. Controlled ablation experiments quantify the effects of GFPN, CTA, and DyHead, while confusion matrices and Grad-CAM visualizations provide additional evidence about missed detections and feature responses. The results show clear gains over the YOLOv11 baseline, especially on the person-focused Person Detection in Water dataset. They also reveal that the improvement is task-specific rather than uniform across every maritime object category.

2. Related Work

2.1. Maritime Search-and-Rescue Target Detection

Maritime target detection has progressed from manual observation to sensor-assisted search, conventional computer vision, and deep-learning-based detection. Early search-and-rescue operations depended on personnel observing the sea from vessels, fixed-wing aircraft, or helicopters, often with binoculars and searchlights. Although straightforward, this approach offered limited coverage and endurance, and its effectiveness varied with observer experience, viewing distance, weather, and sea state.

Radar, automatic identification systems, infrared sensors, and synthetic aperture radar later became important sources of maritime surveillance data. Radar provides broad, all-weather coverage but has limited sensitivity to people in the water because they present a small radar cross-section. Automatic identification systems primarily locate equipped vessels and cannot identify a person without a transponder. Infrared imaging can exploit the temperature difference between a human body and the water, but its quality depends on

water temperature, solar radiation, and sensor resolution. These sensors provide valuable auxiliary information, yet no single conventional sensor reliably localizes people in complex maritime conditions.

Conventional vision methods extract color, edge, texture, and shape descriptors from visible or infrared imagery and combine them with classifiers such as support vector machines and AdaBoost. Their computational cost is often moderate, but their dependence on handcrafted features limits robustness. Waves, foam, sunlight reflections, floating objects, and changing illumination can substantially alter these features and make a person difficult to distinguish from the background.

UAVs have become an important search platform because they can be deployed rapidly, cover large areas, and operate at relatively low cost. A visible-light or infrared camera mounted on a UAV can acquire wide-area imagery and transmit detections and location information to rescue personnel. Varga et al. released SeaDronesSee in 2022, providing maritime UAV imagery captured at different altitudes, viewing angles, and sea states and supporting research on person detection and tracking [2]. Cafarelli et al. released MOBDrone, a large-scale drone-view benchmark dedicated to persons overboard, with more than 125K images collected under varied altitudes, camera angles, and illumination conditions [3]. Building on SeaDronesSee, the Maritime Computer Vision (MaCVi) workshops organized UAV-based maritime object detection and tracking challenges and released the extended SeaDronesSee Object Detection v2 benchmark [4]. Gonçalves and Damas demonstrated real-time on-board detection of rescue targets on a UAV using standard RGB cameras, indicating that lightweight detectors can support autonomous search with minimal human intervention [5]. These data demonstrate that people viewed from a UAV often occupy very few pixels and may be partially submerged, occluded, blurred, or observed in unusual poses. The task is therefore harder than standard pedestrian detection.

Related work in China initially focused on vessels, nearshore targets, and floating objects. As UAV platforms and deep learning developed, attention shifted toward small-object detection, suppression of sea-surface clutter, and lightweight deployment. Liu et al. introduced inverted depthwise separable convolution into YOLOv4 to improve maritime target detection for unmanned surface vehicles [6]. Zhao et al. incorporated mixed-domain attention into YOLOv4 to strengthen vessel features in complex sea backgrounds [7]. Zhou et al. modified the loss function and spatial pyramid pooling structure of YOLOv5 to improve vessel detection [8]. These studies provide useful foundations, but vessels generally have clearer contours and stronger structural cues than people in the water.

Current maritime search systems move from single-sensor processing toward multisensor cooperation, from handcrafted descriptors toward learned representations, and from offline analysis toward real-time inference on UAVs. Three challenges remain central. Flight-altitude variation produces large scale changes and turns a person into an extremely small target at high altitude. Waves, foam, glare, and debris resemble human targets and increase false detections. Body pose, immersion, and location vary continuously and increase missed detections. Improving small-target sensitivity and resistance to background interference under limited computation remains an open problem.

2.2. Deep-Learning-Based Object Detection

Deep object detection develops along several major lines, including two-stage detectors, one-stage detectors, Transformer-based detectors, multiscale feature fusion, and feature enhancement for complex scenes. Girshick et al. introduced R-CNN in 2014, using selective search to generate region proposals and a convolutional neural network to extract their features [9]. Fast R-CNN then shared convolutional features across the full image to reduce repeated computation [10]. Ren et al. proposed Faster R-CNN in 2015 and replaced

selective search with a region proposal network, enabling unified training of proposal generation and detection [11]. Cai and Vasconcelos introduced Cascade R-CNN in 2018, using a sequence of detectors with increasing intersection-over-union thresholds to refine localization [12]. These two-stage methods are often accurate, but proposal generation and multistage prediction increase computational cost and limit real-time deployment on a UAV.

One-stage detectors improve efficiency by predicting classes and boxes directly. Redmon et al. introduced YOLO in 2016 and formulated classification and box regression as a unified end-to-end task [13]. In the same year, Liu et al. proposed SSD, which predicts targets of different sizes from feature maps at several resolutions [14]. Lin et al. introduced RetinaNet in 2017 and used focal loss to address the imbalance between foreground and background examples [15]. The YOLO family subsequently evolved through changes to its backbone, neck, head, and label-assignment strategy. YOLOX introduced anchor-free detection, a decoupled head, and SimOTA assignment in 2021 [16]. YOLOv9 used programmable gradient information to improve information flow in deep networks in 2024 [17]. YOLOv10 combined consistent dual assignment with end-to-end prediction to reduce dependence on non-maximum suppression [18]. Ultralytics later released YOLOv11 with further improvements in accuracy, efficiency, and task coverage [19]. Its balance of speed, size, and deployability makes YOLOv11 a suitable baseline for maritime UAV detection.

Transformer-based detection provides a different approach to global context modeling. Carion et al. proposed DETR in 2020, formulating detection as set prediction and using a Transformer to remove many hand-designed elements of conventional detectors [20]. Deformable DETR restricts attention to a small set of sampling points and improves convergence and small-object detection [21]. Despite their strong global modeling capacity, Transformer detectors can be expensive to train and deploy on constrained airborne devices.

Multiscale fusion addresses the loss of small-object information during repeated downsampling. Lin et al. proposed FPN in 2017, combining high-level semantics with low-level spatial detail through a top–down pathway and lateral connections [22]. PANet adds a bottom–up aggregation path to shorten the transfer of localization cues from shallow to deep layers [23]. BiFPN uses learnable weights to fuse features efficiently across scales [24]. GiraffeDet introduces GFPN, whose extensive cross-scale connections increase information exchange among feature levels [25]. These methods improve multiscale detection, but high-altitude maritime imagery still requires stronger interaction between weak shallow detail and deep semantic context.

Attention mechanisms offer another means of enhancing targets in cluttered scenes. SENet models dependencies among channels [26], CBAM applies channel and spatial attention sequentially [27], and ECA-Net reduces overhead through local cross-channel interaction [28]. Although these mechanisms can suppress some irrelevant background responses, sea glare, foam, and waves remain visually similar to people in the water. Local spatial or channel attention alone may therefore be insufficient.

Detection heads are also changing from fixed structures to adaptive ones. Dai et al. proposed DyHead in 2021, unifying scale-aware, spatial-aware, and task-aware attention within the detection head [29]. This design allows predictions to respond dynamically to changes in scale, position, and task. HPSNet builds on these developments by combining CTA for discriminative feature extraction, GFPN for cross-scale fusion, and DyHead for adaptive prediction.

3. Method

3.1. Overall Architecture

HPSNet uses YOLOv11 as its baseline and consists of an enhanced backbone, a GFPN neck, and a DyHead prediction head, as shown in Figure 1. The backbone contains convolutional layers, C3k2_FCA_CTA blocks, an SPPF block, and a C2PSA block. The C3k2_FCA_CTA block incorporates frequency channel attention (FCA) and CTA into the original C3k2 structure. FCA and CTA strengthen features around people in the water and suppress responses to waves, glare, foam, and floating debris. SPPF expands the receptive field, while C2PSA further refines high-level semantic representations.

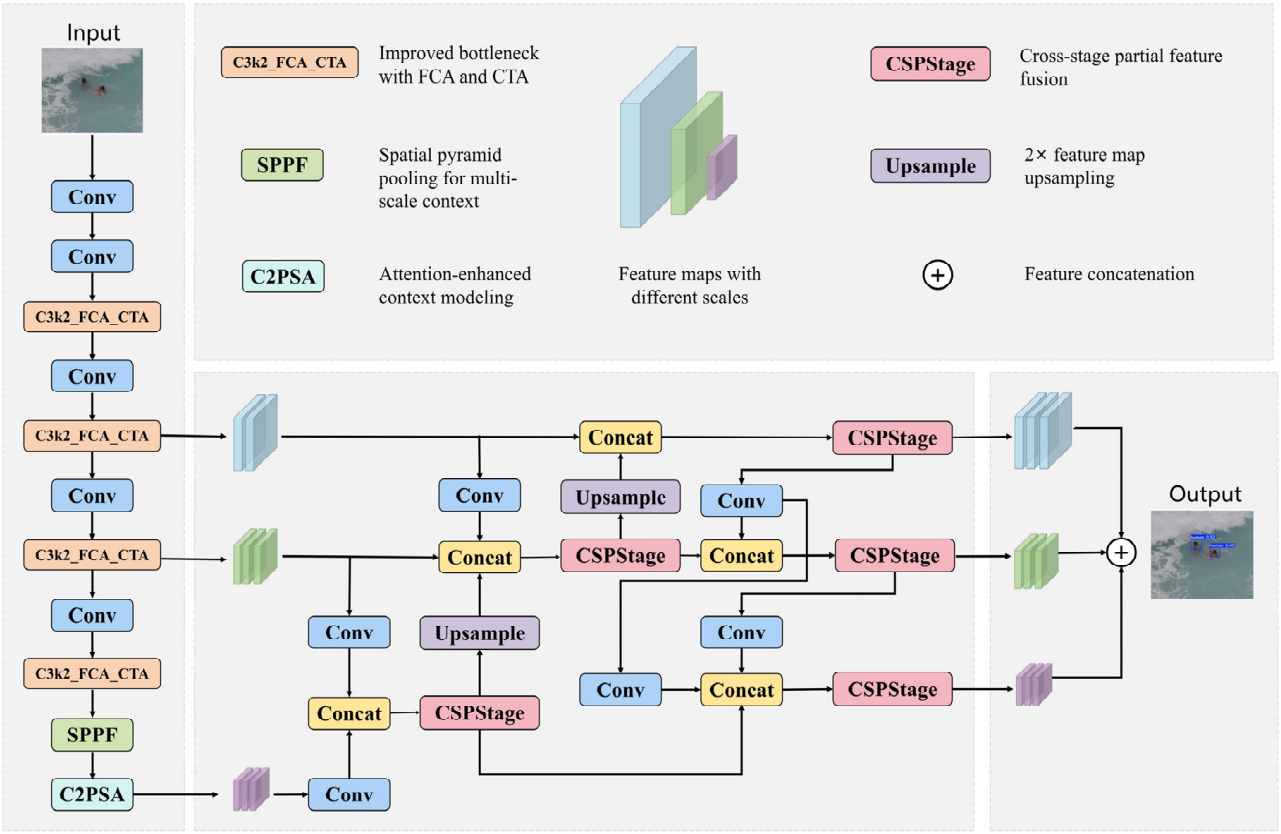


Figure 1. Overall architecture of HPSNet.

The neck replaces the original YOLOv11 fusion structure with GFPN. Through upsampling, downsampling, concatenation, and CSPStage blocks, GFPN creates multiple information paths between shallow, high-resolution features and deep, semantically rich features. The prediction head applies DyHead to the P3, P4, and P5 outputs of GFPN. Its scale-aware, spatial-aware, and task-aware mechanisms adjust the contribution of each feature level and improve robustness to changes in size, pose, and partial occlusion.

3.2. CTA Feature Enhancement

People in maritime UAV images are small, weakly textured, and easily confused with sea-surface patterns. HPSNet therefore introduces CTA into the YOLOv11 C3k2 block to form the C3k2_FCA_CTA feature-enhancement block. CTA models dependencies across channels, strengthens responses that are useful for person-overboard detection, and suppresses interference from waves, strong reflections, foam, and floating objects [30], as illustrated in Figure 2.

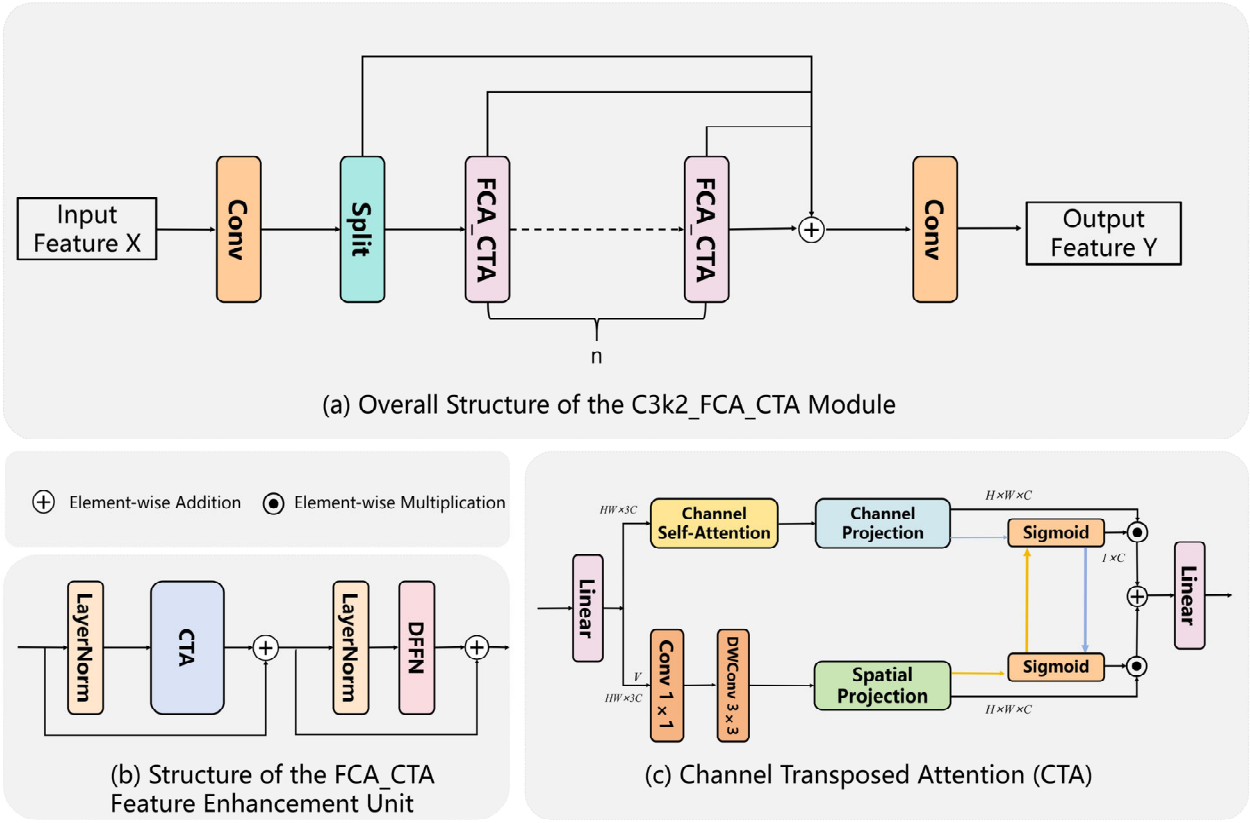


Figure 2. Structure of the CTA module.

Let the input feature map be $X \in R^{B\times C\times H\times W}$, where B, C, H, W denote batch size, channel count, height, and width. The module reshapes the spatial dimensions into a token sequence. It then applies LayerNorm and channel transposed attention with a residual connection:

$$X_1 = Flatten(X) \in R^{B\times H\times W\times C} \tag{1}$$

The residual path retains the original representation and reduces the risk of losing fine target detail. A detail-enhanced feed-forward network (DFFN) then performs nonlinear transformation and feature refinement:

$$X_1 = X_t + DropPath(CTA(LN(X_t))) \tag{2}$$

Finally, $Y \in R^{B\times C\times H\times W}$ is reshaped to $Y = X_1 + DropPath(DFFN(LN(X_1)))$. This design retains local visual cues while increasing global cross-channel interaction, which improves sensitivity to weak targets in complex maritime backgrounds.

3.3. GFPN Multiscale Fusion

A person in the water often occupies few pixels and has weak edges, texture, and shape. Conventional FPN and PAN structures fuse multiple levels, but most information flows between adjacent scales. Fine cues may therefore weaken after repeated downsampling. HPSNet replaces the original YOLOv11 neck with GFPN to provide more extensive cross-scale interaction.

As shown in Figure 3, GFPN repeatedly upsamples, downsamples, and concatenates feature maps from different levels. Let the three backbone outputs be P_3, P_4, and P_5. The high-resolution P_3 retains spatial detail, while P_5 carries stronger semantics and global context. GFPN first compresses the high-level features and fuses them with intermediate features. It then propagates semantic information upward to higher-resolution levels

and combines it with shallow and intermediate features. A subsequent bottom–up path aggregates the enhanced representations again. CSPStage blocks process each fusion stage and produce P_3, P_4, and P_5 for detection. These multiple paths allow the spatial detail needed to locate a small person and the semantics needed to distinguish that person from clutter to interact more fully.

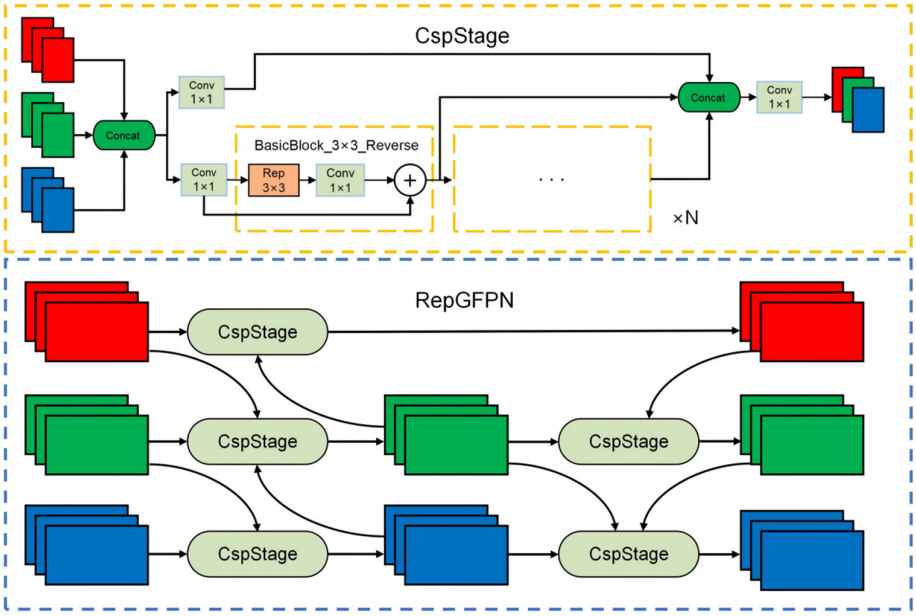


Figure 3. Structure of the GFPN module.

3.4. Dynamic Detection Head

A fixed convolutional head cannot adjust its behavior directly to changes in target scale, position, or appearance. HPSNet therefore uses DyHead, whose structure is illustrated in Figure 4, to model these variations dynamically. Let the multiscale GFPN output be:

$$X = \{X_3,\ X_4,\ X_5\} \tag{3}$$

where X_3, X_4, X_5 correspond to P_3, P_4, and P_5. For level l, a convolution predicts the offsets Δp and modulation mask m_l required by deformable convolution:

$$[\Delta p_l, m_l] = Conv(X_l) \tag{4}$$

DyDCNv2 then samples the current feature dynamically:

$$F_l^{mid} = DyDCN(X_l, \Delta p_l, m_l) \tag{5}$$

When adjacent levels exist, the head aligns and fuses them with the current level. A stride-2 dynamic convolution aligns the higher-resolution feature, while bilinear interpolation upsamples the lower-resolution feature after dynamic convolution. In general form, the fused feature is:

$$F_l = F_l^{mid} + F_l^{low} + F_l^{high} \tag{6}$$

which includes a neighboring branch only when that level exists. Scale attention assigns an adaptive weight to each fused level:

$$S_l = \sigma(Conv(GAP(F_l))) \tag{7}$$

$$\hat{F}_l = F_l \cdot S_l \tag{8}$$

Finally, *DyReLU* provides task-aware nonlinear transformation:

$$Y_l = DyReLU(\hat{F}_l) \tag{9}$$

DyHead thus adjusts spatial sampling, cross-scale fusion, and task responses within a unified head. This dynamic modeling is well suited to small, partially occluded people whose position and pose vary across maritime UAV images.

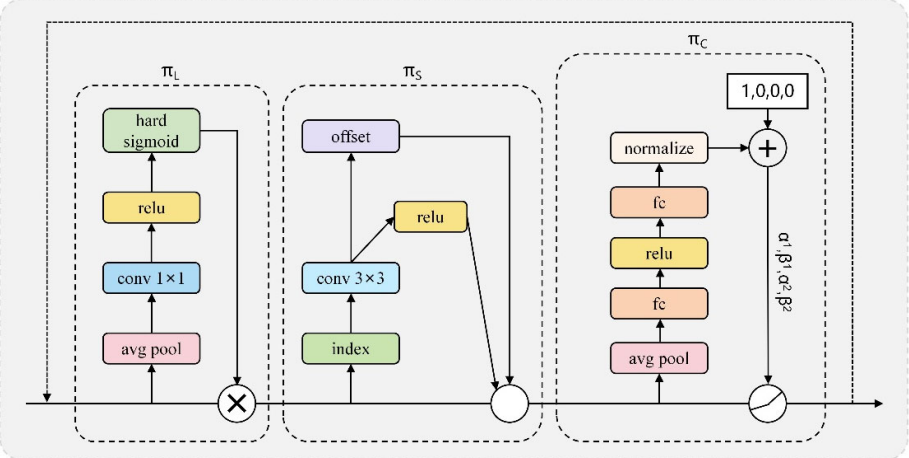


Figure 4. Structure of the DyHead detection module.

4. Experiments and Results

4.1. Datasets and Experimental Design

1. The Person Detection in Water dataset is a large benchmark for visual detection and tracking. It contains top-view UAV images captured at altitudes from 5 to 260 m and viewing angles from 0 to 90 degrees, with acquisition parameters such as altitude and angle recorded for each image. No constraint is imposed on the appearance of the targets: persons wearing ordinary clothing and persons wearing a life jacket are both present in the data and are annotated uniformly as a single person class; so, the detector is not tuned to any particular clothing condition. The images are used at their original acquisition quality and are resized to 640 × 640 for training and inference; no minimum-quality filtering, deblurring, or manual selection of favorable frames is applied. The subset used in this study contains 4281 images and 24,131 annotated objects. Following relabeling to reduce the effects of class imbalance, the original competition split is retained: 2975 images for training, 859 for validation, and 447 for testing. Because subsequent detection performance depends on the quality of the initial imagery, Gaussian noise is added at random to 45% of the validation images to examine robustness to image degradation.
2. The AFO dataset is designed for training machine learning and deep learning models for maritime search and rescue. It contains approximately 2904 manually annotated floating-object instances from aerial UAV video. Many targets are small and difficult to detect. The source split assigns 70% of the data to training, 20% to validation, and 10% to testing. Gaussian noise is also added at random to 45% of the validation images.

4.2. Implementation Details and Evaluation Metrics

The method is implemented in PyTorch2.1 and trained on Ubuntu 18.04. The allocated system provides 15 CPU cores on an AMD EPYC 7642 processor, 80 GB of system memory, and one NVIDIA RTX 4090 GPU with 24 GB of memory. Table 1 lists the main training settings. Neither pretrained weights nor automatic mixed precision is used.

Table 1. Training settings.

Setting	Value
Initial learning rate	0.01
Momentum	0.937
Training epochs	400
Batch size	32
Input resolution	640 × 640
Pretrained weights	None
Automatic mixed precision	Disabled

Performance is evaluated with precision, recall, average precision, mean average precision, parameter count, floating-point operations, and inference time. Precision measures the fraction of predicted positives that are correct, and recall measures the fraction of ground-truth positives that are detected:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

where TP, FP, and FN denote true positives, false positives, and false negatives. For class c, average precision is the area under the precision–recall curve:

$$AP = \int_0^1 P(R)dR \tag{12}$$

Mean average precision averages this quantity over classes:

$$mAP = \frac{1}{n}\sum_{i=1}^{n}(AP)_i \tag{13}$$

This study reports mAP@50 at an intersection-over-union threshold of 0.50 and mAP@50:95 averaged over thresholds from 0.50 to 0.95 in increments of 0.05. Parameter count measures model size, whereas FLOPs estimate the number of floating-point operations required for one forward pass. Inference time is measured in milliseconds per image on the stated hardware.

4.3. Ablation Study

The ablation study uses Person Detection in Water and introduces GFPN, CTA, and DyHead incrementally into the YOLOv11 baseline. Each variant differs from its predecessor by exactly one component, so that the marginal contribution of each stage can be read as a controlled comparison between two adjacent rows rather than as an isolated single-module result. All variants follow the same training and testing protocol. Table 2 reports the results.

Table 2. Ablation results on Person Detection in Water.

Model	Exp.	GFPN	CTA	DyHead	mAP@50	mAP@50:95	P	R	Params (M)	FLOPs (G)
YOLOv11n	1	×	×	×	0.755	0.353	0.792	0.63	2.58	6.3
YOLOv11n + GFPN	2	√	×	×	0.763	0.357	0.79	0.636	3.65	8.2
YOLOv11n + CTA + GFPN	3	√	√	×	0.77	0.363	**0.796**	0.641	3.66	8.4
YOLOv11n + GFPN + DyHead	4	√	×	√	0.76	0.358	0.784	0.616	4.17	9.3
HPSNet	5	√	√	√	**0.782**	**0.375**	0.757	**0.672**	**4.18**	**9.6**

√ indicates that the module is adopted; × indicates that the module is not used. Bold values indicate the best results.

GFPN increases mAP@50 from 75.5% to 76.3%, mAP@50:95 from 35.3% to 35.7%, and recall from 63.0% to 63.6%. This improvement raises the parameter count from 2.58 M to 3.65 M and the computation from 6.3 to 8.2 GFLOPs. Adding CTA to GFPN produces a further gain, reaching 77.0% mAP@50, 36.3% mAP@50:95, 79.6% precision, and 64.1% recall. Relative to GFPN alone, these changes correspond to gains of 0.7, 0.6, 0.6, and 0.5 percentage points, respectively.

Combining GFPN with DyHead gives 76.0% mAP@50 and 35.8% mAP@50:95. This variant improves both mAP measures over the baseline but reduces recall to 61.6%, indicating that DyHead does not provide a uniform benefit without CTA. The full HPSNet model produces the strongest mAP and recall in the ablation study, reaching 78.2% mAP@50, 37.5% mAP@50:95, and 67.2% recall. Relative to YOLOv11, the gains are 2.7, 2.2, and 4.2 percentage points. Precision decreases from 79.2% to 75.7%, however, which indicates a trade-off between retrieving more targets and avoiding false positives. The ablation results therefore support the complementary effect of the three modules on mAP and recall, but they do not show that every metric improves.

In the low-altitude examples in Figure 5, people are relatively clear but remain small and partly similar in color to the surrounding water. YOLOv11 detects the targets, yet some boxes overlap, shift away from the target, or have low confidence. GFPN makes localization more concentrated through stronger multiscale fusion. Adding CTA further focuses the representation on relevant regions, while DyHead improves adaptation to scale and appearance. The complete model produces boxes that align more closely with the targets and provides more stable confidence across the examples.

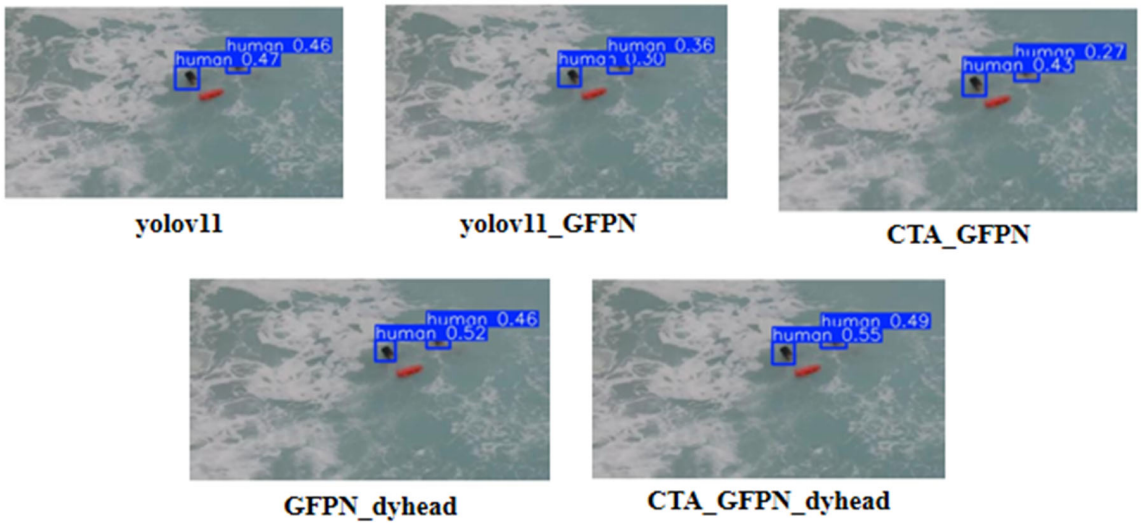


Figure 5. Detection results of the ablation variants in low-altitude scenes.

The high-altitude examples in Figure 6 contain fewer target pixels and stronger interference from waves, foam, and illumination. YOLOv11 finds the people but localizes some boundaries inconsistently and assigns low confidence. GFPN alone does not fully recover the weakened cues. CTA reduces part of the background response, and DyHead improves localization and confidence by adapting to weak targets at different scales. The complete model identifies both people with confidence scores of 0.55 and 0.49, which are higher than those of most other ablation variants in the illustrated example. This qualitative result is consistent with the quantitative gain in recall, although it does not by itself establish general performance.

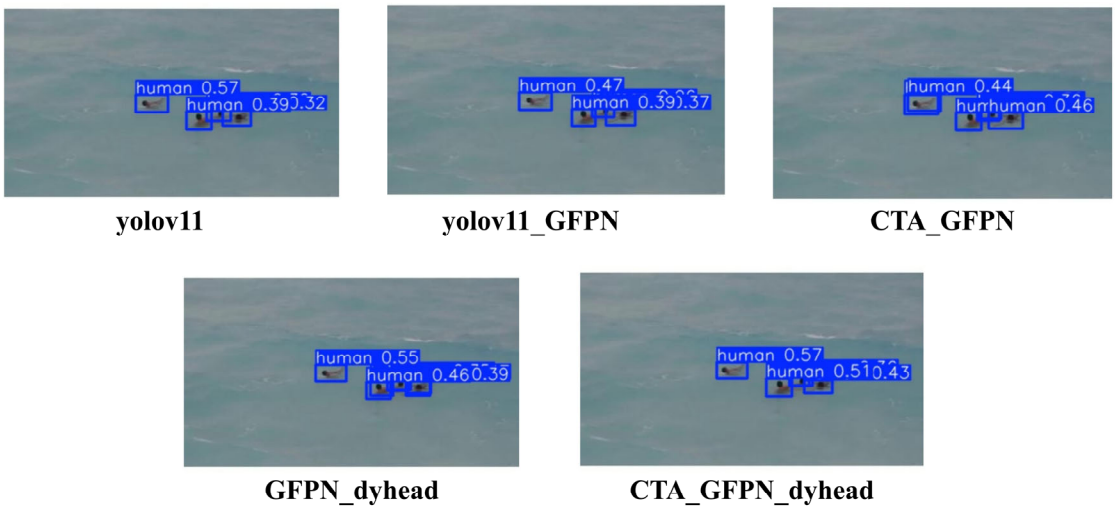


Figure 6. Detection results of the ablation variants in high-altitude scenes.

4.4. Comparison with Representative Detectors

4.4.1. Person Detection in Water

Table 3 compares HPSNet with representative YOLO-family, Transformer-based, and two-stage detectors on the Person Detection in Water test set. All models are trained and evaluated using the same dataset split, input resolution, training schedule, and evaluation protocol.

Table 3. Comparison on Person Detection in Water.

Method	mAP@50	mAP@50:95	Parameters	FLOPs
DETR	52.3%	27.5%	36.7	100.9
Faster-RCNN	49.6%	25.8%	41	208.1
YOLOv8n	75.7%	35.1%	3.01	8.1
YOLOv9t	72.2%	33.7%	2.61	10.4
YOLOv10n	75.7%	34.8%	2.3	6.7
YOLOv11n	75.5%	35.3%	2.58	6.3
YOLOv12n	75.2%	34.6%	2.51	5.8
YOLO26n	76.4%	36.6%	2.37	5.2
YOLOX	76.4%	29.1%	5.03	7.58
SuperYOLO	73.3%	28%	4.82	16.6
DAMO-YOLO	80.1%	36.4%	28.2	61.8
Cascade-RCNN	71.7%	25.4%	69.15	236
HPSNet	78.2%	37.5%	4.18	9.6

HPSNet achieves 78.2% mAP@50 and 37.5% mAP@50:95, improving the YOLOv11 baseline by 2.7 and 2.2 percentage points, respectively. HPSNet outperforms all nano- and tiny-scale YOLO-family detectors (YOLOv8n, YOLOv9t, YOLOv10n, YOLOv12n, YOLO26n, YOLOX, and SuperYOLO) in terms of mAP@50, and attains the highest mAP@50:95 (37.5%) among all evaluated YOLO-family detectors, exceeding the second-ranked YOLO26n (36.6%) by 0.9 percentage points. Only DAMO-YOLO achieves a slightly higher mAP@50 (80.1%), but it does so with 28.2 M parameters and 61.8 GFLOPs—more than six times the parameters and computation of HPSNet—while delivering lower mAP@50:95 (36.4%).

To examine whether the accuracy improvement is obtained at a reasonable computational cost, Figure 7 visualizes the relationship among mAP@50:95, FLOPs, and parameter count for the evaluated YOLO-family and YOLO-derived detectors.

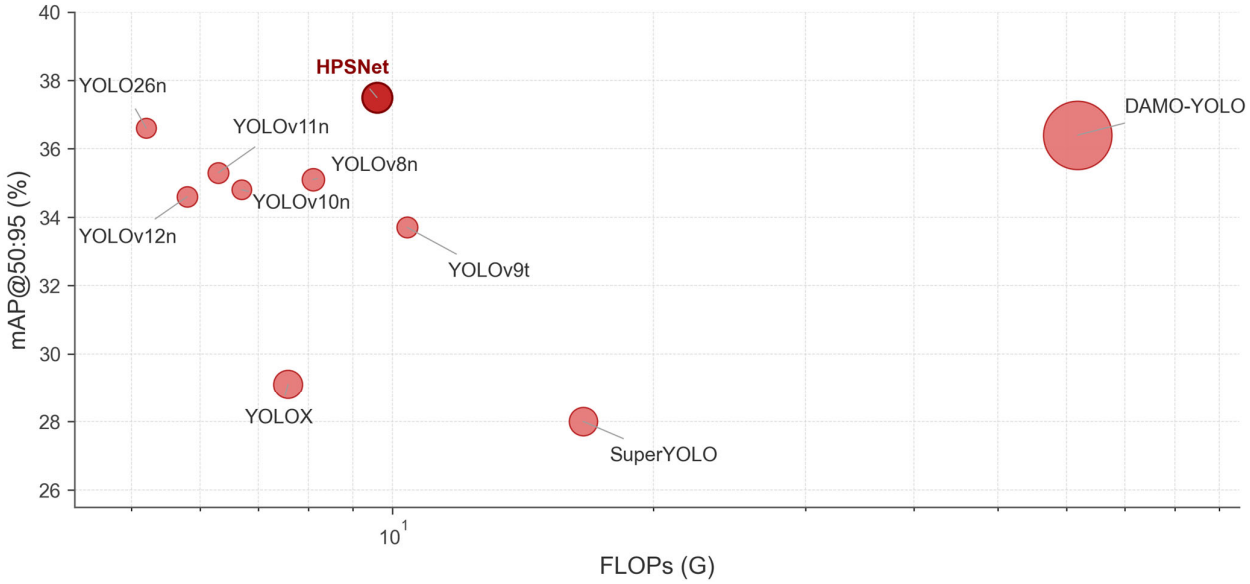


Figure 7. Accuracy–complexity trade-off of YOLO-family detectors on Person Detection in Water.

As shown in Figure 7, YOLO26n and HPSNet form the accuracy–computation Pareto frontier among the evaluated YOLO-family detectors. YOLO26n represents the lowest-computation operating point, requiring 5.2 GFLOPs and achieving 36.6% mAP@50:95. HPSNet provides the highest accuracy, reaching 37.5% mAP@50:95 at 9.6 GFLOPs. All remaining detectors are dominated: each either requires more computation for lower accuracy or attains lower accuracy at comparable cost. HPSNet therefore achieves the best mAP@50:95 among the evaluated YOLO-family detectors while maintaining a moderate scale of 4.18 M parameters and 9.6 GFLOPs.

To further examine the influence of the confidence threshold, the nano-scale detectors YOLOv8n, YOLOv10n, YOLOv11n, YOLO26n, and HPSNet were evaluated using the same test images and inference settings. Predictions were retained at a minimum confidence threshold of 0.001, and true positives were determined at an IoU threshold of 0.5. The confidence threshold was then varied from 0 to 1 to generate the Precision–Recall, F1–Confidence, Recall–Confidence, and Recall–FPPI curves shown in Figure 8.

As shown in Figure 8a, HPSNet maintains a favorable precision–recall relationship over most of the medium- and high-recall regions. The F1–Confidence curves in Figure 8b show that the optimal operating threshold differs among the evaluated detectors, indicating that performance comparisons based on a single common confidence threshold may be incomplete. All curves rise over the low-threshold interval and exhibit a local increase at a confidence value of approximately 0.2. At very low thresholds, a large number of low-confidence false positives are retained; so, precision is low and F1 is suppressed even though recall is close to its maximum. As the threshold increases toward 0.2, these low-confidence false positives are progressively discarded, while most true positives are preserved; so, precision recovers faster than recall declines and F1 therefore increases. Beyond this region, the removal of correct detections begins to dominate and the two terms decrease together. The effect is more pronounced for HPSNet because its recall-oriented design produces a larger population of low-confidence candidate detections. Although HPSNet does not achieve the highest maximum F1 score, it remains competitive over a relatively broad threshold range.

Figure 8c shows that HPSNet generally retains higher recall over a broad range of confidence thresholds, which is consistent with its recall-oriented design. As the threshold increases, the recall of all detectors decreases because more low-confidence detections are re-

moved. The comparatively gradual decline in HPSNet indicates that its recall improvement is not restricted to a single confidence setting.

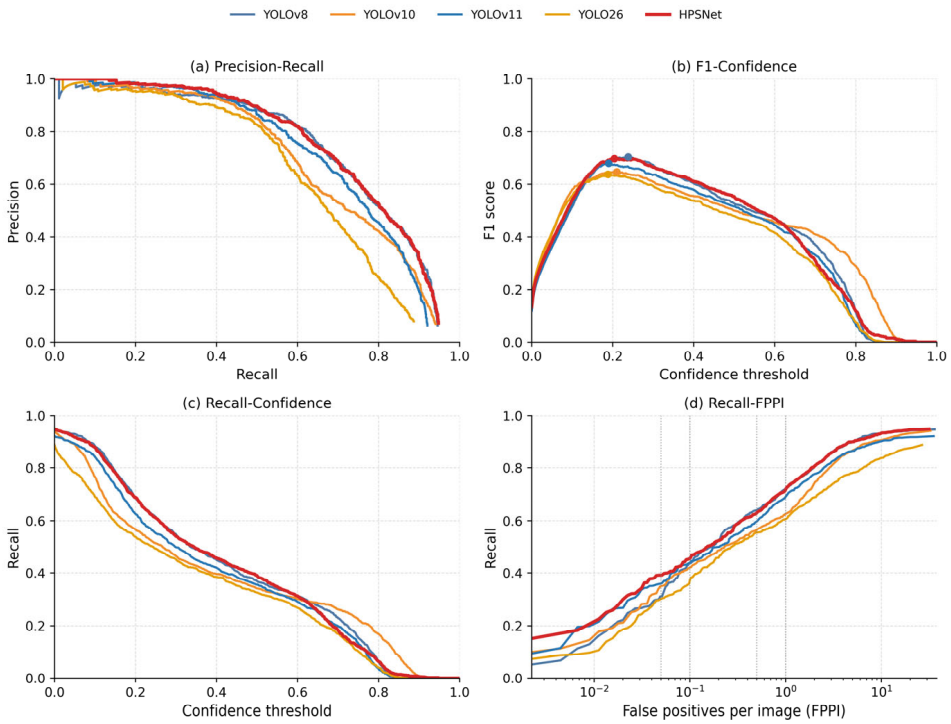


Figure 8. Threshold-sensitivity analysis of selected detectors on Person Detection in Water: (**a**) Precision–Recall, (**b**) F1–Confidence, (**c**) Recall–Confidence, and (**d**) Recall–FPPI curves.

The Recall–FPPI curves in Figure 8d provide a more operationally relevant comparison by controlling the number of false alarms per image. HPSNet achieves comparatively high recall across most of the evaluated FPPI range and generally outperforms the YOLOv11 baseline under the same false-positive constraint. This result indicates that the recall improvement in HPSNet is not obtained solely by accepting an uncontrolled increase in false detections.

Overall, the threshold analysis characterizes HPSNet as a recall-oriented detector rather than a model that uniformly improves every metric. This property is relevant to maritime search and rescue, where missed persons may have more serious consequences than a manageable number of additional alarms. Nevertheless, the operating confidence threshold should be selected according to the false-alarm tolerance of the deployment system.

Relative to the YOLOv11 baseline, HPSNet increases mAP@50:95 from 35.3% to 37.5%, corresponding to a gain of 2.2 percentage points. This improvement is accompanied by an increase in computational cost from 6.3 to 9.6 GFLOPs and an increase in parameter count from 2.58 M to 4.18 M. These changes correspond to increases of approximately 52.4% in FLOPs and 62.0% in parameters. HPSNet should therefore be interpreted as an accuracy- and recall-oriented enhancement in YOLOv11 rather than an unconditionally more efficient detector.

4.4.2. AFO

The AFO experiment examines whether the model remains effective on a multiclass maritime dataset. Table 4 gives the results.

HPSNet achieves the highest mAP in Table 4, with 88.4% mAP@50 and 59.4% mAP@50:95. Relative to YOLOv11, these values represent gains of 0.7 and 1.9 percentage points. HPSNet contains 4.18 M parameters and requires 9.6 GFLOPs. Its measured in-

ference time is 13.4 ms per image on the RTX 4090, which corresponds to approximately 74.6 images per second without accounting for data transfer and other system overhead. This result demonstrates efficient desktop-GPU inference, but deployment performance on an embedded UAV platform requires separate validation.

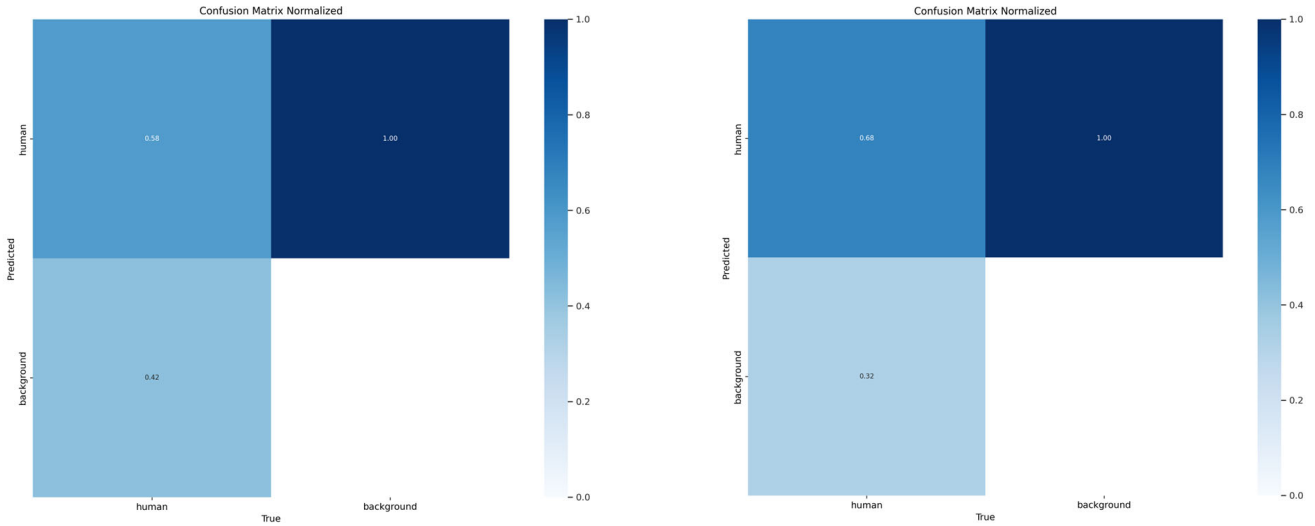
Table 4. Comparison on AFO.

Method	mAP@50	mAP@50:95	Parameters	FLOPs
DETR	57.1%	32%	36.7	100.9
Faster-RCNN	55.3%	28.6%	41	208.1
YOLOv8n	82.2%	50.5%	3.01	8.1
YOLOv9t	84.3%	52.6%	2.61	10.4
YOLOv10n	86.6%	56.4%	2.3	6.7
YOLOv11n	87.7%	57.5%	2.58	6.3
YOLOv12n	84.9%	56.8%	2.51	5.8
YOLO26n	87.2%	55.6%	2.37	5.2
YOLOX	72.4%	28%	5.03	7.58
SuperYOLO	86.5%	54.8%	4.82	16.6
DAMO-YOLO	79.8%	41.9%	28.2	61.8
Cascade-RCNN	76.2%	48.4%	69.15	236
HPSNet	88.4%	59.4%	4.18	9.6

4.5. Visualization Analysis

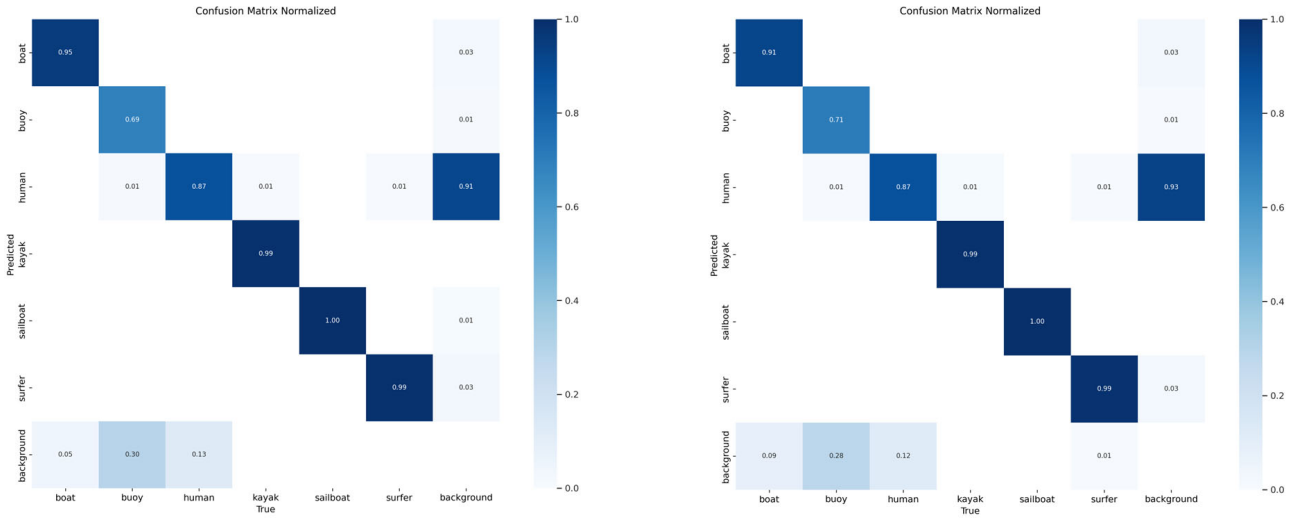
4.5.1. Confusion Matrices

Normalized confusion matrices are generated for YOLOv11 and HPSNet on the same test sets and with the same evaluation settings. In Figures 9 and 10, the horizontal axis denotes the ground-truth class and the vertical axis denotes the predicted class. Diagonal values give the fraction of correctly detected instances. The background row represents ground-truth targets that do not match a prediction, while the background column represents predictions that do not match a ground-truth target. Because normalization is performed by ground-truth class, stronger detection is reflected by a larger diagonal value and a smaller value in the background row.



(**a**) YOLOv11 (**b**) HPSNet

Figure 9. Normalized confusion matrices of (**a**) YOLOv11 and (**b**) HPSNet on the Person Detection in Water test set.



(**a**) YOLOv11 (**b**) HPSNet

Figure 10. Normalized confusion matrices of (**a**) YOLOv11 and (**b**) HPSNet on the AFO test set.

On Person Detection in Water, the correct-detection rate increases from 0.58 for YOLOv11 to 0.68 for HPSNet, while the missed-detection rate decreases from 0.42 to 0.32. These changes correspond to an absolute improvement of 10 percentage points and a relative reduction of approximately 23.8% in missed detections. The result is consistent with stronger representations of small, weakly textured people under waves, foam, illumination changes, and partial immersion.

On AFO, the correct-detection rates of YOLOv11 for the boat, buoy, human, kayak, sailboat, and surfer classes are 0.95, 0.69, 0.87, 0.99, 1.00, and 0.99, respectively. The corresponding HPSNet values are 0.91, 0.71, 0.87, 0.99, 1.00, and 0.99. HPSNet improves the buoy class by 2 percentage points and maintains the values for human, kayak, sailboat, and surfer, but it reduces the boat value by 4 points. The missed-detection rate decreases from 0.30 to 0.28 for buoy and from 0.13 to 0.12 for human. These results indicate a benefit for small maritime targets but also reveal category-specific variation.

Taken together, the matrices show a clear reduction in missed detections on the single-class Person Detection in Water task. On multiclass AFO, most class-level results remain stable and buoy and human misses decrease, but the boat class deteriorates. The three added modules therefore do not benefit all maritime categories equally.

4.5.2. Grad-CAM Analysis

Grad-CAM visualizations examine where each model responds in representative high-altitude, wave-cluttered, and multi-target scenes. In Figure 11, the color changes from blue to yellow and red as the response strength increases. Relative to YOLOv11, HPSNet concentrates more of its high response on people in the water and their immediate surroundings, while its response to waves, foam, and water texture is weaker. This pattern is consistent with the intended functions of the architecture: CTA enhances target-related channels, GFPN retains spatial information for small targets, and DyHead adapts the prediction response to changes in scale and position. The visualization is qualitative and does not establish a causal contribution for any individual module.

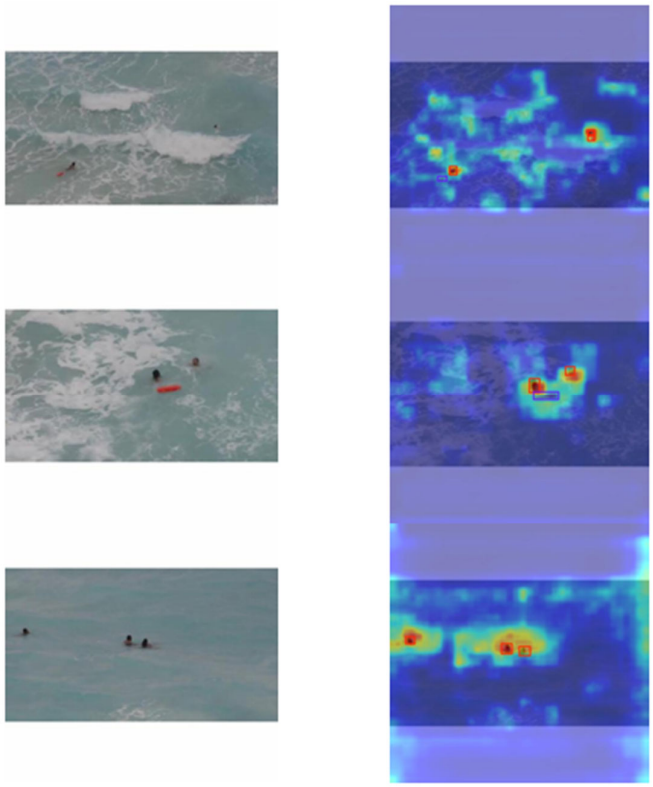


Figure 11. Grad-CAM visualizations in representative maritime scenes.

5. Discussion

The results suggest that effective person-overboard detection depends less on simply enlarging the network than on coordinating feature extraction, multiscale fusion, and adaptive prediction. GFPN preserves spatial detail for small targets, CTA reduces confusing sea-surface responses, and DyHead adjusts predictions across scales and locations. The complete model gives the strongest mAP and recall in the ablation study, which supports a complementary relationship among the three stages. The results also show a precision–recall trade-off: HPSNet raises recall from 63.0% to 67.2% but lowers precision from 79.2% to 75.7%. For search and rescue, higher recall may be desirable because a missed person can be especially costly, but the increase in false alarms still requires consideration in operational use.

The different gains on the two datasets further clarify the scope of the method. HPSNet improves mAP@50 over YOLOv11 by 2.7 percentage points on Person Detection in Water and by 0.7 points on AFO. The former focuses directly on people in complex water backgrounds and is closely aligned with the design goal of HPSNet. AFO includes boat, buoy, human, kayak, sailboat, and surfer, whose scales, shapes, and textures differ substantially. The smaller aggregate gain and the class-level confusion matrix indicate that the method is more specifically suited to people and weak small maritime targets than to every maritime category.

The AFO confusion matrix makes this limitation explicit. HPSNet improves the correct-detection rate for buoy and reduces the missed-detection rate for human, but the correct-detection rate for boat decreases from 0.95 to 0.91. Stronger multiscale fusion and dynamic feature adjustment may change responses that are already effective for objects with complete contours and clear structure. HPSNet should therefore be viewed as a targeted improvement for person-overboard and weak small-object detection, not as a universally beneficial structure for all maritime classes.

The Grad-CAM results qualitatively support this interpretation. YOLOv11 responds more strongly to waves, foam, and bright reflections, whereas HPSNet concentrates more response near people. This change is compatible with enhanced target responses and suppressed background interference, but attention maps alone cannot prove that a particular module causes a performance gain. The quantitative comparisons and controlled ablations provide the primary evidence.

The Pareto analysis in Figure 7 further clarifies the position of HPSNet among the evaluated YOLO-family detectors. HPSNet occupies the accuracy-oriented end of the Pareto frontier: it attains the highest mAP@50:95 (37.5%) at 9.6 GFLOPs, whereas YOLO26n offers the lowest computational cost (5.2 GFLOPs) with slightly lower accuracy (36.6%). Its advantage therefore lies in the best accuracy–cost combination among the comparably sized nano-scale detectors.

The accuracy and recall gains come at a moderate computational cost, as quantified in Section 4.4 (4.18 M parameters and 9.6 GFLOPs). Nevertheless, HPSNet retains a moderate model scale, and its inference time of 13.4 ms per image on an RTX 4090 demonstrates efficient desktop-GPU processing and provides an initial basis for investigating embedded deployment. Future work will evaluate HPSNet on resource-constrained UAV computing platforms, such as NVIDIA Jetson Orin Nano and Orin NX, using batch-size-1 inference at an input resolution of 640 $\times$ 640. FP16 and INT8 TensorRT implementations will be examined in terms of mean and P95 latency, throughput, peak memory consumption, power use, energy per frame, and thermal stability. End-to-end latency, including image preprocessing, data transfer, model inference, and post-processing, will also be measured under continuous video processing. In addition, structured pruning, knowledge distillation, and hardware-aware optimization will be investigated to reduce computational cost while preserving the recall required for person-overboard detection.

6. Conclusions

This paper presents HPSNet, a YOLOv11-based detector for person-overboard detection in maritime UAV imagery. HPSNet integrates CTA into the backbone to enhance target-related features and suppress sea-surface interference, employs GFPN to improve the interaction between shallow spatial details and deep semantic information, and uses DyHead to adapt predictions to variations in target scale, position, and appearance. By coordinating feature enhancement, multiscale fusion, and adaptive prediction at different stages of the detection pipeline, HPSNet is designed to address weak small-target features, complex maritime backgrounds, and substantial appearance variations.

On Person Detection in Water, HPSNet achieves 78.2% mAP@50, 37.5% mAP@50:95, and 67.2% recall, improving the YOLOv11 baseline by 2.7, 2.2, and 4.2 percentage points, respectively. On AFO, it achieves 88.4% mAP@50 and 59.4% mAP@50:95, corresponding to improvements of 0.7 and 1.9 percentage points over YOLOv11. The ablation results indicate that the joint integration of CTA, GFPN, and DyHead improves mAP and recall, while the confusion-matrix analysis shows fewer missed detections for the person and buoy classes. However, the recall improvement in Person Detection in Water is accompanied by a decrease in precision from 79.2% to 75.7%. In addition, HPSNet requires a moderate model scale of 4.18 M parameters and 9.6 GFLOPs. HPSNet should therefore be interpreted as a targeted, recall-oriented enhancement in the YOLOv11 framework rather than a detector that uniformly improves every metric. Its inference time of 13.4 ms per image on an RTX 4090 demonstrates efficient desktop-GPU processing and provides a basis for further embedded-deployment investigation.

Future work will focus on the deployment and optimization of HPSNet on resource-constrained maritime UAV platforms. The model will be evaluated on NVIDIA Jetson-

class devices using batch-size-1 inference and FP16 and INT8 TensorRT implementations. Mean and P95 latency, throughput, peak memory consumption, power use, energy per frame, thermal stability, and end-to-end video-processing latency will be measured to identify suitable accuracy–efficiency operating points for different UAV resource budgets. Structured pruning, knowledge distillation, low-bit quantization, and hardware-aware optimization will be investigated to reduce computational cost while preserving detection recall. Future experiments will also extend the evaluation to low-light conditions, strong glare, sea fog, rain, severe occlusion, and camera shake; conduct cross-dataset validation; and integrate infrared imagery and temporal video information to improve robustness in practical maritime search-and-rescue operations. Thermal imaging, which exploits the temperature contrast between the human body and cold water, is expected to further simplify the recognition of partially submerged persons—for instance, when only the head is visible above the surface—and to accelerate the rescue decision-making process; integrating such modalities is therefore a key direction for future work.

Author Contributions: Conceptualization, Y.R. and G.W.; Methodology, G.W.; Formal analysis, G.W. and L.Z.; Data curation, G.W.; Writing—original draft preparation, G.W.; Writing—review and editing, Y.R.; Resources, X.F. and Y.H. Funding acquisition, X.F. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Agriculture and Rural Affairs of China grant number B050102.

Data Availability Statement: The datasets used in this study are publicly available. The Person Detection in Water dataset is available at https://universe.roboflow.com/drowning-detection/person-detection-in-water/browse (accessed on 14 August 2025), and the AFO dataset is available at https://universe.roboflow.com/cscie25-project/afo---aerial-dataset-of-floating-objects/dataset/7 (accessed on 17 August 2025). Further details are provided in Section 4.1.

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

References

1. Ministry of Agriculture, Fisheries and Fishery Administration Bureau. *China Fishery Vessel Safety Analysis Report (1994–2015)*; China Agriculture Press: Beijing, China, 2017. (In Chinese)
2. Varga, L.A.; Kiefer, B.; Messmer, M.; Zell, A. SeaDronesSee: A maritime benchmark for detecting humans in open water. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 3686–3696.
3. Cafarelli, D.; Ciampi, L.; Vadicamo, L.; Gennaro, C.; Berton, A.; Paterni, M.; Benvenuti, C.; Passera, M.; Falchi, F. MOBDrone: A drone video dataset for man overboard rescue. In *Image Analysis and Processing—ICIAP 2022*; Springer: Cham, Switzerland, 2022.
4. Kiefer, B.; Kristan, M.; Perš, J.; Žust, L.; Poiesi, F.; Andrade, F.; Bernardino, A.; Dawkins, M.; Raitoharju, J.; Quan, Y.; et al. 1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2–7 January 2023.
5. Gonçalves, L.; Damas, B. Automatic detection of rescue targets in maritime search and rescue missions using UAVs. In Proceedings of the 2022 International Conference on Unmanned Aircraft Systems (ICUAS), Dubrovnik, Croatia, 21–24 June 2022; pp. 1638–1643.
6. Liu, T.; Pang, B.; Zhang, L.; Yang, W.; Sun, X. Sea surface object detection algorithm based on YOLOv4 fused with reverse depthwise separable convolution (RDSC) for USV. *J. Mar. Sci. Eng.* **2021**, *9*, 753. [CrossRef]
7. Zhao, Y.R.; Guo, H.M.; Jiao, H.; Zhang, J.W. Application of YOLOv4 integrated with hybrid domain attention mechanism in ship detection. *Comput. Mod.* **2021**, *9*, 75–82. (In Chinese)
8. Zhou, J.; Jiang, P.; Zou, A.; Chen, X.; Hu, W. Ship target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2021**, *9*, 908. [CrossRef]
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

10. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
12. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot MultiBox detector. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 21–37.
15. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
16. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430.
17. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 1–21.
18. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [CrossRef]
19. Ultralytics. Ultralytics YOLO11 Documentation. Available online: https://docs.ultralytics.com/models/yolo11/ (accessed on 14 August 2026).
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
21. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
22. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
23. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
24. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
25. Jiang, Y.; Tan, Z.; Wang, J.; Sun, X.; Lin, M.; Li, H. GiraffeDet: A heavy-neck paradigm for object detection. *arXiv* **2022**, arXiv:2202.04256.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
27. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 3–19.
28. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
29. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
30. Dai, T.; Wang, J.; Guo, H.; Li, J.; Wang, J.; Zhu, Z. FreqFormer: Frequency-aware Transformer for Lightweight Image Super-resolution. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), Jeju, Republic of Korea, 3–9 August 2024; pp. 781–789.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.