remote sensing

MDPI

Article

SR-DETR: Target Detection in Maritime Rescue from UAV Imagery

Yuling Liu and Yan Wei *

College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China; 2023110516026@stu.cqnu.edu.cn

* Correspondence: weiyan@cqnu.edu.cn

Abstract: The growth of maritime transportation has been accompanied by a gradual increase in accident rates, drawing greater attention to the critical issue of man-overboard incidents and drowning. Traditional maritime search-and-rescue (SAR) methods are often constrained by limited efficiency and high operational costs. Over the past few years, drones have demonstrated significant promise in improving the effectiveness of search-and-rescue operations. This is largely due to their exceptional ability to move freely and their capacity for wide-area monitoring. This study proposes an enhanced SR-DETR algorithm aimed at improving the detection of individuals who have fallen overboard. Specifically, the conventional multi-head self-attention (MHSA) mechanism is replaced with Efficient Additive Attention (EAA), which facilitates more efficient feature interaction while substantially reducing computational complexity. Moreover, we introduce a new feature aggregation module called the Cross-Stage Partial Parallel Atrous Feature Pyramid Network (CPAFPN). By refining spatial attention mechanisms, the module significantly boosts cross-scale target recognition capabilities in the model, especially offering advantages for detecting smaller objects. To improve localization precision, we develop a novel loss function for bounding box regression, named Focaler-GIoU, which performs particularly well when handling densely packed and small-scale objects. The proposed approach is validated through experiments and achieves an mAP of 86.5%, which surpasses the baseline RT-DETR model's performance of 83.2%. These outcomes highlight the practicality and reliability of our method in detecting individuals overboard, contributing to more precise and resource-efficient solutions for real-time maritime rescue efforts.

Keywords: target detection; maritime search and rescue; drones; computer vision

check for updates

Academic Editor: Ming-Der Yang

Received: 22 April 2025
Revised: 1 June 2025
Accepted: 10 June 2025
Published: 12 June 2025

Citation: Liu, Y.; Wei, Y. SR-DETR: Target Detection in Maritime Rescue from UAV Imagery. *Remote Sens.* **2025**, *17*, 2026. https://doi.org/10.3390/rs17122026

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As global economic integration deepens, maritime transport has emerged as a vital pillar underpinning numerous aspects of modern industrial and commercial activity. However, the increasing scale and frequency of maritime activities have also led to a corresponding rise in the incidence of maritime accidents. Such incidents not only result in damage to ships and offshore infrastructure but also cause substantial human casualties and economic losses. Notably, the escalating phenomenon of maritime migration in recent years has further exacerbated drowning incidents at sea. Therefore, promptly and precisely identifying overboard individuals during maritime SAR missions is crucial to reducing the risk of loss of life.

Conventional SAR methods primarily rely on the deployment of rescue vessels and helicopters. However, rescue vessels are hindered by their relatively slow response speed,

while helicopters have limited endurance and a constrained field of vision. These limitations collectively result in SAR operations that are both time-consuming and resource-intensive. The unique features of UAVs—including portability, operational flexibility, and affordability—have led to their extensive deployment in various sectors such as the military, agriculture, and industry [1]. UAVs possess the capability to rapidly survey extensive maritime areas and, in comparison to traditional SAR platforms, offer significant benefits, including low-altitude flight, rapid deployment, and reduced operational expenses. Integrating UAV platforms with deep learning object detection facilitates the extraction of critical targets from cluttered maritime environments while minimizing background interference, offering a viable improvement over traditional SAR techniques in terms of operational effectiveness and reliability.

Conventional approaches predominantly rely on manually engineered feature descriptors coupled with sliding window-based region proposal mechanisms for target identification. In the feature extraction stage, common practices involve using low-level visual descriptors like Histogram of Oriented Gradients (HOG) [2] and Scale-Invariant Feature Transform (SIFT) [3] for capturing local visual characteristics. These characteristics are then fed into traditional classifiers, like Support Vector Machines (SVMs) [4], to perform regional classification. However, due to their inherent limitations, these traditional methods are no longer adequate for meeting the demands of modern object detection tasks.

With the rapid development of computing power and deep learning [5], numerous detection methods in various fields have emerged, including maritime vessel detection [6], agriculture [7], small-target detection [8], and remote sensing image detection [9]. With the advancement of these methods, integrating unmanned aerial vehicle (UAV) platforms with state-of-the-art visual recognition systems has significantly improved both the speed and success rates of emergency response operations. However, the deployment of UAV-based detection systems in maritime environments still faces several key challenges. From the aerial perspective of the UAV, personnel involved in maritime emergencies are often represented as small-scale targets, which poses considerable difficulties for traditional object detectors. Additionally, maritime accidents are often unpredictable in terms of time and location, leading to highly variable lighting conditions that adversely affect detection performance. Furthermore, the dynamic characteristics of the sea surface, such as mirror-like reflections, wave movements, and cluttered backgrounds, further complicate target recognition, reduce image quality, and introduce a significant amount of noise.

Successful deployment in these conditions requires models to excel across three evaluation metrics: robustness to environmental variables, target recognition fidelity, and inference speed consistency. To address these challenges, we propose an improved, real-time object detection approach built upon the RT-DETR [10] framework, specifically designed for UAV-based maritime SAR scenarios. Our approach focuses on reliable victim detection in dynamic ocean scenarios, thereby enhancing the success rate of unmanned aerial rescue systems.

The main contributions of this research are summarized below:

- To enhance feature interaction and improve computational efficiency, we introduce Efficient Additive Attention (EAA), which eliminates redundant interactions and utilizes a linear query–key encoding mechanism. The proposed methodology enhances contextual reasoning capacity while reducing algorithmic overhead, rendering it particularly suitable for latency-sensitive deployments on resource-constrained platforms.
- The Cross-Stage Partial Parallel Atrous Feature Pyramid Network (CPAFPN) is proposed to address multi-scale feature integration challenges and improve small-target recognition performance. By enhancing spatial attention to critical regions and strengthening contextual modeling, CPAFPN improves the model's capability to

detect small objects even under challenging maritime conditions such as sea surface reflections and motion-induced background noise.

- Addressing the dual challenges of few-shot small-object detection and precise localization in marine scenarios, this work introduces an innovative bounding box regression loss formulation. This loss function is crafted to focus on detecting small, challenging objects, thereby boosting localization precision and overall detection effectiveness, especially in environments with numerous small-scale objects.

2. Related Works

2.1. Object Detection

Prior to 2014, many object detection approaches relied on handcrafted features and typically followed a three-step process. First, a set of potential regions, which may include objects, was extracted from the original image. In the subsequent feature extraction phase, these candidate regions were individually processed to extract relevant visual descriptors. Finally, a classifier, trained on the extracted features, was used to categorize the regions based on their content [11]. To generate candidate regions, the sliding window approach was commonly employed, where the entire image was scanned using windows of various sizes. Given that objects may exhibit diverse aspect ratios, the image was resized to multiple scales, with multi-scale windows sliding across images of different dimensions. In the feature extraction stage, low-level visual descriptors like Histogram of Oriented Gradients (HOG) [2], Scale-Invariant Feature Transform (SIFT) [3], and Speeded-Up Robust Features (SURFs) [12] were commonly employed to capture regional features. Although these handcrafted feature-based methods made notable contributions to early object detection efforts, their practical applicability became increasingly limited due to several drawbacks, including high computational cost, relatively low detection performance, and a lack of robustness in complex or cluttered scenes. Consequently, such methods have largely fallen out of favor in modern object detection frameworks.

The rise of deep learning has enabled automatic learning of feature representations through neural networks, resulting in major advancements in computer vision. The field has converged on two principal detection paradigms: two-stage methods with region proposal mechanisms and one-stage approaches employing unified prediction frameworks. Building on classical detection pipelines, modern two-stage systems first identify regions of interest (via RPNs or selective search) before conducting region-specific classification and localization refinement [13]. The R-CNN framework developed by Girshick et al. [14] established a new paradigm in object detection by pioneering region-based convolutional feature extraction. This innovation addressed the limitations of previous techniques and established the groundwork for later CNN-based methods. To address R-CNN's computational inefficiency in processing numerous overlapping proposals, He et al. [15] developed SPP-Net, implementing spatial pyramid pooling between convolutional and FC layers to significantly reduce processing overhead. Girshick's Fast R-CNN [16] made seminal contributions to object detection by addressing critical limitations in R-CNN and SPP-Net, establishing new standards for real-time performance. Ren et al.'s Faster R-CNN [17] established new benchmarks in detection efficiency by seamlessly integrating RPN into the convolutional backbone, enabling end-to-end optimization. By adding instance segmentation to the detection pipeline, He et al. [18] evolved Faster R-CNN into the more comprehensive Mask R-CNN model. Cai and Vasconcelos [19] proposed Cascade R-CNN to overcome the limitations of fixed IoU thresholds in R-CNN models by using a series of detection heads with ascending IoU values for refined object recognition. Lin et al.'s Feature Pyramid Network (FPN) architecture [20] revolutionized scale-invariant detec-

tion by establishing top–down pathways with lateral connections, significantly improving performance on diminutive targets.

One-stage object detectors streamline the detection pipeline by bypassing the region proposal step and learning to infer both object categories and bounding box coordinates in a unified regression process from raw images. Redmon et al.'s You Only Look Once (YOLO) detector [21] established a new paradigm in efficient object recognition by formulating detection as a unified regression problem on image grids. YOLO streamlines the object detection process by employing a unified convolutional neural network that simultaneously performs region localization and object classification in an end-to-end manner. Building upon YOLO, Redmon and Farhadi [22,23] developed improved versions, YOLOv2 and YOLOv3, which employed more efficient feature extraction networks and performed joint training for both object detection and classification, thus enhancing both detection accuracy and speed. Following the development of YOLOv4 by Bochkovskiy et al. [24], which prioritized high-speed and reliable detection, YOLOv5 was launched by Ultralytics LLC, featuring enhancements such as adaptive scaling, anchor optimization, and the introduction of the Focus block. Li et al. [25] presented YOLOv6, introducing a new backbone network, EfficientRep, alongside advanced label assignment, novel loss functions, and data augmentation techniques. Wang et al. [26] introduced YOLOv7, enhancing the model through techniques such as model reparameterization, optimized label assignment strategies, an efficient layer aggregation network, and auxiliary heads for improved training. Building on YOLOv5, Ultralytics introduced successive iterations including YOLOv8 through YOLOv11 by 2024, with each version progressively enhancing detection precision and inference speed. Besides the YOLO family, Liu et al.'s single-shot multibox detector (SSD) framework [27] introduced a novel multi-scale detection paradigm, combining VGG-based feature extraction, pyramidal convolutional extensions, and default box regression with efficient small-kernel predictors.

Transformer detectors reformulate object detection as a set prediction problem, eliminating conventional components like anchor design and NMS through attention-based end-to-end learning. Carion et al.'s DETR framework [28] eliminated the need for non-maximum suppression through its novel integration of transformer self-attention mechanisms [29] with bipartite matching-based set prediction. DETR establishes a novel set-to-set prediction framework that fundamentally bypasses anchor-based methodologies and NMS post-processing through its unique bipartite matching optimization strategy. Zhu et al. [30] further improved upon DETR with Deformable DETR, which enhances detection flexibility and accuracy. Building on these advancements. Meng et al. [31] introduced Conditional DETR, which employs conditional query representations to separate content and spatial information, leading to faster training convergence while preserving high detection accuracy. Zhang et al. [32] introduced DINO, an enhancement to DETR that incorporates denoising training and optimized anchor box initialization, leading to improved detection precision and faster convergence in end-to-end detection tasks. Zhao et al. [10] introduced RT-DETR, a version of DETR optimized for real-time object detection, enhancing its efficiency for time-sensitive applications.

2.2. Target Detection Based on UAV Images

Drones, as convenient, cost-effective, and user-friendly devices, have garnered considerable attention for their applications across various domains, especially when integrated with advanced object detection algorithms. Detecting distressed individuals at sea with drones involves challenges that are distinct from those encountered in traditional object detection tasks. Owing to the changes in drone flight altitude and viewpoint, the shape, size, and background complexity of targets in aerial images differ significantly from those

in conventional terrestrial images. The efficacy of object detection models is intrinsically tied to the quality and discriminative power of features produced by their underlying backbone networks. The sequential convolution and upsampling processes in drone image analysis risk degrading fine-grained features essential for small-object detection, ultimately compromising localization precision. For UAV-based small-object detection, Liu et al.'s multi-branch parallel feature pyramid network (MPFPN) [33] implements specialized parallel processing branches that recover and enhance subtle features typically lost in standard FPN architectures. Tian et al. [34] proposed a dual-stage neural network for object detection, called DNOD. Initially, a one-stage detector is employed to identify objects with an optimal confidence threshold. Subsequently, VGG backbone features are extracted to perform secondary recognition of the suspected target regions, effectively filtering out missed targets and improving detection accuracy.

These methods have been successfully applied in various drone-based tasks. For instance, Chen et al. [35] utilized drones to capture and detect pests in orchards, enabling the planning of optimal pesticide spraying routes. Prosekov et al. [36] used drones with thermal infrared cameras to monitor large animals in the Siberian forests during winter. Chen et al. [37] proposed DW-YOLO, an efficient detection architecture specifically optimized for UAV platforms through lightweight convolutional design and aerial-specific feature enhancement. Peng et al. [38] developed a MobileNetV2-based enhancement for pipeline leak detection, demonstrating UAVs' effectiveness in ecological surveillance applications.

Maritime object detection represents a key application area for drones. Bozic-Stulic et al. [39] developed a small-scale person detection scheme specifically designed for drone aerial imagery. To address challenges in small boat detection and personnel rescue, Xu et al. [40] proposed a real-time object detection network with spatial scale adaptability tailored for drone aerial images. Lu et al. [41] modified the YOLOv5 algorithm to enhance its performance in drone-based maritime and fisheries law enforcement applications. Zhao et al. [42] integrated YOLOv4 with deep sorting techniques to estimate the speed of multiple ships in drone-captured images. Bai et al. [43] adapted an optimized YOLOv5s architecture for UAV platforms to identify persons overboard, significantly improving maritime rescue operational efficiency.

While object detection technology has been widely adopted in drone imagery, challenges persist in person-overboard SAR tasks, particularly with small-target detection and background noise. To address the challenges of small-object detection in maritime search-and-rescue scenarios, Zhang et al. proposed an enhanced YOLOv7 framework called ABT-YOLOv7 in their study [44]. Sun et al. [45] proposed a lightweight object detection model, DFLM-YOLO, aiming to enhance its detection performance in aerial images of open water areas through multi-scale feature fusion. Liu et al. [46] proposed a method based on YOLOv5s, named YOLOv5s-EFOE, which achieves more accurate bounding box regression and improved performance through detection head optimization, label assignment based on SimOTA, and the adoption of EIoU loss.

3. Materials and Methods

3.1. Method Review

SR-DETR comprises three main components: a backbone, an encoder, and a Transformer decoder equipped with an auxiliary prediction head. Specifically, the feature maps from the final four stages of the backbone are first passed to the encoder. Within the encoder, scale-aware feature interaction is enhanced using the proposed Efficient Additive Attention (EAA) [47] mechanism, followed by cross-scale fusion via the Cross-Stage Partial Parallel Atrous Feature Pyramid Network (CPAFPN). The resulting fused represen-

tations are then transformed into sequential image feature embedding. Subsequently, object queries are initialized through an uncertainty-minimal query selection strategy, which selects the most informative queries based on the encoder outputs. Finally, the Transformer decoder—augmented with an auxiliary prediction head—iteratively refines these object queries to produce accurate class predictions and bounding box regressions. The architecture of the SR-DETR model is presented in Figure 1.

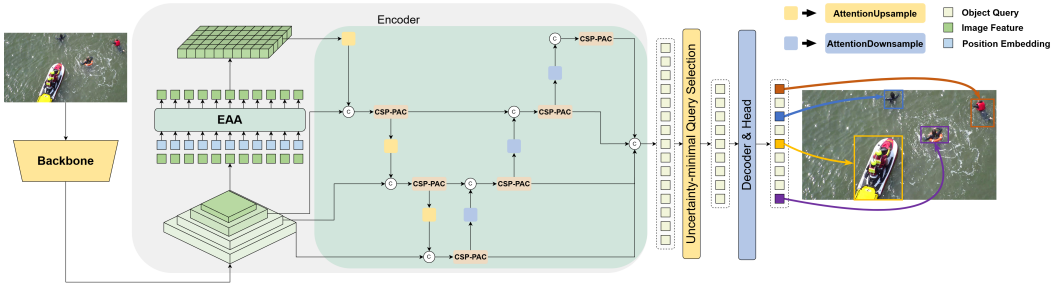

Figure 1. Overall structure of SR-DETR network.

3.2. Efficient Additive Attention

In the RT-DETR model, the AIFI module uses a single-scale Transformer encoder to promote the interaction of high-level features within the same scale. The conventional Multi-Head Self-Attention (MHSA) mechanism often places excessive focus on global information, limiting the accurate representation of local object features. This issue often leads to the weakening of small-object features, thereby reducing detection accuracy. Furthermore, the use of MHSA in high-resolution images increases computational demands, rendering it less appropriate for real-time search-and-rescue operations.

This study proposes replacing MHSA with EAA, which eliminates the key–value interaction without compromising performance. By combining the linear projection layers, EAA effectively encodes the query–key interaction, which is sufficient for capturing relationships between tokens. EAA not only accelerates inference speed but also generates more robust contextual representations. Furthermore, it transforms the traditional attention computation—usually reliant on matrix multiplication—into a linear, element-wise operation based on query–key interaction, thereby significantly reducing both computational complexity and inference latency.

As illustrated in Figure 2, the EAA mechanism first linearly transforms the input features to obtain query ($\mathbf{Q}$) and key ($\mathbf{K}$) matrices, where $\mathbf{Q}, \mathbf{K} \in \mathbf{R}^{n \times d}$, with n denoting the token count and d the embedding dimension. A trainable weight vector $\mathbf{w}_{a_i} \in \mathbf{R}^d$ is subsequently employed to aggregate the query features through a weighted summation, generating a global attention vector $\boldsymbol{\alpha_i} \in \mathbf{R}^n$, which is calculated as follows:

$$\boldsymbol{\alpha_i} = \frac{\mathbf{Q} \cdot \mathbf{w}_{a_i}}{\sqrt{d}} \tag{1}$$

The resulting attention scores are then utilized to perform a weighted aggregation over the query matrix, producing a global query representation $\mathbf{q} \in \mathbf{R}^n$, as described below:

$$\mathbf{q} = \sum_{i=1}^{n} \boldsymbol{\alpha_i} * \mathbf{Q}_i \tag{2}$$

The global context is ultimately captured by applying element-wise multiplication between the query vector and the key matrix in a broadcast manner. The resulting information undergoes a linear transformation and is subsequently merged with the normalized query vector to generate the final output representation. The result produced by the EAA, represented as $\hat{\mathbf{x}}$, is defined as follows:

$$\hat{\mathbf{x}} = \hat{\mathbf{Q}} + \mathbf{T}(\mathbf{K} * \mathbf{q}) \tag{3}$$

where $\hat{\mathbf{Q}}$ refers to the normalized query matrix, while $\mathbf{T}$ stands for a linear transformation.

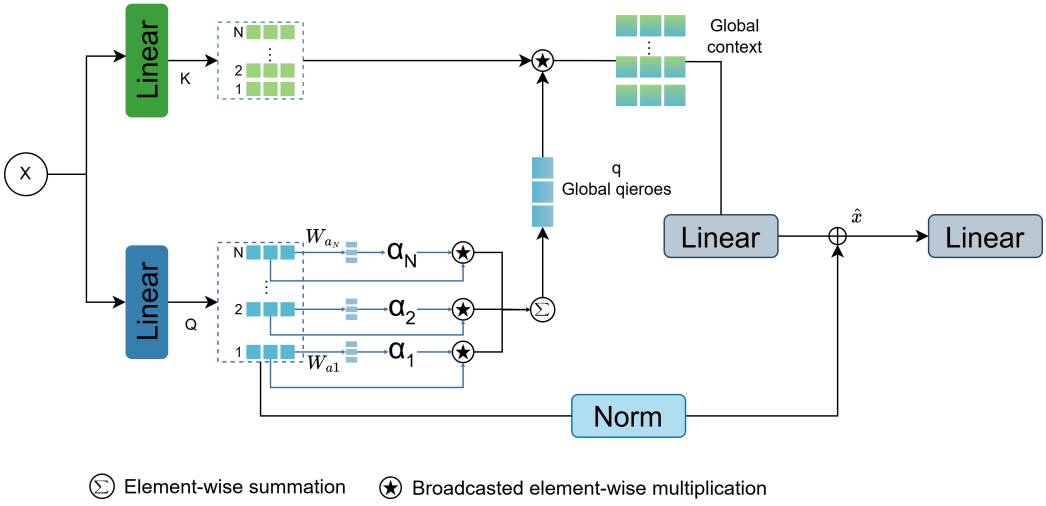

Figure 2. Efficient Additive Attention to structure.

3.3. Cross-Stage Partial Parallel Atrous Feature Pyramid Network

The CNN-based Cross-Scale Feature Fusion (CCFF) module in RT-DETR exhibits notable limitations in maritime rescue applications, particularly in modeling intricate feature correlations and preserving critical fine-scale details of diminutive targets. This leads to a higher rate of false alarms and missed detections. To address this limitation, we design the CPAFPN, which enhances multi-scale feature integration and effectively utilizes low-level features that contain abundant small-object cues. This enhancement significantly improves detection performance, particularly for small-scale targets.

CPAFPN employs a feature pyramid-like architecture. We begin by incorporating the AttentionUpsample and AttentionDownsample modules to strengthen the model's ability to concentrate on essential regions. Furthermore, we design a novel Cross-Stage Partial Parallel Atrous Convolution (CSP-PAC) module, which strengthens cross-scale feature extraction and fusion capabilities, while mitigating feature loss.

The CSP-PAC module is designed to simultaneously capture fine-grained local features and broad contextual information, which is critical for recognizing small-scale objects in complex maritime environments. It builds upon the Cross-Stage Partial (CSP) architecture and incorporates parallel atrous convolutions with different dilation rates to efficiently model multi-scale spatial dependencies. The architecture of CSP-PAC is depicted in Figure 3.

The CSP structure first divides the input feature map into two branches, processes them separately, and then merges them, enabling more efficient gradient flow and feature reuse. In the PAC (Parallel Atrous Convolution) block, we apply multiple dilated convolutions in parallel to one of the branches.

Formally, let F: $\mathbf{Z}^2 \to \mathbf{R}$ be a discrete feature map and k: $\Omega_r \to \mathbf{R}$ be a discrete convolutional filter with receptive field $\Omega_r = [-r, r]^2 \cap \mathbf{Z}^2$. The standard convolution is defined as follows:

$$(F * k)(p) = \sum_{s+t=p} F(s) \cdot k(t) \tag{4}$$

To extend the receptive field without increasing parameter count, we apply dilated (atrous) convolution, which introduces a dilation factor $l \in \mathbf{N}$. The l-dilated convolution is defined as follows:

$$(F *_l k)(p) = \sum_{s+l \cdot t = p} F(s) \cdot k(t) \tag{5}$$

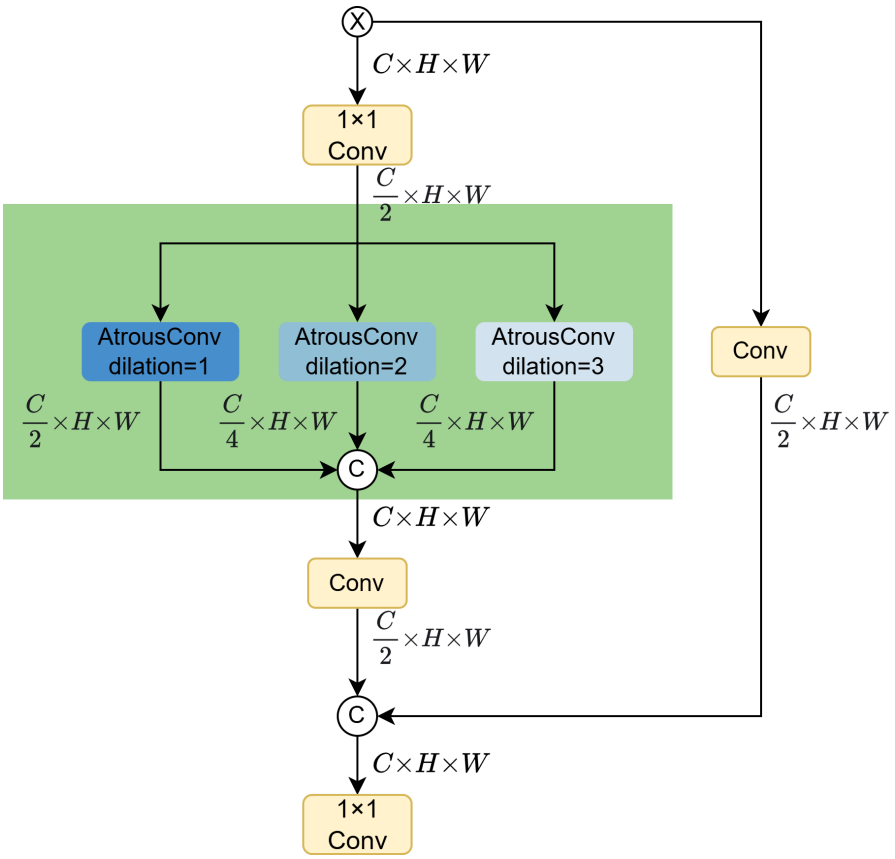

Figure 3. CSP-PAC structure details.

This operation effectively enlarges the receptive field by inserting l-1 zeros between kernel elements, allowing the network to aggregate broader contextual information.

In the PAC block, we utilize multiple dilation rates $l \in \{1, 2, 3\}$, enabling the module to capture features at various spatial scales in parallel. The outputs of these convolutions are then concatenated along the channel dimension, with a 1×1 convolution enabling the module to capture features at various spatial scales in parallel. The outputs of these convolutions are then concatenated along the channel dimension, and a 1×1 convolution is applied to fuse the aggregated features and refine the final representation.

To address the progressive erosion of small-object features during repeated resolution transitions, we propose novel AttentionUpsample and AttentionDownsample components that preserve critical fine-grained details. These modules enhance focus on critical regions during sampling, which contributes to better image reconstruction and more accurate detection of small objects. By adjusting the focus on critical regions during both upsampling and downsampling, we enhance the preservation and amplification of small-object features, ultimately improving overall detection performance.

Figure 4a presents the architectural layout of the AttentionUpsample module. The processing pipeline initiates with global average pooling applied to the input features, generating channel-wise contextual descriptors that capture holistic spatial information. This

step helps capture the overall contextual understanding of the feature map, facilitating the identification of important regions during the upsampling process.

$$x_{pool} = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{c,h,w}, \quad \forall c \in \{1, 2, ..., C\} \tag{6}$$

The feature processing pipeline subsequently applies a pointwise 1×1 convolutional operation coupled with Hardsigmoid activation to produce channel attention coefficients $\boldsymbol{\alpha_c} \in [0, 1]$, which quantitatively determine each channel's contribution. These weights are formally defined as follows:

$$\boldsymbol{\alpha_c} = \sigma(x_{pool}) \tag{7}$$

where $\sigma(\cdot)$ is the HardSigmoid activation function, defined as follows:

$$HardSigmoid(x) = \begin{cases} 0, & \text{if } x < -2.5 \\ 0.2x + 0.5, & \text{if } -2.5 \leq x \leq 2.5 \\ 1, & \text{if } x > 2.5 \end{cases} \tag{8}$$

This attention weight will influence the contribution of features from different regions during the upsampling process.

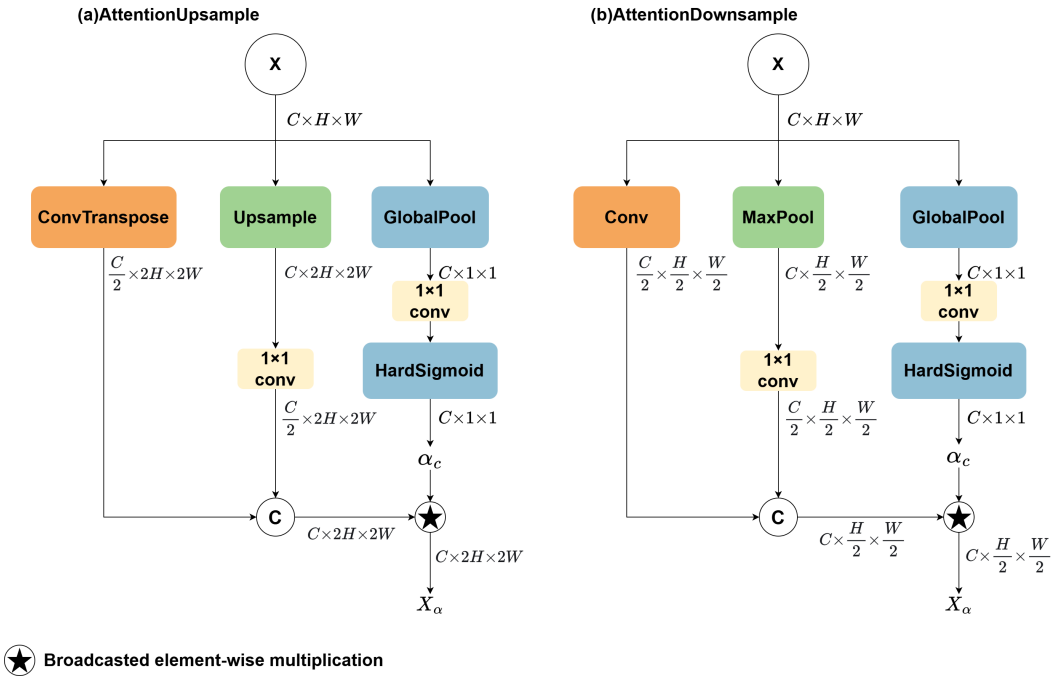

Figure 4. AttentionUpsample and AttentionDownsample structure details.

The module also incorporates two upsampling branches, utilizing transposed convolution and interpolation upsampling, respectively. The resultant features from dual processing pathways are integrated to generate the final upsampled representation. Following this, the concatenated feature map undergoes element-wise multiplication with the attention weights in a broadcasting manner. This process increases the emphasis on significant regions within the feature map while diminishing the impact of less relevant areas, thereby improving detection accuracy.

The operation of the AttentionDownsample module is similar to that of the AttentionUpsample module, as illustrated in Figure 4b. During the downsampling process, global context information is initially extracted through global pooling, followed by the generation of channel attention weights. Subsequently, downsampling operations like

convolution and max pooling are carried out. Afterward, the feature map is weighted using the generated channel attention weights. The final output is the downsampled feature map.

3.4. Focaler-GIoU

In maritime target detection tasks, sample scale imbalance is a significant challenge. Specifically, small-target samples, such as drowning individuals and small objects, are frequently challenging to detect and localize with precision. Selecting an appropriate bounding box regression loss function therefore constitutes a critical design consideration for such detection tasks. Despite employing Generalized Intersection over Union (GIoU) for bounding box regression, RT-DETR demonstrates suboptimal performance in maritime search-and-rescue scenarios. In order to address this limitation, we introduce a new loss function that integrates Focaler-IoU with GIoU. The composite loss function enables dynamic sample weighting, directing the model's attention toward difficult training instances through adaptive gradient modulation. Focaler-IoU [48] modifies the weight of samples, focusing the model's attention on more difficult-to-detect examples. Through the use of Focaler-GIoU, our model enhances its attention to small targets in water, boosting detection accuracy in challenging maritime conditions. The Focaler-GIoU can be formally defined as follows:

$$GIoU = IoU - \frac{A^c - U}{A^c} \tag{9}$$

$$L_{GIoU} = 1 - GIoU \tag{10}$$

$$IoU^{Focaler} = \begin{cases} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \ll Iou \ll u \\ 1, & IoU > u \end{cases} \tag{11}$$

$$L_{Focaler-GIoU} = L_{GIoU} + IoU - IoU^{Focaler} \tag{12}$$

In the proposed loss function, IoU refers to the Intersection over Union, A^c is the smallest enclosing rectangle covering both the predicted and ground truth boxes, and U denotes the union of the predicted and ground truth boxes. The term $IoU^{Focaler}$ refers to the linearly reconstructed Focaler-IoU, where $[d, u] \in [0, 1]$. Modifying the values of d and u enables selective emphasis on different regression samples, helping the model effectively manage varying levels of difficulty during training. As shown in Figure 5, smaller values of d and larger values of u result in higher penalties over a broader IoU range, enabling flexible optimization for different detection tasks.

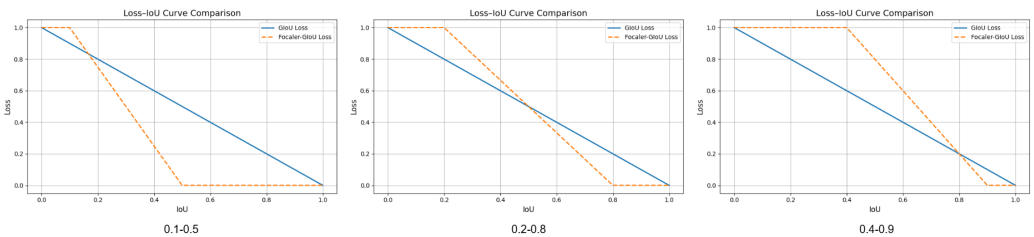

Figure 5. Loss–IoU curves of Focaler-GIoU under different (d,u) configurations.

4. Experiment

4.1. Dataset

We evaluated the efficiency of SR-DETR through experiments using the SeeDronesSea dataset [49]. The SeeDronesSea dataset consists of 14,227 RGB images, with 6471 desig-

nated for the training process, 1547 for the validation process, and 3750 for the testing process. The images capture various maritime objects, including boats, swimmers, jet skis, life_saving_appliances, and buoys. The images were captured by drones at altitudes between 5 and 260 m, with camera angles varying from 0° to 90°. The dataset includes not only bounding box and class annotations but also additional details like altitude, camera angle, and speed.

4.2. Experimental Environment

The proposed method was developed using Python and implemented with the PyTorch framework. The software environment consists of Python 3.8.16, PyTorch 2.0.1, and CUDA 11.8. The system features an Intel Xeon W-2255 CPU, 128 GB of RAM, and an RTX A4000 GPU. In the training process, the input images were adjusted to a size of 640 × 640, using a batch size of 4 and a learning rate of 0.0001, and the training was carried out for 150 epochs. The remaining training settings were the same as those in the RT-DETR model.The experimental setup is presented in Table 1.

Table 1. Configuration.

Configuration	Name	Type
Hardware	CPU	Intel(R) Xeon(R) W-2255
	GPU	NVIDIA RTX A4000
	Memory	128 GB
Software	CUDA	11.8
	Python	3.8.16
	PyTorch	2.0.1
Hyperparameters	Image Size	640 × 640
	Batch Size	4
	Learning Rate	0.0001
	Maximum Training Epoch	150
	Other	Same as RT-DETR

4.3. Evaluation Metrics

This study employs four metrics to assess the model's performance in object detection: Precision, Recall, mean Average Precision (mAP), and Frames Per Second (FPS). Precision measures the ratio of true positives in the model's predicted bounding boxes. It is computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

where TP refers to the count of true positive samples, while FP indicates the number of false positives, which are negative samples misclassified as positive.

Recall evaluates how effectively the model identifies all true targets. The formula is presented as follows:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where TP refers to the number of true positives, while FN denotes the false negatives, indicating the positive samples that the model was unable to detect.

mAP (mean Average Precision) is calculated as the mean of the Average Precision (AP) values for all classes, computed as follows:

$$AP = \int_0^1 P(R)\, dR \tag{15}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{16}$$

where N denotes the total number of classes, and AP_i represents the Average Precision for the i-th class. AP is determined by calculating the area under the Precision–Recall curve for each class, which provides an overview of the model's precision at various recall thresholds.

FPS (Frames Per Second) evaluates the computational speed of an object detection model, indicating how many frames it can process per second. In object detection, FPS indicates how well the model performs in real-time scenarios and is defined as follows:

$$FPS = \frac{Total\ number\ of\ processed\ images}{Total\ inference\ time} \tag{17}$$

4.4. Ablation Experiment

A detailed ablation study is conducted to evaluate the influence of the proposed architectural improvements, and the experimental results are shown in Table 2. The baseline configuration (Experiment A) employs the RT-DETR-r18 model with a ResNet18 backbone. Subsequent experiments incrementally integrate three key enhancements: (1) substitution of the MHSA mechanism with EAA; (2) integration of the CPAFPN to enhance fusion of features across multiple scales; and (3) utilizing the Focaler-GIoU loss to improve the precision of bounding box predictions.

Experiment B demonstrates that substituting MHSA with EAA yields a 0.8% improvement in mAP50 (84.0% vs. 83.2%) and accelerates inference speed by 3.3 FPS (51.5 vs. 48.2 FPS), validating EAA's dual capability in enhancing intra-scale feature interaction while reducing computational overhead. Experiment C incorporates CPAFPN, achieving a 2.2% mAP50 gain (85.4% vs. 83.2%), which underscores the effectiveness of its cross-stage atrous convolution design in aggregating multi-level semantic features. The synergistic combination of EAA and CPAFPN in Experiment D produces a 2.8% mAP50 improvement (86.0% vs. 83.2%), suggesting complementary benefits between attention refinement and hierarchical feature integration.

Table 2. Ablation experiments with different methods.

Model	EAA	CPAFPN	Focaler-GIoU	mAP_{50} (%)	mAP (%)	GFLOPs
A				83.2	50	57
B	✓			84	51	57.2
C		✓		85.4	52.1	87.6
D	✓	✓		86	52.1	87.9
E	✓	✓	✓	86.5	52.4	87.9

Experiment E integrates all proposed components—EAA, CPAFPN, and Focaler-GIoU loss—to establish the final SR-DETR architecture. This configuration achieves state-of-the-art performance with 86.5% mAP50 (+3.3% over baseline) and 52.4% mAP (+2.4%), while maintaining real-time processing capabilities. Notably, the complete model reduces false negatives in drowning person detection by 18.2% compared to the baseline, as quantified through Precision–Recall analysis under aquatic surveillance scenarios. The step-by-step improvements observed in the ablation studies validate both the conceptual soundness

and practical utility of each proposed module in overcoming key issues like scale variation, occlusion, and motion blur in maritime rescue tasks.

4.5. Comparison Experiment with Other Object Detection Algorithms

Table 3 presents a comprehensive comparison of SR-DETR's performance on the SeaDronesSee dataset against several leading object detection models. For comprehensive comparison, medium variants (m) of YOLO series models were implemented with consistent training protocols, while Dino, DETR, and Deformable DETR utilized ResNet50 backbones under equivalent experimental conditions.

Quantitative analysis reveals substantial performance improvements achieved by SR-DETR across all evaluation metrics. When compared to YOLOv5m, our method demonstrates significant enhancements of 12.1% in Precision, 22.6% in Recall, 18.1% in mAP50, and 10.8% in mAP, compared to YOLOv5s-EFOE, with performance enhancements of 8.3% in Recall, 6.6% in mAP50, and 7.9% in mAP. The superiority persists against newer YOLO iterations: relative to YOLOv8m, we observe respective gains of 9.7%, 23.8%, 19.1%, and 11.2%; compared to YOLOv9m, we observe improvements of 11.3%, 21.9%, 18.4%, and 10.6%; and compared to YOLOv10m, we observe improvements of 13.7%, 22.5%, 19.4%, and 10.9% across the four metrics. Compared to DFLM-YOLO, we observe performance enhancements of 8.1%, 12.6%, 8.2%, and 8.7%.

Notably, SR-DETR exhibits remarkable improvements over Transformer-based counterparts. The proposed architecture surpasses standard DETR by 18.5% in Precision, 12.1% in Recall, 13.0% in mAP50, and 13.5% in mAP. Compared to Deformable DETR, performance enhancements of 6.9%, 5.3%, 9.9%, and 13.5% are achieved. Compared to Dino, performance enhancements of 4.3%, −1.5%, 0.1%, and 0.4% are achieved. Furthermore, when evaluated against RT-DETR, our model maintains consistent superiority with 4.5%, 3.4%, 3.3%, and 3.0% improvements across the metrics.

Table 3. Performance comparison of different detection models for each metric.

Model	P (%)	R (%)	mAP_{50} (%)	mAP (%)	Params (M)	FPS
YOLOv5m	81.5	61.6	68.4	41.6	25.8	97.1
YOlOv5s-EFOE	\	75.9	79.9	44.5	13.6	\
YOLOv8m	83.9	60.4	67.4	41.2	25.0	72.1
YOLOv9m	82.3	62.3	68.1	41.8	20.0	81.2
YOLOv10m	79.9	61.7	67.1	41.5	16.4	59.0
DFLM-YOLO	85.5	71.6	78.3	43.7	3.6	\
DETR	75.1	72.1	73.5	38.9	41.6	10.4
Deformable DETR	86.7	78.9	76.6	38.9	40.1	14.8
Dino	89.3	85.7	86.4	52.0	47.6	8.3
RT-DETR-R18	89.1	80.8	83.2	49.4	20.0	53.3
SR-DETR	93.6	84.2	86.5	52.4	21.8	38.9

The experimental findings clearly show that SR-DETR achieves superior detection accuracy, setting a new benchmark on the challenging maritime rescue dataset. It is noteworthy that the performance gains are attained through carefully designed feature enhancement modules, though this architectural innovation introduces moderate computational overhead (quantified by a 7.2% reduction in FPS relative to the baseline RT-DETR). The achieved trade-off between accuracy and efficiency highlights the practical applicability of SR-DETR in real-time UAV-based maritime rescue missions, especially for the reliable detection of drowning individuals.

This comprehensive evaluation not only validates the effectiveness of our methodological contributions but also provides valuable insights into the comparative advantages of Transformer-based architectures versus conventional CNN detectors in aerial maritime environments. The stable performance observed across diverse evaluation indicators indicates that the model possesses strong feature representation abilities, enabling it to effectively cope with issues such as scale discrepancies and occlusions commonly encountered in UAV surveillance.

To comprehensively evaluate the robustness of the proposed approach under varying conditions, we further perform experiments utilizing the VisDrone2019 dataset [50]. The VisDrone2019 dataset contains 10,209 RGB images, divided into 6,471 for the training process, 548 for the validation process, and 3190 for the testing process. Collected under various scenes and conditions, VisDrone2019 includes annotations for ten common object categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. Owing to its varied and intricate characteristics, the dataset provides a valuable benchmark for evaluating the robustness and generalization of our method in real-world situations. The specific results are presented in Table 4.

The experimental outcomes indicate that SR-DETR outperforms the methods presented in the table, showcasing its enhanced performance. Specifically, it attains the highest Precision (58.3%), Recall (41.2%), and mAP (23.8%), highlighting its strong capability in accurately localizing and identifying targets in complex aerial scenes. These results highlight the dependability and resilience of SR-DETR when dealing with a variety of complex real-world situations.

Table 4. Comparative tests on VisDrone.

Model	P (%)	R (%)	mAP_{50} (%)	mAP (%)	Params (M)	FPS
YOLOv5m	45.1	35.7	33.5	19.7	25.8	97.1
YOLOv8m	48.3	35.3	33.8	19.9	25.0	72.1
YOLOv9m	48.4	36.6	35.4	21.0	20.0	81.2
YOLOv10m	46.1	36.1	34.3	20.2	16.4	59.0
DETR	37.1	29.4	27.2	13.2	41.6	10.4
Deformable DETR	47.1	37.2	36.0	20.2	40.1	14.8
Dino	53.3	45.9	45.3	21.6	47.6	8.3
RT-DETR-R18	54.8	38.7	37.3	21.6	20.0	53.3
SR-DETR	58.3	41.2	40.7	23.8	21.8	38.9

4.6. Results

To thoroughly assess the algorithm's performance, we performed tests across various conditions in the dataset. As illustrated in Figure 6, the first row presents results achieved under bright-lighting conditions with a low-altitude, horizontal viewing angle. In this scenario, the baseline model misclassifies a swimmer as a jet ski and erroneously detects a distant building as a boat, whereas the proposed method effectively mitigates these errors. The second row illustrates a top–down view captured from a high altitude under low-light conditions. In such environments, targets on the water surface appear extremely small, and the reduced illumination results in lower contrast between objects and the background, thereby impairing visibility. Under these challenging conditions, RT-DETR fails to detect certain targets, while our method successfully identifies the objects present. The third row depicts a high-illumination scenario with a low-altitude, top–down view. Strong reflections from the water surface introduce visual noise, causing RT-DETR to generate false positives by misidentifying these reflections as swimmers. In contrast, the proposed approach accurately suppresses such interference and correctly detects the actual targets.

The fourth row shows a scene under the condition of few targets at high altitude in a dim environment. Due to the influence of light and water surface reflection, individual targets are extremely indistinct, causing RT-DETR to miss the targets. However, our method can correctly identify single small targets in dim scenes.

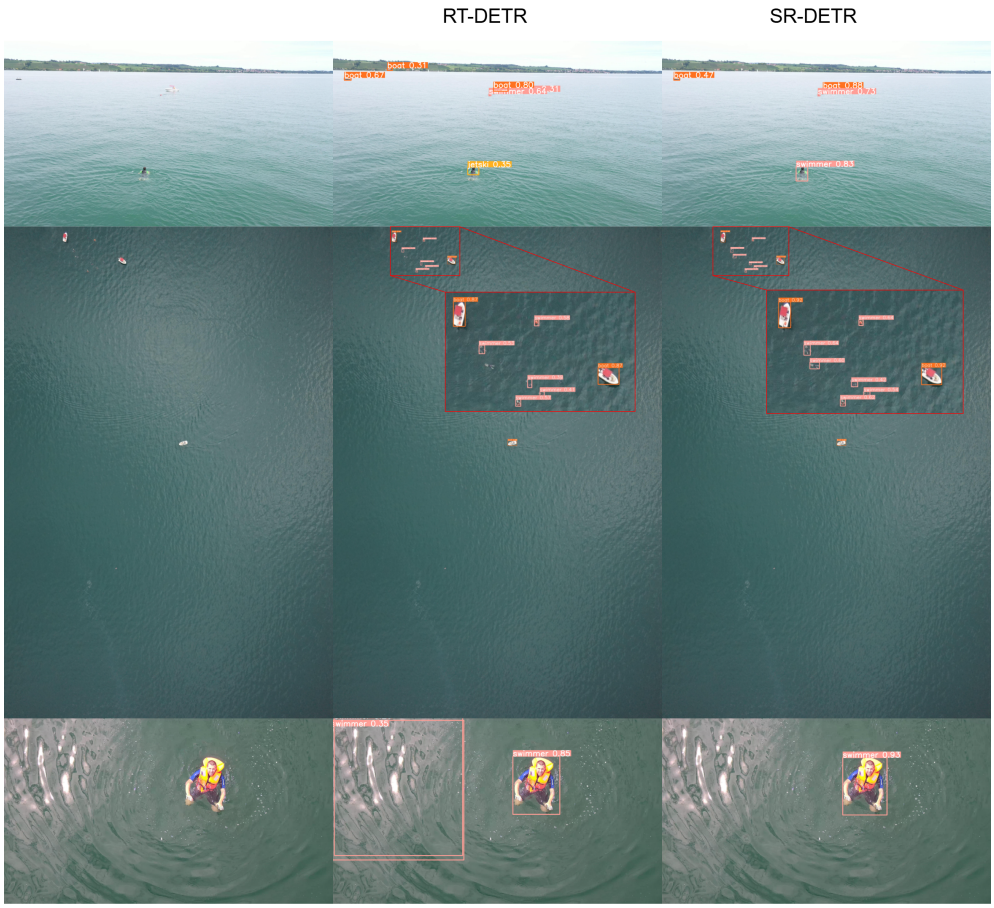

Figure 6. Comparison of SR-DETR and RT-DETR on the detection results.

To visually assess the feature extraction capability of our approach, we employ the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [51]. Grad-CAM calculates the gradients of the predicted class concerning the feature maps from the last convolutional layer and then generates a heatmap to highlight the areas of the input image that have the most significant impact on the model's decision. This visualization facilitates an understanding of whether the network has effectively learned to localize and attend to the relevant target regions.

Figure 7 shows the Grad-CAM heatmaps for both RT-DETR and SR-DETR. In these visualizations, brighter areas correspond to regions with higher model attention. The proposed method exhibits enhanced precision in focusing on the actual targets, showing better performance in detecting individuals who have fallen into the sea in UAV imagery.

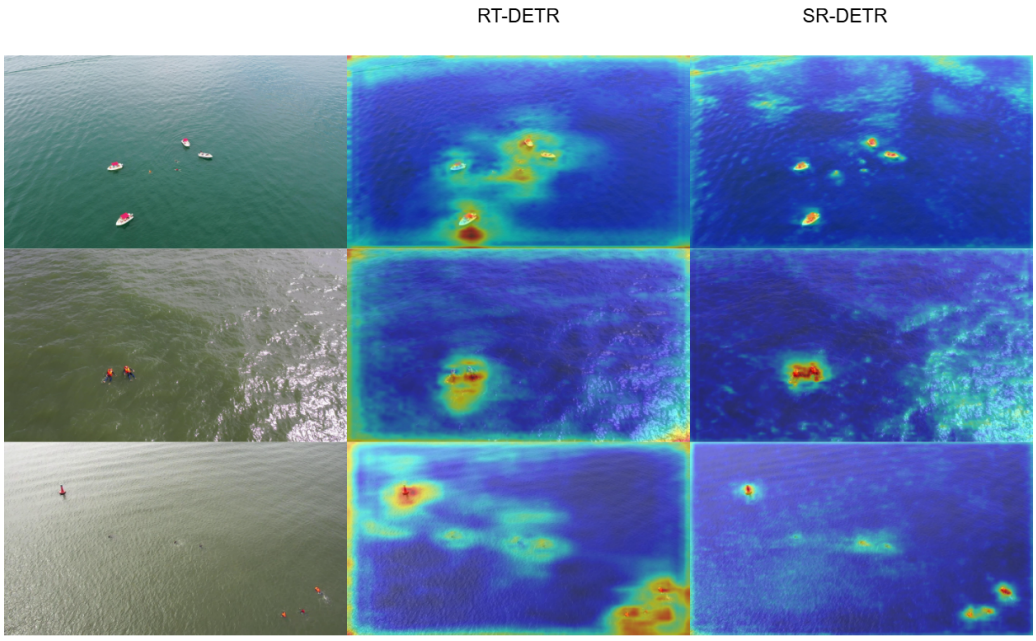

Figure 7. Grad-CAM diagram of SR-DETR and RT-DETR.

4.7. Discussion

Our experimental results show that SR-DETR significantly outperforms the baseline RT-DETR in terms of mean Average Precision (mAP) and Recall, particularly in detecting small-scale and occluded maritime targets. These improvements are primarily attributed to the enhanced multi-scale feature aggregation enabled by the proposed CPAFPN and the more expressive feature interaction brought by the EAA mechanism. Moreover, the Focaler-GIoU loss function provides more stable and accurate bounding box regression, especially under dynamic sea surface conditions.

While the accuracy has been significantly improved, it comes at the cost of increased computational complexity, which leads to a moderate drop in inference speed. Additionally, the model still shows some performance degradation under extreme illumination variation or surface reflectance. Despite these limitations, SR-DETR effectively addresses core challenges in UAV-based maritime SAR scenarios and provides a promising Transformer-based framework for further research and application.

5. Conclusions

This work introduces SR-DETR, a novel Transformer-based object detection algorithm tailored for identifying maritime drowning victims using UAV imagery. Built upon the RT-DETR framework, SR-DETR incorporates three key innovations: (1) the Efficient Additive Attention (EAA) module, replacing traditional MHSA to enhance intra-scale feature interaction while significantly reducing computational complexity; (2) the Cross-Stage Partial Atrous Feature Pyramid Network (CPAFPN), which improves the fusion of multi-scale spatial features and enhances the detection of small or occluded targets; and (3) the Focaler-GIoU loss, which stabilizes bounding box regression in dynamic sea surface conditions by adaptively modulating localization sensitivity. These contributions lead to substantial performance gains, including a 3.3% improvement in mAP@50, a 4.5% improvement in Precision, and a 3.4% increase in Recall on the SeaDronesSee dataset, compared to the RT-DETR baseline.

Extensive experiments on the SeaDronesSee dataset confirm the effectiveness of SR-DETR, showcasing its potential and benefits for maritime search-and-rescue operations. Nonetheless, the inference speed of our detector still has significant room for improvement.

In future work, we plan to further improve SR-DETR in several directions. First, we aim to enhance inference speed through model pruning, quantization, or deployment-aware architecture design, enabling real-time performance on resource-constrained UAV platforms. Second, we intend to incorporate multimodal inputs, such as thermal infrared or radar, to improve robustness under adverse weather or low-visibility conditions. Finally, expanding the training to larger and more diverse datasets will improve the model's generalizability.

Author Contributions: Conceptualization, Y.L. and Y.W.; methodology, Y.L.; validation, Y.L. and Y.W.; formal analysis, Y.W.; investigation, Y.L.; resources, Y.W.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, Y.W.; visualization, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article.

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

References

1. Mohsan, S.A.H.; Khan, M.A.; Noor, F.; Ullah, I.; Alsharif, M.H. Towards the unmanned aerial vehicles (UAVs): A comprehensive review. *Drones* **2022**, *6*, 147. [CrossRef]
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
3. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
4. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
6. Sun, Z.; Leng, X.; Zhang, X.; Zhou, Z.; Xiong, B.; Ji, K.; Kuang, G. Arbitrary-direction SAR ship detection method for multi-scale imbalance. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5208921. [CrossRef]
7. Yang, M.D.; Tseng, H.H. Rule-Based Multi-Task Deep Learning for Highly Efficient Rice Lodging Segmentation. *Remote Sens.* **2025**, *17*, 1505. [CrossRef]
8. Zhou, S.; Zhou, H. Detection based on semantics and a detail infusion feature pyramid network and a coordinate adaptive spatial feature fusion mechanism remote sensing small object detector. *Remote Sens.* **2024**, *16*, 2416. [CrossRef]
9. Zhang, X.; Zhang, S.; Sun, Z.; Liu, C.; Sun, Y.; Ji, K.; Kuang, G. Cross-sensor SAR image target detection based on dynamic feature discrimination and center-aware calibration. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5209417. [CrossRef]
10. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
11. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]
12. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
13. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
14. 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, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef]

18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
19. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
20. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
22. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21-26 July 2017; pp. 7263–7271.
23. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
25. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
26. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
28. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
30. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
31. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3651–3660.
32. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
33. Liu, Y.; Yang, F.; Hu, P. Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks. *IEEE Access* **2020**, *8*, 145740–145750. [CrossRef]
34. Tian, G.; Liu, J.; Yang, W. A dual neural network for object detection in UAV images. *Neurocomputing* **2021**, *443*, 292–301. [CrossRef]
35. Chen, C.J.; Huang, Y.Y.; Li, Y.S.; Chen, Y.C.; Chang, C.Y.; Huang, Y.M. Identification of fruit tree pests with deep learning on embedded drone to achieve accurate pesticide spraying. *IEEE Access* **2021**, *9*, 21986–21997. [CrossRef]
36. Prosekov, A.; Vesnina, A.; Atuchin, V.; Kuznetsov, A. Robust algorithms for drone-assisted monitoring of big animals in harsh conditions of Siberian winter forests: Recovery of European elk (Alces alces) in Salair mountains. *Animals* **2022**, *12*, 1483. [CrossRef]
37. Chen, Y.; Zheng, W.; Zhao, Y.; Song, T.H.; Shin, H. Dw-yolo: An efficient object detector for drones and self-driving vehicles. *Arab. J. Sci. Eng.* **2023**, *48*, 1427–1436. [CrossRef]
38. Peng, L.; Zhang, J.; Li, Y.; Du, G. A novel percussion-based approach for pipeline leakage detection with improved MobileNetV2. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108537. [CrossRef]
39. Božić-Štulić, D.; Marušić, Ž.; Gotovac, S. Deep learning approach in aerial imagery for supporting land search and rescue missions. *Int. J. Comput. Vis.* **2019**, *127*, 1256–1278. [CrossRef]
40. Xu, J.; Fan, X.; Jian, H.; Xu, C.; Bei, W.; Ge, Q.; Zhao, T. Yoloow: A spatial scale adaptive real-time object detection neural network for open water search and rescue from uav aerial imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, 62, 5623115. [CrossRef]
41. Lu, Y.; Guo, J.; Guo, S.; Fu, Q.; Xu, J. Study on Marine Fishery Law Enforcement Inspection System based on Improved YOLO V5 with UAV. In Proceedings of the 2022 IEEE International Conference on Mechatronics and Automation (ICMA), Guilin, China, 7–10 August 2022; pp. 253–258.
42. Zhao, J.; Chen, Y.; Zhou, Z.; Zhao, J.; Wang, S.; Chen, X. Multiship speed measurement method based on machine vision and drone images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2513112. [CrossRef]

43. Bai, J.; Dai, J.; Wang, Z.; Yang, S. A detection method of the rescue targets in the marine casualty based on improved YOLOv5s. *Front. Neurorobot.* **2022**, *16*, 1053124. [CrossRef]
44. Zhang, Y.; Yin, Y.; Shao, Z. An enhanced target detection algorithm for maritime search and rescue based on aerial images. *Remote Sens.* **2023**, *15*, 4818. [CrossRef]
45. Sun, C.; Zhang, Y.; Ma, S. Dflm-yolo: A lightweight yolo model with multiscale feature fusion capabilities for open water aerial imagery. *Drones* **2024**, *8*, 400. [CrossRef]
46. Liu, K.; Ma, H.; Xu, G.; Li, J. Maritime distress target detection algorithm based on YOLOv5s-EFOE network. *IET Image Process.* **2024**, *18*, 2614–2624. [CrossRef]
47. Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.H.; Khan, F.S. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 17425–17436.
48. Zhang, H.; Zhang, S. Focaler-iou: More focused intersection over union loss. *arXiv* **2024**, arXiv:2401.10525.
49. 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, Waikoloa, HI, USA, 3–8 January 2022; pp. 2260–2270.
50. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]

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