Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Open-Water UAV Small-Object Detection via Re-parameterized Convolutions and Context- Guided Fusion

Yuping Shui

suiypcindy@163.com

Changchun University

Xinlei Zhi

Changchun University

Research Article

Keywords: UAV imagery, small-object detection, end-to-end detection, re-parameterization, multi-scale feature fusion

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8792984/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Open-Water UAV Small-Object Detection via Re-parameterized Convolutions and Context-Guided Fusion

1st Yuping Shui*
School of Computer Science and Technology
Changchun University
Chang Chun, China
suiypcindy@163.com*

2nd Xinlei Zhi
School of Computer Science and Technology
Changchun University
Chang Chun, China
17613273741@163.com

Abstract

Object detection for open-water search and rescue is a critical entry point in the perception pipeline. Constrained by UAV payload and onboard computer, existing detectors already offer competitive speed and accuracy; however, they still struggle in real open-water scenes. Swimmers and life-saving devices typically appear as tiny objects with weak texture and blurred boundaries, while wave patterns and specular reflections continuously generate high-response background pseudo-structures, leading to both missed detections and false alarms. To address these challenges, we develop an efficient end-to-end detector tailored to open-water scenarios, improving tiny-object visibility and background suppression without relying on complex post-processing. Our network first introduces a re-parameterizable multi-branch convolutional representation and cross-layer reuse in the backbone to strengthen fine-grained edge/shape cues while preserving a compact inference path. We then employ a channel-partitioned subspace self-attention strategy in high-level semantic encoding to impose necessary long-range dependencies at low overhead and to reduce cross-region spurious associations triggered by sea-surface pseudo-structures. Building on this, we further design a context-guided selective cross-scale feature pyramid that injects semantic information into high-resolution branches in a more controlled manner, mitigating the cumulative amplification of noise along the multi-scale propagation pathway. We

evaluated the effectiveness of the proposed method through experiments. Compared with RT-DETR-R18, our approach improves mAP@0.5 and mAP@0.5:0.95 by 4.73 and 6.34 percentage points, respectively, while reducing the model size by approximately 19.9%. These results indicate that the proposed design achieves a more favorable accuracy and efficiency trade-off by simultaneously boosting detection performance and lowering model complexity.

Keywords: UAV imagery; small-object detection; end-to-end detection; re-parameterization; multi-scale feature fusion

1 Introduction

In open-water search and rescue operations[1], unmanned aerial vehicles (UAVs) can swiftly capture nadir-view imagery across extensive maritime areas, often serving as the initial entry point in the perception pipeline[2]. Despite significant advancements in real-time detection frameworks regarding the speed–accuracy trade-off[3], these systems still face considerable challenges under actual sea-surface conditions. Swimmers and life-saving equipment typically manifest in aerial imagery as diminutive, low-textured, and low-contrast targets[4], which exhibit significant scale variation due to fluctuations in flight altitude and camera pitch. Additionally, wave patterns, foam, and specular reflections generate persistent high-response background pseudo-structures that frequently trigger false positives and obscure genuine targets, resulting in both missed detections and false alarms[5].

Class imbalance, scale imbalance, spatial imbalance, and imbalance among optimization objectives are widely acknowledged as critical factors that constrain detection accuracy and training stability [6].In open-water UAV imagery, these imbalances manifest more acutely: the prevalence of small objects and significant scale drift render fine-grained cues particularly susceptible to attenuation due to hierarchical down sampling. Additionally, the predominance of sea-surface backgrounds and the pronounced responses of pseudo-structures are intensified during multi-scale interactions and propagated across feature levels, resulting in more persistent false detections that are consequently more challenging to suppress [7]. Therefore, addressing open-water SAR detection cannot be achieved solely by increasing model size; it necessitates architectural designs that simultaneously enhance representations of tiny objects, enforce controllability in cross-scale semantic injection, and mitigate background pseudo-structures, all while adhering to the computational and latency constraints of edge deployment.

To alleviate the training and inference challenges posed by high-resolution inputs and the frequent occurrence of small objects, previous research has investigated various strategies. Tiling and uniform cropping transform large images into more manageable local views, thereby reducing the computational load per pass[8]. Region-focused pipelines initially coarsely localize areas with a high concentration of targets and subsequently conduct fine-grained detection to enhance the effective resolution for densely packed small objects[9]. Additional methods employ adaptive zooming or search mechanisms to assign more suitable observation scales, thereby improving the detectability of small objects within expansive scenes[10]. From an inference standpoint, cascaded sparse queries have been implemented to circumvent unnecessary feature updates, thereby accelerating high-resolution detection of small objects while minimizing redundant computations[11]. Global–local self-adaptive frameworks integrate cropping with local enhancement to achieve a balance between accuracy and computational cost[12]. While these approaches provide valuable insights for small-object detection under constrained computational resources, their effectiveness diminishes significantly in open-water environments, where background pseudo-structures are both prominent and spatially widespread. Specifically, pipelines that depend on region filtering or repeated cropping may be easily misled by high-response sea-surface patterns, while unconstrained cross-scale injections or global interaction can inadvertently propagate noise, leading to more persistent false detections. Consequently, there is a pressing need for a more rigorously controlled information-flow mechanism that can enhance recall while minimizing false positives under real-time constraints.

More critically, under conditions of strong sea-surface pseudo-structures, extremely small targets, and pronounced scale drift, global interaction without explicit constraints is prone to redundant attention and spurious cross-region associations that are triggered—and then amplified—by background peaks[13]. Likewise, multi-scale semantic injection without an effective denoising mechanism can transmit sea-surface responses as pseudo semantics across feature levels, increasing the persistence of false alarms and undermining localization reliability[14].To this end, we have constructed an efficient end-to-end detection network for open waters: under the premise of keeping the inference overhead controllable, we simultaneously enhance the representation of small targets, cross-scale alignment and sea surface interference suppression through structural design. Compared with the RT-DETR-R18 baseline [48], our method achieves a better accuracy–efficiency trade-off on SeaDronesSee, dataset improving detection performance by about 4.7 percentage points while reducing model complexity. Our main contributions

are as follows:

(1) RMConv, a re-parameterizable multi-branch convolutional representation unit, is designed to enhance the backbone network's representation of small-object details and local structures at a lower inference cost;

(2) A lightweight CPSA (Channel-Partitioned Subspace Self-Attention strategy) is proposed to remedy the issue that "relying only on local textures is unreliable" in sea-surface scenes via more controlled global interaction, while avoiding the efficiency loss caused by redundant attention;

(3) A context-guided feature fusion and gated enhancement mechanism (CGFM) is constructed to highlight target-consistent features and suppress the propagation of sea-surface pseudo-structures during cross-scale aggregation, thereby improving detection accuracy.

2 Related Work

This section provides a structured review and analysis of prior work on object detection from the UAV perspective in open-water scenarios. We focus on two major research lines: CNN-based real-time detection methods and Transformer-centric end-to-end detection approaches.

2.1 Traditional Method

First-stage CNN detectors, exemplified by YOLO, have long been at the forefront of edge-side deployments. Their advantages stem[15] from a well-established ecosystem of dense prediction paradigms and convolutional operators, clear inference pathways, and the ability to achieve satisfactory accuracy within stringent delay constraints. YOLOX[16] improves training stability and enhances accuracy limits through its Anchor-free paradigm, decoupled detection heads, and more effective label allocation. YOLOv7[17] systematically integrates its architecture and training enhancements, ensuring robust competitiveness in real-time detection tasks. In environments characterized by a high density of small objects and numerous candidate predictions, traditional YOLO-style methods often depend on non-maximum suppression (NMS) for candidate filtering. When utilizing high-resolution inputs or high-recall configurations, NMS can lead to increased latency and redundant computations, potentially exacerbating the accumulation of false positives. To mitigate this deployment challenge[18], YOLOv10[19] explicitly prioritizes "Real-Time End-to-End" performance and investigates NMS-free training in conjunction with component-level efficiency optimizations, with the goal of minimizing reliance on post-processing and enhancing overall efficiency. YolooW[20] integrates scale-adaptive

mechanisms into its architectural design and incorporates targeted treatments for disturbances such as sea-surface glare, thereby enhancing detection stability in open-water conditions. FPN facilitates multi-scale feature fusion through top-down pathways and lateral connections, establishing itself as one of the most widely adopted pyramid paradigms in general detection[21]. Building on this foundation, PAN introduces a bottom-up pathway to enhance cross-layer information flow[22]. EfficientDet further advances this concept by proposing weighted bidirectional fusion structures, such as BiFPN, to improve fusion efficiency and scalability[23]. However, in scenarios like open waters, [24]where the background pseudo-structure exhibits a pronounced response, the absence of selective constraints in multi-scale fusion may lead to the amplification of background responses. This amplification can propagate across layers during the semantic injection process, resulting in more persistent false detections and location drifts. RepVGG[25] exemplifies a paradigm in which the training phase involves multiple branches, while the inference phase is effectively consolidated into a single-path convolution. This approach enables the model to achieve enhanced representational capabilities while preserving the simplicity of the deployment computational graph[27]. Lightweight Attention modules, such as Coordinate Attention[26], can introduce direction-aware feature recalibration with minimal overhead and are frequently employed to optimize effective feature allocation in mobile or edge-side environment.

2.2 End-to-end detection method based on Transformer

End-to-end Transformer[28]-based detectors redefine object detection as a set prediction problem, enforcing one-to-one supervision through bipartite matching. This approach significantly diminishes dependence on handcrafted anchor designs and post-processing techniques, such as non-maximum suppression (NMS), resulting in a more cohesive training and inference pipeline. Deformable DETR[29] substitutes dense global attention with sparsely sampled deformable attention. This method aggregates information from a limited set of critical sampling points across multi-scale features, thereby significantly reducing computation and improving both convergence and small-object performance. In addressing slow convergence and inadequate localization priors, Conditional DETR[30] enhances the relationship between queries and spatial positions through conditional cross-attention, which accelerates training. Meanwhile, DAB-DETR[31] explicitly parameterizes queries as dynamic anchor-box coordinates to incorporate more direct localization priors, thereby improving training stability. DINO[32] further enhances the practicality and performance potential of end-to-end detection by integrating

denoising training and query initialization, demonstrating more stable optimization behavior across various settings .Despite the significant advancements in efficiency and convergence achieved by end-to-end Transformer-based approaches in streamlining the detection pipeline, challenges persist in open-water conditions characterized by strong sea-surface pseudo-structures, extremely small targets, and significant scale drift. Failure to effectively constrain global interaction can lead to spurious cross-region associations triggered by strong background responses, resulting in redundant computation. Moreover, without dedicated denoising mechanisms accompanying multi-scale semantic injection, sea-surface activations may spread as "pseudo semantics" across feature levels, heightening the risk of false alarms and compromising localization reliability[33]. These findings underscore the importance of not simply enhancing global interaction or fusion strength for real-time end-to-end detection in open-water settings. Instead, the focus should be on limiting the interaction scope and information channels with manageable costs to ensure that context modeling and multi-scale injection specifically contribute to denoising and consistency learning.

3 Proposed Framework and Architecture

In open-water search-and-rescue operations, challenges in object detection stem from a combination of factors, including inadequate representation for small targets and significant interference from sea-surface pseudo-structures. To address these issues, we have developed an effective detection framework for open-water search and rescue utilizing RT-DETR, as depicted in Fig. 1. This framework comprises a re-parameterizable RMBB backbone, a CPSA module at the high-level semantic encoding stage, and a CGFM neck. The RMBB enhances detailed structural representations while maintaining a concise inference graph. The CPSA module establishes essential long-range dependency constraints through controlled global interaction, and the CGFM facilitates context-guided, selective cross-scale information injection to mitigate noise propagation across the multi-scale pathway.

3.1 Backbone network

3.1.1 RMBB structure

In open-water environments, swimmers and life-saving gear often display poor texture, low contrast, and blurry boundaries. Meanwhile, wave patterns and shiny reflections consistently produce high-frequency, high-intensity "pseudo-structures," causing significant background interference in feature space[34]. This creates a dual challenge: firstly, the subtle edge and shape details crucial for small targets

are near the noise threshold in shallow layers, easily dampened or erased by subsequent downsampling and nonlinear changes, heightening the risk of missing distant point-like targets. Secondly, pseudo-structures on the sea surface can be magnified by local convolutional operations, spreading across higher layers through inter-layer data transmission and fusion, resulting in increased false positives and localization drift. To address this challenge, we introduce a re-parameterizable multi-branch convolutional cross-stage backbone unit (RMBB) into the network backbone. The core RMConv operator employs parallel branches during training to expand the local receptive field and enhance directional sensitivity, thereby enhancing the detection of faint boundaries, delicate structures, and local shape features. These branches are folded into a single convolutional path to streamline the computation graph and prevent additional runtime overhead. The RMBB approach integrates cross-stage partial connections to enforce structural constraints on operator stacking: a portion of channels follows a shallow path with minimal alterations to maintain information fidelity and gradient stability, while the remaining channels take a deeper path for detailed enhancement to reinforce target responses and suppress the spread of background pseudo-structures. By combining the principles of "rich representation during training, compact structure during inference" and "partial fidelity, partial enhancement," the signal-to-noise ratio (SNR) of backbone outputs is enhanced, offering a more dependable input for subsequent global modeling and multi-scale fusion. The architecture of RMBB is depicted in Figure 1.

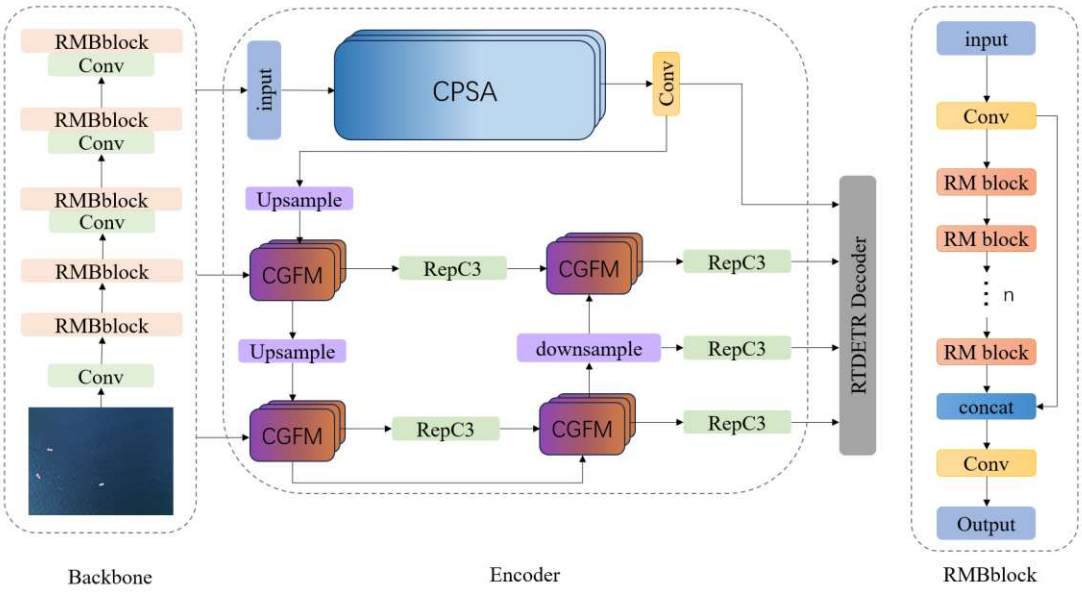


Figure 1 The overall architecture

(1) RMConv: A Re-parameterizable Multi-Branch Convolutional Representation Module

During training, RMConv adopts a parallel multi-branch topology to increase local representational capacity, as illustrated in Figure 2. Given an input feature map x, RMConv consists of two 3×3 convolution branches to learn complementary local texture/edge patterns; a 1×1 convolutional channel branch to strengthen channel mixing and response calibration; and an identity (BN) branch that is enabled when the stride is 1 and the input/output channel dimensions are identical, providing a low-perturbation shortcut to stabilize gradient propagation. The outputs of all branches are summed element-wise and then passed through a nonlinear activation function (SiLU in our implementation). The training-time forward process can be written as:

$$y = SiLU\Big(B_{3\times3}^{(1)}(x) + B_{3\times3}^{(2)}(x) + B_{1\times1}(x) + B_{id}(x)\Big). \quad (1)$$

where y denotes the output feature map. $B_{3\times3}^{(1)}$ and $B_{3\times3}^{(2)}$ represent the two 3×3 convolution (with BN) branches, which provide complementary local spatial modeling; $B_{1\times1}(\cdot)$ convolution (with BN) branch for channel mixing and response calibration; and $B_{id}(.)$ denotes the identity (BN) branch, which stabilizes information and gradient propagation via a low-perturbation shortcut. During inference, to avoid the extra operators and memory-access overhead introduced by multi-branch execution, RMConv equivalently folds all branches (including BN) into a single 3×3 convolution layer. For an arbitrary Conv+BN branch, let the convolution weight be w and the bias be b, and let the BN parameters be $(\gamma, \beta, \mu, \sigma^2, \varepsilon)$. The fused equivalent convolution parameters $(\hat{W}, \hat{b})$ are given by:

$$\widehat{W} = \frac{\gamma}{\sqrt{\sigma^2+\varepsilon}} \cdot W, \hat{b} = \beta + \frac{\gamma}{\sqrt{\sigma^2+\varepsilon}} \cdot (b - \mu). \quad (2)$$

We directly sum the fused $(\hat{W}, \hat{b})$ of the two 3×3 branches; for the 1×1 branch, $\hat{W}$ is embedded into the center of a 3×3 kernel by zero-padding the remaining positions and then added to the summation.The identity (BN) branch can be regarded as $W = I$ and $b = 0$, fused in the same manner via Eq.(2), and included in the summation as well. This yields a single equivalent convolution (W_{cq}, b_{cq}), which folds the integrated advantages of the multi-branch training design into a compact

single-path computation graph during inference without changing output equivalence, thereby maintaining favorable computational complexity and latency characteristics.

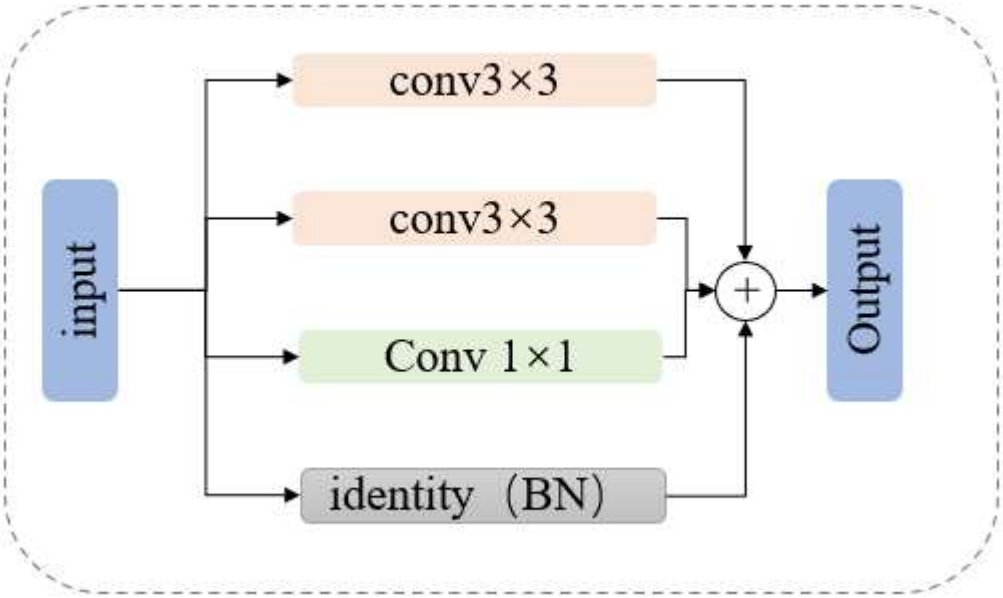


Figure 2 RMConv structure

(2) RMBB: Backbone Construction with Cross-Stage Partial Connections

Using RMConv as an isolated local enhancement unit is still insufficient to compensate for the loss of fine details caused by deep stacking. In open-water scenarios, the responses of tiny targets are often already close to the noise floor at shallow layers; if all channels undergo successive nonlinear transformations and downsampling, fine-grained edge/shape cues are more likely to be "drowned out" by sea-surface pseudo-structures during inter-layer propagation[35]. This typically manifests in later decoding and regression as reduced recall and localization drift. To alleviate this issue, a cross-stage partial connection (CSP) paradigm is adopted in the backbone to impose a structured constraint on RMConv stacking: features are split along the channel dimension into a shallow path and a deep path. The shallow path preserves higher information fidelity with fewer transformations and provides a stable shortcut for gradient flow, whereas the deep path cascades several RMBlocks to learn fine-grained discriminative transformations. The RMBlock structure is shown in Figure 3.RMBlock. Given an input feature z, RMBlock consists of two RMConv layers with SiLU activations, and employs a residual connection to stabilize information and gradient propagation. Without significantly increasing the

complexity of the inference path, this design strengthens the modeling of faint boundaries and local shape cues, and mitigates the tendency of deep-layer rewriting to overwrite weak target information.

$$(x_s, x_d) = Spli\,t(x)\,, x_s \in R^{\rho C\times H\times W}, x_d \in R^{(1-\rho)C\times H\times W} \tag{3}$$

where $\rho \in (0,1)$ denotes the channel allocation ratio.The deep path extracts more discriminative fine-grained representations through a transformation branch composed of multiple RMBlocks:

$$x'_d = f_d(x_d)\,, f_d(.) = RMBloc\,k^{\circ N}(.)\,, \tag{4}$$

Finally,the outputs from the shallow and deep paths are concatenated along the channel dimension, followed by a 1×1 projection for feature fusion and dimensional alignment:

$$y = Pro\,j(Conca\,t(x_s, x'_d))\,, \tag{5}$$

where $Proi(.)$ denotes a 1×1 convolution used to fuse information from the two paths and restore the target channel dimensionality. The key inductive bias of this design is that the network no longer forces all channels to be rewritten at an equal depth; instead, it adopts a division of labor between partial fidelity preservation and partial enhancement, so that—within the same scale level—it simultaneously maintains stable low-frequency structures and high-frequency details strengthened by RMConv. This structure improves the visibility of long-range tiny targets and stabilizes localization, while reducing the dominance of background pseudo-structures in deeper representations, thereby providing a more reliable input foundation for subsequent global modeling and selective cross-scale fusion.

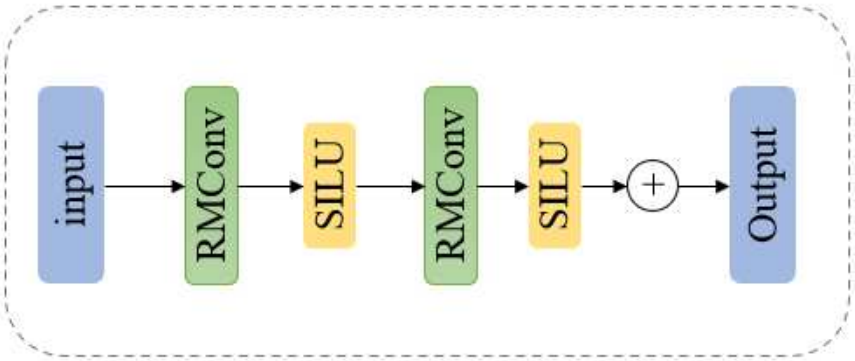


Figure 3 Rmblcok structure

3.2 CPSA EncoderLayer

In open-water scenarios, tiny targets often lack sufficient local texture and color cues to substantiate their presence, while sea-surface glare, wave patterns, and foam can readily produce target-like, high-

intensity structural responses within local windows[37], leading to a large number of isolated high-confidence false positives. Introducing moderate global dependency modeling helps exploit broader contextual consistency, enabling the model to distinguish pseudo-structures that exhibit strong local responses but lack semantic support from true target regions that are semantically coherent and physically plausible in space[38]. Motivated by this, we place global interaction on the high-level semantic pathway and follow the efficiency principle of the RT-DETR hybrid encoder: global modeling is performed only on the lowest-resolution feature level to obtain the necessary long-range constraints while preserving real-time performance. The core of global modeling lies not in whether attention is introduced, but in how its interaction scope and information channels are constrained. To this end, we propose a Channel-Partitioned Subspace Attention strategy (CPSA), which reformulates global interaction from "full-channel, multi-head rewriting" into subspace, single-head correction. Specifically, given the top-level semantic feature $x \in R^{B\times C\times H\times V}$, we split it along the channel dimension into an attention subspace x_a and a bypass subspace x_b. Single-head self-attention is computed only on x_a to obtain controlled long-range correction signals, while x_b serves as a low-perturbation bypass for stable transmission of local representations; the two are then concatenated and fused via a lightweight projection to produce the output. The overall formulation is given by:

$$CPS\,A(x) = Pro\,j(Conca\,t(A\,t\tan(x_a)\,,x_b))\,. \tag{6}$$

The CPSA encoder layer's architecture is depicted in Figure 4. Initially, normalization is applied within the attention subspace to stabilize the feature distribution. Subsequently, is generated through a convolution operation, followed by single-head self-attention. A convolutional FFN, along with residual updates, is then utilized to maintain consistency with the baseline in the training process. This

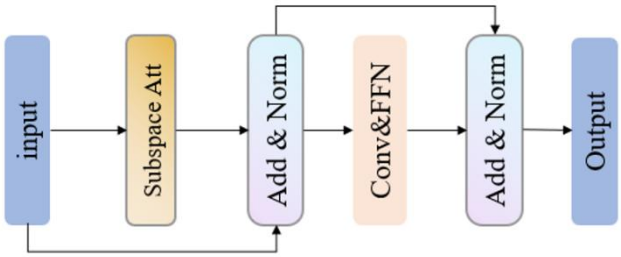


Figure 1 CPSA encoder layer's architecture

approach ensures that the replacement remains localized and easier to attribute. The primary

advantages of this design include restricting attention computation to a smaller representational subspace, eliminating multi-head redundancy, and reducing unnecessary correlation estimation at the source. Additionally, the bypass subspace establishes a stable information channel, thereby decreasing the risk of amplifying sea-surface pseudo-structures through full-channel association. Essentially, CPSA prioritizes sufficient global context over maximizing interaction strength, enhancing computational efficiency. It functions as a lightweight corrector for high-level semantics, offering robust semantic support for subsequent cross-scale fusion and decoder matching.

3.3 CGFM

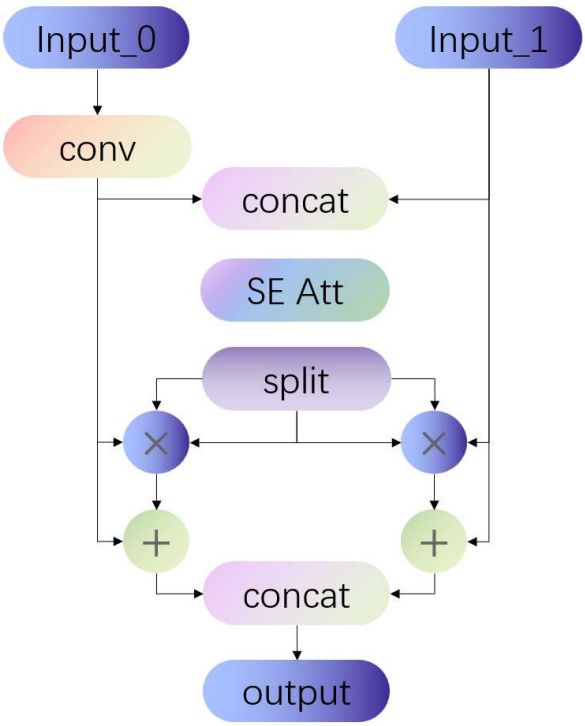


Figure 5 CGFM structure

In open-water scenarios, multi-scale fusion exhibits significantly more pronounced adverse effects. While the high-resolution component retains the edge and shape details essential for detecting small targets, sea-surface wave patterns and specular highlights can provoke equally strong and persistent responses at the same scale. When conventional FPN/PAN-style unconditional top-down summation is employed, these background pseudo-structures are often inadvertently semanticized

during the semantic injection process and subsequently propagated across scales. This results in a more coherent—and consequently more challenging to suppress—manifestation of accumulated false positives or localization drift. Thus, the primary challenge in open-water environments is not simply "insufficient fusion," but rather "noise amplification and cross-scale diffusion introduced by fusion." In response to this observation, we redefine cross-scale interaction within the pyramid as context-constrained feature injection. Rather than directly incorporating high-level features into low-level ones, we first treat high-level features as contextual priors. This approach generates a set of modulation weights that selectively filter cross-scale information, allowing only the semantic components aligned with target evidence to be injected into the high-resolution pathway. To facilitate this controlled injection, we have developed a Context-Guided Fusion Module (CGFM). The CGFM architecture is illustrated in Figure 5. Given two adjacent-level features—a low-level high-resolution feature $F_l \in R^{B\times C\times H\times W}$ and a high-level low-resolution feature $F_h \in R^{B\times C\times \frac{H}{2}\times \frac{W}{2}}$.CGFM starts by upsampling F_h to achieve spatial alignment and applying a lightweight channel mapping to match the channel dimensions. The aligned features are subsequently concatenated along the channel axis, after which a Squeeze-and-Excitation SE channel attention[39] module generates branch-specific channel modulation weights from global context; the weights are then split and applied to their corresponding branches, thereby realizing a “filter-before-fuse” interaction. To mitigate detail attenuation caused by reweighting, residual passthroughs are introduced in both branches to preserve the original fine details; the modulated features are thereafter concatenated and projected to produce the fused output. This procedure can be abstracted as follows:

$$F_{out} = CGF\,M(F_l, F_h) = \phi(F_l + M(F_h) \odot \psi(U_P(F_h))) \tag{4}$$

where $\mathrm{Up}(\cdot)$ denotes upsampling for spatial resolution matching,$\psi(\cdot)$ is a lightweight channel-alignment mapping,and $M(F_h) \in [0,1]^{B\times C\times H\times W}$ denotes the modulation weights generated from high-level context (implemented via an SE block followed by a sigmoid).$\odot$ indicates element-wise modulation,and $\phi(\cdot)$ denotes the post-fusion projection mapping.The key insight is that cross-scale information injection is no longer an unconditional summation;instead,contextual weights are first used to filter and denoise the injected information,making high-level semantics behave more like a corrective term rather than a rewriting term over low-level details.Because modulation is applied

before fusion,even if sea-surface pseudo-structures elicit strong responses at high resolution, they are less likely to be amplified into stable semantics through cross-scale interaction; conversely, true target responses are more likely to be consistently reinforced under the constraint of contextual consistency. Ultimately, Consequently,the decoder is provided with multi-scale representations of improved semantic coherence and more tightly regulated noise propagation, thereby better meeting the dual requirements of cross-scale alignment and background suppression in open-water tiny-target detection.

4 Experiments and results analysis

4.1 Datasets and evaluation metrics

4.1.1 Dataset preparation

Experiments are primarily conducted on the SeaDronesSee dataset, a public open-water search-and-rescue dataset consisting of UAV-captured images in open-water settings with challenging sea-surface backgrounds containing distractions like wave textures, foam wakes, and specular reflections. The official training and evaluation split is strictly adhered to, with the training set utilized for model training, while all quantitative results and ablation studies are based on the validation set. Due to the unavailability of annotations for the official test set, test-set performance is not reported to prevent uncontrollable evaluation discrepancies and reproducibility issues. To analyze the scale characteristics of this task, we also examine the width-height distribution of bounding boxes, depicted in Figure 6. The analysis shows that the majority of boxes are concentrated in the small-size category, displaying

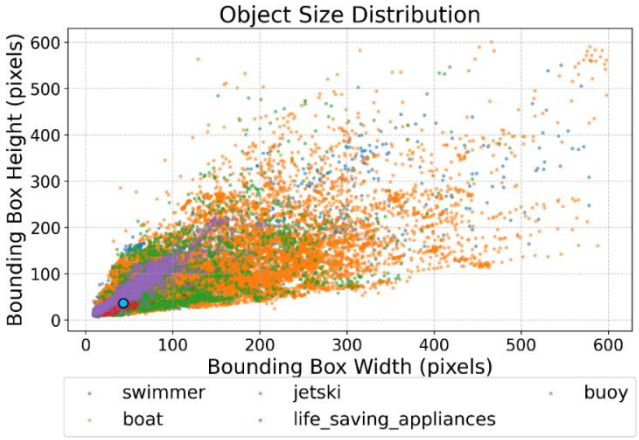


Figure 2 Category-wise bounding-box size distribution on SeaDronesSee

a long-tailed distribution. This observation corresponds with the practical imaging features of open-water scenarios, where distant point-like targets are prevalent, leading to significant scale variations. This phenomenon underscores the task's high sensitivity to detailed representations and cross-scale alignment at the data level.

4.1.2 Evaluation metrics

We adopt the standard IOU-based evaluation protocol widely used in object detection. Let true positives (TP), false positives (FP), and false negatives (FN) denote the numbers of correct detections, false alarms, and missed detections, respectively. The detection precision (Precision) and recall (Recall) are defined as:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

In particular, TP assignment is jointly determined by class agreement and an IOU threshold: a predicted box is counted as a TP only when its category matches that of the corresponding ground-truth box and their intersection-over-union is no smaller than a preset threshold τ (i.e., IOU$\geq\tau$). For a predicted box Bp and a ground-truth box Bg , IOU is defined as:

$$Io\,U\left(B_p, B_g\right) = \frac{\left|B_p \cap B_g\right|}{\left|B_p \cup B_g\right|} \tag{7}$$

To provide a unified assessment that jointly reflects precision and recall, we adopt the COCO-style average precision metric. For each class, AP is computed as the area under the precision–recall (PR) curve; correspondingly, the mean average precision (MAP) is defined as the average of AP over all classes:

$$mAP = \frac{1}{|C|}\sum_{c\in C} AP_c \tag{8}$$

To quantify model size and computational overhead, we report both the parameter count and computational cost. Taking a standard convolutional layer as an example, given a $k \times k$ kernel with C_{in} input channels and C_{out} output channels, the number of parameters can be written as:

$$Params = k^2 \cdot C_{in} \cdot C_{out} \tag{9}$$

Correspondingly, the floating-point operation count (FLOPs) can be approximated as:

$$FLOPs \approx 2 \cdot k^2 \cdot C_{in} \cdot C_{out} \cdot H \cdot W, \tag{10}$$

4.2 Implementation of the proposed method

To ensure reproducibility and stability, all experiments are conducted on a unified computing platform. The operating system is Ubuntu 20.04.4 LTS, and the deep learning stack is based on PyTorch 2.0.1 with torch vision 0.15.2. The hardware comprises two Intel Xeon Gold 6226R CPUs and three NVIDIA GeForce RTX 3090 GPUs (24 GB memory per card). Our method is built upon the public RT-DETR-R18 implementation: apart from replacing the backbone with RMBB, introducing CPSA in the high-level semantic encoder, and substituting pyramid fusion nodes with CGFM to construct a context-guided selective cross-scale feature pyramid network., all other components—including training data processing, matching/loss definitions, and decoder settings—are kept identical to the baseline to ensure fair and attributable comparisons. Key training hyperparameters and data-loading configurations are summarized in Table 1.

Table 1 training hyperparameters and data-loading configurations

Parameter	Settings
epochs	300
Batch	16
Imgsz	640
Workers	8
optimizer	AdamW
Lr0	0.0001
Momentum	0.9
Weighdecay	0.0001

4.3 Ablation research and analysis

4.3.1Ablation analysis of backbone networks

To systematically evaluate the marginal contributions and synergistic effects of our three structural modifications in open-water scenarios, we employ an ablation design that integrates both single-factor and multi-factor variants, using RT-DETR-R18 as the baseline model (A). Under consistent training conditions, we sequentially introduce the three key mechanisms and construct the corresponding variants:B (RMBB) substitutes the backbone with RMBB, incorporating re-parameterization into a single-path convolution at inference, with complexity reported accordingly; C (CPSA) replaces the baseline attention encoding unit with channel-partitioned subspace attention in the high-level semantic encoder; and D (CGFM) substitutes the pyramid fusion nodes with CGFM which imposes contextual constraints on cross-scale semantic injection to mitigate noise propagation. Based on these modifications, we assess A, A+B, A+C, A+D, the three pairwise combinations A+B+C, A+B+D, A+C+D, and the complete model A+B+C+D (Ours). As shown in Table 2 All models are trained on the same dataset and evaluated under a unified protocol on the official validation set.

Table 2 Ablation Experiment Results

Model	mAP0.5/%	mAP0.5:0.95/%	GFLOPS/G	Model size/MB
A	80.63	47.38	58.20	40.5
A+B	81.97	48.38	54.36	36.03
A+C	82.34	49.56	53.27	37.51
A+D	82.85	48.89	53.64	37.87
A+B+C	83.90	50.36	47.62	35.24
A+B+D	84.46	51.74	46.58	35.37
A+C+D	84.63	52.19	45.47	34.26
A+B+C+D	85.36	53.72	42.62	32.45

From the perspective of individual components, all three modifications yield consistent gains, albeit with different emphases. RMBB (A+B) improves accuracy while substantially reducing computation and model size, indicating that its primary benefit stems from more efficient fine-grained

representations and a more compact inference graph. (A+C) delivers a more pronounced boost in mAP@0.5: 0.95, suggesting that controlled high-level global dependency modeling helps improve localization quality and suppress false positives. (A+D) shows a clearer improvement in mAP@0.5, reflecting that context-constrained cross-scale semantic injection directly enhances target detectability.Moreover, the pairwise combinations generally achieve larger accuracy gains while continuing to reduce complexity. Among the three two-way variants, (A+C+D) attains the highest mAP@0.5: 0.95, indicating that the combination of controlled global constraints and controlled cross-scale fusion provides stronger complementary suppression against pseudo-structural disturbances in open-water scenes. Ultimately, the full model A+B+C+D (Ours) reaches 85.36% mAP@0.5 and 53.72% mAP@0.5: 0.95, while reducing GFLOPs from 58.20G to 42.62G and model size from 40.50MB to 32.45MB, validating that the three mechanisms deliver coherent synergistic benefits in both accuracy and efficiency.

4.4 Comparative experiment analysis

To further contextualize the proposed method across different detection paradigms, we benchmark the final model (Ours) against a range of representative detectors. The comparison set spans one-stage dense detection frameworks (RetinaNet[40], FCOS[41], CenterNet[42], VFNet[43], YOLOv5m[44], YOLOv7-tiny[45], YOLOv8m[46], and YOLOX[16]), a two-stage framework (Faster R-CNN[47]), and end-to-end Transformer-based detectors (Deformable DETR[29] and RT-DETR-R18). All methods are evaluated under the same data split and a unified evaluation protocol. We report mAP@0.5 and mAP@0.5: 0.95, together with computational complexity (GFLOPs) and model size (MB), to reflect the overall accuracy–efficiency trade-off. The results are summarized in Table 3.

Table 3 Comparison of experimental results

Model	mAP0.5/%	mAP0.5:0.95/%	GFLOPS/G	Model size/MB
RetinaNet	55.95	27.35	205.24	36.15
FCOS	65.03	38.66	196.76	31.84
CenterNet	69.17	38.78	51.02	14.21
FasterR-CNN	73.48	44.08	206.67	41.13
VFNet	72.84	43.17	48.38	32.72
YOLOX	81.43	45.69	26.86	9.0
Yolov5m	75.66	42.36	49.03	21.2
Yolov7-tiny	74.35	43.57	13.27	6.03
Yolov8m	77.84	43.16	78.77	25.9
DeformableDETR	76.52	42.85	196	40
RTDETR-R18	80.63	47.38	58.20	40.5
Our	85.36	53.72	42.62	32.45

Taken together, our method achieves the strongest overall performance on SeaDronesSee-dataset Compared with the end-to-end baseline RT-DETR-R18, Ours improves the more stringent mAP@0.5:0.95 by six percentage points, while reducing computational cost and model size, yielding a more favorable accuracy–efficiency trade-off. These results indicate that our structural refinements not only enhance detection accuracy, but also effectively eliminate redundant computation and parameters, yielding a more favorable trade-off between representational capacity and complexity. Relative to conventional CNN-based detectors, Ours exhibits a more pronounced advantage under the stricter metric. Notably, RetinaNet and FCOS achieve substantially lower mAP@0.5:0.95 on this dataset, indicating that under cluttered sea-surface backgrounds and an extreme tiny-target distribution, relying solely on a generic dense detection paradigm makes it difficult to reconcile recall with localization stability. By contrast, one-stage methods such as the YOLO family and YOLOX remain reasonably competitive on mAP@0.5, yet they still lag behind Ours on mAP@0.5: 0.95. This discrepancy suggests that, beyond improving “detectability,” the more critical issue in open-water scenarios is suppressing false positives induced by sea-surface pseudo-structures and alleviating the instability of bounding-box regression. By enforcing controlled global constraints and controlled cross-scale

semantic injection, our approach improves cross-scale feature consistency and strengthens regression robustness, thereby yielding larger and more stable gains under stringent evaluation.

5 Conclusion

To tackle the significant challenges posed by strong background interference on sea surfaces, including pseudo-structures created by wave textures, foam, and specular reflections, as well as pronounced target scale drift, a high proportion of tiny targets, occlusion, and inter-class appearance similarity, we propose three complementary architectural enhancements to the end-to-end real-time RT-DETR-R18 framework. These enhancements are integrated into a unified model to bolster representation robustness and localization stability in open-water environments. Experimental results indicate that our model achieves superior detection accuracy while simultaneously reducing computational costs and model size, thus providing a more favorable accuracy–efficiency trade-off. Ablation studies and cross-method comparisons collectively validate the effectiveness of the proposed approach under complex sea-surface backgrounds and extreme tiny-target conditions. Notably, the performance remains more stable under the stringent mAP@0.5:0.95 criterion, demonstrating that the proposed controlled global modeling and controlled cross-scale semantic injection effectively mitigate the influence of sea-surface pseudo-structures and enhance the robustness of bounding-box regression. Future research can enhance stability and generalization under complex sea states. To mitigate localization bias for extremely small targets and decrease missed detections in densely occluded environments, it is beneficial to implement finer-grained scale-binning evaluations and to develop corresponding architectural and training strategies tailored to each scale regime. Simultaneously, to address extreme conditions such as abrupt illumination shifts and severe specular glare, the incorporation of targeted robustness-oriented training schemes, along with uncertainty-aware modeling, could improve reliable detection across a wider array of maritime UAV scenarios.

Acknowledgment

This research received no external funding.

- Availability of data and materials The data sets generated during and analysed during the current study are available from the corresponding author on reasonable request.

- No potential conflict of interest was reported by the authors.

References

[1] Huang, J., Han, N., Jiang, B. Q., Cao, Y., & Han, X. T. (2022). Water search and rescue (SAR) for ship accidents in China: analysis of 12 years' data. *Annals of translational medicine*, *10*(22), 1207.

[2] Geng, W., Yi, J., & Cheng, L. (2025). An efficient detector for maritime search and rescue object based on unmanned aerial vehicle images. *Displays*, *87*, 102994.

[3] Cai, Y., Luan, T., Gao, H., Wang, H., Chen, L., Li, Y., ... & Li, Z. (2021). YOLOv4-5D: An effective and efficient object detector for autonomous driving. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-13.

[4] Islam, M. J., Xia, Y., & Sattar, J. (2020). Fast underwater image enhancement for improved visual perception. *IEEE robotics and automation letters*, *5*(2), 3227-3234.

[5] Zhou, F., Wu, Y., Dai, Y., Wang, P., & Ni, K. (2019). Graph-regularized laplace approximation for detecting small infrared target against complex backgrounds. *IEEE Access*, *7*, 85354-85371.

[6] Wang, Y., Liu, W., & Xing, W. (2021). Balanced-RetinaNet: solving the imbalanced problems in object detection. *Journal of Electronic Imaging*, *30*(3), 033009-033009.

[7] Li, S., Li, C., Wang, Z., Jia, Z., Zhu, J., Cui, X., & Liu, J. (2024). Pas: a scale-invariant approach to maritime search and rescue object detection using preprocessing and attention scaling. *Intelligent Service Robotics*, *17*(3), 465-477.

[8] Fan, P., Diao, Y., Li, F., Zhao, W., & Chen, Z. (2024). SRSegNet: Super-resolution-assisted small targets polyp segmentation network with combined high and low resolution. *Journal of King Saud University-Computer and Information Sciences*, *36*(3), 101981.

[9] Zhong, S., Zhou, H., Zheng, Z., Ma, Z., Zhang, F., & Duan, J. A. (2024). Hierarchical attention-guided multiscale aggregation network for infrared small target detection. *Neural Networks*, *171*, 485-496.

[10] Zhou, H., Chen, P., Li, Y., & Wang, B. (2023). Enhanced detection method for small and occluded targets in large-scene synthetic aperture radar images. *Journal of Marine Science and Engineering*, *11*(11), 2081.

[11] Xu, X., Zhan, W., Jiang, Y., Zhu, D., Chen, Y., Guo, J., ... & Liu, Y. (2024). A single-frame infrared small target detection method based on joint feature guidance. *Complex & Intelligent Systems*, *10*(4), 5281-5300.

[12] Chen, B., Huang, S., Liang, Z., Chen, W., & Pan, B. (2019). A fractional order derivative based active contour model for inhomogeneous image segmentation. *Applied Mathematical Modelling*, *65*, 120-136.

[13] Li, Y., Lv, X., Liu, K., & Zhao, S. (2014). Fractal-based weak target detection within sea clutter. *Acta Oceanologica Sinica*, *33*(9), 68-72.

[14] Lin, H., Shi, Z., & Zou, Z. (2017). Maritime semantic labeling of optical remote sensing images

with multi-scale fully convolutional network. *Remote sensing*, *9*(5), 480.

[15] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. *Procedia computer science*, *199*, 1066-1073.

[16] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

[17] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7464-7475).

[18] Sun, K., Li, Z., Zheng, Y., Kuo, H. W., Lee, K. P., & Tang, K. T. (2023). An area-efficient accelerator for non-maximum suppression. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *70*(6), 2251-2255.

[19] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., & Han, J. (2024). Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, *37*, 107984-108011.

[20] Xu, J., Fan, X., Jian, H., Xu, C., Bei, W., Ge, Q., & Zhao, T. (2024). Yoloow: A spatial scale adaptive real-time object detection neural network for open water search and rescue from uav aerial imagery. *IEEE transactions on geoscience and remote sensing*, *62*, 1-15.

[21] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[22] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759-8768).

[23] Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781-10790).

[24] Song, S., Jia, Z., Yang, J., & Kasabov, N. (2022). Salient detection via the fusion of background-based and multiscale frequency-domain features. *Information Sciences*, *618*, 53-71.

[25] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13733-13742).

[26] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13713-13722).

[27] Pan, R., Wang, Z., Wei, Y., Gao, H., Ou, G., Cao, C. C., ... & Chen, W. (2022). Towards efficient visual simplification of computational graphs in deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*, *30*(7), 3359-3373.

[28] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Cham: Springer International Publishing.

[29] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

[30] Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., ... & Wang, J. (2021). Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3651-3660).

[31] Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., ... & Zhang, L. (2022). Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.

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

[33] Zuo, X., Jiang, J., Shen, J., & Yang, W. (2025). Improving underwater semantic segmentation with underwater image quality attention and muti-scale aggregation attention. *Pattern Analysis and Applications*, *28*(2), 1-12.

[34] Qu, Q., Liu, W., Wang, J., Li, B., Liu, N., & Wang, Y. L. (2023). Enhanced CNN-based small target detection in sea clutter with controllable false alarm. *IEEE Sensors Journal*, *23*(9), 10193-10205.

[35] Zhang, M., Dong, L., Zheng, H., & Xu, W. (2021). Infrared maritime small target detection based on edge and local intensity features. *Infrared Physics & Technology*, *119*, 103940.

[36] Varga, L. A., Kiefer, B., Messmer, M., & Zell, A. (2022). Seadronessee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2260-2270).

[37] Bai, X., Xu, S., Zhu, J., Guo, Z., & Shui, P. (2023). Floating small target detection in sea clutter based on multifeature angle variance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 9422-9436.

[38] Robertson, A. E., Kelly, C., Buzzy, M., & Kalidindi, S. R. (2023). Local–global decompositions for conditional microstructure generation. *Acta Materialia*, *253*, 118966.

[39] Arabiat, M. O. H. A. M. M. A. D., Abuowaida, S. U. H. A. I. L. A., Al-Momani, A. D. A. I., Alshdaifat, N. A. W. A. F., & Chan, H. Y. (2024). Depth estimation method based on residual networks and se-net model. *Journal of Theoretical and Applied Information Technology*, *102*(3), 866-871.

[40] Cheng, X., & Yu, J. (2020). RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-11.

[41] Tian, Z., Shen, C., Chen, H., & He, T. (2020). FCOS: A simple and strong anchor-free object detector. *IEEE transactions on pattern analysis and machine intelligence*, *44*(4), 1922-1933.

[42] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6569-6578).

[43] Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). Varifocalnet: An iou-aware dense object detector. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8514-8523).

[44] Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., ... & Dave, P. (2020). ultralytics/yolov5: v3. 0. *Zenodo*.

[45] Cheng, P., Tang, X., Liang, W., Li, Y., Cong, W., & Zang, C. (2023, September). Tiny-YOLOv7: Tiny object detection model for drone imagery. In *International conference on image and graphics* (pp. 53-65). Cham: Springer Nature Switzerland.

[46] Hussain, M. (2024). Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision. *arXiv preprint arXiv:2407.02988*.

[47] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, *39*(6), 1137-1149.

[48] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16965-16974).