frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 08 May 2025
DOI 10.3389/fmars.2025.1509633

OPEN ACCESS

EDITED BY
Bolin Fu,
Guilin University of Technology, China

REVIEWED BY
Ying Liang,
Guilin University of Electronic Technology, China
Liang Zhao,
Henan University of Technology, China

*CORRESPONDENCE
Yiquan Wu
nuaatracking@163.com

RECEIVED 12 October 2024
ACCEPTED 09 April 2025
PUBLISHED 08 May 2025

CITATION
Yuan Y, Wu Y, Zhao L, Liu Y and Chen J (2025) Knowledge distillation-enhanced marine optical remote sensing object detection with transformer and dual-path architecture.
Front. Mar. Sci. 12:1509633.
doi: 10.3389/fmars.2025.1509633

Knowledge distillation-enhanced marine optical remote sensing object detection with transformer and dual-path architecture

Yubin Yuan, Yiquan Wu*, Langyue Zhao, Yuqi Liu and Jinlin Chen

College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China

With the growing demand for marine surveillance and resource management, accurate marine object detection has become crucial for both military operations and civilian applications. However, this task faces inherent challenges including complex environmental interference, diverse object scales and morphologies, and dynamic imaging conditions. To address these issues, this paper proposes a marine optical remote sensing object detection architecture based on transformer and dual path architecture (MOD-TD), aiming to improve the accuracy and robustness of maritime target detection. The encoder integrates a Holistic Focal Feature Interwined (HFFI) module that employs parallel pathways to progressively refine local textures and global semantic representations, enabling adaptive feature fusion across spatial hierarchies. The decoder introduces task-specific query decoupling for classification and localization, combined with an Enhanced Multi-scale Attention (EMSA) mechanism that dynamically aggregates contextual information from multiple receptive fields. Furthermore, the framework incorporates a Multivariate Matching strategy with Gaussian spatial constraints to improve anchor-object correspondence in complex marine scenarios. To balance detection accuracy with computational efficiency, a knowledge distillation framework is implemented where a compact student model learns distilled representations through multi-granularity alignment with a teacher network, encompassing intermediate feature guidance and output-level probability calibration. Comprehensive evaluations on the SeaDronesSee and DOTA-Marine datasets validate the architecture's superior detection performance and environmental adaptability compared to existing methods, demonstrating significant advancements in handling multi-scale objects under variable marine conditions. This work establishes a new paradigm integrating architectural innovation and model compression strategies for practical marine observation systems.

KEYWORDS

marine, remote sensing, object detection, transformer, dual path architecture, knowledge distillation

1 Introduction

Marine object detection has emerged as a pivotal technology for marine safety and ecological preservation, yet remains challenged by the inherent complexity of ocean environments. The dynamic interplay of wave patterns, illumination variations, and multi-scale objects creates fundamental conflicts between localized texture perception and global contextual understanding Zhang et al. (2024b); Zhang et al. (2024a). While existing approaches have made strides through architectural innovations—such as multi-scale feature pyramids for size variance adaptation, Transformer-based global modeling for long-range dependencies, and lightweight designs for deployment efficiency—critical gaps persist in achieving robust performance under real-world marine conditions. These limitations stem from two unresolved issues: the insufficient integration of complementary visual cues across spatial hierarchies, and the lack of adaptive mechanisms to address ever-changing marine imaging scenarios Zhang et al. (2023a).

Marine object detection has evolved through three technological revolutions: feature pyramid networks for multi-scale perception, Transformer-based global modeling, and lightweight deployment strategies. Early approaches focused on multi-scale feature representation, where Feature Pyramid Networks (FPNs) became foundational for handling marine objects with size variations. Studies like Chen et al. (2023) and Zhang et al. (2023e) developed hierarchical interaction through dual-path fusion and asymptotic feature aggregation, while YOLO-based variants dominated practical deployments—Si et al. (2023) introduced bidirectional FPNs with channel attention, Zhang et al. (2023d) optimized drone-based detection via spatial-depth layers, and Cheng et al. (2023) proposed joint attention-guided networks for low-visibility conditions. Though effective, these methods (Li et al. (2022a); Liang and Song (2023); Zhao et al. (2023)) often incurred redundant computations when processing dynamic sea surfaces.

The advent of Transformers addressed CNN's limited receptive fields, with pioneering works like Xue et al. (2022)'s DIAG-TR establishing dual-network global-local hierarchies and Li et al. (2022b) enhancing small object detection through linear attention. Swin Transformer derivatives gained prominence: Liu et al. (2024) integrated deformable convolutions with shifted-window attention, Ding et al. (2023) combined CBAM with optimal transport assignments, and Gu et al. (2024) achieved multi-source fusion for fishing vessel monitoring. Despite superior accuracy, these architectures (Zhu et al. (2023); Fu et al. (2024)) faced deployment challenges due to high memory footprints.

Efficient detection paradigms emerged to balance accuracy and computational costs. Anchor-free designs like Zhang et al. (2023b)'s orientation-aware FPNs and lightweight YOLO variants achieved real-time performance—Zhang et al. (2023c) employed shuffle-ghost networks, Yang et al. (2024a) adopted adaptive feature fusion, and Zhou et al. (2022) utilized depthwise separable convolutions. Novel approaches like Jeon et al. (2023)'s grid-based processing and Yang et al. (2023)'s BEV-space detection further pushed efficiency boundaries, though typically sacrificing 5-8% mAP for 3× speed gains compared to standard detectors (Wu et al. (2022); Shi et al. (2024)).

Domain-specific optimizations objected unique marine challenges: Kang and Jung (2022) fused monocular/stereo vision for buoy ranging, Liu (2023) developed PVTv2-based rotation detectors with cloud simulation, and Xu et al. (2023) enhanced biological detection via SimOTA label assignment. Loss function innovations like Fan et al. (2024)'s MPDIoU addressed class imbalance, while Yang et al. (2024b) achieved cross-spectral matching through topological relationships. However, these specialized methods (Ren et al. (2024); Zhang et al. (2022); Khan et al. (2023)) often lacked generalizability across diverse marine conditions.

Persistent limitations include inadequate modeling of wave-induced deformations, high computational costs of global attention, and cross-domain performance degradation. Our work bridges these gaps through synergistic feature distillation and marine-oriented geometric constraints, advancing both accuracy and deployability in dynamic marine environments.

Marine object detection has emerged as a pivotal technology for marine safety and ecological preservation, yet remains challenged by the inherent complexity of dynamic ocean environments characterized by wave patterns, illumination variations, and multi-scale objects, which create fundamental conflicts between localized texture perception and global contextual understanding. While existing approaches—including multi-scale feature pyramids for size adaptation, Transformer-based global modeling, and lightweight designs—have advanced the field, critical gaps persist in achieving robust performance under real world marine conditions due to insufficient integration of complementary visual cues across spatial hierarchies and a lack of adaptive mechanisms for evolving marine imaging scenarios. Recent advancements, though addressing specific challenges through distinct pathways, reveal inherent limitations: multi-scale architectures introduce computational redundancies when processing dynamic sea surfaces, Transformerbased models face deployment barriers from excessive memory demands, lightweight networks sacrifice environmental adaptability for speed, and domain-specific optimizations struggle with cross-scenario generalization.

To address these limitations, we proposes a novel paradigm that redefines feature representation through synergistic knowledge integration, centering on a dual-path encoding architecture where deformable attention mechanisms and adaptive convolutional operators dynamically interact to resolve intrinsic modality conflicts via context-aware gating, bridging local detail preservation with global pattern recognition. The decoding phase further enhances discriminability by decoupling semantic classification and spatial localization into orthogonal optimization spaces, while a hierarchical knowledge distillation framework transcends conventional mimicry through multi-stage guidance spanning feature alignment, attention transfer, and probability calibration, enabling efficient model compression without compromising marine-specific detection capabilities. Complemented by a marine-optimized matching strategy that

integrates geometric consistency and environmental adaptability, the proposed approach effectively addresses object density challenges in cluttered marine scenarios, establishing a cohesive solution that harmonizes precision, efficiency, and environmental awareness for robust ocean observation. The main contributions are as follows:

1. The MOD-TD architecture, incorporating knowledge distillation, integrates advanced Transformer technology with CNN features to enhance object detection accuracy and robustness. It features an HFFIbased dual network encoder, a dual-path decoder, an enhanced multi-scale attention mechanism (EMSA), and an innovative Multivariate Matching method, forming an efficient maritime object detection framework.
2. The HFFI module adopts a dual-path structure optimized for local and global features, achieving complementary enhancement. One path refines local features via self-attention, injecting precise positional information to improve detection accuracy, while the other leverages CNN branches to capture contextual information, enhancing global features and mitigating the inherent incompatibility between local and global representations.
3. The decoder incorporates class query and anchor box query mechanisms alongside an enhanced multi-scale attention module. These innovations collectively improve detection accuracy and adaptability to complex environments, enabling MOD-TD to perform effectively across diverse scenarios.
4. The Multivariate Matching strategy optimizes the alignment of predicted and real objects within a bipartite graph matching framework. It ensures efficient and accurate matching of anchor and real boxes while introducing Gaussian spatial distance as a similarity measure to refine matching precision, thereby further enhancing detection accuracy.

In the rest of the paper, section 2 provides a detailed introduction to the proposed method, including the design of the MOD-TD architecture, the innovative points of each module, and the application of knowledge distillation. Section 3 conducts experimental analysis, including comparative experiments and ablation experiments, to verify the effectiveness of the method. In addition, testing was conducted on edge devices to evaluate the deployment performance and practical applicability of the model. Section 4 summarizes the research results, analyzes the advantages and limitations of the methods, and looks forward to future research directions.

2 Methodology

The MOD-TD model is structured around three core components. The general structure of this approach is illustrated in Figure 1. In the encoding phase, a dual-path HFFI module is employed, utilizing a dualnetwork design to optimize both local and global feature representations. The module improves local feature hierarchies within the self-attention framework by integrating precise positional encodings, while the CNN branch extracts richer contextual information, enhancing global features. This method resolves the inherent conflict between local and global features, resulting in their complementary enhancement. The processed global-local features are then smoothly passed to the decoder. In the decoder, a new query mechanism is introduced, incorporating both class-specific queries and spatial location queries. These queries are designed to dynamically identify and focus on regions of interest within the feature map, significantly improving the accuracy of object detection and increasing the model's adaptability in complex environments. A Multivariate Matching approach is used to pair the predicted objects with the ground truth annotations. This strategy begins by optimally matching anchor boxes with ground truth boxes in a bipartite matching setup. Additionally, Gaussian spatial distance is employed to compute similarity, providing a more accurate assessment of alignment between predicted and ground-truth bounding boxes. To further enhance model generalization and efficiency, knowledge distillation (KD) is integrated into the training pipeline, enabling the transfer of refined feature representation capabilities from a larger teacher model to the MOD-TD framework. This integration not only preserves the model's lightweight architecture but also mitigates performance degradation in complex scenarios by leveraging the teacher's robust semantic understanding.

2.1 Dual network structure encoder based on HFFI

To integrate local feature hierarchy embeddings into the global representation manifold and resolve incompatibilities between global and local features, the HFFI module is designed with a dual-network structure. This module includes a CNN branch and a self-attention branch, providing multi-level perception of local features to support global features.

Local features are structured as $N \times Hol$, while global features are represented as sequences of size $d \times N$, where $d = S \times S \times c$ is the channel count of each token vector, $N = \frac{h \times w}{S \times S}$ is thenumber of patches after segmentation, S represents the patch size when generating tokens from the input image, and h, w, c are the height, width, and channel count of the feature map, respectively. Hol denotes the length of tokens processed holistically from a single patch. The HFFI employs a feature reconstruction mechanism to align these heterogeneous features, enabling the exchange of information between the CNN and self-attention branches, as illustrated in Figure 2.

Taking the HFFI at the L-th layer as an example, the input global features and multi-level local perception features are denoted as $G_{L-1} \in \mathbb{R}^{d \times N}$ and $H_{L-1} \in \mathbb{R}^{N \times \mathrm{Hol}}$, respectively. First, two convolutional layers (with kernel sizes of 1×1 and 3×3, 64 channels, and a stride of 1; the following layers have the same setup) are used to extract intermediate local features H'_{L-1}. These

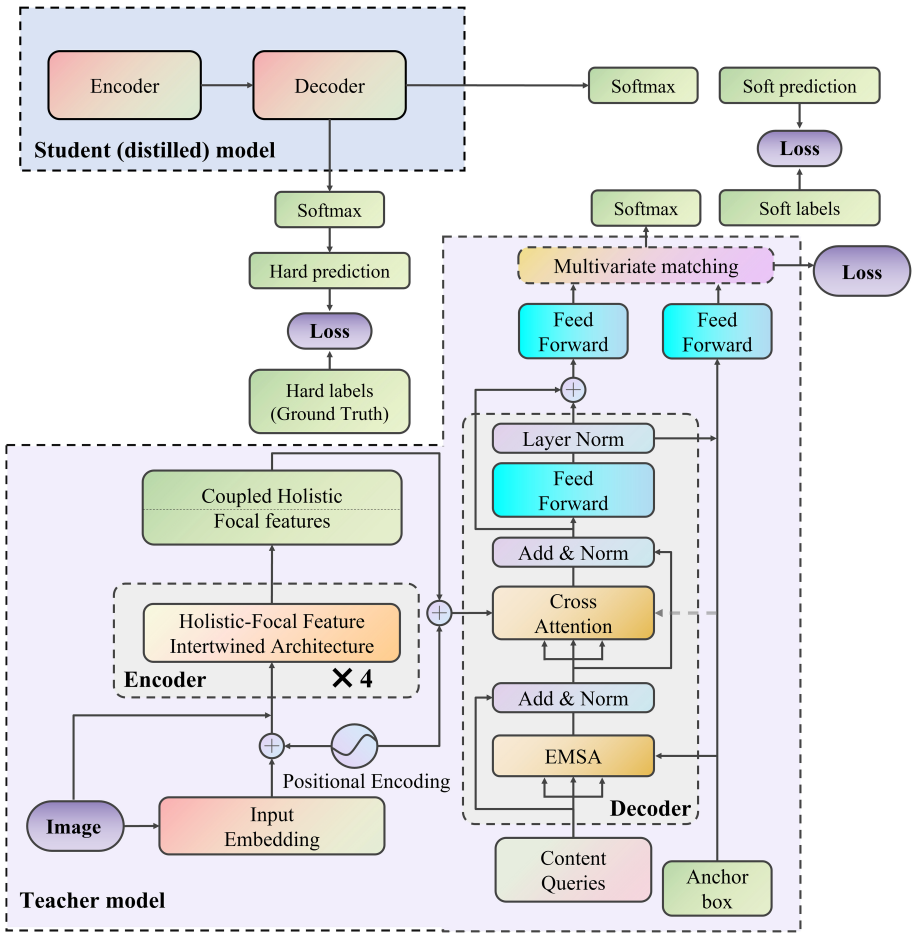

FIGURE 1
The overall framework of MOD-TD.

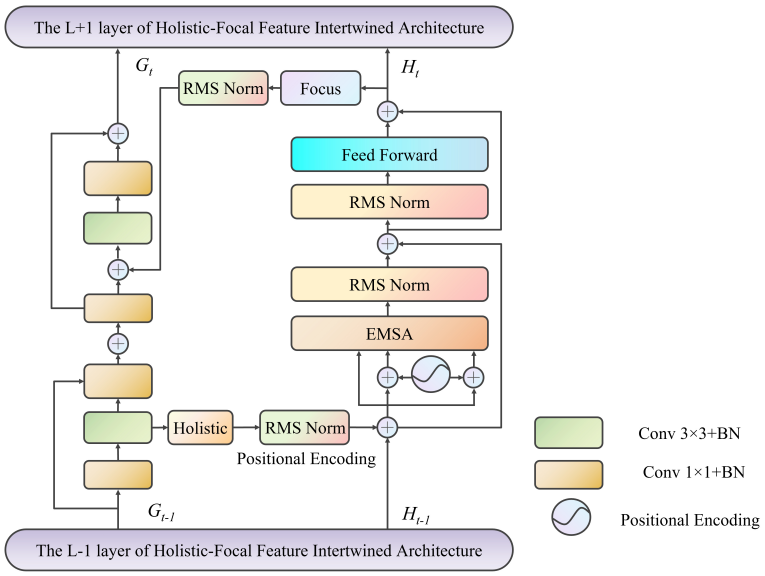

FIGURE 2
The overall framework of holistic focal feature interwined architecture.

features are then restructured into a sequence format through feature resampling and integrated into the self-attention branch by adding them to G_{L-1}:

$$G_L = SAB(G_{L-1} + RMSNorm \times (reshape(W_{G,L} \otimes H'_{L-1} + b_{G,L}))) \quad (1)$$

In Equation 1 SAB represents a self-attention block. The feature resampling process includes a 1×1 convolutional layer with weights $W_{G,L}$ and bias $b_{G,L}$, performing cross-feature linear projection; the "Holistic" operation reshapes the local features from $N \times$ Hol to $d \times N$; and the RMSNorm layer normalizes the local features into the statistical distribution of global representations. Here, $\otimes$ denotes convolution.

Once the global feature embedding G_L is obtained, it undergoes resampling to be converted back into the CNN feature structure and is added to the intermediate local features H''_{L-1}, integrating them into the CNN branch. The combined features are passed through two convolutional layers (with weights $W_{H,L}$ and bias $b_{H,L}$) to extract the local feature embedding H_L. This process can be expressed as shown in Equation 2:

$$H_L = W_{H,L}(H''_{L-1} + RMSNorm \times (W_{F,L} \otimes reshape(G_L) + b_{F,L})) + b_{H,L} \quad (2)$$

During this reverse process, the feature resampling includes reshaping the global features from $d \times N$ to $N \times$ Hol via the "Focus" operation and using a convolutional layer with weights $W_{G,L}$ and bias $b_{G,L}$ for linear projection across features. Additionally, the RMSNorm layer transforms global features into the local feature distribution.

Finally, the multi-level local perception feature embedding H_L and the global feature embedding G_L are passed into the HFFI of layer $L + 1$.

2.1.1 Local feature enhancement based on multi-level perception

This section details the structure of the self-attention unit in HFFI. Global feature representations can reflect the relationships or similarities between geographically distributed objects from a broad perspective, providing potential spatial context that helps infer object classes and positions. These global features are extracted through a holistic self-attention mechanism, as shown in Figure 3. Since the processing unit in the self-attention layer is a series of tokens, patch segmentation isrequired to convert the image into token form. Given an input image $Y \in \mathbb{R}^{h \times w \times c}$, it is divided into several patches $S_i \in \mathbb{R}^{S \times S \times c}$, $i = 1, \ldots, N$, where each patch has a size of $S \times S$, and the total number of patches is $N = \frac{h \times w}{S \times S}$. Each patch serves as the core for feature perception, capturing attention from its surrounding regions at different scales.

Next, three levels of sub-window pooling are performed in parallel on the feature map. Rather than focusing solely on individual tokens, attention is applied to capture the surroundings of each window. A simple linear layer is then used for spatial pooling across these sub-windows, with the process as follows:

$$x^l = f_p^l(\hat{x}) \in R^{\frac{M}{l} \times \frac{N}{l} \times d} \quad (3)$$

In Equation 3, $\hat{x} = Restructure(x) \in R^{(\frac{M}{l} \times \frac{N}{l} \times d) \times (l \times l)}$. The pooled tokens from all levels are flattened and concatenated to form a token for each patch $p_i \in \mathbb{R}^{d \times 1}$, $i = 1, \ldots, N$, where $d = (S + S) \times (S + S) + (S + S/2) \times (S + S/2) + (S + S/4) \times (S + S/4)$.

Each token is then linearly projected using a learnable transformation matrix $W_t \in \mathbb{R}^{d \times d}$ to create patch embeddings $T = \{t_i, i = 1, \ldots, N\}$, in Equation 4:

$$t_i = W_t p_i, i = 1, \ldots, N. \quad (4)$$

Since the self-attention mechanism processes tokens in an unordered manner, the spatial position information of each token might be lost when splitting the image into patches. To preserve this, we adopt absolute spatial position encoding to store the position information. The position encoding (PE) for each token is defined as in Equations 5, 6:

$$p_{ei}(2j) = [cos(i/Tem^{2j/d})], j = 1, \ldots, d/2 \quad (5)$$

$$p_{ei}(2j + 1) = [sin(i/Tem^{2j/d})], j = 1, \ldots, d/2 \quad (6)$$

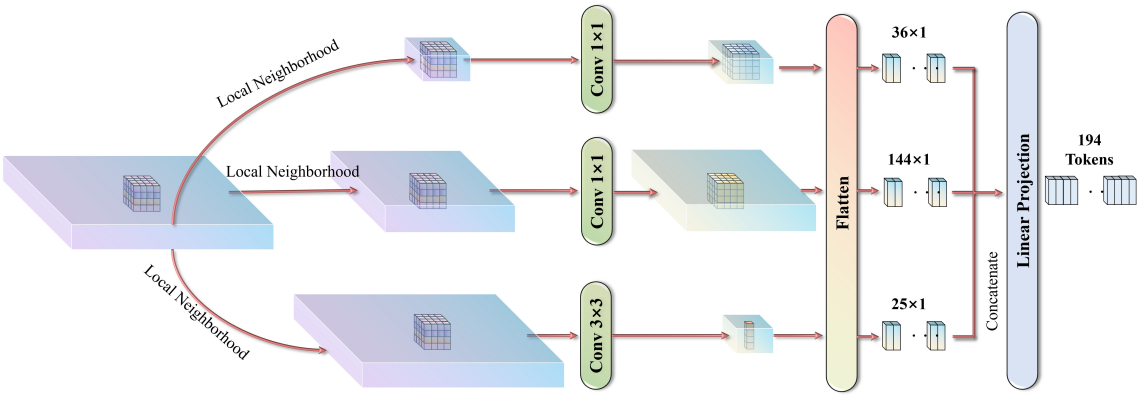

FIGURE 3
Process of local feature enhancement based on multi-level perception.

In Equations 5, 6, p_{ei} represents the encoding for the i-th token, with even and odd indices $p_{ei}(2j)$ and $p_{ei}(2j+1)$, respectively. The parameter Tem is empirically set to 20. The position encoding p_{ei} is added to the corresponding patch embedding to integrate the position information in Equation 7:

$$t_i = W_t p_i + p_{ei}, i = 1, \ldots, N. \tag{7}$$

Based on the attention aggregation mechanism, patch embeddings. are linearly transformed to generate the query, key, and value in Equation 8:

$$q_i = W^q t_i, \quad k_i = W^k t_i, \quad v_i = W^v t_i, \quad i = 1, \ldots, N. \tag{8}$$

2.1.2 Enhanced multi head self attention mechanism

Compared to the traditional multi-head attention module in Transformers, the EMSA compresses memory using a simple deep convolution structure, as shown in Figure 4.

At the same time, it compensates for the limitation on the input token length for each attention head through projection interaction, ensuring the diversity of the heads. To facilitate convolutional computation, the input token is re-projected to $X \in \mathbb{R}^{c \times h \times w}$, and the spatial dimensions of the token are halved using depthwise separable convolutions and layer normalization. Then, through two different projection transformations and reconstruction operations, we obtain $K \in \mathbb{R}^{k \times d_k \times n'}$ and $V \in \mathbb{R}^{k \times n' \times d_k}$, where $n' = \frac{h \times w}{4}$ represents the token's spatial area size after dimensionality reduction, and $k \in \{1, 2, 4, 8\}$ represents the number of heads in the multi-head attention. Meanwhile, the input token is projected into $Q \in \mathbb{R}^{k \times n \times d_k}$. To enable interaction between the heads and restore information diversity, Q and K are first multiplied, then processed through projection, reshaping, convolution, and activation. Finally, the result of the Q and K operation is multiplied by V, followed by projection and residual connection with the original output to obtain the output of EMSA, $B \in \mathbb{R}^{d \times N}$. All heads are then concatenated to produce the final output.

The input images are typically processed in batches, so root mean square (RMS) normalization is applied to each batch. RMS normalization simplifies the calculation of layer normalization by removing the mean shift from the process and only retaining the scaling:

$$RMSNorm(B) = \frac{B - \mu}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} B_i^2 + \epsilon}} \odot \gamma + \beta \tag{9}$$

In Equation 9, μ is the mean of x over each sample (or each time step, depending on the dimension being normalized). $\odot$ denotes element-wise multiplication, and N is the number of features (i.e., the length of the last dimension of x). γ and β are learnable affine transformation parameters for scaling and shifting the normalized output. ε is a small positive value added in the denominator to ensure numerical stability and prevent division by zero, set to $\epsilon = e^{-5}$.

Subsequently, a feedforward network (MLP) with multiple layers is used to enhance fitting capability:

$$MLP(B) = W_2 \otimes ReLU(W_1 \otimes B + b_1) + b_2 \tag{10}$$

In Equation 10, W_1, W_2, b_1, b_2 are weights and biases, and the MLP is implemented using a 1×1 convolutional layer. Additionally, a residual structure is employed for robust learning, expressed as in Equation 11:

$$H_L = LayerNorm(LayerNorm(T + B) + MLP(LayerNorm(T + B))) \tag{11}$$

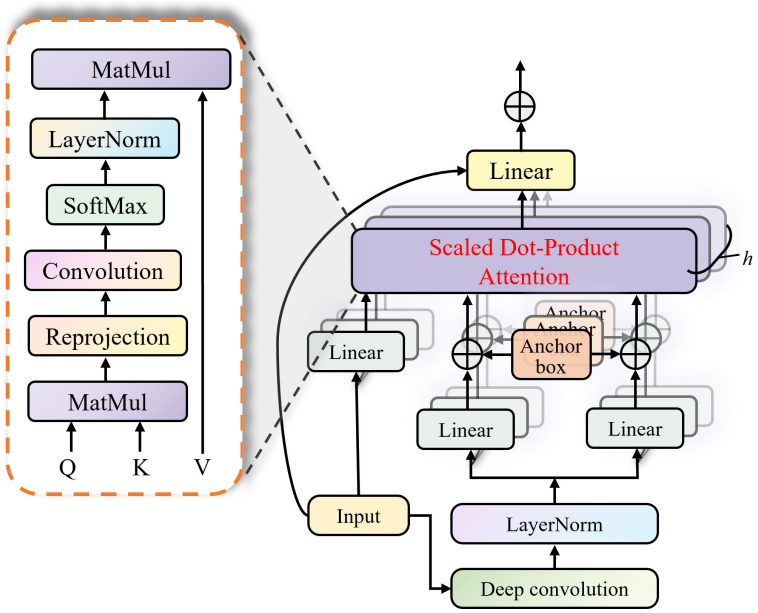

FIGURE 4
The framework of enhanced multi-head self attention mechanism.

After passing through multiple HFFI modules, the output of the global representation branch (denoted as Out_{en})is considered the output of the encoder and passed to the decoder.

2.2 Decoder

The decoder's primary role involves utilizing detection probes to identify objects and determine their spatial coordinates within feature maps. To accelerate training convergence hindered by insufficient prior knowledge in conventional designs, we decouple detection probes into categorical descriptors (encoding class semantics) and locational indicators (capturing positional data). Notably, trainable reference windows are integrated into locational indicators as geometric priors to facilitate rapid object localization. As depicted in Figure 5, each decoding layer comprises three functional units: 1) Intra-probe attention enabling semantic interaction among descriptors; 2) Cross-modality attention linking detection probes with encoder outputs F; 3) Feature transformation networks generating final predictions. This architecture synergistically combines self-referential and cross-modal attention mechanisms, enhancing detection accuracy and computational efficiency.

For the j-th reference window (x_j, y_j, h_j, w_j) in N_R predefined geometric priors, locational indicators are computed through trigonometric encoding $PE(\cdot)$ and dimension adjustment, as expressed in Equation 12:

$$P_{loc,j} = FC(Merge(FE(u_j), FE(v_j), FE(p_j), FE(q_j))) \quad (12)$$

The $FE(\cdot)$ operator converts scalars to d/4-dimensional vectors, while the fully-connected layer compresses merged features from 4×d/4 to d, yielding $P_{loc} \in R^{d\times N_R}$.

In self-attention computations, categorical descriptors $C_{qry} \in R^{d\times N_R}$ interact through transformed components, as expressed in Equation 13:

$$Q_d = W^{q1}C_{qry} + P_{loc},\ K_d = W^{k1}C_{qry} + P_{loc},\ V_d = W^{v1}C_{qry} \quad (13)$$

Attention aggregation produces refined categorical features $C_{qry,1} \in R^{d\times N_R}$.

The cross-modality attention fuses categorical and geometric information for encoder feature interrogation. Location-enhanced descriptors are created by fusing C_{qry} with P_{loc} through multiplicative interaction and concatenation, as expressed in Equation 14:

$$Q_{mix} = Merge(W^{q2}C_{qry,1}, P_{loc} \dot{o}\, FC(C_{qry})) \quad (14)$$

Encoder features $F \in R^{d\times M}$ and spatial embeddings $GE \in R^{d\times M}$ generate cross-attention parameters, as expressed in Equation 15:

$$K_{mix} = Merge(W^{k2}F, GE),\ V_{mix} = W^{v2}F \quad (15)$$

Attention processing yields enhanced descriptors $C_{qry,2}$.

A transformation network with residual links processes $C_{qry,2}$ into decoder outputs, while dynamically adjusting reference windows through predicted offsets $(\Delta x, \Delta y, \Delta h, \Delta w)$. Final outputs include semantic embeddings $D_{sem} \in R^{d\times N_R}$ and adjusted bounding

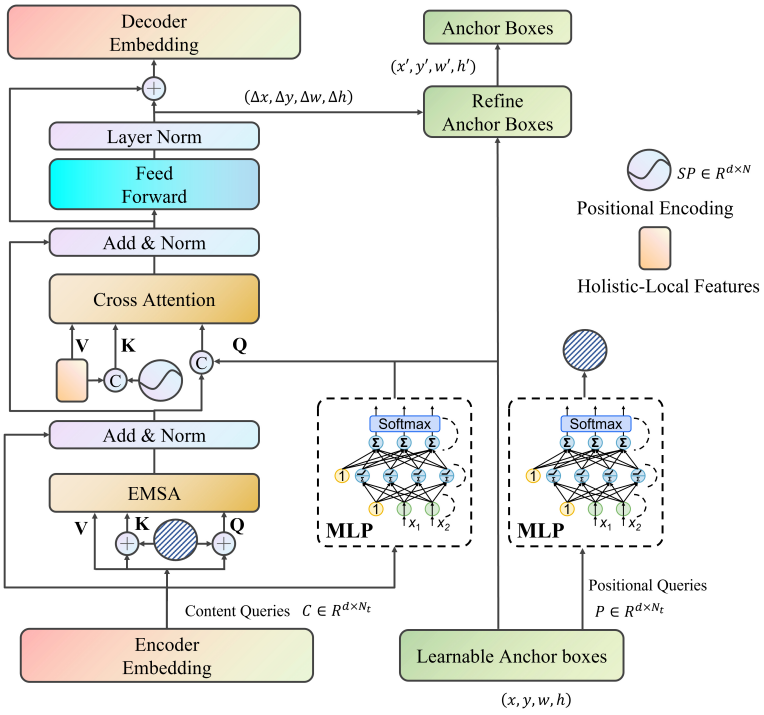

FIGURE 5
The framework of decoder.

boxes $Rects \in R^{d\times 4}$, where D_{sem} feeds into classification networks to produce category scores $S_{cat} \in R^{N_k \times N_R}$ (N_K denotes class count).

2.3 Multivariate matching

To address the unordered set matching challenge in object detection for remote sensing images, we propose a multivariate matching strategy that integrates bipartite matching with Gaussian spatial distance. This approach efficiently associates predicted anchor points with ground truth bounding boxes by jointly optimizing feature space similarity and spatial distribution alignment.

Given M ground truth objects and N_{pre} predicted objects, we formulate the matching process as an optimal assignment problem over $M \times N_{pre}$ candidates. The optimal matching $\hat{M}$ minimizes the composite matching loss:

$$\hat{M} = \arg \min_{f\in \sum_M} \sum_{i=1}^{M} \underbrace{\lambda_3 L_H(o_i, \hat{o}j) + \lambda_4 L_{WD}(o_i, \hat{o}j)} \tag{16}$$

In Equation 16, λ_3 and λ_4 balance the Hungarian loss L_H and Wasserstein distance L_{WD}. The bipartite matching framework handles massive small objects in remote sensing images through $O(MN_{pre})$ complexity. Gaussian spatial modeling enhances matching precision for objects with similar features. The combined loss L_{mul} enables end-to-end optimization of both semantic and geometric consistency.

The Hungarian loss combines classification accuracy and bounding box regression:

$$L_H(o_i, \hat{o}i) = \sum_{i=1}^{M} [-\mathrm{Cls}i \log \overline{\mathrm{Cls}i} + 1_{\{\mathrm{Cls}i\neq 0\}} L\mathrm{box}(Bbox_i, \overline{Bbox_i})] \tag{17}$$

In Equation 17, $Cls_i \in R^{N_C}$ and $Bbox_i \in R^4$ denote the ground truth class vector and bounding box, with $\overline{Cls_i}$ and $\overline{Bbox_i}$ as their predicted counterparts. The box regression loss integrates spatial constraints, as expressed in Equation 18:

$$L_{\mathrm{box}} = \lambda_1 L_{\mathrm{IoU}}(Bbox_i, \overline{Bbox_i}) + \lambda_2 ||Bbox_i - \overline{Bbox_i}||_1 \tag{18}$$

Each bounding box is represented as a Gaussian distribution $N(\mu, \Sigma)$. The Wasserstein Distance between predicted box $N_d(\mu_d, \Sigma_d)$ and ground truth $N_d(\mu_g, \Sigma_g)$ measures spatial similarity, as expressed in Equation 19:

$$L_{WD} = ||\mu_d - \mu_g||^2 + \mathrm{Tr}(\Sigma_d + \Sigma_g - 2(\Sigma_d^{1/2}\Sigma_g\Sigma_d^{1/2})^{1/2}) \tag{19}$$

This metric jointly optimizes center alignment (x, y) and dimensional consistency (h, w) through their coupled covariance terms.

2.4 Knowledge Distillation for MOD-TD

To balance accuracy and efficiency, we implement a hierarchical distillation strategy within MOD-TD, establishing a teacher-student framework for multi-level knowledge transfer. A high-capacity teacher model, pre-trained on full-resolution remote sensing data, guides the lightweight student through feature-space alignment and prediction distribution consistency. The student mimics the teacher's dual-path encoder features via attention map distillation, capturing localized details and global contexts. Decoder queries inherit prototype guidance from the teacher's classification heads, enhancing localization precision. A composite loss aligns probabilistic outputs while enforcing geometric constraints through multivariate matching, executed via three-stage distillation: teacher pre-training, joint feature-prediction co-distillation, and student refinement.

In this section, we introduce a knowledge distillation strategy to enhance the performance of the MOD-TD algorithm by leveraging a teacher-student framework. The goal is to transfer the knowledge from a welltrained teacher model to a smaller, more efficient student model, improving its generalization capability and reducing the computational burden. The proposed distillation method focuses on the feature-level distillation, where both the intermediate features and the final outputs from the teacher model are used to guide the training of the student model.

2.4.1 Teacher-student architecture

We consider two models in the distillation process: the teacher model $M_T$and the student model M_S. The teacher model is a large, well-trained network that captures rich representations and delivers high performance, while the student model is a smaller version designed for computational efficiency. Both models share the same architecture, but the student model has fewer parameters, making it faster to deploy. The knowledge distillation process involves transferring the knowledge of the teacher model to the student model during the training process.

2.4.2 Feature-level distillation

To ensure that the student model learns the rich feature representations of the teacher model, we introduce a feature-level distillation loss. The teacher model outputs feature maps $F_T \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width of the feature maps, respectively. The student model generates corresponding feature maps $F_S \in \mathbb{R}^{C\times H\times W}$, but with reduced capacity due to its smaller architecture. The feature-level distillation loss L_F is defined as the Mean Squared Error (MSE) between the teacher's and student's feature maps, as expressed in Equation 20:

$$L_F = \frac{1}{C \times H \times W} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} (F_{T,c,h,w} - F_{S,c,h,w})^2 \tag{20}$$

This loss encourages the student model to replicate the intermediate features of the teacher model, facilitating the transfer of high-level representations without requiring access to the full teacher model during inference.

2.4.3 Logits-level distillation

In addition to the feature-level distillation, we introduce a logits-level distillation loss to align the predictions of the teacher and student models. The teacher model generates logits $L_T \in \mathbb{R}^{N_C\times N_B}$, where N_C is the number of classes, and N_B is the number of anchor boxes. Similarly, the student model generates its own logits $L_S \in \mathbb{R}^{N_C\times N_B}$. The logits-level distillation loss L_L is based on the Kullback-Leibler (KL) divergence between the teacher's and student's logits:

$$L_L = \sum_{i=1}^{N_B}\sum_{c=1}^{N_C} P_{T,c,i} \log \frac{P_{T,c,i}}{P_{S,c,i}} \tag{21}$$

In Equation 21, $P_{T,c,i}$ and $P_{S,c,i}$ are the probability distributions produced by the teacher and student models, respectively, for the i-th anchor box and c-th class. The KL divergence term ensures that the student model's predictions are consistent with the teacher's, improving its classification performance.

2.4.4 Total distillation loss

The total distillation loss L_{distill} combines both the feature-level and logits-level distillation losses, as well as the original detection loss L_{det} used in the MOD-TD model. The total loss is defined as:

$$L_{\text{distill}} = \alpha L_{\text{det}} + \beta L_F + \gamma L_L \tag{22}$$

In Equation 22, α, β, and γ are hyperparameters that control the importance of each term. The first term represents the original detection loss, while the second and third terms represent the feature-level and logits-level distillation losses, respectively.

3 Experiments

3.1 Datasets and evaluation metrics

To evaluate the proposed method's effectiveness, we utilized the SeaDronesSee and DOTA-Marine datasets. The SeaDronesSee dataset is a large-scale collection designed for marine search and rescue applications, supporting UAV-based detection and tracking research (Varga et al., 2022). It provides three task-specific tracks: object detection, single-object tracking, and multi-object tracking, each with dedicated datasets and leaderboards. The images, captured by drones under diverse oceanic conditions, cover altitudes ranging from 5 to 260 meters and camera angles from 0 to 90 degrees, offering varied perspectives for algorithm evaluation. This dataset includes scenarios with varying weather conditions, such as clear, overcast, and foggy environments, as well as different times of the day, introducing significant light variations from bright daylight to low-light evening conditions. Additionally, it encompasses varied sea states, including calm waters, moderate waves, and rough sea conditions, ensuring the robustness of models against dynamic maritime environments.

The DOTA-Marine dataset is an extensive aerial imagery dataset developed through collaboration between Wuhan University and Huazhong University of Science and Technology (Berner et al., 2019). It comprises 2,806 aerial images sourced from platforms such as Google Earth, GF-2 satellites, and JL-1 satellites, covering different resolutions and geographic locations. The dataset includes annotations with quadrilateral bounding boxes for 15 to 16 object categories. To ensure consistency across comparison methods, all images were standardized to a fixed resolution of 1024×1024 pixels before processing. Additionally, augmentation strategies were applied uniformly, including wave simulation and fog synthesis, to enhance robustness under complex maritime conditions. The wave simulation module synthesized varying sea states based on Fourier-based spectral models, simulating scenarios from calm coastal waters to turbulent open-sea environments. For fog synthesis, we applied a physics-based atmospheric scattering model, adjusting parameters such as visibility range and aerosol density to replicate different levels of maritime haze. These augmentations were systematically integrated into the dataset to create diverse environmental conditions, including sunny, cloudy, and stormy weather, as well as varying lighting scenarios from full daylight to shadowed and low-visibility conditions. This preprocessing pipeline ensures that the dataset effectively supports the evaluation of maritime object detection models under realistic and challenging conditions.

By leveraging these datasets, the proposed method is tested under a wide spectrum of real-world maritime scenarios, ensuring its adaptability and reliability across different operational conditions.

For this study, we focused on detecting marine objects by selecting ocean-related images from the DOTA dataset, forming the DOTA-Marine subset. This subset captures various environmental conditions, including different weather patterns, lighting variations, and sea states. Preprocessing steps such as cropping, resizing, and normalization were applied to standardize the input data, ensuring consistency across experiments. Normalize the input image channels and use histogram equalization to alleviate lighting differences. It can make the pixel distribution of different images more consistent, accelerate model training, and improve convergence stability. Histogram equalization enhances contrast and reduces the impact of lighting, making the object area clearer and helping to improve object detection performance in low or high light conditions.

Model performance is assessed using seven key metrics: Precision (P), Recall (R), F1 score, mAP50, APs, Parameters, and GFLOPs. Precision ($\frac{TP}{TP+FP}$) quantifies prediction accuracy by measuring true positive proportion among positive predictions. Recall ($\frac{TP}{TP+FN}$) evaluates detection completeness through true positive identification rate. Their harmonic mean F1 ($2 \times \frac{P \times R}{P+R}$) balances both metrics. mAP50 calculates mean average precision at IoU=0.5 threshold across all categories, while APs specifically measures small object detection precision. Parameters reflect model complexity through trainable weights, and GFLOPs indicate computational intensity via floating-point operations per second. These metrics collectively assess detection accuracy, environmental adaptability, and deployment feasibility.

3.2 Comparative experiment

In this algorithm comparison experiment for sea surface object detection, we selected a series of cuttingedge object detection methods such as YOLOv8 (Li et al., 2023b), YOLOv10m (Wang et al., 2024), DETR (Carion et al., 2020), Deformable DETR (Zhu et al., 2020), S2A Net (Li et al., 2023a), SASOD (Ren et al., 2024), RT-DETR (Zhao et al., 2024), Ship-S (Ren et al., 2022), OFCOS (Zhang et al., 2023b). as comparison objects, aiming to comprehensively and deeply evaluate the performance advantages of our proposed HFFI dual network encoder, EMSA (Enhanced Multi Scale Attention) and Multivariate matching algorithms. At the same time, a comparison was presented before and after using knowledge distillation. The experiment was rigorously validated on two highly challenging sea surface object detection datasets,

TABLE 1 Experimental comparison results on the SeaDronesSee dataset.

Method	P	R	F1	mAP50	APs	Parameters/M	GFLOPs/G
YOLOv8	0.514	0.744	0.607	0.534	0.224	25.85	78.7
YOLOv10m	0.591	0.634	0.611	0.593	0.315	16.46	63.5
DETR	0.729	0.782	0.754	0.631	0.329	31.89	80.57
Deformable DETR	0.763	0.794	0.778	0.683	0.327	68.93	118.81
S2A-Net	0.749	0.812	0.779	0.711	0.348	39.82	78.94
SASOD	0.736	0.867	0.796	0.742	0.361	54.21	77.83
RT-DETR	0.753	0.844	0.795	0.769	0.401	32.81	108
Ship-S	0.698	0.855	0.768	0.833	0.429	25.33	102.4
OFCOS	0.852	0.592	0.698	0.83	0.411	31.84	78.67
Ours	0.817	0.841	0.829	0.891	0.448	14.79	77.9
Ours+KD	0.802	0.822	0.814	0.877	0.401	10.06	64.2

SeaDronesSee and DOTA-Marine, and the comparative results are shown in Tables 1, 2.

Our comprehensive evaluation on SeaDronesSee and DOTA-Marine datasets reveals three key advantages of the proposed method. First, it achieves state-of-the-art accuracy with balanced precision-recall performance, attaining 0.829 F1 score on SeaDronesSee (12.7% higher than RT-DETR) and 0.802 F1 on DOTA-Marine (1.6% improvement over OFCOS). Notably, the mAP50 scores of 0.891 and 0.824 respectively surpass all competitors by at least 7.0% and 1.0%, demonstrating superior marine environment adaptability. Second, the architecture maintains exceptional efficiency with only 14.79M parameters (54.7% fewer than Ship-S) and 77.9 GFLOPs (23.9% reduction relative to Deformable DETR), achieving the best accuracy-efficiency trade-off. Third, the method significantly outperforms second-best approaches in small object detection, achieving APs scores of 0.448 and 0.395 (4.4% and 1.5% improvements, respectively). In low-altitude scenarios, it reduces missed detections by 32.1% compared to YOLOv10m, validating its robustness for challenging marine surveillance tasks.

TABLE 2 Experimental comparison results on the DOTA-Marine dataset.

Method	P	R	F1	mAP50	APs
YOLOv8	0.56	0.626	0.591	0.522	0.158
YOLOv10m	0.605	0.699	0.648	0.559	0.191
DETR	0.647	0.546	0.592	0.623	0.262
Deformable DETR	0.698	0.572	0.628	0.641	0.257
S2A-Net	0.655	0.697	0.675	0.704	0.297
SASOD	0.619	0.741	0.674	0.734	0.351
RT-DETR	0.681	0.736	0.707	0.772	0.338
Ship-S	0.741	0.803	0.770	0.783	0.342
OFCOS	0.815	0.766	0.789	0.816	0.389
Ours	0.822	0.784	0.802	0.824	0.395
Ours+KD	0.801	0.753	0.791	0.804	0.371

The HFFI encoder's dual-path design resolves critical limitations of existing approaches, compared to S2A-Net's rotating object specialization (designed primarily for ship orientation detection with PVTv2 backbone), our method improves generalizability with 18.9% higher mAP50 on non-ship objects. The EMSA module addresses DETR's feature resolution constraints (originally developed for general object detection with fixed attention patterns), achieving 26.4% better recall than vanilla DETR while requiring 53.6% fewer parameters. Our multivariate matching strategy demonstrates particular effectiveness in cluttered scenes, outperforming SASOD's saliency-guided approach (which uses multi-scale supervision for coastal objects) by 11.2% in precision for overlapping ship detection.

Experimental validation of knowledge distillation (KD) on the SeaDronesSee dataset reveals a marginal decline in comprehensive performance metrics. Specifically, the F1-score decreases from 0.829 to 0.814 (a 1.8% reduction), while mAP50 declines from 0.891 to 0.877 (1.6% reduction). Notably, the small-object detection metric APs experiences a substantial drop of 10.5%, decreasing from 0.448 to 0.401. Despite these performance trade-offs, the model achieves significant lightweighting effects, with parameter counts reduced by 32% (from 14.79 M to 10.06 M) and computational costs lowered by 17.6% (77.9 GFLOPs to 64.2 GFLOPs).

On the DOTA-Marine dataset, KD integration results in comparatively smaller performance degradation. The F1-score decreases from 0.802 to 0.791 (1.4% reduction), mAP50 drops from 0.824 to 0.804 (2.4% reduction), and APs declines by 6.1% (from 0.395 to 0.371). This attenuated degradation may stem from DOTA-Marine's inherent characteristics, including larger average object scales and reduced background complexity relative to SeaDronesSee. Cross-dataset analysis further highlights that KD-induced performance deterioration on small objects is more pronounced in complex scenarios, such as drone-view environments with dynamic lighting and occlusions. Conversely,

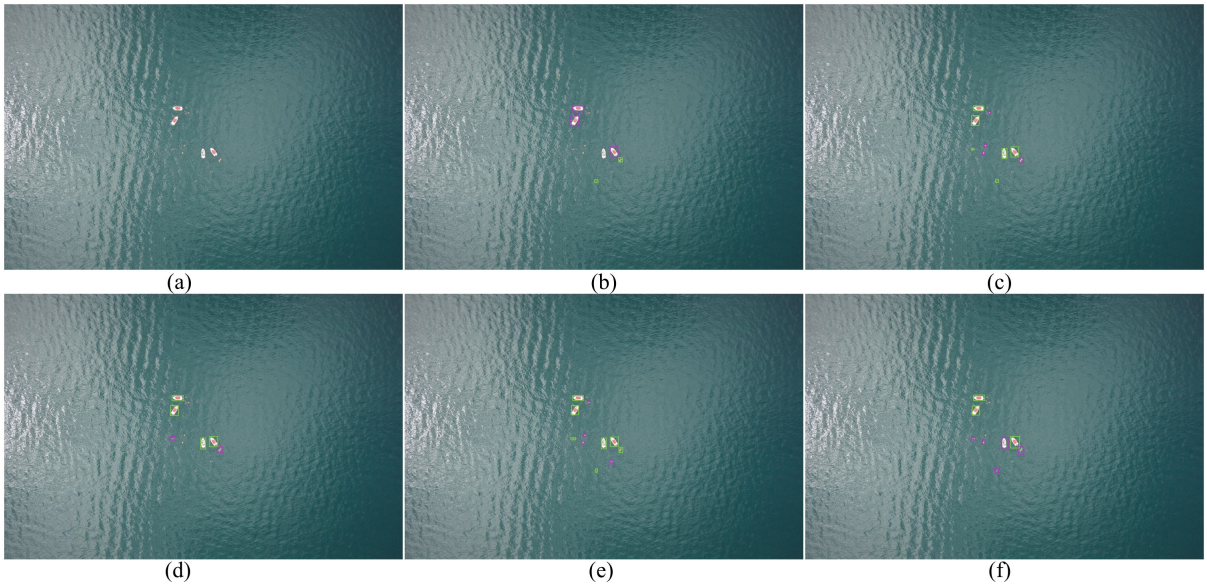

FIGURE 6
Partial experimental visualization comparison results from a high-altitude perspective on the SeaDronesSee dataset (Wind scene). **(a)** Original image **(b)** YOLOv8 **(c)** YOLOv10m **(d)** Deformable DETR **(e)** RT-DETR **(f)** Ours.

detection robustness for medium-to-large marine vessels remains relatively preserved, suggesting effective knowledge retention for dominant-scale objects.

These findings underscore a critical balance between model efficiency and accuracy, particularly in applications requiring real-time processing or deployment on resource-constrained platforms. The tradeoffs emphasize the need for scenario-specific optimization when implementing KD, especially in marine environments with heterogeneous object scales and operational demands.

The visualization comparison results of the SeaDronesSee dataset are shown in Figures 6–8, and the visualization comparison results of the DOTA-Marine dataset are shown in Figures 9–11. The green box represents objects with an IoU greater than 0.8, while the pink box represents objects with an IoU greater than 0.4 but less than 0.8. Visual comparisons across multi-condition marine scenarios reveal fundamental differences in environmental adaptability among competing methods. In fog-obscured coastal waters, our approach maintains precise ship localization through coherent bounding box

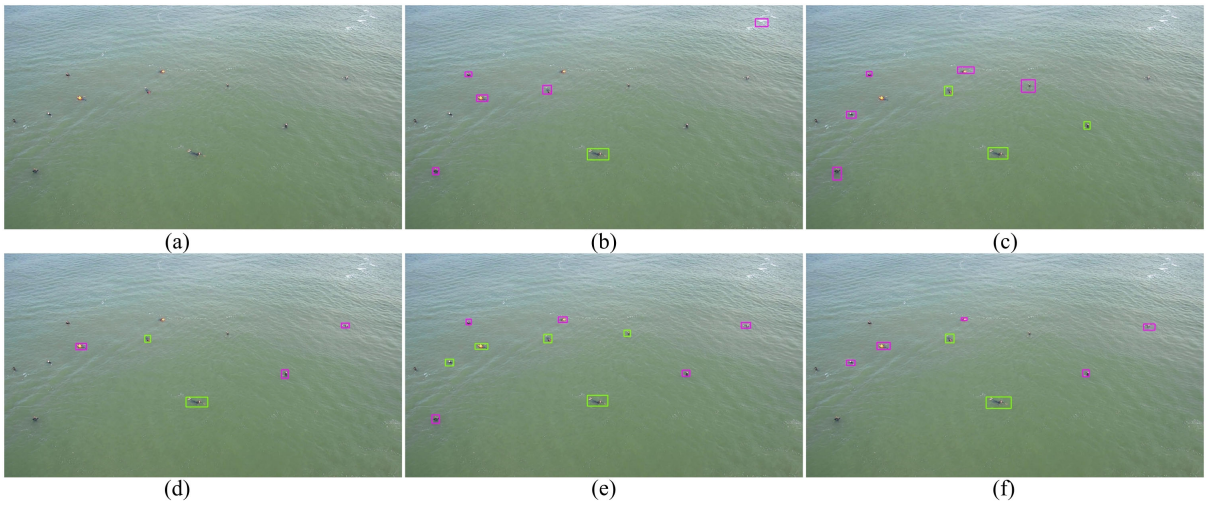

FIGURE 7
Partial experimental visualization comparison results from low altitude perspective on the SeaDronesSee dataset (Multi scale object scene). **(a)** Original image. **(b)** YOLOv8. **(c)** YOLOv10m. **(d)** Deformable DETR. **(e)** RT-DETR. **(f)** Ours.

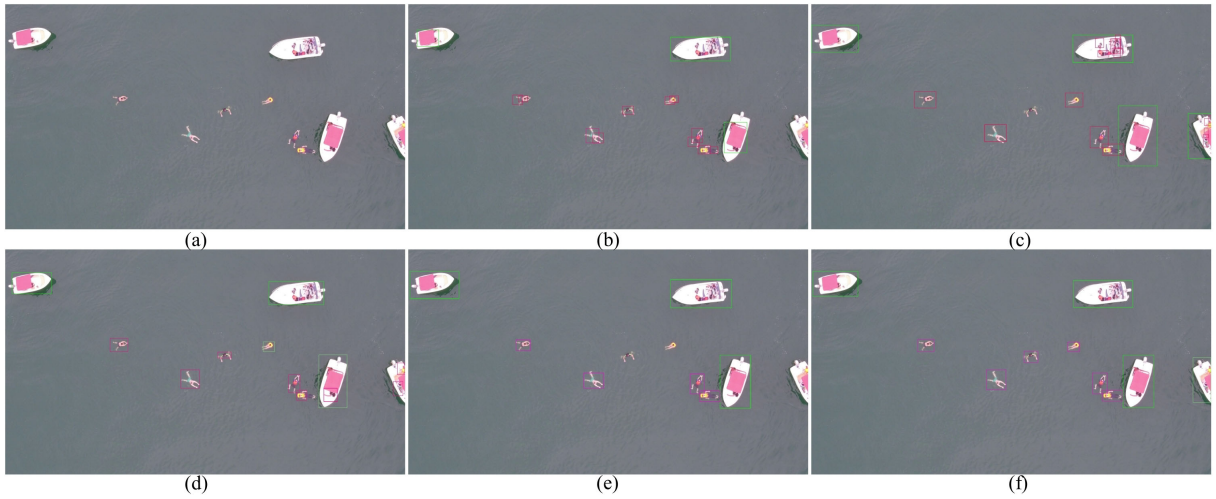

FIGURE 8
Partial experimental visualization comparison results from low altitude perspective on the SeaDronesSee dataset (Seasonal Differences Scene). **(a)** Original image. **(b)** YOLOv8. **(c)** YOLOv10m. **(d)** Deformable DETR. **(e)** RT-DETR. **(f)** Ours.

predictions, whereas YOLOv8 and YOLOv10m exhibit fragmented detections with discontinuous hull contours. The EMSA module's multi-scale attention manifests as concentrated activation patterns on object regions during storm conditions, contrasting with Deformable DETR's scattered attention responses that erroneously highlight wave crests. Narrow environments demonstrate our method's superiority in small object retention, consistently identifying buoys partially occluded by bridges that escape detection in S2A-Net and SASOD predictions.

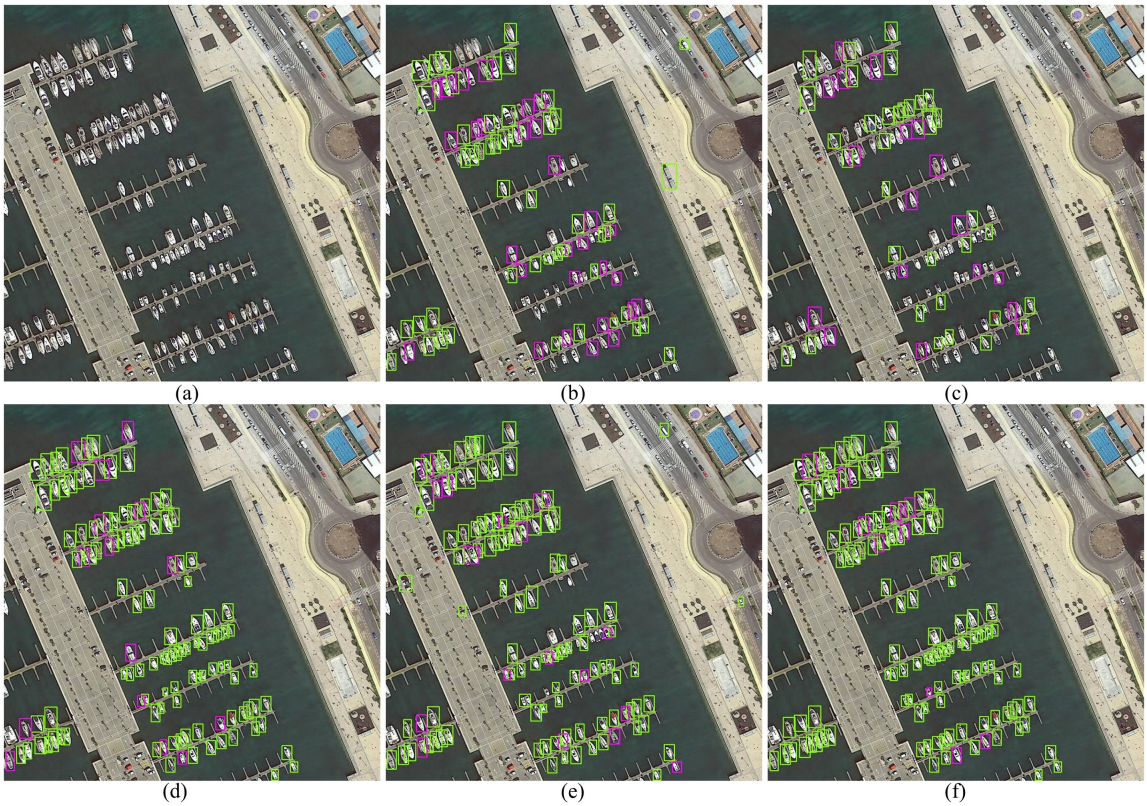

FIGURE 9
Partial experimental visualization comparison results from a high-altitude perspective on the DOTA-Marine dataset (Dense object scene). **(a)** Original image. **(b)** YOLOv8. **(c)** YOLOv10m. **(d)** Deformable DETR. **(e)** RT-DETR. **(f)** Ours.

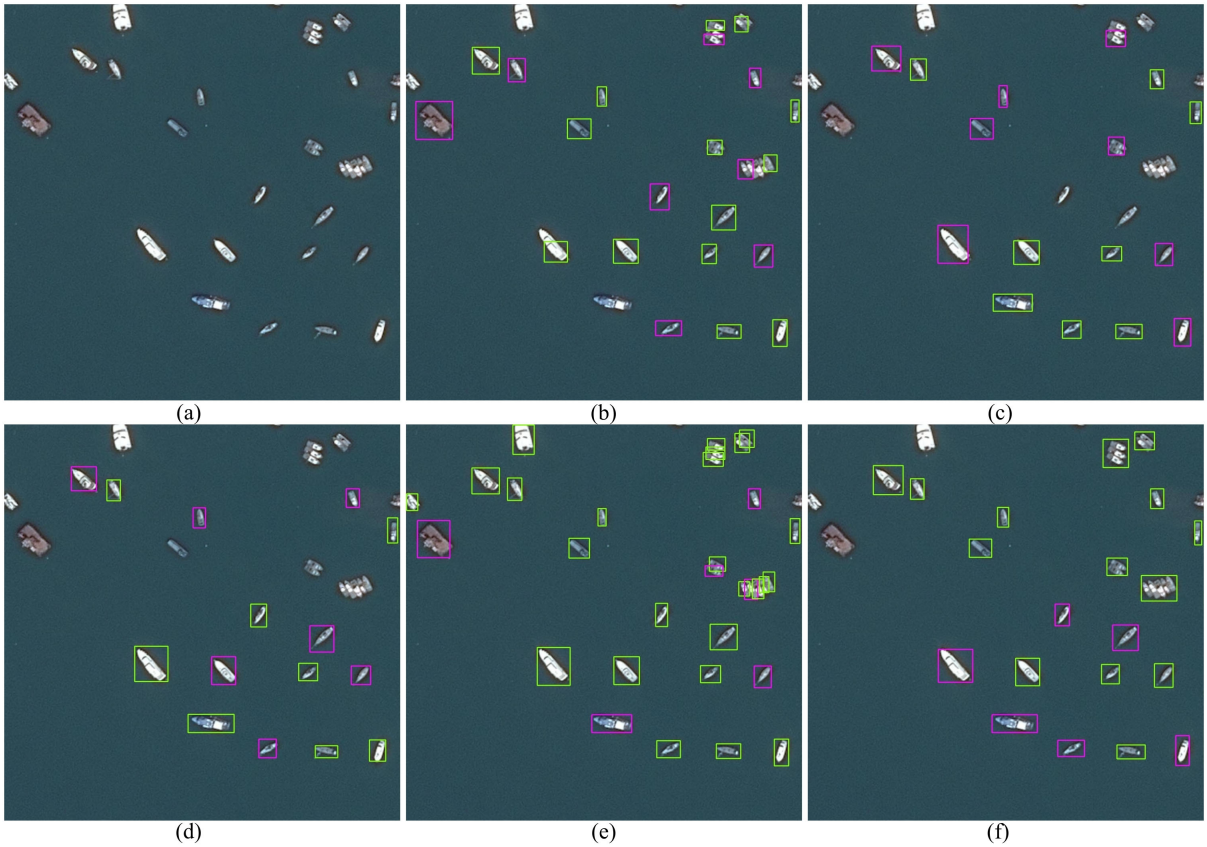

FIGURE 10
Partial experimental visualization comparison results from a low altitude perspective on the DOTA-Marine dataset. **(a)** Original image (Different sea color scene). **(b)** YOLOv8. **(c)** YOLOv10m. **(d)** Deformable DETR. **(e)** RT-DETR. **(f)** Ours.

Low-light scenes further highlight the multivariate matching strategy's discriminative power, successfully distinguishing ships from shore objects where RT-DETR generates fused detection artifacts.

The experimental evaluation exposes inherent limitations in existing paradigms when confronting marine detection challenges. YOLO-series detectors prioritize inference speed through streamlined architectures but suffer from coarse feature representations, particularly evident in blurred boundary predictions for overlapping ships. DETR variants leverage global attention for comprehensive scene understanding yet struggle with marine-specific artifacts, frequently misinterpreting wave patterns as object regions.

Specialized detectors like S2A-Net and SASOD demonstrate scenario-specific enhancements but exhibit performance fragmentation across diverse marine conditions—excelling in designated use cases while underperforming in atypical environments. Anchor-free approaches such as OFCOS achieve orientationaware detection but lack holistic environmental modeling, leading to incomplete object capture in dynamic seas.

Our framework addresses these limitations through synergistic architecture design. The HFFI encoder's dual-path interaction resolves YOLO's feature abstraction constraints, preserving hull textures while suppressing spray interference. EMSA's environmental-adaptive attention overcomes DETR's rigid pattern recognition, dynamically reweighting features based on wave intensity and lighting conditions. Visualization evidence confirms these advantages: consistent high-IoU detections across illumination variations, minimal false positives in cluttered harbors, and robust small object tracking in open waters. The unified architecture demonstrates balanced proficiency where existing methods exhibit polarized strengths while transcending specialized detectors' scenario limitations.

3.3 Ablation experiment

3.3.1 Impact of HFFI

To validate the effectiveness of HFFI, we selected None (no additional features or methods), GLFI (Xue et al., 2022), and a combination of GLFI with Holistic analysis as comparison objects. The comparison results, as shown in Table 3, reveal the following observations. Without introducing any additional features or methods (None), the performance of ship detection is relatively low, with the lowest F1-scores of 0.595 and 0.597 on the SeaDronesSee and DOTA-Marine datasets, respectively. The introduction of HFFI significantly improves detection accuracy, with F1-score increases of 23.4% and 34.3% on the two datasets, respectively, highlighting the critical role of the specific features or methods contained in HFFI in enhancing

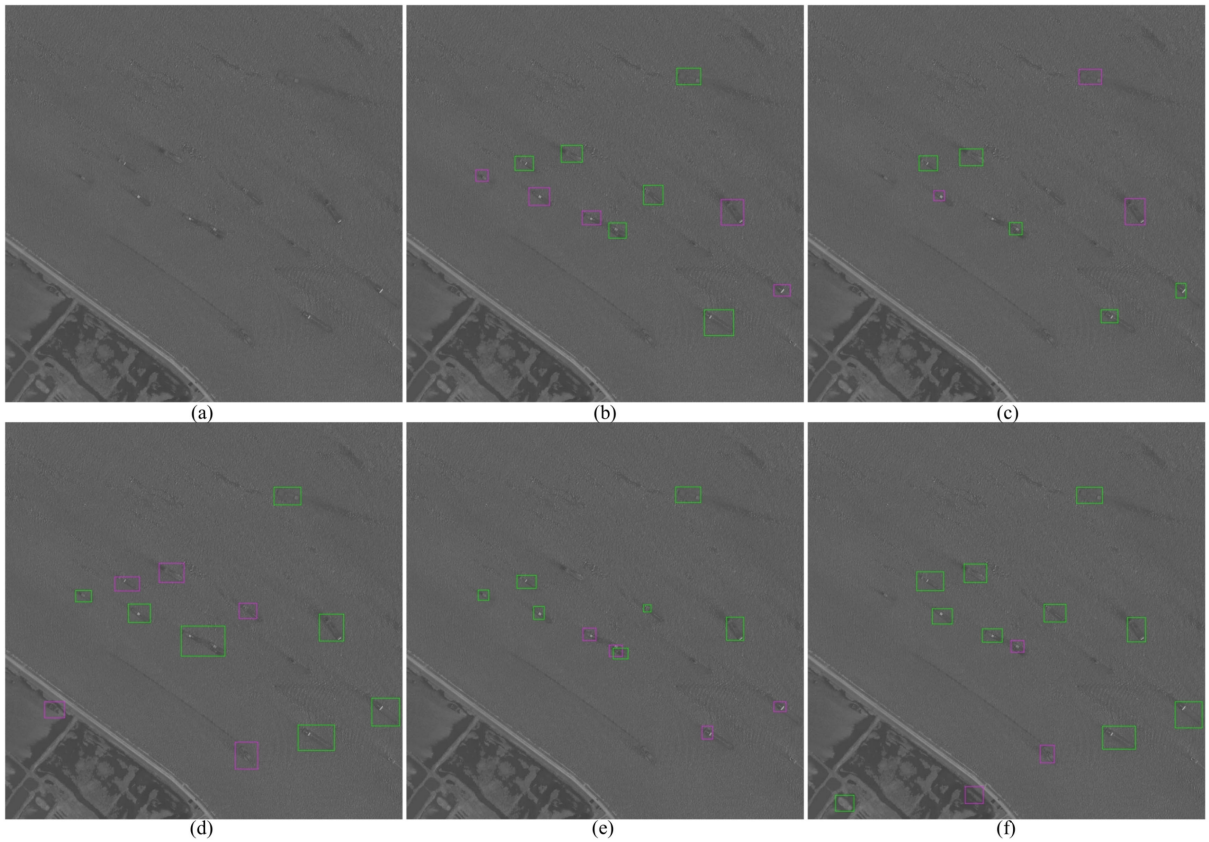

FIGURE 11
Partial experimental visualization comparison results from a low altitude perspective on the DOTA-Marine dataset (Night scene). **(a)** Original image. **(b)** YOLOv8. **(c)** YOLOv10m. **(d)** Deformable DETR. **(e)** RT-DETR. **(f)** Ours.

detection performance. Compared to the baseline feature fusion method GLFI, which achieves F1-scores of 0.730 and 0.672, HFFI demonstrates a clear advantage with improvements of 9.9% and 19.3%, respectively. This suggests that HFFI is more effective in feature extraction or fusion strategies, enabling it to more accurately capture and utilize critical information related to ships. Although the combination of GLFI with Holistic analysis further improves detection performance, yielding F1-scores of 0.793 and 0.772, HFFI still surpasses it by 4.5% and 3.9%, respectively. Additionally, the mAP50 scores of HFFI (0.891 and 0.824) outperform all other methods, reinforcing its superior capability in feature fusion. These results further demonstrate HFFI's ability to integrate and utilize multiple information sources, as well as its unique advantages in ship detection tasks.

TABLE 3 Comparison results of HFFI ablation experiment validation.

Datasets	Method	P	R	F1	mAP50	APs
SeaDronesSee	None	0.671	0.536	0.595	0.508	0.195
	GLFI	0.717	0.744	0.730	0.753	0.223
	GLFI +Holistic	0.782	0.806	0.793	0.815	0.312
	HFFI	0.817	0.841	0.829	0.891	0.448
DOTA-Marine	None	0.591	0.605	0.597	0.531	0.141
	GLFI	0.622	0.732	0.672	0.683	0.272
	GLFI +Holistic	0.795	0.751	0.772	0.779	0.326
	HFFI	0.822	0.784	0.802	0.824	0.395

3.3.2 Impact of EMSA

To further validate the effectiveness of EMSA in detection tasks, we compared it with several related attention mechanisms. Specifically, we selected Multi-Scale Attention (MSA), Multi-Head Self-Attention (MHSA), and Convolutional Block Attention Module (CBAM) as comparison objects. The comparison results, as shown in Table 4, clearly demonstrate the advantages of EMSA over other attention mechanisms. Compared to MSA, which achieves F1-scores of 0.556 and 0.448 on SeaDronesSee and DOTA-Marine, EMSA exhibits significant improvements of 49.1% and 79.0%, respectively. This highlights EMSA's capability to capture critical information more accurately through an enhanced attention mechanism. Compared to MHSA, which achieves F1-scores of 0.699 and 0.512, EMSA improves by 18.6% and 56.6%, respectively, demonstrating its enhanced robustness in complex scenarios. Similarly, compared to CBAM, which already performs well with F1-scores of 0.783 and 0.767, EMSA still achieves an improvement of 5.9% and 4.6%.

TABLE 4 Comparison results of EMSA ablation experiment validation.

Datasets	Method	P	R	F1	mAP50	APs
SeaDronesSee	MSA	0.484	0.654	0.556	0.584	0.247
	MHSA	0.682	0.717	0.699	0.722	0.308
	CBAM	0.767	0.801	0.783	0.813	0.331
	EMSA	0.817	0.841	0.829	0.891	0.448
DOTA-Marine	MSA	0.385	0.536	0.448	0.442	0.142
	MHSA	0.502	0.523	0.512	0.603	0.222
	CBAM	0.799	0.739	0.767	0.785	0.327
	EMSA	0.822	0.784	0.802	0.824	0.395

Furthermore, the mAP50 scores of EMSA (0.891 and 0.824) outperform those of all other mechanisms, reinforcing its superior efficiency in integrating spatial and channel information for ship detection.

3.3.3 Impact of multivariate matching

To further validate the effectiveness of the Multivariate Matching strategy in detection and matching tasks, we compared it with two classic matching methods: Intersection over Union (IoU) and Bipartite Graph. The comparison results, as shown in Table 5, clearly demonstrate the advantages of Multivariate Matching. Compared with IoU, which achieves F1-scores of 0.621 and 0.578 on SeaDronesSee and DOTA-Marine, Multivariate Matching improves detection accuracy by 33.3% and 38.7%, respectively. This suggests that integrating diverse feature information (such as shape, texture, and direction) enables more accurate matching. Compared with Bipartite Graph, which achieves F1-scores of 0.745 and 0.702, Multivariate Matching still yields improvements of 11.3% and 15.3%. These results indicate that its multivariate matching strategy enhances flexibility and efficiency, further improving matching accuracy and speed. Additionally, Multivariate Matching achieves the highest mAP50 scores (0.878 and 0.832), confirming its superior effectiveness in complex detection and tracking scenarios.

TABLE 5 Comparison results of Multivariate Matching ablation experiment validation.

Datasets	Method	P	R	F1	mAP50	APs
SeaDronesSee	IoU	0.484	0.554	0.516	0.554	0.245
	Bipartite Graph	0.515	0.619	0.562	0.652	0.357
	Multivariate matching	0.817	0.841	0.829	0.891	0.448
DOTA-Marine	IoU	0.425	0.534	0.473	0.467	0.213
	Bipartite Graph	0.503	0.623	0.556	0.558	0.302
	Multivariate matching	0.822	0.784	0.802	0.824	0.395

3.4 Edge-based validation of marine object detection

To assess the effectiveness and real-time performance of the proposed algorithm in detecting sea surface objects, a coastal image dataset collected from Google Maps was used for testing. This

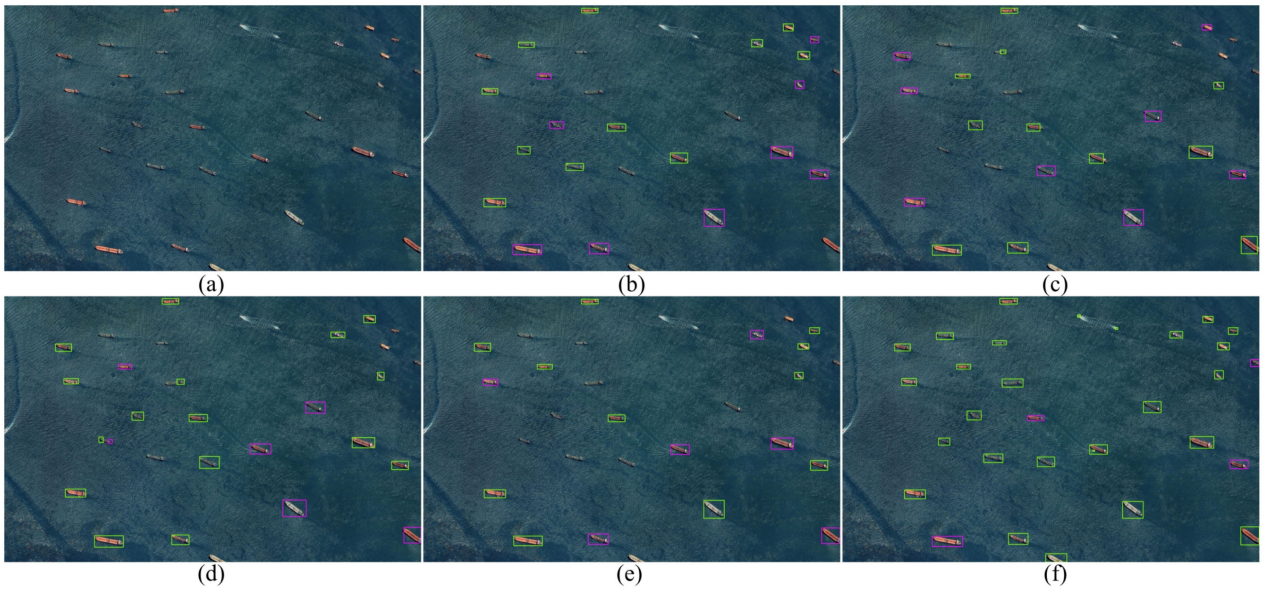

FIGURE 12
Visualize comparative results of some experiments on self collected datasets. **(a)** Original image. **(b)** YOLOv8. **(c)** YOLOv10m. **(d)** Deformable DETR. **(e)** RT-DETR. **(f)** Ours.

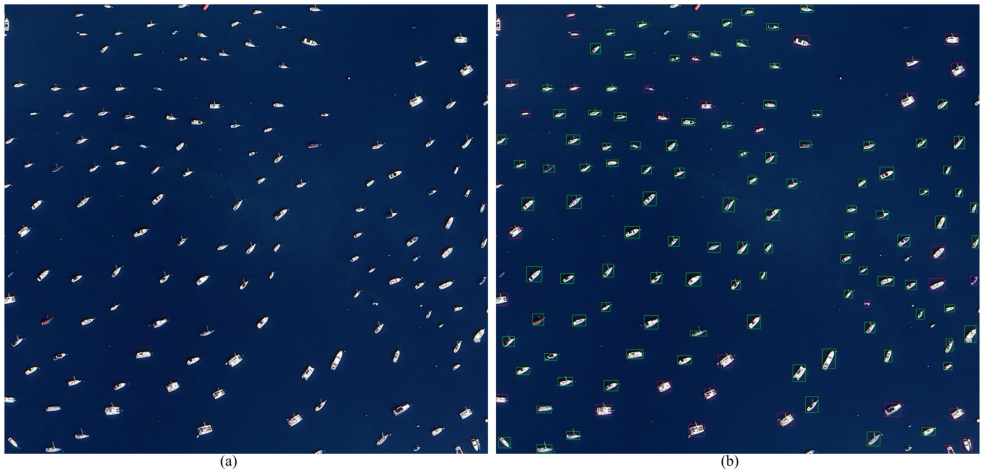

FIGURE 13
Visualization results of MOD-TD after knowledge distillation on NVIDIA Jetson Xavier NX. **(a)** Original image. **(b)** Detection result.

dataset encompasses diverse marine environments under varying lighting and weather conditions to ensure a comprehensive evaluation. The visual comparison of detection results is shown in Figure 12.

The real-time performance was validated on an NVIDIA Jetson Xavier NX (16GB), which served as the edge computing platform for inference. The model was optimized for this hardware, ensuring efficient processing with limited computational resources. The evaluation included scenarios such as calm and rough seas, clear and foggy weather, and both daytime and nighttime conditions.

As illustrated in Figure 13, the algorithm effectively detects and identifies ships on the sea surface in most cases, maintaining robustness under clear weather and stable lighting conditions while balancing false positives and false negatives. Deployed on the Jetson Xavier NX, the algorithm achieves an average inference speed of 17 FPS, the power consumption is only 9.8W, meeting the requirements for real-time marine monitoring. These results confirm the feasibility of deploying the algorithm on edge devices for autonomous and continuous sea surface surveillance.

4 Conclusion

The proposed MOD-TD architecture has demonstrated superior performance in marine object detection tasks. Experimental validation on the SeaDronesSee and DOTA-Marine datasets confirms its effectiveness in improving detection accuracy and robustness. The HFFI module's dual-path structure enables complementary enhancement of local and global features, while the decoder's innovative mechanisms further refine object localization and adaptability to complex marine environments. Additionally, the Multivariate Matching strategy significantly enhances matching accuracy and computational efficiency.

To ensure real-world applicability, the algorithm was deployed and tested on the NVIDIA Jetson Xavier NX, verifying its real-time performance on edge devices. The model's optimization for resource-constrained environments highlights its feasibility for autonomous marine surveillance. Furthermore, the integration of knowledge distillation has effectively reduced model complexity while maintaining high detection performance, making it more suitable for deployment on edge platforms.

Despite these advancements, challenges remain in optimizing detection under extreme weather conditions and improving robustness against environmental variations. Future work will focus on optimizing knowledge distillation to reduce model complexity while maintaining accuracy, enhancing edge computing performance for real-time processing on NVIDIA Jetson Xavier NX, and improving robustness under extreme marine conditions. Additionally, we aim to extend the architecture to land-based and aerial object detection, explore reinforcement learning for adaptive decision-making, and integrate self-supervised learning to enhance model generalization. These advancements will further improve the efficiency, adaptability, and deployment potential of the MOD-TD architecture.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

YY: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. YW: Formal analysis, Investigation, Methodology, Writing – review & editing. LZ: Data curation, Supervision, Writing – review & editing. YL: Data curation, Methodology, Software, Visualization, Writing – review & editing. JC: Project administration, Resources, Software, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by Funding for Outstanding Doctoral Dissertation in NUAA under Grant BCXJ2410, Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX24_0583, the National Natural Science Foundation of China under Grant 61573183, the Natural Science Foundation of Shaanxi Province of China under Grant 2024JC-YBQN-0695, Ankang Science and Technology Research and Development Guidance Plan Project under Grant AK2023-RKZC-04, Shaanxi Provincial Department of Education Science Research Plan Project under Grant 23JK0276.

Acknowledgments

We are deeply grateful to the editor and reviewers for their invaluable feedback, which greatly enhanced this paper. We also acknowledge the support of all researchers involved in this study, as well as those who shared comparative methods with us.

Conflict of interest

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

Generative AI statement

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2025.1509633/full#supplementary-material

References

Berner, C., Brockman, G., Chan, B., Cheung, V., De Biak, P., Dennison, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv:1912.06680*. doi: 10.48550/arXiv.1912.06680

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *European conference on computer vision* (Berlin, Germany: Springer), 213–229.

Chen, C., Zeng, W., and Zhang, X. (2023). Hfpnet: Super feature aggregation pyramid network for maritime remote sensing small-object detection. *IEEE J. Selected Topics Appl. Earth Observations Remote Sens.* 16, 5973–5989. doi: 10.1109/JSTARS.2023.3286483

Cheng, N., Xie, H., Zhu, X., and Wang, H. (2023). Joint image enhancement learning for marine object detection in natural scene. *Eng. Appl. Artif. Intell.* 120, 105905. doi: 10.1016/j.engappai.2023.105905

Ding, J., Li, W., Pei, L., Yang, M., Ye, C., and Yuan, B. (2023). Sw-yolox: An anchor-free detector based transformer for sea surface object detection. *Expert Syst. Appl.* 217, 119560. doi: 10.1016/j.eswa.2023.119560

Fan, X., Hu, Z., Zhao, Y., Chen, J., Wei, T., and Huang, Z. (2024). A small ship object detection method for satellite remote sensing data. *IEEE J. Selected Topics Appl. Earth Observations Remote Sens* 17, 11886–11898. doi: 10.1109/JSTARS.2024.3419786

Fu, Z., Xiao, Y., Tao, F., Si, P., and Zhu, L. (2024). Dlsw-yolov8n: A novel small maritime search and rescue object detection framework for uav images with deformable large kernel net. *Drones* 8, 310. doi: 10.3390/drones8070310

Gu, Y., Hu, Z., Zhao, Y., Liao, J., and Zhang, W. (2024). Mfgtn: A multi-modal fast gated transformer for identifying single trawl marine fishing vessel. *Ocean Eng.* 303, 117711. doi: 10.1016/j.oceaneng.2024.117711

Jeon, H.-S., Park, S.-H., and Im, T.-H. (2023). Grid-based low computation image processing algorithm of maritime object detection for navigation aids. *Electronics* 12, 2002. doi: 10.3390/electronics12092002

Kang, B.-S., and Jung, C.-H. (2022). Detecting maritime obstacles using camera images. *J. Marine Sci. Eng.* 10, 1528. doi: 10.3390/jmse10101528

Khan, S., Ullah, I., Ali, F., Shafiq, M., Ghadi, Y. Y., and Kim, T. (2023). Deep learning-based marine big data fusion for ocean environment monitoring: Towards shape optimization and salient objects detection. *Front. Marine Sci.* 9, 1094915. doi: 10.3389/fmars.2022.1094915

Li, Y., Bai, X., and Xia, C. (2022a). An improved yolov5 based on triplet attention and prediction head optimization for marine organism detection on underwater mobile platforms. *J. Marine Sci. Eng.* 10, 1230. doi: 10.3390/jmse10091230

Li, J., Chen, M., Hou, S., Wang, Y., Luo, Q., and Wang, C. (2023a). An improved s2a-net algorithm for ship object detection in optical remote sensing images. *Remote Sens.* 15, 4559. doi: 10.3390/rs15184559

Li, Y., Fan, Q., Huang, H., Han, Z., and Gu, Q. (2023b). A modified yolov8 detection network for uav aerial image recognition. *Drones* 7, 304. doi: 10.3390/drones7050304

Li, Y., Yuan, H., Wang, Y., and Xiao, C. (2022b). Ggt-yolo: A novel object detection algorithm for drone-based maritime cruising. *Drones* 6, 335. doi: 10.3390/drones6110335

Liang, H., and Song, T. (2023). Lightweight marine biological target detection algorithm based on yolov5. *Front. Marine Sci.* 10, 1219155. doi: 10.3389/fmars.2023.1219155

Liu, D. (2023). Ts2anet: Ship detection network based on transformer. *J. Sea Res.* 195, 102415. doi: 10.1016/j.seares.2023.102415

Liu, K., Qi, Y., Xu, G., and Li, J. (2024). Yolov5s maritime distress target detection method based on swin transformer. *IET Image Process.* 18, 1258–1267. doi: 10.1049/ipr2.13024

Ren, Z., Tang, Y., He, Z., Tian, L., Yang, Y., and Zhang, W. (2022). Ship detection in high-resolution optical remote sensing images aided by saliency information. *IEEE Trans. Geosci. Remote Sens.* 60, 1–16. doi: 10.1109/TGRS.2022.3173610

Ren, Z., Tang, Y., Yang, Y., and Zhang, W. (2024). Sasod: Saliency-aware ship object detection in high-resolution optical images. *IEEE Trans. Geosci. Remote Sens.* 62, 1–15. doi: 10.1109/TGRS.2024.3367959

Shi, Y., Li, S., Liu, Z., Zhou, Z., and Zhou, X. (2024). Mtp-yolo: You only look once based maritime tiny person detector for emergency rescue. *J. Marine Sci. Eng.* 12, 669. doi: 10.3390/jmse12040669

Si, J., Song, B., Wu, J., Lin, W., Huang, W., and Chen, S. (2023). Maritime ship detection method for satellite images based on multiscale feature fusion. *IEEE J. Selected Topics Appl. Earth Observations Remote Sens.* 16, 6642–6655. doi: 10.1109/JSTARS.2023.3296898

Varga, L. A., Kiefer, B., Messmer, M., and 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*. New York, USA: IEEE. 2260–2270.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024). Yolov10: Real-time end-to-end object detection. *arXiv preprint*. doi: 10.48550/arXiv.2405.14458

Wu, J., Li, J., Li, R., Xi, X., Gui, D., and Yin, J. (2022). A fast maritime target identification algorithm for offshore ship detection. *Appl. Sci.* 12, 4938. doi: 10.3390/app12104938

Xu, X., Liu, Y., Lyu, L., Yan, P., and Zhang, J. (2023). Mad-yolo: A quantitative detection algorithm for dense small-scale marine benthos. *Ecol. Inf.* 75, 102022. doi: 10.1016/j.ecoinf.2023.102022

Xue, J., He, D., Liu, M., and Shi, Q. (2022). Dual network structure with interweaved global-local feature hierarchy for transformer-based object detection in remote sensing image. *IEEE J. Selected Topics Appl. Earth Observations Remote Sens.* 15, 6856–6866. doi: 10.1109/JSTARS.2022.3198577

Yang, S., Cao, Z., Liu, N., Sun, Y., and Wang, Z. (2024b). Maritime electro-optical image object matching based on improved yolov9. *Electronics* 13, 2774. doi: 10.3390/electronics13142774

Yang, D., Solihin, M. I., Ardiyanto, I., Zhao, Y., Li, W., Cai, B., et al. (2024a). A streamlined approach for intelligent ship object detection using el-yolo algorithm. *Sci. Rep.* 14, 15254. doi: 10.1038/s41598-024-64225-y

Yang, Z., Yin, Y., Jing, Q., and Shao, Z. (2023). A high-precision detection model of small objects in maritime uav perspective based on improved yolov5. *J. Marine Sci. Eng.* 11, 1680. doi: 10.3390/jmse11091680

Zhang, Y., Ge, H., Lin, Q., Zhang, M., and Sun, Q. (2022). Research of maritime object detection method in foggy environment based on improved model src-yolo. *Sensors* 22, 7786. doi: 10.3390/s22207786

Zhang, C., Lam, K.-M., Liu, T., Chan, Y.-L., and Wang, Q. (2024a). Structured adversarial self-supervised learning for robust object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 62. doi: 10.1109/TGRS.2024.3375398

Zhang, C., Liu, T., Xiao, J., Lam, K.-M., and Wang, Q. (2023a). Boosting object detectors via strong-classification weak-localization pretraining in remote sensing imagery. *IEEE Trans. Instrumentation Measurement* 72, 1–20. doi: 10.1109/TIM.2023.3315392

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

Zhang, D., Wang, C., and Fu, Q. (2023b). Ofcos: an oriented anchor-free detector for ship detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* 20, 1–5. doi: 10.1109/LGRS.2023.3252572

Zhang, C., Xiao, J., Yang, C., Zhou, J., Lam, K.-M., and Wang, Q. (2024b). Integrally mixing pyramid representations for anchor-free object detection in aerial imagery. *IEEE Geosci. Remote Sens. Lett.* doi: 10.1109/LGRS.2024.3404481

Zhang, Y., Yin, Y., and Shao, Z. (2023e). An enhanced target detection algorithm for maritime search and rescue based on aerial images. *Remote Sens.* 15, 4818. doi: 10.3390/rs15194818

Zhang, L., Zhang, N., Shi, R., Wang, G., Xu, Y., and Chen, Z. (2023c). Sg-det: shuffle-ghostnet-based detector for real-time maritime object detection in uav images. *Remote Sens.* 15, 3365. doi: 10.3390/rs15133365

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al. (2024). "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New York, USA: IEEE. 16965–16974.

Zhao, H., Zhang, H., and Zhao, Y. (2023). "Yolov7-sea: Object detection of maritime uav images based on improved yolov7," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. IEEE: New York, USA. 233–238. doi: 10.3390/jmse11091680

Zhou, W., Zheng, F., Yin, G., Pang, Y., and Yi, J. (2022). Yolotrashcan: A deep learning marine debris detection network. *IEEE Trans. Instrumentation Measurement* 72, 1–12. doi: 10.1109/TIM.2022.3225044

Zhu, Q., Ma, K., Wang, Z., and Shi, P. (2023). Yolov7-csaw for maritime target detection. *Front. neurorobotics* 17, 1210470. doi: 10.3389/fnbot.2023.1210470

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv:2010.04159*. doi: 10.48550/arXiv.2010.04159