Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.087611

ARTICLE

Edge-Oriented Infrared Ship Pattern Recognition in Complex Maritime Scenes via Deep Feature Enhancement and Teacher-Guided Distillation

Hongliang Tian[1], Chenying Pei[1,*], Jin Lei[2], Xiaoke Liu[1] and Xin Ma[3]

[1]Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education (Northeast Electric Power University), Jilin, China

[2]School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China

[3]Micro Engineering and Micro Systems Laboratory, School of Mechanical and Aerospace Engineering, Jilin University, Changchun, China

*Corresponding Author: Chenying Pei. Email: 2202400392@neepu.edu.cn

Received: 19 June 2026; Accepted: 28 August 2026; Published: 28 September 2026

ABSTRACT: Infrared ship detection is an important deep learning-based pattern recognition task for maritime visual perception, where accurate target recognition under complex thermal backgrounds is essential for intelligent monitoring and real-time decision support. However, low target-background contrast, sea-wave thermal textures, coastline heat-source interference, and specular thermal reflections in infrared maritime imaging weaken discriminative ship patterns and reduce recognition reliability in complex scenes. To address these challenges, we propose an edge-oriented infrared ship detection method for real-time maritime monitoring. The proposed method reconstructs the feature pyramid by integrating a Wavelet-Frequency Enhancement Module (WFEM) with a Dynamic Multi-Scale Feature Fusion Module (DMS-FFM), thereby enhancing thermal clutter suppression, small-target pattern representation, and multi-scale feature interaction in dense maritime environments. A Task-Conditioned Unified Detection Head (TCUDH) is further introduced to improve localization robustness through shared representation learning and task-conditioned feature modulation, while retaining structural extensibility for related visual prediction tasks. In addition, a Teacher-Guided Dual-level Structured Knowledge Distillation (TDSKD) strategy improves student classification and localization through prediction-structure and geometric consistency distillation. Experimental results demonstrate that the proposed method achieves excellent detection performance across multiple datasets while maintaining low computational overhead, with 3.71M parameters, 22.2 giga floating-point operations (GFLOPs), and 234.7 frames per second (FPS) on an NVIDIA GeForce RTX 3090 graphics processing unit (GPU). Moreover, the proposed model achieves an average effective inference throughput of 28.23 FPS on a Jetson Orin Nano 8 GB edge device, demonstrating its potential for real-time edge-side pattern recognition and visual perception in infrared maritime monitoring tasks.

KEYWORDS: Infrared ship detection; deep learning; pattern recognition; computer vision; target detection; edge deployment; maritime visual perception

1 Introduction

Vision-based pattern recognition plays an important role in intelligent monitoring systems because it enables automatic target recognition, scene understanding, and real-time decision support from visual data. With the development of machine learning and deep learning, visual perception models have been widely applied to complex target recognition tasks in remote sensing, transportation, industrial inspection, and maritime monitoring. In maritime scenarios, infrared ship detection serves as a key automated target

recognition task for understanding ship distribution and supporting timely visual analysis under low-visibility conditions. Compared with visible-light imaging, infrared imaging does not rely on external illumination, making it suitable for nighttime and low-visibility maritime monitoring. However, infrared maritime images are often affected by low resolution, sparse texture, low target-background contrast, and severe background thermal noise. In such scenes, small and weak ships usually appear as localized thermal anomalies with blurred boundaries, while sea-wave thermal textures, coastline heat sources, and specular thermal reflections introduce complex clutter and spectral interference. This challenge becomes more severe in port and nearshore scenes, where densely moored ships are prone to adjacent-target aliasing, increasing the difficulty of multi-scale feature alignment and precise localization. Therefore, achieving effective thermal clutter suppression, discriminative feature representation, high localization accuracy, and real-time edge inference remains a critical challenge for deep learning-based infrared ship pattern recognition in complex maritime scenes.

Recent advances in computer vision have promoted the development of visual perception methods for complex infrared scenarios. Among existing object detection architectures, You Only Look Once version 11 (YOLOv11) [1] and Faster-RCNN [2] represent the one-stage and region-proposal-based two-stage detection paradigms, respectively. Considering the real-time and computational constraints of edge-side maritime monitoring, the compact YOLOv11n variant is selected as the baseline model in this study. The applicability of efficient one-stage detectors to maritime monitoring has also been demonstrated in satellite-based imagery. For example, one study constructed a diverse ship detection dataset covering open-ocean, near-coastal, and port scenes and systematically compared YOLOv5, YOLOv7, YOLOv8, and YOLOv9. Through diverse data augmentation techniques and atmospheric-scattering suppression, the resulting YOLOv9-based detector achieved strong small-ship detection performance under challenging conditions involving cloud interference, ship wakes, complex coastal backgrounds, illumination variations, and substantial target-scale variations, thereby demonstrating the practical value of modern YOLO detectors for intelligent maritime surveillance and time-sensitive ship recognition [3]. Building on this broader maritime-monitoring context, shore-based infrared monitoring involves modality-specific imaging mechanisms and sources of interference. In complex nearshore infrared scenes, although one-stage detectors offer high efficiency, they remain susceptible to factors such as sea-wave thermal clutter, background heat-source interference, large target scale variation, dense occlusion, and the low contrast of small and weak targets, which can lead to insufficient feature separation and localization representation, thereby degrading detection performance. Related research on densely moored and multi-scale targets predominantly focuses on optimizing multi-scale representations by enhancing feature pyramids or introducing attention mechanisms; for example, some studies dynamically filter critical regions via a bi-level routing attention mechanism to improve multi-scale localization accuracy in cluttered backgrounds [4], while others combine multi-scale edge fusion with dynamic task alignment to strengthen edge representations and boost multi-scale detection performance [5], and still others construct learnable feature pyramids and adaptive task-aligned detection heads to enhance scale robustness in dense scenes [6]. Various strategies have been explored in related studies to address the missed detection of low-contrast, small, and weak targets in infrared scenes. For instance, some works enhance bounding-box regression via dynamic task decoupling or improved loss functions to reduce the miss rate [7,8], while others strengthen small-target feature representation through spatial semantic enhancement and feature reconstruction [9], and still others extract edge and point features using multi-directional gradient operators to enhance small-target perception [10]. Meanwhile, to better balance noise suppression and edge-detail preservation, several studies improve target localization accuracy through dual-branch feature alignment [11] or through the combination of Chinese-knot convolution and multi-scale overlapping attention [12]. Although these methods improve the saliency of weak small

targets and multi-scale representation to a certain extent, they usually introduce more parameters and higher computational cost, which limits their applicability in real-time edge-side maritime monitoring. To alleviate these deployment limitations, recent studies on edge-side intelligent monitoring have increasingly explored lightweight model selection, hardware-aware acceleration, and energy-efficient system design. For example, some studies improve sea-surface target perception for unmanned surface vehicles by incorporating lightweight spatial pyramid pooling, enhanced multi-scale path aggregation, and fused-attention convolution into YOLO, thereby strengthening multi-scale feature extraction and fusion while reducing model complexity in complex marine environments [13], while others benchmark YOLOv6–YOLO11 across Intel and NVIDIA edge platforms and employ quantization and pruning to improve inference speed and power efficiency for unmanned surface vehicles [14], and still others introduce Ghost modules and Transformer-enhanced feature extraction into YOLOv5 to reduce model size and improve detection efficiency, and validate the resulting detector on ship videos collected by an unmanned surface vehicle under different marine conditions [15]. These studies indicate that practical embedded vision systems should be evaluated comprehensively in terms of detection accuracy, inference latency, hardware-specific performance, power consumption, and deployment adaptability. Based on these advances, the present study focuses on the additional challenges associated with deploying real-time infrared ship detection in maritime environments, including thermal clutter, weak target contrast, and constrained onboard computing resources.

To improve the robustness and deployability of infrared ship detection in complex maritime environments, an edge-oriented infrared ship detection system is proposed from two aspects: network design and hardware deployment. At the algorithmic level, the model improves vision-based perception capability through thermal clutter suppression, multi-scale feature modeling in dense scenes, high-quality localization, and training optimization. At the hardware level, the system establishes an edge-side deployment loop over a local-area network built by a mobile hotspot (Fig. 1). In this loop, the Jetson Orin Nano 8GB and a Lenovo tablet are collaboratively connected. Since the Tianyan X3 thermal imager does not provide an independent software development kit (SDK), the tablet serves as a relay terminal that forwards the thermal stream to the Jetson device through the real-time streaming protocol (RTSP), thereby enabling the complete chain of infrared acquisition, inference, and visualization in real-world outdoor monitoring scenarios. The main contributions of this study are as follows:

(1) A deep learning-based infrared ship pattern recognition framework is proposed for complex maritime scenes. The framework improves vision-based perception reliability under low contrast, thermal clutter interference, and dense target distribution, while supporting real-time maritime monitoring on edge devices.

(2) For feature modeling, Wavelet-Frequency Enhancement Module (WFEM) and Dynamic Multi-Scale Feature Fusion Module (DMS-FFM) are combined to enhance feature representation under complex infrared backgrounds and improve cross-scale semantic alignment and feature interaction in dense scenes.

(3) Task-Conditioned Unified Detection Head (TCUDH) is introduced to improve prediction robustness through shared representation learning and task-conditioned feature modulation, strengthening localization representation for low-contrast, elongated, and densely distributed ship targets.

(4) Teacher-guided Dual-level Structured Knowledge Distillation (TDSKD) is designed to enhance student classification and localization through prediction-structure consistency and object-level geometric consistency, without increasing inference-stage computational cost.

(5) Extensive experiments on multiple maritime datasets demonstrate the effectiveness and cross-scene robustness of the proposed method. RTSP-based field operation on the Jetson Orin Nano 8GB platform achieves an average effective inference throughput of 28.23 FPS, further verifying its engineering feasibility for real-time edge-side maritime visual perception.

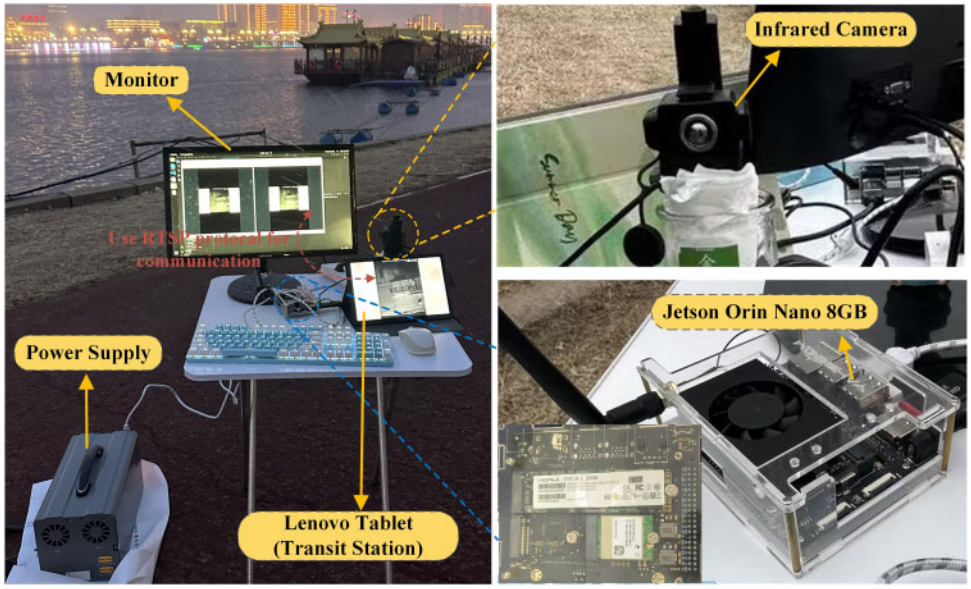

Figure 1: Edge deployment.

2 Methodology

Before introducing the individual components, we first provide an overview of the proposed network in Fig. 2. Fig. 2a presents the detailed network architecture and defines the principal module abbreviations, whereas Fig. 2b provides a simplified workflow from YOLOv11n-based multi-level feature extraction through DMSFPN to TCUDH. Built upon YOLOv11n, the proposed method improves infrared ship detection in complex maritime scenes through the coordinated optimization of feature modeling, object prediction, and training supervision, rather than relying solely on isolated module-level enhancements. These designs aim to improve vision-based perception reliability under thermal clutter interference, weak target responses, and dense target distributions. For feature modeling, a dynamic multi-scale feature pyramid network (DMSFPN), composed of WFEM and DMS-FFM, is constructed to suppress maritime thermal clutter, enhance feature representation, and facilitate cross-scale feature interaction. For object prediction, TCUDH is introduced to improve localization robustness and representation quality for low-contrast, elongated, and densely distributed infrared ship targets through shared representation learning and task-conditioned feature modulation. During training, TDSKD is further employed to improve detection performance without incurring any additional inference cost.

2.1 WFEM

In complex infrared maritime imaging, global periodic interference (such as thermal textures of ocean waves, thermal sources and reflections along the coastline) tends to smooth the high-frequency singularities of weak targets, resulting in smooth edges of weak targets, the local thermal anomalies being submerged, and the detection of small targets being compromised. Existing frequency-aware methods either enhance tiny-object spectral signatures while suppressing high-frequency background noise or exploit wavelet-domain decomposition for red–green–blue (RGB)–infrared feature fusion [16,17]. Although these methods improve frequency-aware representation, they are designed primarily for general tiny-object detection or multimodal RGB–infrared fusion and do not explicitly coordinate frequency-domain clutter suppression, wavelet-based boundary preservation, and aliasing mitigation during cross-scale feature-pyramid reconstruction. To address these limitations, WFEM (Fig. 3) integrates Gaussian-modulated dual-path downsampling and Haar-based cross-scale reconstruction with adaptive Fourier-wavelet-spatial fusion. Unlike methods that

perform frequency filtering or wavelet enhancement independently, WFEM jointly suppresses spectral clutter, preserves weak target boundaries, and maintains cross-scale feature consistency under complex thermal backgrounds.

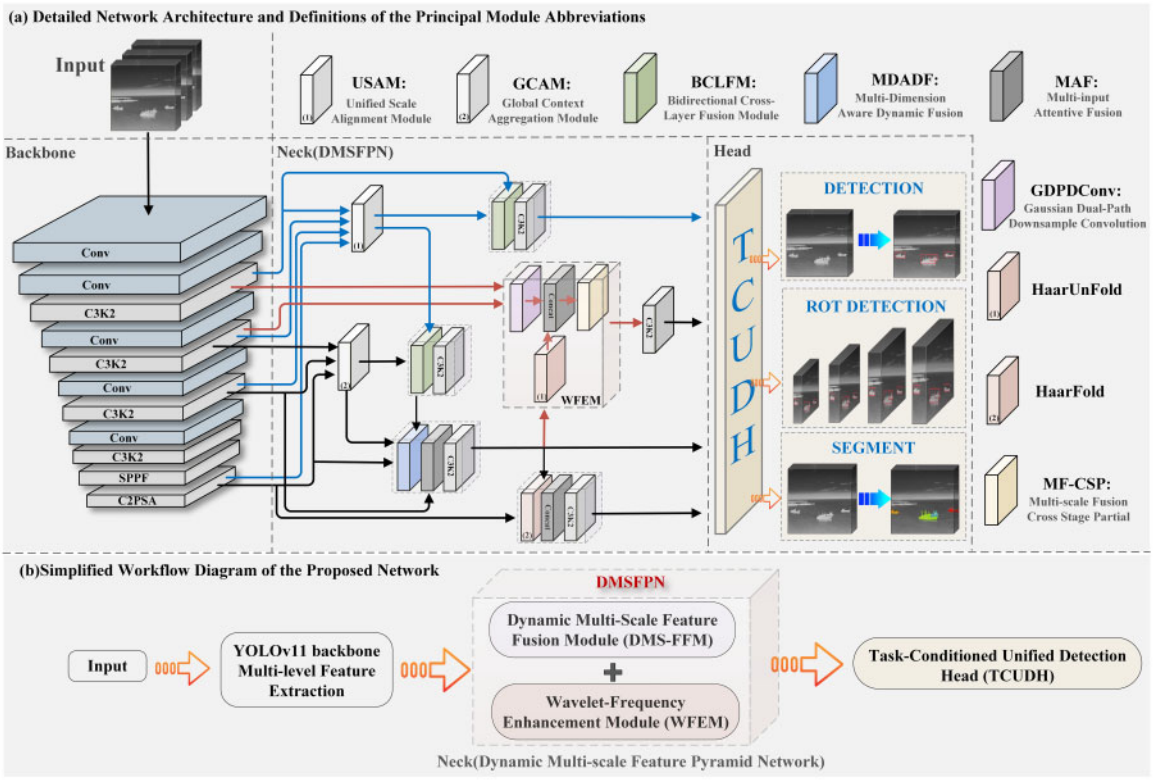

Figure 2: Overall architecture of the proposed network: (**a**) detailed network architecture and definitions of the principal module abbreviations; (**b**) simplified workflow of the proposed network.

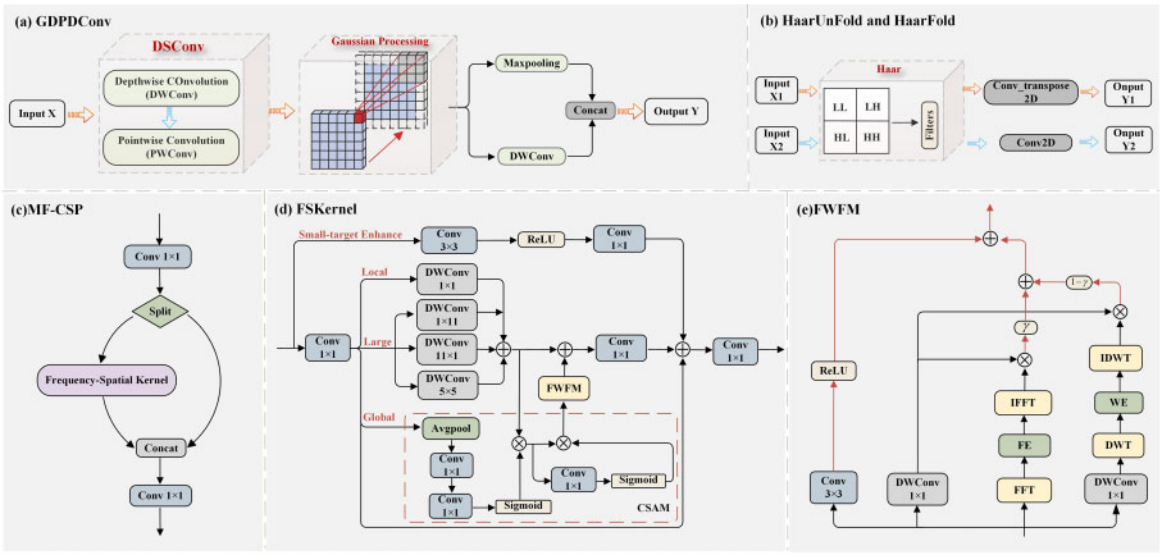

Figure 3: Schematic diagram of the internal module of WFEM.

WFEM first introduces the Gaussian Dual-Path Downsample Convolution (GDPDConv, Fig. 3a) to mitigate low-quality spectral aliasing during downsampling. It employs depthwise separable convolution to project features into a high-dimensional space while encoding local spatial context, and further introduces a Gaussian modulation mechanism, in which a fixed 5 × 5 Gaussian kernel $W(p)$ weights the local neighborhood features $F_c\,(i+p)$ to obtain the output $Y_c(i)$ at spatial position i and channel c (Eq. (1)).

$$W(p)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{\|p\|^2}{2\sigma^2}\right),\ Y_c(i)=\sum_{p\in\mathcal{N}}F_c\,(i+p)\cdot W(p) \tag{1}$$

here, p represents the relative spatial offset within the local neighborhood $\mathcal{N}$, $W(p)$ denotes the Gaussian weight at position p, and σ controls the spatial distribution of the Gaussian kernel. The modulated features then undergo dual-branch downsampling to achieve both anti-aliasing and detail preservation. Furthermore, WFEM utilizes HaarUnFold (Fig. 3b) to replace traditional interpolation upsampling, maintaining frequency-domain consistency in cross-scale reconstruction and mitigating edge blurring in infrared imaging.

GDPDConv alleviates downsampling aliasing, but infrared weak and small targets are still prone to being lost in deep networks. The proposed Multi-scale Fusion Cross Stage Partial module (MF-CSP, Fig. 3c) employs a frequency-spatial collaborative enhancement mechanism to perform semantic decoupling, separating target signals from complex background textures, while residual connections are incorporated to ensure the lossless propagation of weak gradients from small targets. In the spatial perception phase, Frequency-Spatial Kernel (FSKernel, Fig. 3d) employs multi-kernel depthwise separable convolutions, including asymmetric 1 × 11 and 11 × 1 kernels, to capture the elongated topological structures of ships. Given the input feature Z, it generates the multi-scale aggregated spatial feature Z_{agg} (Eq. (2)).

$$Z_{agg}=f^{dw}_{1\times11}(Z)\oplus f^{dw}_{11\times1}(Z)\oplus f^{dw}_{5\times5}(Z)\oplus f^{dw}_{1\times1}(Z) \tag{2}$$

here, $\oplus$ represents element-wise addition, and $f^{dw}_{k\times k}(\cdot)$ denotes depthwise separable convolution with a kernel size of $k\times k$, which aggregate spatial context from the horizontal, vertical, and local dimensions, respectively. After being recalibrated by the Channel-Spatial Attention Module (CSAM), the aggregated features are subsequently fed into the Frequency-Wavelet Fusion Model (FWFM, Fig. 3e), where dual-spectrum denoising is performed. A Fast Fourier Transform (FFT) is first applied to map the spatial feature F_{in} into the frequency domain S, from which the amplitude spectrum A and phase spectrum φ are extracted (Eq. (3)). A high-pass filter is then employed (Eq. (4)) to reweight and enhance the amplitude spectrum (Eq. (5)), while the phase spectrum is kept unchanged to preserve the structural layout of ship targets. The enhanced complex spectrum is transformed back into the spatial domain via the Inverse Fast Fourier Transform (IFFT). By taking the real component $\Re$ of the reconstructed result, the frequency-domain denoised spatial feature F_f is obtained (Eq. (6)).

$$S=\mathcal{F}\,(F_{in})=A\cdot\exp\,(j\varphi) \tag{3}$$

$$H_{HPF}\,(u,v)=1-\exp\left(-\frac{(u-u_0)^2+(v-v_0)^2}{2D_0^2}\right) \tag{4}$$

$$A_{en}=A\otimes(1+\lambda_E\otimes H_{HPF}) \tag{5}$$

$$F_f=\Re\left(\mathcal{F}^{-1}\,(A_{en}\cdot\exp\,(j\varphi))\right) \tag{6}$$

here, $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the FFT and IFFT, respectively; (u_0,v_0) represents the center of the frequency domain coordinate, and D_0 is the cutoff frequency radius. λ_E, initialized as 0.5, is a channel-wise learnable

enhancement factor, and ⊗ denotes element-wise multiplication. Meanwhile, the wavelet branch decomposes the input feature F_{in}, after 1×1 channel mixing, through a Haar-based discrete wavelet transform (DWT) into a concatenated tensor X_w composed of four wavelet sub-bands: low-low (LL), low-high (LH), high-low (HL), and high-high (HH). Among them, LL represents the low-frequency approximation component, whereas LH, HL, and HH represent the high-frequency detail components. Depthwise separable convolution combined with the rectified linear unit (ReLU) activation function ρ is then applied to enhance interaction across all four wavelet sub-bands, after which the enhanced feature $X_{w_{en}}$ is reconstructed through an inverse discrete wavelet transform (IDWT) to obtain F_w (Eq. (7)).

$$X_{w_{en}} = f_{1\times1}^{conv}\left(\rho\left(f_{3\times3}^{dw}\left(X_w\right)\right)\right), F_w = IDWT\left(X_{w_{en}}\right) \tag{7}$$

Considering the substantial variations in noise distribution across different infrared scenarios, such as nearshore and remote-sensing scenes, FWFM employs a dynamic adaptive balancing mechanism (Eq. (8)), whereby a learnable scalar coefficient $\gamma \in [0,1]$ is introduced to adjust the fusion weights of the Fourier branch F_f and the wavelet-based edge-preserving branch F_w, while a parallel spatial branch is incorporated to extract high-frequency details.

$$F_{FWFM} = \gamma \cdot F_f + (1-\gamma) \cdot F_w + \rho\left(f_{3\times3}^{conv}\left(F_{in}\right)\right) \tag{8}$$

Through the dynamic adaptive balancing mechanism in FWFM, WFEM adaptively weights the Fourier and wavelet branches while incorporating the parallel spatial branch, thereby suppressing complex infrared background clutter and enhancing the high-frequency structural information of weak targets.

2.2 DMS-FFM

In infrared maritime scenarios, ship targets often exhibit large scale variations, dense mooring, and mutual interference from adjacent thermal responses, leading to edge attenuation, semantic aliasing between adjacent ships, and cross-layer representation misalignment. Existing studies have improved multi-scale ship detection by combining Codebook-based weather weakening, uniform-resolution MFBlock fusion, and re-parameterized information injection to strengthen distant small-target representation [18], while others introduce high-resolution P2 features and dynamic spatial-channel attention fusion to improve low-angle small-ship detection under resource-constrained conditions [19]. Although effective, these methods primarily focus on uniform-resolution feature aggregation, information injection into the small-target detection head, or attention-based feature reweighting, without jointly modeling the coupled effects of scale mismatch, cross-layer semantic imbalance, and directional-topology degradation caused by overlapping thermal responses in dense infrared scenes. Therefore, we propose DMS-FFM (Fig. 4), which uses the intermediate feature level as the alignment reference and reconstructs feature interaction into a progressive fusion paradigm comprising scale alignment, large-kernel global context aggregation, reverse-gated bidirectional compensation, direction-aware topology reconstruction, and multi-input attentive calibration. Different from existing uniform-resolution aggregation or attention-based reweighting strategies, DMS-FFM jointly alleviates scale mismatch, semantic aliasing, and long-axis structural degradation caused by overlapping thermal responses, thereby enhancing the consistency and discriminability of multi-scale features in complex infrared maritime scenarios.

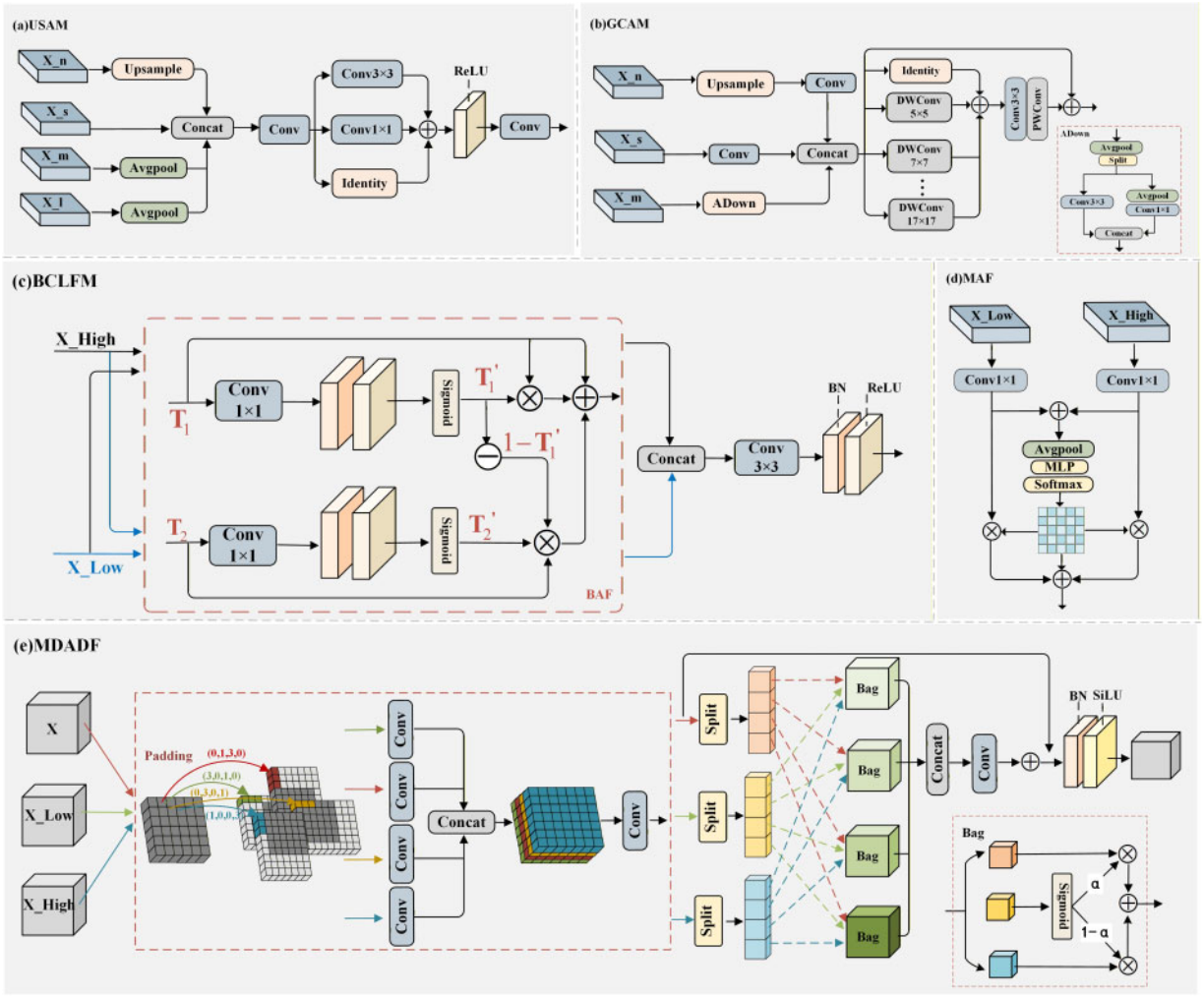

Figure 4: Schematic diagram of the internal module of DMS-FFM.

The Unified Scale Alignment Module (USAM, Fig. 4a) is first introduced to mitigate spatial misalignment across feature levels, in which the spatial resolution of the intermediate scale X_s is used as the reference to align features from different levels into a unified coordinate space through adaptive average pooling $\mathcal{P}(\cdot)$ and bilinear upsampling $\mathcal{U}(\cdot)$ (Eq. (9)), and an embedded enhancement mapping is further applied to reduce inter-layer redundancy and strengthen effective semantic interaction (Eq. (10)).

$$Z_u = \text{Concat}\left(\mathcal{P}\left(X_l; H, W\right), \mathcal{P}\left(X_m; H, W\right), X_s, \mathcal{U}\left(X_n; H, W\right)\right) \tag{9}$$

$$F_u = \Phi_u(Z_u) = \phi_o(\mathcal{R}(\phi_e(Z_u))) \tag{10}$$

here, $\Phi_u(\cdot)$ denotes the embedded enhancement mapping, while $\phi_e(\cdot)$ and $\phi_o(\cdot)$ represent the input and output projection convolutions, respectively, and $\mathcal{R}(\cdot)$ indicates the re-parameterized convolution enhancement operation. Since scale alignment alone remains insufficient to capture the broad contextual information in infrared maritime scenes, the Global Context Aggregation Module (GCAM, Fig. 4b) is introduced, where the concatenated feature Z_g is processed by multi-branch large-kernel convolutions $D_{k_i}(\cdot)$ to enlarge the receptive field and enhance contextual feature representation (Eq. (11)), alleviating semantic incompleteness caused by open backgrounds and sparse target distribution.

$$\hat{Z}_g = Z_g + \sum^{N} D_{k_i}\left(Z_g\right), \quad k_i \in \{5, 7, \ldots, 17\}, \tag{11}$$

In dense scenes, where the direct fusion of shallow and deep features tends to cause a semantic-detail imbalance, the Bidirectional Cross-Layer Fusion Module (BCLFM, Fig. 4c) is introduced to enable more effective cross-layer feature integration. The two input features T_1 and T_2 are first aligned and recalibrated by 1×1 convolutions to produce two attention weights, $T_1^{'}$ and $T_2^{'}$ (Eq. (12)), based on which T_1 is selected as the main branch for attention enhancement, while complementary information from the secondary input T_2 is injected through a reverse gating coefficient $1 - T_1^{'}$ (Eq. (13)); the two BAF-calibrated features are then fused to generate the aligned output feature Y (Eq. (14)), improving the balance between detail preservation and semantic consistency. Here, δ represents the Sigmoid function, and $\otimes$ denotes channel-wise multiplication.

$$T_1' = \delta\left(Conv_{1\times1}\left(T_1\right)\right), \quad T_2' = \delta\left(Conv_{1\times1}\left(T_2\right)\right) \tag{12}$$

$$BAF\left(T_1, T_2\right) = T_1' \otimes T_1 + T_2' \otimes T_2 \otimes \left(1 - T_1'\right) + T_1 \tag{13}$$

$$Y = Conv_{3\times3}\left(Concat\left(BAF\left(T_1, T_2\right), BAF\left(T_2, T_1\right)\right)\right) \tag{14}$$

BCLFM alleviates the imbalance between shallow and deep features, but cross-layer calibration alone remains insufficient for recovering the long-axis contours and local topological structures of ship targets, motivating the introduction of the Multi-Dimension Aware Dynamic Fusion module (MDADF, Fig. 4e) to compensate for the limited ability of conventional square convolutions to perceive directional structures in dense scenes. By employing horizontal and vertical directional convolutions with 1 × k and k × 1 kernels together with asymmetric padding (Eq. (15)), it constructs a star-shaped receptive field, and the resulting direction-aware features are then fused (Eq. (16)).

$$F_g = \begin{cases} Conv_{(1,\mathrm{k})}\left(Pad_{g(X)}\right), & g = 0, 1 \\ Conv_{(\mathrm{k},1)}\left(Pad_{g(X)}\right), & g = 2, 3 \end{cases} \tag{15}$$

$$Y = \mathrm{Conv}_{2\times2}\left(\mathrm{Concat}\left(F_0, F_1, F_2, \mathrm{F}_3\right)\right) \tag{16}$$

here, Pad_g denotes the g-th asymmetric padding operation. Features aligned with the axial contours of ship targets are partitioned into multiple channel blocks (bags). For each block, a gating mechanism generates the attention weight ε to adaptively fuse the existing shallow-detail feature l_i and deep-semantic feature h_i through a weighted sum, producing the aggregated feature m_i' as given in Eq. (17), where $i \in [1, 2, 3, 4]$. The complete feature $\hat{F}_m$ is then restored through channel-wise recombination and convolutional mapping, enhancing spatial topological sensitivity while preserving semantic consistency (Eq. (18)).

$$\varepsilon = \delta\left(\mathrm{m}_i\right), \quad m_i' = \varepsilon l_i + \left(1 - \varepsilon\right) h_i \tag{17}$$

$$F_m' = \left[m_1', m_2', m_3', m_4'\right], \hat{F}_m = \alpha\left(\beta\left(Conv\left(F_m'\right)\right)\right) \tag{18}$$

here, α and β denote the Sigmoid Linear Unit (SiLU) activation and batch normalization, respectively. In the output stage, the Multi-input Attentive Fusion module (MAF, Fig. 4d) applies branch-level channel-wise recalibration to features from different sources, where the two input features are first added to form the global guidance feature F_s, based on which adaptive branch weights A are generated through global average pooling and a multi-layer perceptron (Eq. (19)), and used for channel-wise weighting and summation (Eq. (20)), which enables an adaptive balance between shallow detail representations and deep semantic representations.

$$A = \mathrm{Softmax}_{branch}\left(\mathrm{MLP}\left(\mathrm{GAP}\left(F_s\right)\right)\right), A = \left\{A_{low}, A_{high}\right\}, A_k \in \mathbb{R}^{B\times C\times1\times1} \tag{19}$$

$$F_{out} = A_{low} \otimes F_{low} + A_{high} \otimes F_{high} \tag{20}$$

DMS-FFM alleviates scale mismatch, semantic aliasing, and topological degradation in dense infrared maritime scenes through progressive feature reconstruction.

2.3 TCUDH

Conventional detection heads often suffer from limited geometric localization accuracy and weak boundary delineation when applied to infrared ship targets with elongated contours, weak edge responses, and non-uniform thermal radiation distributions. Existing detection-head designs improve prediction efficiency through decoupled classification and regression branches [20], while dynamic heads adaptively recalibrate features across scale, spatial, and task dimensions [21], and scale- and occlusion-aware methods strengthen difficult-target responses through receptive-field enhancement and attention-based feature compensation [22]. Although effective, these methods primarily focus on branch separation or adaptive feature reweighting and do not explicitly model the multidirectional boundary differences and task-dependent representation requirements of low-contrast and elongated infrared ships. Therefore, we propose TCUDH (Fig. 5), which performs channel alignment on multi-scale features and employs quad-directional difference convolution (QDConv) to encode center, horizontal, vertical, and angular difference cues into a shared enhanced representation. On this basis, a bounded task-conditioned modulation mechanism generates objective-specific scaling, bias, gating, and lightweight residual adaptation. Different from conventional decoupled or attention-enhanced heads, TCUDH coordinates weak-boundary enhancement and controlled task adaptation within a compact shared prediction structure, thereby improving localization robustness while retaining structural extensibility for related visual prediction tasks.

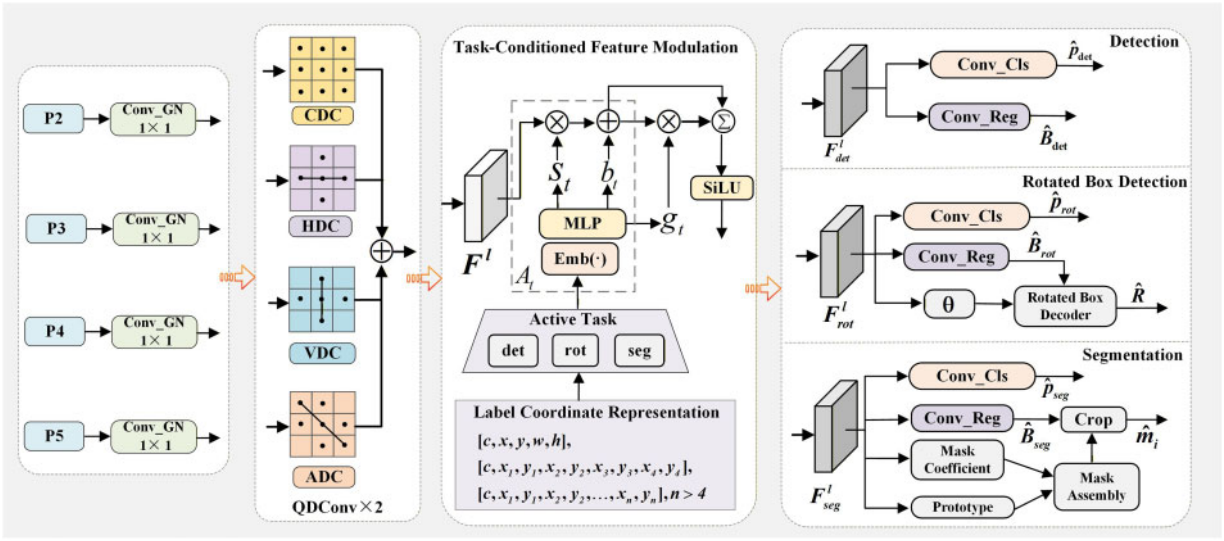

Figure 5: TCUDH.

For the input feature at the l-th scale, 1 × 1 channel alignment and QDConv are applied to obtain the shared enhanced feature F^l (Eq. (21)), where QDConv is uniformly formulated over a 3 × 3 local spatial grid Ω by representing each neighboring pixel $u = (x, y)$, $x, y \in \{-1, 0, 1\}$ through differential encoding with the corresponding position-specific weight $w_m(u)$ (Eq. (22)).

$$F^l = \alpha\left(Q_2\left(\alpha\left(Q_1\left(\mathrm{Conv}_{1\times1}\left(\mathrm{X}^l\right)\right)\right)\right)\right) \tag{21}$$

$$Q\left(Z(\mathrm{p})\right) = \sum_{m\in\mathcal{M}}\sum_{u\in\Omega} w_m(u)\left(Z\left(\mathrm{p}+u\right) - Z\left(p + r_m\right)\right), \mathcal{M} = \{c, h, v, a\} \tag{22}$$

here, $\alpha(\cdot)$ denotes the SiLU activation function, and r_m represents the reference position determined by the corresponding difference mode. The reference positions for center difference r_c, horizontal difference r_h, vertical difference r_v, and angular difference r_a are $(0,0)$, $(x,-y)$, $(-x,y)$, and $(-x,-y)$, respectively. These four difference branches operate in parallel over the local neighborhood to enhance the representation of target directional structures and weak boundaries.

After obtaining the shared enhanced feature F^l, a task-conditioned modulation mechanism is introduced, where the task identifier $t \in \{\text{det}, \text{rot}, \text{seg}\}$ is first mapped into a task embedding vector e_t, from which the scale modulation term s_t, bias modulation term b_t, and gating response g_t are derived and jointly integrated with a lightweight residual adapter to construct the task-conditioned feature F_t^l (Eq. (23)).

$$\begin{cases} e_t = Emb(t), \quad \left[\hat{s}_t, \hat{b}_t, \hat{g}_t\right] = MLP\left(e_t\right) \\ s_t = 1 + 0.25tanh\left(\hat{s}_t\right), \quad b_t = 0.25tanh\left(\hat{b}_t\right), \quad g_t = \delta\left(\hat{g}_t\right), \\ F_t^l = \alpha\left(s_t \otimes F^l + b_t + g_t \otimes A_t\left(F^l\right)\right) \end{cases} \tag{23}$$

where, $Emb(\cdot)$ denotes the task embedding lookup, $\delta(\cdot)$ is the Sigmoid function, and $A_t(\cdot)$ represents the task-related lightweight residual adapter. By constraining the magnitudes of s_t and b_t through bounded mapping, strong conditional perturbations that may disrupt the distribution of shared features can be avoided. For the task-conditioned feature F_t^l, a unified detection projection is applied to generate the classification output C_t^l and the distance distribution regression output R_t^l. The spatial dimensions of the multi-scale outputs are then flattened and concatenated using $\mathrm{Vec}(\cdot)$, yielding the unified distance distribution R_t and classification response C_t (Eq. (24)).

$$\begin{cases} \mathrm{R}_t^l = s_l \mathcal{C}_{\text{reg}}\left(\mathrm{F}_t^l\right), \quad \mathrm{C}_t^l = \mathcal{C}_{\text{cls}}\left(\mathrm{F}_t^l\right) \\ \mathrm{R}_t = \mathrm{Concat}^l\left(\mathrm{Vec}\left(\mathrm{R}_t^l\right)\right), \quad \mathrm{C}_t = \mathrm{Concat}^l\left(\mathrm{Vec}\left(\mathrm{C}_t^l\right)\right), \quad \hat{\mathrm{d}}_t = \mathrm{DFL}\left(\mathrm{R}_t\right) \end{cases} \tag{24}$$

here, $\mathcal{C}_{reg}$ and $\mathcal{C}_{cls}$ denote the regression projection layer and classification projection layer, respectively, and s_l is the learnable scaling factor at the l-th scale. $\mathrm{DFL}(\cdot)$ denotes the distribution focal learning decoding operator. Within TCUDH, a shared prediction representation is used for classification and regression. The DFL operator first decodes the unified distance distribution R_t into continuous distance estimates $\hat{d}_t$. Subsequently, $\mathrm{dist2bbox}\left(\hat{\mathrm{d}}_t, \mathrm{a}\right)$ converts these distance estimates into bounding boxes relative to the anchor points a, and the resulting boxes are scaled by the stride tensor τ to obtain the predicted bounding boxes $\hat{B}$. Meanwhile, the class confidences $\hat{p}$ are independently obtained by applying the Sigmoid function $\delta(\cdot)$ to the classification response C_t (Eq. (25)).

$$\hat{B} = \mathrm{dist2bbox}\left(\hat{\mathrm{d}}_t, \mathrm{a}\right) \otimes \tau, \quad \hat{p} = \delta\left(\mathrm{C}_t\right) \tag{25}$$

For the rotated detection task, an additional rotation parameter θ is predicted (Eq. (26)). By mapping the original angle $\hat{\theta}$ into a continuous interval $[-\pi/4, 3\pi/4]$, numerical discontinuities near the parameter boundaries for high-aspect-ratio targets can be alleviated. The distance distribution regression result, rotation parameter, and anchor information are then jointly decoded to obtain the final rotated bounding boxes (Eq. (27)).

$$\hat{\theta} = Concat\left(\mathrm{Vec}\mathcal{C}_{ang,l}\left(\mathrm{F}_{rot}^l\right)\right), \theta = \left(\delta\left(\hat{\theta}\right) - 0.25\right) \cdot \pi \tag{26}$$

$$\hat{R} = dist2rbox\left(\hat{d}_{rot}, \theta, a\right) \otimes \tau \tag{27}$$

where, $\mathcal{C}_{ang,l}(\cdot)$ denotes the angle prediction branch at the l-th scale and $dist2rbox(\cdot)$ denotes the rotated bounding-box decoding function based on boundary distances and rotation parameters.

For the segmentation task, the task-conditioned segmentation feature F_{seg}^{l} is processed by two dedicated segmentation branches. The prototype generation branch produces the prototype mask set P, while the scale-specific mask coefficient prediction branches generate the instance coefficients M. The operator Crop$(\cdot)$ is then used to perform spatial cropping and resampling according to the predicted bounding boxes, yielding the final segmentation result $\hat{m}_i$ for the i-th instance and providing explicit region-level prediction within TCUDH (Eq. (28)).

$$\mathrm{P} = \mathcal{P}\left(\mathrm{F}_{seg}^{l}\right), \mathrm{M} = \mathrm{Concat}_{l}\left(\mathrm{Vec}\left(\mathcal{C}_{\mathrm{mask},l}\left(\mathrm{F}_{seg}^{l}\right)\right)\right), \hat{m}_i = \mathrm{Crop}\left(\sum_{k=1}^{n_m} m_{ik}\mathrm{P}_k, \hat{B}_i\right) \tag{28}$$

here, $\mathcal{P}\left(\mathrm{F}_{seg}^{l}\right)$, $\mathrm{P} = \{\mathrm{P}_k\}_{k=1}^{n_m}$, and $\mathrm{M} = [m_{ik}]$ denote the prototype generation branch, prototype mask set, and instance mask coefficients, respectively, while $\mathcal{C}_{mask,l}(\cdot)$ denotes the mask coefficient prediction branch at the l-th scale. TCUDH improves localization robustness for infrared ship detection through task-conditioned unified prediction, while preserving structural extensibility for segmentation-related prediction.

2.4 TDSKD

Knowledge distillation (KD) aims to transfer the representation capability and decision distribution of a teacher model to a lightweight student model. Recent detector-oriented knowledge distillation methods transfer task-specific prediction knowledge through cross-head prediction mimicking [23], apply masked generative distillation to lightweight infrared object detection [24], or use channel- and spatial-attention knowledge distillation to enhance foreground and effective-background feature transfer in aerial object detection [25]. Although effective, these methods mainly emphasize prediction imitation or feature-level knowledge transfer, without jointly constraining the fine-grained classification–localization structure and object-level rotated geometry of infrared ships. To address these limitations, we propose TDSKD (Fig. 6), which combines prediction-structure consistency with teacher-guided object-level geometric consistency to jointly transfer classification responses, localization distributions, and rotated-box geometry. Different from existing prediction- or feature-oriented distillation methods, TDSKD coordinates prediction-level and object-level guidance for weak, elongated, and densely distributed infrared ships without introducing additional inference overhead.

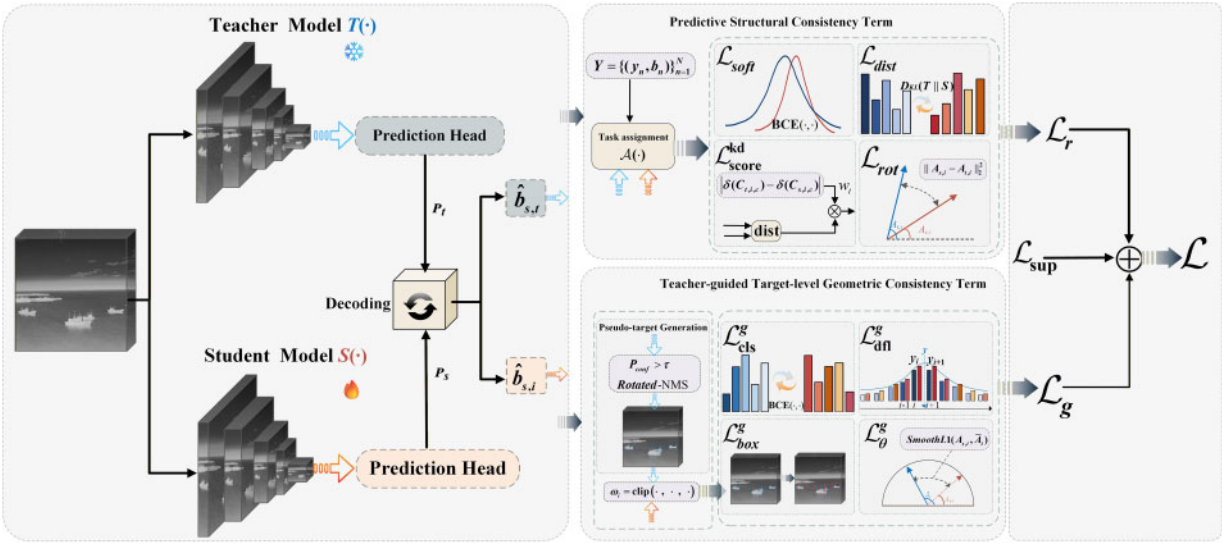

Figure 6: TDSKD.

Given an input image x, the teacher network $T(\cdot)$ and the student network $S(\cdot)$ use the boundary distance distribution $D^{(l)}$, classification logits $C^{(l)}$, and rotation angle-related outputs $A^{(l)}$ to uniformly represent the prediction results of the student and the teacher at the l-th detection layer (Eq. (29)). The set of predictions across all scales is denoted as $P_s = \left\{P_s^{(l)}\right\}_{l=1}^{L}$, $P_t = \left\{P_t^{(l)}\right\}_{l=1}^{L}$, where $l \in \{1, \cdots, L\}$, and L is the total number of detection scales.

$$P_s^{(l)} = \left(D_s^{(l)}, C_s^{(l)}, A_s^{(l)}\right), \quad P_t^{(l)} = \left(D_t^{(l)}, C_t^{(l)}, A_t^{(l)}\right) \tag{29}$$

Since the detection head is parameterized based on Distribution Focal Loss, the boundary regression results must be further decoded from the discrete distance distribution into continuous distance estimates. For any anchor point a_i, the corresponding continuous distance estimates for the student and teacher, $\overline{d}_{s,i}$ and $\overline{d}_{t,i}$, are obtained using the maximum discrete order R and the probability response at the r-th position $\mathrm{Softmax}(\cdot)_r$ (Eq. (30)), based on which the decoded rotated bounding box results for the student and teacher, $\hat{b}_{s,i}$ and $\hat{b}_{t,i}$, are further derived by utilizing the rotated box solver operator $\mathrm{Dec}_{\mathrm{obb}}(\cdot)$ together with the anchor point location and angle outputs (Eq. (31)).

$$\overline{d}_{s,i} = \sum_{r=0}^{R-1} r \cdot \mathrm{Softmax}\left(D_{s,i}\right)_r, \quad \overline{d}_{t,i} = \sum_{r=0}^{R-1} r \cdot \mathrm{Softmax}\left(D_{t,i}\right)_r \tag{30}$$

$$\hat{b}_{s,i} = \mathrm{Dec}_{\mathrm{obb}}\left(\overline{d}_{s,i}, A_{s,i}, a_i\right), \quad \hat{b}_{t,i} = \mathrm{Dec}_{\mathrm{obb}}\left(\overline{d}_{t,i}, A_{t,i}, a_i\right) \tag{31}$$

From a unified optimization perspective, the overall training objective $\mathcal{L}$ comprises the student model's original rotated-bounding-box detection supervision loss $\mathcal{L}_{\mathrm{sup}}$ and the distillation loss $\mathcal{L}_{\mathrm{kd}}$, with $\mathcal{L}_{\mathrm{kd}}$ scaled by the current mini-batch size B (Eq. (32)). Here, B denotes the number of samples in the current mini-batch, while $\mathcal{L}_r$ and $\mathcal{L}_g$ represent the prediction-structure consistency term and the teacher-guided object-level geometric consistency term, respectively.

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + B\mathcal{L}_{\mathrm{kd}}, \mathcal{L}_{\mathrm{kd}} = 2\mathcal{L}_r + \mathcal{L}_g \tag{32}$$

The prediction-structure consistency term is constructed by establishing positive- and negative-sample assignments according to the student model's current predictions and the ground-truth annotations. Specifically, the task assignment operator $\mathcal{A}(\cdot)$ takes the student classification probability vectors $\delta(C_s)$, the decoded rotated bounding boxes $\hat{B}_s$, and the ground-truth annotation set $Y = \{(y_n, b_n)\}_{n=1}^{N}$ as inputs, producing the assigned target boxes B^*, target scores S^*, and foreground mask M (Eq. (33)).

$$(B^*, S^*, M) = \mathcal{A}\left(\delta\left(C_s\right), \hat{B}_s, Y\right) \tag{33}$$

here, y_n and b_n represent the class label and the corresponding ground truth box of the i-th target, and B^*, S^*, and M denote the target box, target score, and foreground mask assigned to the anchor, while $M_i = 1$ indicates that the i-th anchor is a positive sample. After obtaining the foreground mask M_i, we compute the class response difference $\Delta_{i,c}$ for the c-th class using the classification probability vectors $\delta(C_{t,i,c})$ and $\delta(C_{s,i,c})$ of the teacher and student, and define the corresponding structural distillation weight w_i (Eq. (34)), based on which the scaled geometric score distillation term $\mathcal{L}_{\mathrm{score}}^{\mathrm{kd}}$ is further obtained using the weight w_i, the rotated bounding box overlap quality $\mathrm{CIoU}\left(\hat{b}_{s,i}, \hat{b}_{t,i}\right)$, and the normalization factor Ω (Eq. (35)).

$$\Delta_{i,c} = \left|\delta\left(C_{t,i,c}\right) - \delta\left(C_{s,i,c}\right)\right|, \quad w_i = M_i \cdot \max_c \Delta_{i,c} \tag{34}$$

$$\mathcal{L}_{\text{score}}^{\text{kd}} = \frac{7.5}{\Omega} \sum_i w_i \left(1 - \text{CIoU}(\hat{b}_{s,i}, \hat{b}_{t,i})\right) \tag{35}$$

To align the fine-grained distribution structures of the teacher and the student in the boundary distance space, we perform a vectorization operation $\text{vec}(\cdot)$ on the foreground locations to obtain the discrete boundary vectors $z_{s,i} = \text{vec}(D_{s,i})$, $z_{t,i} = \text{vec}(D_{t,i})$, $i \in \{M_i = 1\}$. Based on these distributions, we redefine the distribution structural alignment term using the KL divergence and the distribution distillation weight $\tilde{w}_i$ of the i-th foreground location (Eq. (36)).

$$\mathcal{L}_{\text{dist}} = \frac{T^2}{\sum_i \tilde{w}_i} \sum_i \tilde{w}_i \text{KL}\left(\text{Softmax}\left(\frac{z_{t,i}}{T}\right) \text{Softmax}\left(\frac{z_{s,i}}{T}\right)\right) \tag{36}$$

In Eq. (36), T denotes the temperature coefficient used to control the smoothness of the teacher and student distance distributions during KL-divergence-based alignment. Beyond the distance distribution, the teacher's classification response also contains crucial soft structural information. Therefore, the soft classification distillation term is computed using the binary cross-entropy function $\text{BCE}(\cdot,\cdot)$ and the corresponding teacher supervision probability $p_{t,i,c}$ (Eq. (37)). For rotation detection, we utilize the angle outputs of both the teacher and the student, alongside the squared Euclidean distance $\|\cdot\|_2^2$, to obtain the angular response consistency term $\mathcal{L}_{\text{rot}}$ (Eq. (38)).

$$\mathcal{L}_{\text{soft}} = \frac{1}{\Omega} \sum_i M_i \sum_c BCE\left(C_{s,i,c}, p_{t,i,c}\right) \Delta_{i,c} \tag{37}$$

$$\mathcal{L}_{\text{rot}} = \frac{1}{N_a} \sum_i \| A_{s,i} - A_{t,i} \|_2^2 \tag{38}$$

where, N_a denotes the total number of angle prediction units. Based on the aforementioned steps, we obtain the final prediction structural consistency loss $\mathcal{L}_r$ (Eq. (39)).

$$\mathcal{L}_r = \mathcal{L}_{\text{score}}^{kd} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{soft}} + \mathcal{L}_{\text{rot}} \tag{39}$$

After obtaining the prediction structural consistency term, we further introduce teacher-guided object-level geometric consistency learning. Based on the classification output of the teacher, we utilize the maximum class confidence $q_{t,i} = \max_c \delta(C_{t,i,c})$ and the corresponding predicted category $\ell_{t,i} = \arg\max_c \delta(C_{t,i,c})$ to construct the teacher pseudo-object set $\hat{Y}_t$. For candidate locations that satisfy the high-confidence condition, this set is generated through pre-screening $\text{TopK}_{\text{pre}}(\cdot)$, rotated non-maximum suppression $\text{NMS}_\theta(\cdot)$, and a final retention strategy $\text{TopK}(\cdot)$ (Eq. (40)); here, τ denotes the confidence threshold for the teacher pseudo-objects.

$$\hat{Y}_t = \text{TopK}\left(\text{NMS}_\theta\left(\text{TopK}_{\text{pre}}\left\{\left(\ell_{t,i}, \hat{b}_{t,i}, q_{t,i}\right) \mid q_{t,i} \geq \tau\right\}\right)\right) \tag{40}$$

here, $\hat{Y}_t$ is converted into a structured target tensor $\Phi(\cdot)$ using the target packing operator $\left(\tilde{L}, \tilde{B}, \tilde{Q}, \tilde{M}\right) = \Phi\left(\tilde{Y}_t\right)$. Subsequently, the teacher pseudo-targets are mapped into the student anchor space via the task assignment operator $\mathcal{A}_{\text{rot}}$ designed for rotation detection (Eq. (41)).

$$\left(\overline{B}, \overline{S}, \overline{M}\right) = \mathcal{A}_{\text{rot}}\left(\delta(C_s), \hat{B}_s, \tilde{Y}_t\right) \tag{41}$$

where, $\tilde{L}$, $\tilde{B}$, $\tilde{Q}$, and $\tilde{M}$ denote the class label tensor, rotated bounding box tensor, confidence tensor, and valid mask of the teacher pseudo-targets, respectively. To identify locations where the teacher is confident but the student has a weak response, we calculate the student's maximum class confidence as $q_{s,i} = \max_c \delta\left(C_{s,i,c}\right)$ and define the significant advantage as $g_i = \mathrm{ReLU}\left(q_{t,i} - q_{s,i} - m\right)$, where denotes the confidence margin. The student's predicted category is separately defined as $\ell_{s,i} = \arg\max_c \sigma(C_{s,i,c})$, and the category inconsistency indicator is calculated as $\chi_i = 1\left(\ell_{t,i} \neq \ell_{s,i}\right)$. Together with the maximum assigned pseudo-target score $s_i = \max_c \overline{S}_{i,c}$, these terms are used to calculate the difficulty weight ω_i in Eq. (42).

$$\omega_i = \mathrm{clip}\left(\left(1 + \rho\left(g_i + \chi_i\right)\right)\left(1 + s_i\right), 1, 1 + 2\rho\right) \tag{42}$$

here, ρ denotes the hard sample reinforcement coefficient, and $\mathrm{clip}\left(\cdot, 1, 1 + 2\rho\right)$ represents clipping the weights to a specified range. To mitigate the interference from low-value positive samples, we obtain the hard sample mask by utilizing the assigned positive sample mask and the vector $\omega \odot \overline{M}$, which is derived after applying hard weights to the positive samples (Eq. (43)).

$$\hat{M}_i = \overline{M}_i \cdot 1\left(i \in \mathrm{TopK}_{K_h}\left(\omega \odot \overline{M}\right)\right) \tag{43}$$

where, $\mathrm{TopK}_{K_h}(\cdot)$ represents the top K_h samples with the largest weights. After hard sample reinforcement, the corresponding classification distillation term $\mathcal{L}^g_{\mathbf{cls}}$ is obtained by combining the assigned pseudo-target scores $\overline{S}_{i,c}$ and the object-level anchor weights $\tilde{a}_i$ (Eq. (44)). Based on this, utilizing the assigned pseudo-target boxes $\overline{B}$, the target scores after hard reinforcement $\overline{S} \odot \omega$, and the rotated bounding box joint regression operator $\Psi_{rbox}(\cdot)$, the object-level rotated bounding box regression term $\mathcal{L}^g_{\mathrm{box}}$ and the distribution focal regression term $\mathcal{L}^g_{\mathrm{dfl}}$ are obtained (Eq. (45)). Meanwhile, based on the assigned pseudo-target angle $\overline{A}_i$ and the confidence $\tilde{Q}_{n,j}$ of the j-th teacher pseudo-target in the n-th sample, the object-level angular consistency term $\mathcal{L}^g_\theta$ and its overall intensity modulation coefficient are obtained (Eq. (46)).

$$\tilde{a}_i = \max_c \overline{S}_{i,c} \cdot \omega_i \cdot \hat{M}_i, \quad \mathcal{L}^g_{\mathrm{cls}} = \frac{\sum_i \tilde{a}_i \sum_c \mathrm{BCE}\left(C_{s,i,c}, \overline{S}_{i,c}\right)}{\sum_i \tilde{a}_i} \tag{44}$$

$$\left(\mathcal{L}^g_{\mathrm{box}}, \mathcal{L}^g_{\mathrm{dfl}}\right) = \Psi_{\mathrm{rbox}}\left(D_s, \hat{B}_s, \overline{B}, \overline{S} \odot \omega, \hat{M}\right) \tag{45}$$

$$\mathcal{L}^g_\theta = \frac{\sum\limits_{i:\hat{M}_i=1} \tilde{a}_i SmoothL1\left(A_{s,i}, \overline{A}_i\right)}{\sum\limits_{i:\hat{M}_i=1} \tilde{a}_i}, \quad \psi = \frac{1}{B}\sum_{n=1}^{B} \max_j \tilde{Q}_{n,j} \tag{46}$$

where, ψ represents the average confidence of the teacher pseudo-targets within the current batch (B), which is used to modulate the overall geometric distillation intensity. Based on the above, the teacher-guided object-level geometric consistency term $\mathcal{L}_g$ is obtained (Eq. (47)).

$$\mathcal{L}_g = \psi\left(\mathcal{L}^g_{\mathrm{cls}} + \mathcal{L}^g_{\mathrm{box}} + \mathcal{L}^g_{\mathrm{dfl}} + 0.5\mathcal{L}^g_\theta\right) \tag{47}$$

3 Experimental Details

3.1 Dataset

The proposed model is primarily evaluated on the Infrared Ship Database (IR-Ship) [26]. Supplementary validation is conducted on the Infrared Ship Detection Dataset (ISDD) [27] and the Singapore Maritime Dataset (SMD) [28]. ISDD is a pure infrared remote-sensing dataset, whereas SMD contains both visible-light and near-infrared imagery. IR-Ship follows the fixed partition adopted in this study, ISDD is divided into training, validation, and test subsets using an 8:1:1 ratio, and SMD uses the predefined training, validation, and test subsets adopted in our experiments.

IR-Ship Dataset: Publicly released by Yantai Raytron Technology Co., Ltd. through its Infrared Open Source Platform, this long-wave infrared maritime ship dataset contains a total of 8002 images of varying resolutions across seven ship target categories: liners (LR), bulk carriers (BC), warships (WS), sailboats (SB), canoes (CE), container ships (CS), and fishing ships (FS). This dataset covers diverse scenes, different time periods, and various imaging scales, reflecting practical challenges in complex infrared maritime monitoring, such as background thermal interference, scale variations, and differences in target morphology.

ISDD Dataset: A three-band fused infrared remote sensing ship dataset, comprising a single category, 1284 images, and 3061 ship instances. This dataset features a high-altitude top-down perspective, where the target scales are small and texture information is limited. It is well-suited for evaluating a model's detection capability for dim, small, and low-texture targets.

SMD Dataset: SMD is a multimodal maritime dataset containing onshore and onboard visible-light sequences as well as near-infrared (NIR) sequences acquired in Singapore waters. In this study, 6350 annotated frames are used across nine categories: Boat, Buoy, Ferry, Flying bird-plane, Kayak, Other, Sail boat, Speed boat, and Vessel-ship. "Flying bird-plane" denotes a combined category of flying birds and airplanes in the adopted annotation configuration. SMD is used for supplementary validation of the model under different maritime scenes and imaging modalities.

3.2 Experimental Environment and Evaluation Indicators

The configuration of the experimental environment is presented in Table 1. YOLOv11n is adopted as the baseline model, and all models share the same training settings: an input image size of 640 × 640, 220 training epochs, a batch size of 16, and 16 workers. In the base training phase, the Stochastic Gradient Descent optimizer is utilized, with the initial learning rate set to 0.01, the momentum coefficient to 0.937, and the weight decay coefficient to 0.0005. For the knowledge distillation experiments, an M-scale model with the corresponding structure is employed as the teacher model.

Table 1: Experimental configuration.

Experimental Environment	Details
Operating system	Ubuntu 20.04
GPU	RTX 3090 (24 GB)
CPU	20 vCPU AMD EPYC 7642 48-Core Processor
Language	Python 3.8.0
GPU calculate platform	CUDA 11.8
Deep learning framework	PyTorch 2.0.0

We perform quantitative evaluations from two perspectives, detection accuracy and computational efficiency, to assess model performance. Detection accuracy metrics encompass Precision (P), Recall (R), the F1-score, Average Precision (AP), and mean Average Precision (mAP). Computational efficiency is assessed by inference speed (Frames Per Second), parameter count (Params), and computational load (Giga Floating-point Operations, GFLOPs), which together characterize deployment overhead and real-time capability. P, R, and $F1$ are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{48}$$

$$R = \frac{TP}{TP + FN} \tag{49}$$

$$F1 = \frac{2PR}{P + R} \tag{50}$$

AP represents the average precision for an individual class, which corresponds to the geometrical area under the P-R curve (Eq. (51)), mAP is defined as the mean of AP across all classes (Eq. (52)). The primary detection metrics used throughout the experiments are mAP50 and mAP50-95. Specifically, mAP50 evaluates detection performance at an IoU threshold of 0.50, whereas mAP50-95 averages the AP values over IoU thresholds from 0.50 to 0.95 with a step size of 0.05, thereby providing a more comprehensive assessment of localization accuracy.

$$AP = \int_0^1 p(r)dr \tag{51}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{52}$$

Furthermore, we adopt the standard Common Objects in Context (COCO) evaluation protocol and categorize targets by pixel area S into small ($S < 32^2$), medium ($32^2 \le S < 96^2$), and large ($S \ge 96^2$) classes. We calculate the precision metrics AP^s, AP^m, and AP^l and the recall metrics AR^s, AR^m, and AR^l averaged over IoU thresholds ranging from 0.50 to 0.95, in order to assess the model's balance between detection and localization across various scales. Minor numeric differences may arise in COCO indices because of bounding-box representation conversions, coordinate quantization, and floating-point precision; these do not affect the overall performance comparison.

4 Experimental Analysis

4.1 Ablation Experiment

Detailed ablation experiments were conducted to quantitatively evaluate the individual and combined contributions of WFEM, DMS-FFM, TCUDH, and TDSKD, with the results summarized in Table 2. The baseline YOLOv11n achieves mAP50 and mAP50-95 values of 0.906 and 0.636, respectively, with 2.46M parameters, 6.3 GFLOPs, and an inference speed of 1107.7 FPS. When introduced individually, WFEM, DMS-FFM, and TCUDH improve mAP50-95 to 0.648, 0.653, and 0.766, corresponding to absolute gains of 0.012, 0.017, and 0.130 over the baseline, respectively. WFEM suppresses sea-wave thermal clutter and shoreline heat-source interference through frequency-domain filtering and wavelet-based edge preservation, whereas DMS-FFM improves cross-scale representation and feature interaction for densely distributed ships. Among the three architectural components, TCUDH provides the largest individual improvement, particularly in high-quality localization under strict IoU thresholds. From a computational perspective, introducing WFEM increases the model complexity from 2.46M to 3.06M parameters and from 6.3 to 12.8 GFLOPs, corresponding to increments of 0.60M parameters and 6.5 GFLOPs, respectively. This additional cost mainly arises from its frequency-domain processing and multi-branch wavelet feature enhancement. DMS-FFM increases the model complexity to 3.42M parameters and 13.7 GFLOPs, corresponding to increments of 0.96M parameters and 7.4 GFLOPs, mainly because of its cross-scale feature alignment and dynamic fusion operations. The resulting inference speeds of the WFEM and DMS-FFM configurations are 404.7 and 423.4 FPS, respectively. In contrast, TCUDH replaces the original detection head with a compact shared prediction structure rather than introducing an additional prediction branch. It therefore reduces the model complexity by 0.29M parameters and 0.2 GFLOPs, resulting in 2.17M parameters, 6.1 GFLOPs, and an inference speed

of 1094.4 FPS. The pairwise combinations WFEM+DMS-FFM, WFEM+TCUDH, and DMS-FFM+TCUDH achieve mAP50-95 values of 0.658, 0.775, and 0.777, respectively, corresponding to improvements of 0.022, 0.139, and 0.141 over the baseline. Integrating all three architectural components produces mAP50 and mAP50-95 values of 0.968 and 0.790, respectively, with 3.71M parameters, 22.2 GFLOPs, and an inference speed of 235.3 FPS. Because these components operate at different feature levels and TCUDH replaces the original detection head, the complexity of the complete architecture was measured directly rather than estimated by simply summing the isolated module increments. On the same complete inference architecture, TDSKD further increases mAP50 and mAP50-95 from 0.968 and 0.790 to 0.971 and 0.804, corresponding to additional gains of 0.003 and 0.014, respectively. Because TDSKD is applied only during training, it does not change the inference-stage parameter count or GFLOPs. Consequently, the final model retains 3.71M parameters and 22.2 GFLOPs while achieving an overall mAP50-95 improvement of 0.168 over the baseline. As shown in Fig. 7, the final model produces PR and F1-score curves closest to those of the teacher model and maintains higher precision in the high-recall range, together with a higher peak F1-score and a wider stable range across confidence thresholds. These results further demonstrate that the proposed components provide complementary improvements in detection accuracy, localization quality, and perception stability under complex infrared maritime conditions.

Table 2: Detailed ablation and computational-complexity analysis of the individual and combined components.

WFEM	DMS-FFM	TCUDH	TDSKD	mAP50	mAP50-95	ΔmAP50-95	P	R	Params	GFLOPs	FPS
Baseline				0.906	0.636	–	0.906	0.859	2.46M	6.3	1107.7
√				0.919	0.648	+0.012	0.920	0.868	3.06M	12.8	404.7
	√			0.937	0.653	+0.017	0.920	0.888	3.42M	13.7	423.4
		√		0.950	0.766	+0.130	0.914	0.897	2.17M	6.1	1094.4
√	√			0.941	0.658	+0.022	0.911	0.903	4.05M	21.4	242.2
√		√		0.954	0.775	+0.139	0.933	0.904	2.77M	12.6	400.2
	√	√		0.960	0.777	+0.141	0.926	0.926	3.09M	14.5	402.3
√	√	√		0.968	0.790	+0.154	0.944	0.936	3.71M	22.2	235.3
√	√	√	√	0.971	0.804	+0.168	0.945	0.944	3.71M	22.2	234.7

Note: ΔmAP50-95 denotes the absolute improvement relative to the baseline.

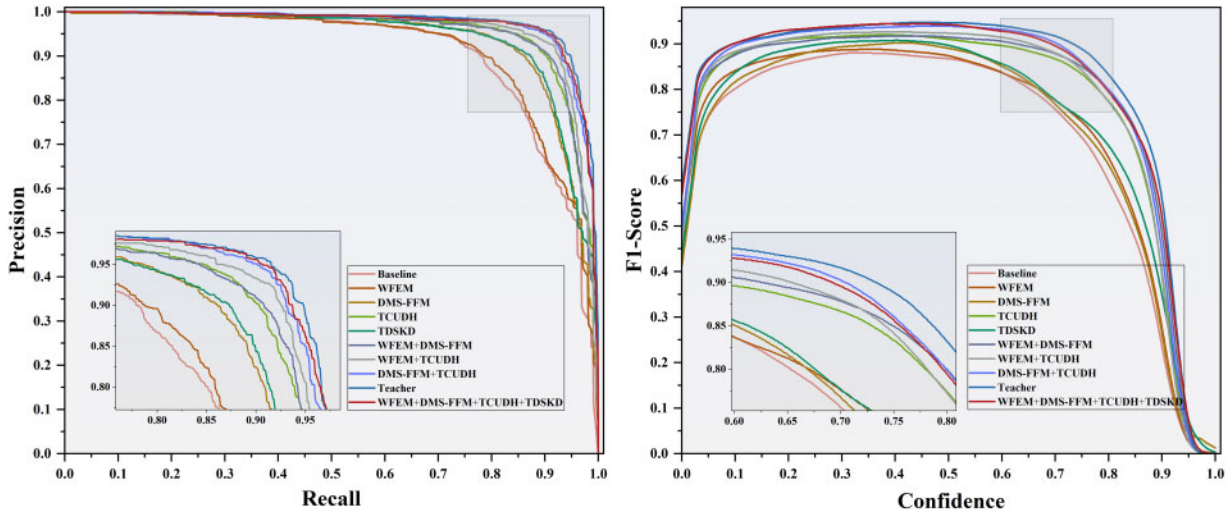

Figure 7: Comparison of PR and F1 curves.

A comparative experiment was conducted to examine whether simpler knowledge-distillation methods could achieve gains comparable to those of TDSKD. Specifically, TDSKD was compared with L2-based logit matching and Channel-Wise Distillation (CWD). The same teacher model, student model, dataset split, initialization, and training settings were used in all distillation experiments, with only the distillation strategy changed. As shown in Table 3, both L2-based logit matching and CWD improve mAP50-95, confirming that teacher supervision is beneficial. However, TDSKD achieves the largest improvement, increasing mAP50-95 from 0.790 to 0.804 and providing a more balanced precision–recall performance. The clearer advantage in mAP50-95 indicates that the prediction-structure and object-level geometric constraints of TDSKD are more effective for high-quality localization than conventional response- or feature-level matching.

Table 3: Comparison with conventional knowledge-distillation strategies.

Distillation Strategy	Distillation Level	P	R	F1	mAP50	mAP50-95	ΔmAP50-95
No distillation	–	0.944	0.936	0.940	0.968	0.790	–
L2-based logit matching	Response level	0.948	0.934	0.941	0.968	0.796	+0.006
CWD	Feature level	0.940	0.942	0.941	0.969	0.797	+0.007
TDSKD	Dual-level structured	0.945	0.944	0.944	0.971	0.804	+0.014

Statistical uncertainty in the reported detection results was further assessed through a nonparametric image-level bootstrap analysis on the validation set. The 632 validation images were sampled with replacement to construct 1000 bootstrap replicates, with each replicate containing the same number of images as the original validation set. The evaluation metrics were recalculated for every replicate, and the 2.5th and 97.5th percentiles of the resulting distributions were used as the lower and upper bounds of the 95% confidence intervals, respectively. As summarized in Table 4, the 95% confidence intervals of the baseline are 0.890–0.921 for mAP50 and 0.619–0.656 for mAP50-95. After integrating WFEM, DMS-FFM, and TCUDH, the corresponding intervals become 0.960–0.976 and 0.780–0.809. For the final model with TDSKD, the intervals are 0.961–0.977 and 0.787–0.814, respectively. The point estimates reported in Table 2 fall within their corresponding confidence intervals. These results quantify the uncertainty associated with validation-set sampling and indicate that the reported performance remains stable across the resampled validation sets.

Table 4: Image-level bootstrap 95% confidence intervals for the representative configurations.

Configuration	mAP50	mAP50 95%CI	mAP50-95	mAP50-95 95%CI
Baseline	0.906	0.890–0.921	0.636	0.619–0.656
WFEM + DMS-FFM + TCUDH	0.968	0.960–0.976	0.790	0.780–0.809
WFEM + DMS-FFM + TCUDH + TDSKD	0.971	0.961–0.977	0.804	0.787–0.814

Note: The confidence intervals were obtained from 1000 nonparametric image-level bootstrap replicates of the 632-image validation set.

To qualitatively evaluate the effects of the proposed components under complex infrared maritime conditions, Fig. 8 compares missed and false detections across representative ablation configurations, where green boxes indicate correct detections, red boxes indicate missed detections, and blue boxes indicate false detections. While Table 2 quantitatively reports both the individual and combined configurations, Fig. 8 focuses on configurations in which WFEM, DMS-FFM, and TCUDH are individually incorporated into the baseline, together with the full model, to maintain visual clarity. The baseline is susceptible to sea-wave

thermal clutter and complex shoreline heat sources, resulting in background false positives (e.g., Image 4) and missed detections of distant, dim, and small targets (e.g., Image 3). When WFEM is individually incorporated into the baseline, periodic thermal clutter is suppressed, the local high-frequency structures of dim targets are enhanced, and the shoreline false alarms in Image 4 are reduced. When DMS-FFM is individually incorporated into the baseline, it strengthens multi-scale feature alignment and cross-scale interaction, alleviating the feature aliasing caused by overlapping thermal responses from adjacent ships and reducing missed detections in dense scenes (e.g., Image 1). By integrating the complementary effects of the three architectural components and TDSKD, the full model suppresses false responses caused by sea-surface and shoreline thermal interference (e.g., Images 2 and 5), improves the detection of low-contrast small targets (e.g., Image 3), and more reliably separates adjacent ships under dense occlusion (e.g., Image 1).

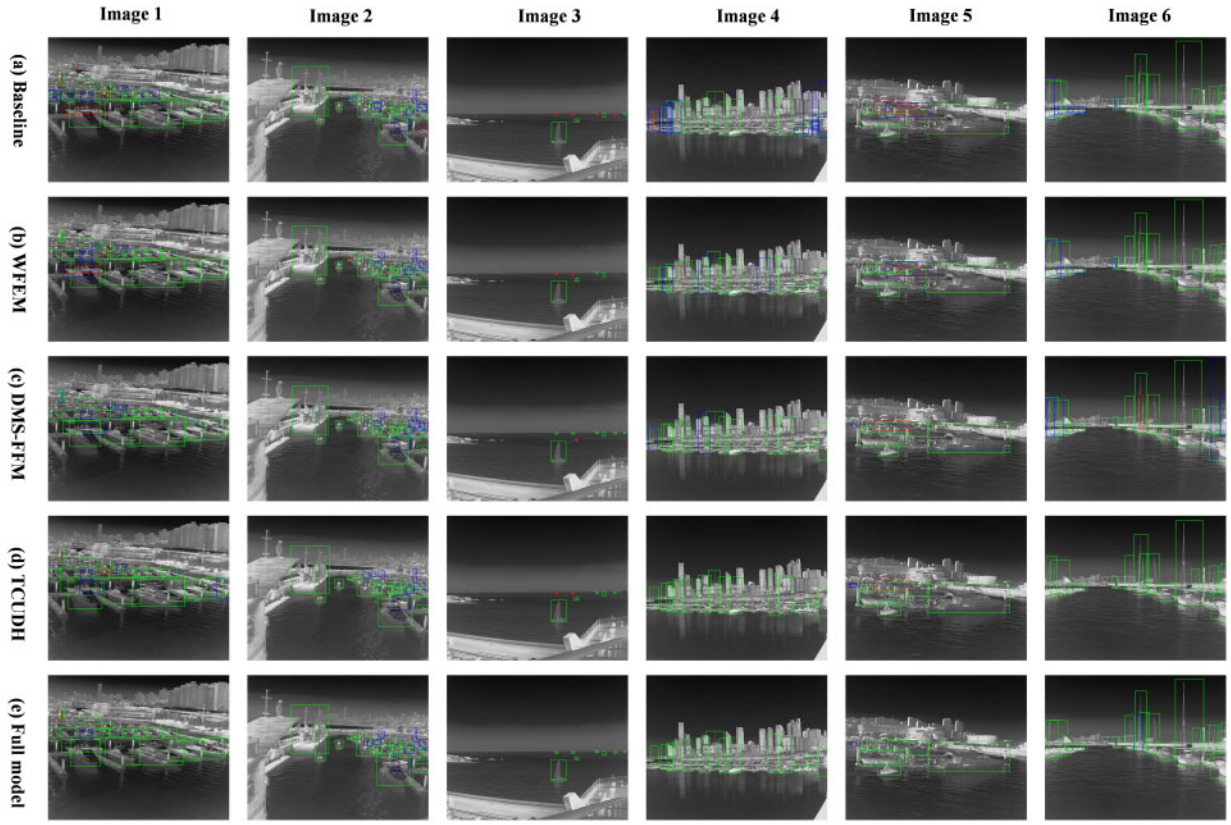

Figure 8: Visual comparison of missed and false detections for individual component ablations and the full model.

4.2 Comparative Experiment

Comparative experiments were conducted in complex infrared maritime scenarios to evaluate whether the proposed model can maintain reliable visual perception from three aspects: feature pyramid structure, detection head design, and overall detection architecture. Furthermore, visualization results were used to analyze the model's discriminative mechanism and scene adaptability.

We compare it with various types of feature pyramid structures to verify the effectiveness of DMSFPN, as shown in Table 5. BiFPN employs learnable weights to adaptively integrate features through bidirectional cross-scale paths [29]. AFPN progressively fuses adjacent low-level features and subsequently incorporates higher-level features, thereby reducing the semantic gaps between non-adjacent feature levels [30]. Slim-Neck introduces GSConv into the neck to reduce computational cost while maintaining feature-fusion

capability [31]. ASFPN integrates Scale Sequence Feature Fusion, a Triple Feature Encoder, and a Channel and Position Attention Mechanism to strengthen cross-scale semantic interaction, enhance target-relevant spatial and channel responses, and improve the representation of multi-scale and weak targets under complex backgrounds; it is therefore included as an attention-enhanced multi-scale fusion baseline [32]. GFPN strengthens information interaction across different feature levels through generalized cross-scale fusion paths [33], whereas CGFPN employs context-guided feature recalibration and fusion to enhance the complementary representation of multi-scale features [34]. As can be seen from the Fig. 9, DMSFPN exhibits a superior overall trend in both the PR and F1-score curves, maintaining high precision and stability even in the higher recall and high-confidence ranges. Meanwhile, it achieves scores of 0.941, 0.724, and 0.658 on mAP50, mAP75, and mAP50-95, respectively. Its overall performance is superior to that of the other compared structures, demonstrating that the proposed method has advantages in complex thermal background suppression, multi-scale target representation, and high-quality localization.

Table 5: Comparison of different feature pyramids network.

Model	mAP50	mAP75	mAP50-95	F1	AP^s	AP^m	AP^l	AR^s	AR^m	AR^l
BiFPN	0.904	0.684	0.633	0.868	0.414	0.647	0.732	0.520	0.703	0.791
FDPN	0.906	0.689	0.633	0.878	0.427	0.651	0.717	0.491	0.708	0.779
Hyper	0.908	0.700	0.642	0.875	0.415	0.657	0.751	0.493	0.713	0.799
GFPN	0.911	0.691	0.645	0.879	0.431	0.657	0.749	0.521	0.714	0.787
MAFPN	0.912	0.697	0.645	0.883	0.418	0.660	0.748	0.516	0.711	0.796
CSFCN	0.914	0.689	0.641	0.881	0.417	0.656	0.750	0.524	0.708	0.793
MSGA	0.917	0.702	0.646	0.884	0.434	0.657	0.746	0.518	0.715	0.790
RCFPN	0.919	0.701	0.649	0.886	0.441	0.664	0.751	0.532	0.724	0.791
RCSOSA	0.924	0.714	0.655	0.891	0.450	0.659	0.738	0.528	0.722	0.786
CGFPN	0.929	0.703	0.649	0.893	0.449	0.657	0.738	0.572	0.724	0.796
AFPN2345	0.930	0.704	0.647	0.895	0.452	0.663	0.712	0.576	0.719	0.780
Slimneck2345	0.931	0.704	0.651	0.897	0.463	0.667	0.754	0.569	0.719	0.795
ASFPN	0.933	0.711	0.644	0.90	0.458	0.662	0.738	0.563	0.720	0.785
DMSFPN	0.941	0.724	0.658	0.911	0.474	0.672	0.735	0.578	0.728	0.806

We conducted comparative experiments in complex infrared maritime scenarios to verify the applicability of TCUDH, as shown in Table 6. EfficientHead, introduced in YOLOv6, employs lightweight decoupled classification and regression branches to improve prediction efficiency [20]. The YOLOv9 decoupled detection head (YOLOv9-DDetect) similarly separates classification and bounding-box regression and is included as a representative general-purpose decoupled head [35]. DyHead jointly applies scale-aware, spatial-aware, and task-aware attention to adaptively recalibrate detection features [21]. The SEAM-based head derived from YOLO-FaceV2, which was originally developed for detecting small and occluded faces, employs separated and enhancement attention to strengthen feature responses under partial occlusion; it is therefore included as a cross-domain attention-enhanced head baseline [22]. In addition, the OBB head predicts oriented bounding boxes to better accommodate elongated targets with arbitrary orientations. As shown in the Fig. 10, TCUDH is overall closer to the optimal envelope in both the PR and F1-score curves, maintaining higher precision and a wider stable high-F1 plateau even in the high recall and high-confidence ranges. Furthermore, as evidenced by the multi-dimensional metrics, TCUDH achieves 0.950, 0.859, and 0.766 on mAP50, mAP75, and mAP50-95, respectively, outperforming general detection heads. The quantitative

results indicate that TCUDH not only improves localization quality for slender targets but also delivers more balanced overall performance in dense small-target detection and cross-scale target representation.

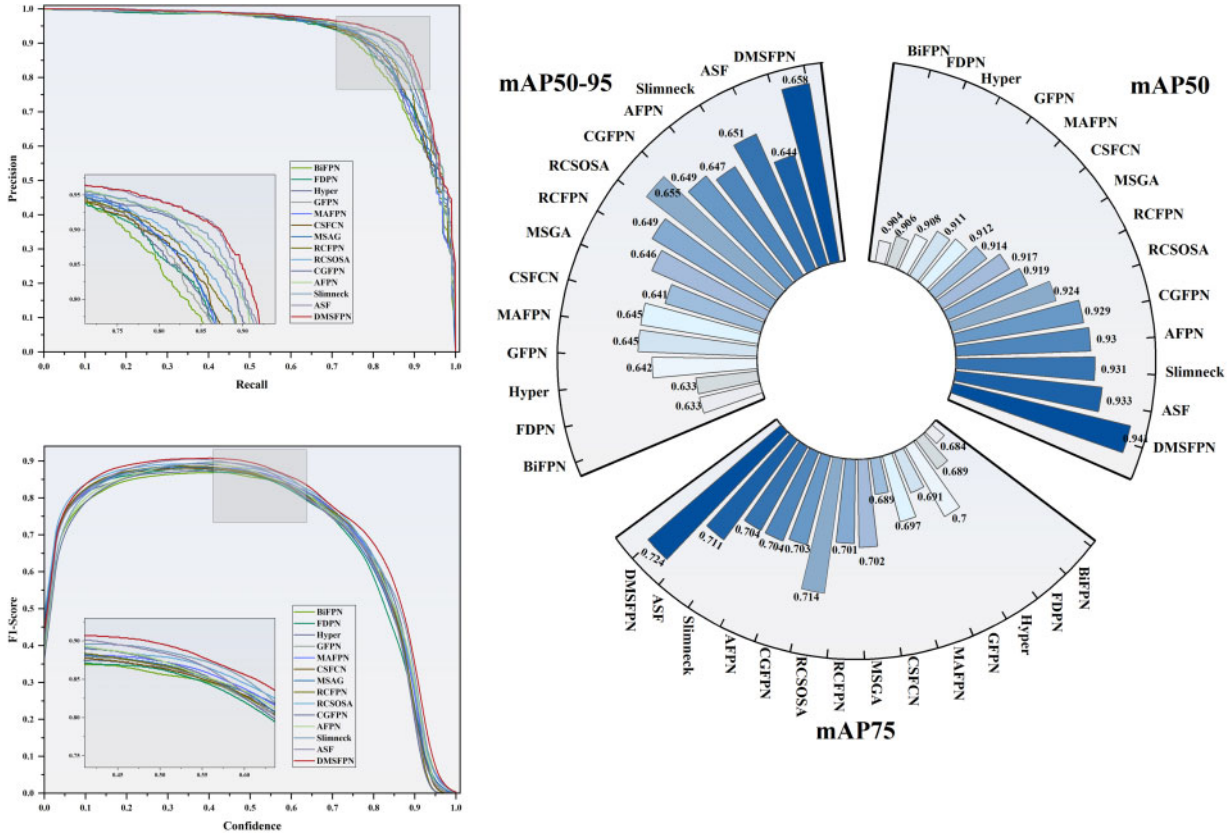

Figure 9: Comparison of feature pyramids.

Table 6: Comparison of different detection heads.

Model	mAP50	mAP75	mAP50-95	F1	AP^s	AP^m	AP^l	AR^s	AR^m	AR^l
EfficientHead	0.913	0.685	0.613	0.881	0.418	0.655	0.739	0.522	0.710	0.794
LQEHead	0.904	0.699	0.636	0.878	0.429	0.652	0.738	0.512	0.712	0.796
LSCSBD	0.913	0.695	0.641	0.887	0.439	0.652	0.730	0.509	0.707	0.808
RSCD	0.908	0.689	0.634	0.877	0.426	0.649	0.731	0.522	0.710	0.797
DyHead	0.913	0.698	0.643	0.880	0.435	0.659	0.726	0.502	0.713	0.780
IDetect	0.900	0.693	0.636	0.866	0.421	0.650	0.743	0.517	0.706	0.791
DDetect	0.911	0.706	0.645	0.879	0.434	0.657	0.735	0.505	0.711	0.790
Detect_ASFF	0.909	0.701	0.643	0.883	0.405	0.654	0.741	0.503	0.708	0.797
WorldDetect	0.917	0.710	0.651	0.886	0.447	0.663	0.743	0.565	0.709	0.783
TADDH	0.925	0.719	0.652	0.889	0.448	0.665	0.753	0.521	0.717	0.802
Atthead	0.907	0.704	0.639	0.902	0.426	0.652	0.746	0.515	0.709	0.784
SEAMHead	0.911	0.697	0.638	0.878	0.403	0.654	0.745	0.500	0.712	0.786
Nmsfree	0.915	0.688	0.632	0.881	0.439	0.643	0.710	0.561	0.711	0.780
OBB	0.947	0.850	0.762	0.852	0.486	0.721	0.738	0.641	0.772	0.809
TCUDH	0.950	0.859	0.766	0.920	0.504	0.743	0.756	0.595	0.791	0.815

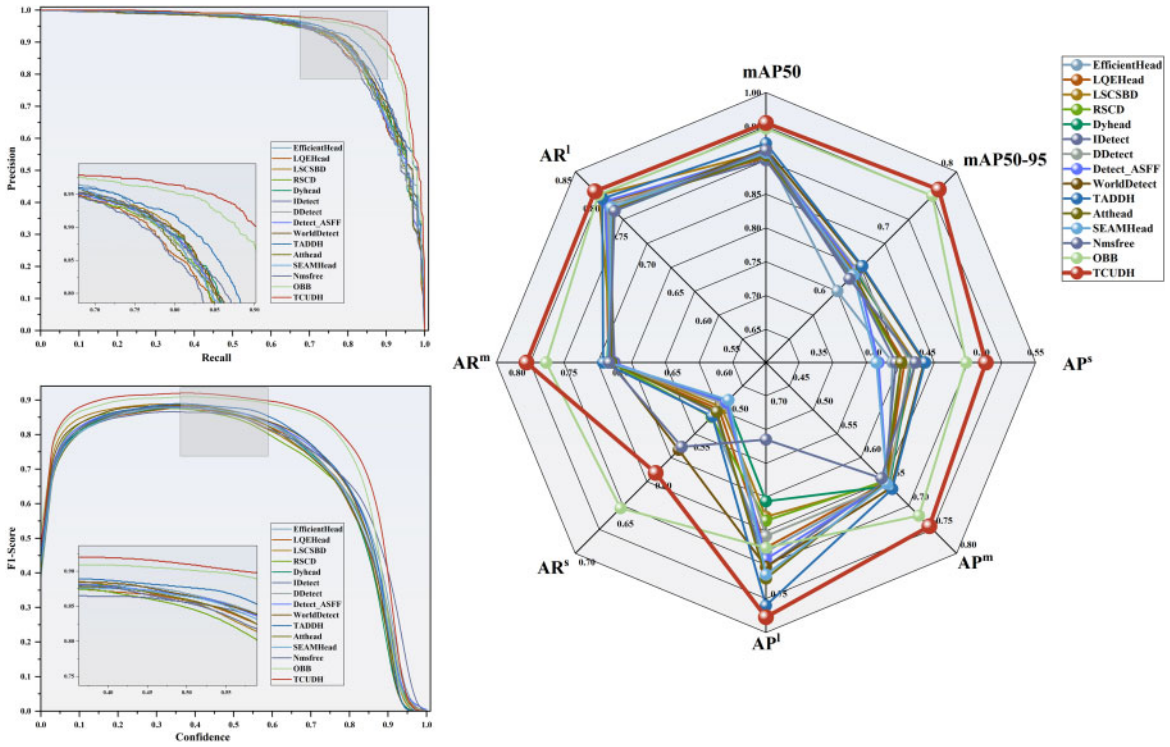

Figure 10: Comparison of detection heads.

We further provide qualitative visualization results on the SeaShips [36] and HRSID [37] datasets to illustrate the extensibility of TCUDH to segmentation-related prediction (Fig. 11). Compared to the baseline, TCUDH achieves more complete coverage for large-scale ships in SeaShips (e.g., Images 1, 4, and 5), and exhibits more concentrated responses to small and distant targets (e.g., Images 2, 3, and 6); For the HRSID dataset, TCUDH provides more complete representation of nearshore ship regions (e.g., Images 1 and 2), while yielding clearer contour delineation for slender, small, and densely distributed targets (e.g., Images 3, 4, 5, and 6). These qualitative results suggest that the task-conditioned design of TCUDH is compatible with extension to segmentation-related prediction.

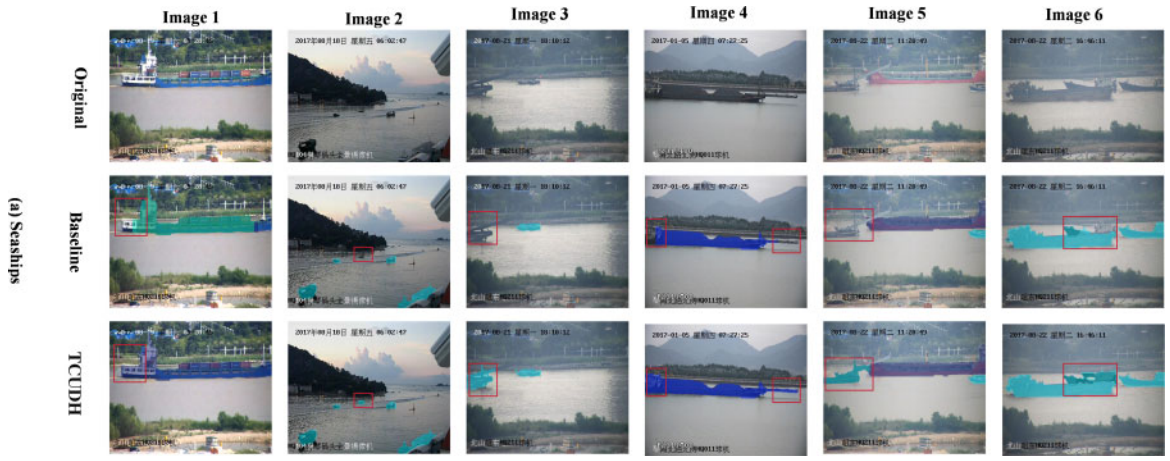

Figure 11: (Continued)

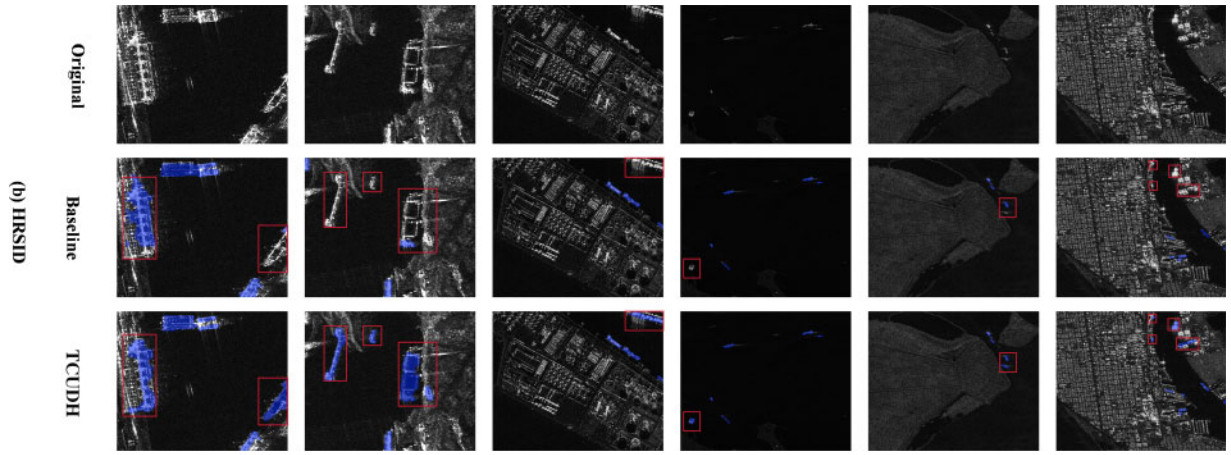

Figure 11: Visualization of the segmentation of TCUDH.

The performance and complexity of the proposed model were compared with those of various YOLO-series models and their representative variants, as presented in Table 7. YOLOv5-YOLO26 exhibit limited differences in detection performance at low thresholds, but their localization capability is constrained at high IoU thresholds, reflecting an inadequacy in high-quality localization for distant, small-scale, and slender targets. A principal reason is that their feature extraction and prediction structures do not explicitly address sea-wave thermal clutter, shoreline heat-source interference, weak target boundaries, or the adjacent-target aliasing frequently encountered in infrared maritime scenes. Among the general-purpose enhanced detectors, Gold-YOLO employs a gather-and-distribute mechanism to strengthen multi-scale feature interaction, while DAMO-YOLO adopts an Efficient-RepGFPN neck to improve cross-level feature aggregation. Mamba-YOLO introduces an SSM-based ODMamba backbone together with RG blocks to model long-range dependencies while reinforcing local feature representation. In the present experiments, these mechanisms provide moderate improvements in mAP50-95 over conventional lightweight YOLO baselines. Nevertheless, their additional feature aggregation or global modeling operations increase model complexity, while their feature-enhancement mechanisms are not specifically designed to distinguish weak infrared ship responses from structured thermal background interference. Infrared-specific detectors provide stronger evidence of the benefit of task-oriented feature modeling. GT-YOLO combines attention-based feature fusion, SPD-Conv, and Soft-NMS to preserve small-target information and reduce missed detections under dense occlusion, explaining its relatively high mAP50-95 of 0.745. YOLO-IRS introduces self-attention and KAN-based nonlinear feature modeling to improve ship representation in complex infrared backgrounds and obtains an mAP50-95 of 0.663. However, these methods mainly focus on spatial-domain feature enhancement or post-processing and do not jointly address thermal-clutter suppression, cross-scale alignment, weak-boundary localization, and teacher-guided geometric consistency. CAA-YOLO and EGISD-YOLO achieve higher mAP50-95 values of 0.828 and 0.841, respectively. CAA-YOLO introduces a high-resolution P2 feature layer, long-range contextual fusion, and combined attention to preserve shallow spatial details while suppressing noise interference, whereas EGISD-YOLO employs Dense-CSP feature reuse, contextual channel attention, edge-guided fusion, and an additional small-target prediction head to strengthen weak-boundary and small-target localization. These specialized designs explain their stronger performance under strict IoU thresholds. However, their accuracy improvements are accompanied by higher resource requirements. CAA-YOLO contains 98.3M parameters, requires 131.9 GFLOPs, and operates at 42 FPS, while EGISD-YOLO operates at 83 FPS with a reported model size of 16.6M, making them less favorable for resource-constrained edge deployment. In comparison, the proposed model achieves mAP50

and mAP50-95 values of 0.971 and 0.804, respectively, with 3.71M parameters, 22.2 GFLOPs, and an inference speed of 234.7 FPS. Although its mAP50-95 is lower than those of CAA-YOLO and EGISD-YOLO, it provides a more favorable balance among detection accuracy, localization quality, model complexity, and inference efficiency. As shown in Fig. 12, the proposed model also reaches convergence earlier and exhibits greater stability in the training curves, outperforming most comparison models in the overall F1-score and PR curves. These results indicate that the proposed method maintains strong fine-grained localization and real-time detection capabilities under reduced resource overhead, making it suitable for edge-side infrared maritime monitoring.

Table 7: Comparison of different YOLO detection algorithms.

Model	mAP50	mAP50-95	P	R	F1	Param	GFLOPs	FPS	Size
YOLOv5	0.909	0.635	0.910	0.860	0.884	2.39M	7.1	1204	5.0M
YOLOv6	0.897	0.635	0.907	0.845	0.875	4.04M	11.8	1159.4	8.3M
YOLOv8	0.910	0.644	0.906	0.861	0.883	2.87M	8.1	1130.6	6.0M
YOLOv9	0.910	0.640	0.909	0.849	0.878	1.88M	7.6	848.6	4.4M
YOLOv10	0.916	0.633	0.890	0.860	0.875	2.16M	6.5	1060.7	5.5M
YOLOv11	0.906	0.636	0.906	0.859	0.882	2.46M	6.3	1107.7	5.2M
YOLOv12 [38]	0.909	0.636	0.904	0.858	0.880	2.51M	5.8	634.6	5.2M
YOLOv13 [39]	0.911	0.639	0.924	0.849	0.885	2.33M	6.1	494.4	5.2M
YOLO26 [40]	0.920	0.645	0.909	0.856	0.882	2.27M	5.2	1129.6	5.2M
Gold-YOLO [41]	0.917	0.648	0.916	0.864	0.889	5.70M	10.3	502.2	11.8M
EGISD-YOLO [42]	0.937	0.841	0.963	0.912	0.937	–	–	83	16.6M
Mamba-YOLO [43]	0.924	0.655	0.924	0.868	0.895	5.71M	13.6	125.9	12.1M
DAMO-YOLO [33]	0.915	0.647	0.912	0.868	0.889	5.49M	11.2	728	11.6M
CAA-YOLO [44]	0.945	0.828	–	–	–	98.3M	131.9	42	–
FBRT-YOLO [45]	0.928	0.659	0.919	0.867	0.892	2.57M	19.8	434.8	5.4M
R_YOLO [46]	0.958	0.686	0.934	0.927	0.930	17.5M	56.8	128	–
YOLO-IRS [47]	0.928	0.663	0.934	–	–	3.96M	11.6	–	–
MAF-YOLO [48]	0.919	0.637	0.897	0.862	0.879	2.31M	8.5	416.4	6.4M
GT-YOLO [49]	0.954	0.745	–	–	–	8.7M	34.4	152	–
SOD-YOLO [50]	0.933	0.646	0.909	0.887	0.898	2.63M	10.8	547.8	5.0M
Our	0.971	0.804	0.945	0.944	0.944	3.71M	22.2	234.7	9.7M

To provide a more intuitive analysis of the differences in the models' focus regions under complex infrared scenarios, we employ the Grad-CAM method to present a visual comparison of heatmaps across different models (Fig. 13). It can be observed that the Baseline, YOLO12, YOLO13, and SOD-YOLO are susceptible to interference from shore-based high-thermal-radiation buildings (e.g., Image 4), sea surface thermal reflections (e.g., Image 2), or background clutter, which leads to dispersed heatmaps, misdirected focus on non-ship regions, and targets being overwhelmed by the thermal background. In contrast, the heatmaps of our model exhibit a higher degree of semantic focus. Whether in dense mooring scenarios (e.g., Image 1), for distant infrared dim and small targets (e.g., Image 3), or against complex shoreline thermal backgrounds, the highlighted regions cover the main bodies of the ships while suppressing background noise responses, demonstrating that the proposed model possesses more robust discriminative capabilities and superior target representation quality under complex infrared thermal interference conditions.

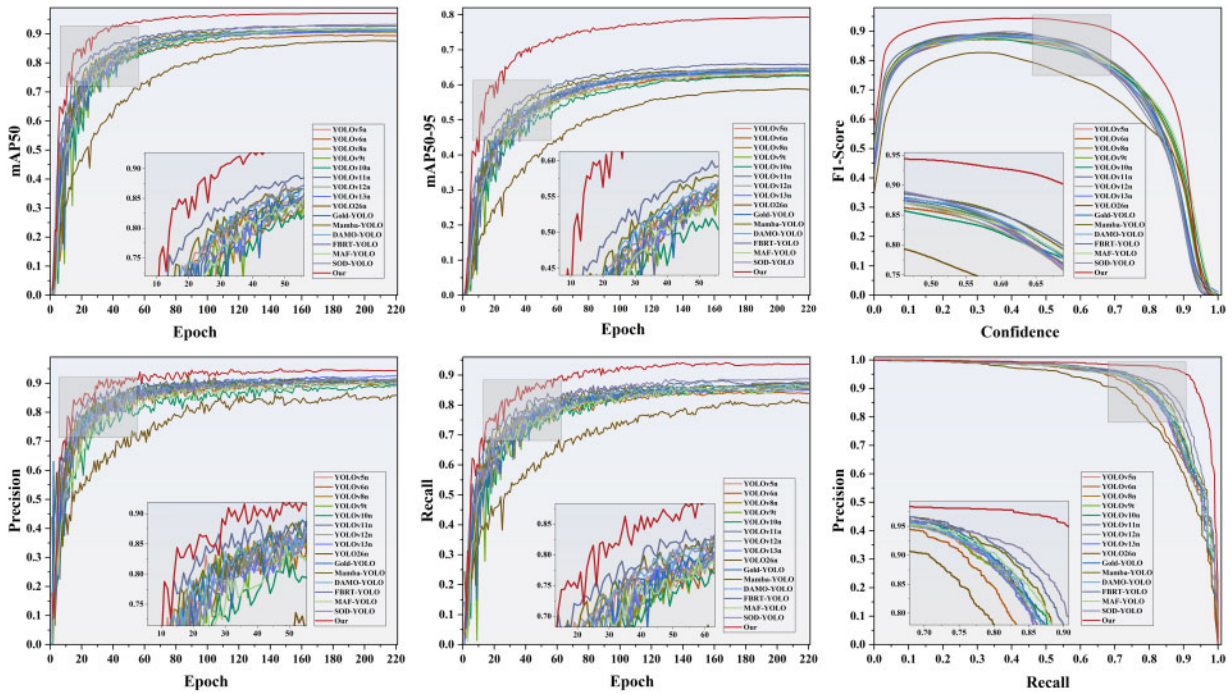

Figure 12: Comparison of different YOLO detection algorithms.

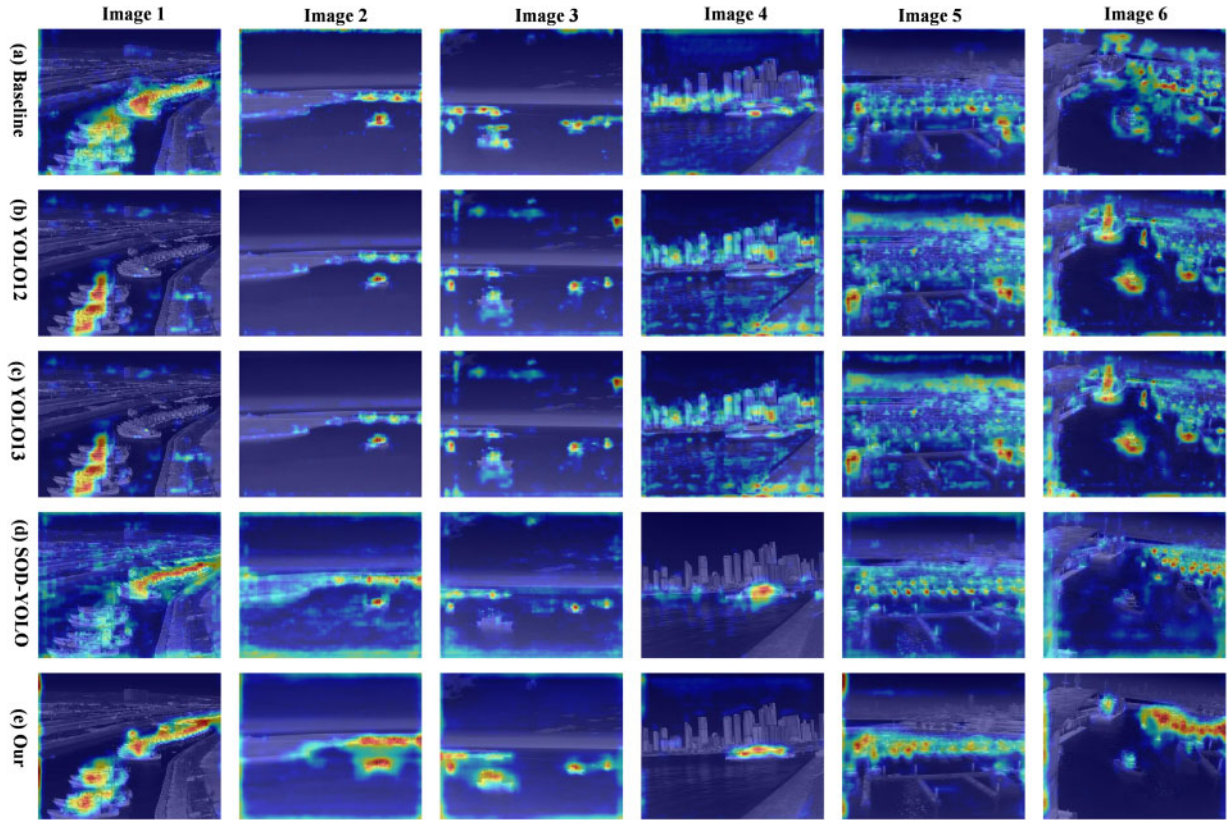

Figure 13: Visualization of heat map.

4.3 Comparison of Detection Paradigms and Challenging-Case Analysis

We evaluated the overall competitiveness of the proposed model against different categories of object detection algorithms, including one-stage methods (e.g., TOOD), two-stage methods (e.g., Cascade-RCNN), and Transformer-based methods (e.g., Deformable-DETR), with results reported in Table 8. Among the one-stage detectors, ATSS, PAA, GFL, VFNet, and TOOD generally outperform RetinaNet and FCOS in mAP50-95. Specifically, PAA adaptively separates positive and negative samples by fitting a probabilistic model to the joint classification and localization losses and additionally predicts IoU-based localization quality. GFL jointly represents classification confidence and localization quality and models bounding-box coordinates as flexible probability distributions. Together with the adaptive sample-assignment, IoU-aware scoring, and task-aligned learning mechanisms used by the other enhanced one-stage detectors, these designs reduce the inconsistency between classification confidence and bounding-box quality. In particular, TOOD achieves an mAP50-95 of 0.621 through task-aligned prediction. However, these methods rely primarily on general-purpose backbones and feature pyramids and do not explicitly suppress structured infrared thermal clutter or preserve the weak boundaries of small ship targets. Their model sizes and computational costs also remain relatively high; for example, TOOD requires 32.03M parameters and 169.98 GFLOPs. Among the two-stage detectors, Mask-RCNN extends Faster-RCNN by adding a parallel mask-prediction branch while retaining region-based classification and bounding-box regression. Because Table 8 evaluates object detection rather than instance segmentation, only its bounding-box detection output is included in the comparison. Cascade-RCNN employs a sequence of detection heads trained with progressively increasing IoU thresholds to iteratively refine candidate boxes. In the present experiments, Cascade-RCNN achieves the highest mAP50-95 of 0.644 among the competing two-stage detectors. Its advantage over Faster-RCNN and Mask-RCNN can therefore be attributed to the progressive refinement of candidate boxes under stricter IoU criteria. However, repeated region proposal processing and multi-stage box refinement increase its complexity to 69.29M parameters and 208.89 GFLOPs. Moreover, distant and low-contrast infrared ships may be insufficiently represented during the initial proposal and region-feature extraction stages, limiting the subsequent refinement of small and weak targets. Transformer-based detectors benefit from global-context modeling and query-based prediction, which contributes to relatively strong target recognition and large-target representation. For example, DINO achieves an mAP50 of 0.929 and an AP^{l} of 0.756. Nevertheless, its mAP50-95 remains at 0.609, indicating that strong global representation does not necessarily translate into sufficiently accurate localization for weak-boundary and densely distributed infrared ships. Deformable DETR restricts cross-attention to a sparse set of sampling points around multi-scale reference locations, thereby alleviating the dense-attention and convergence limitations of the original DETR. Conditional DETR introduces conditional spatial queries to improve object localization, whereas DAB-DETR formulates decoder queries as dynamic anchor boxes that are iteratively updated across decoder layers. DDQ-DETR further improves end-to-end detection by selecting distinct queries from dense candidates. Despite these attention and query-design improvements, the compared Transformer-based detectors are not specifically designed to distinguish weak infrared ship responses from sea-surface and shoreline thermal interference, while their backbone and decoding structures still introduce considerable computational overhead. DINO, for example, contains 47.55M parameters and requires 238.59 GFLOPs.

Table 8: Comparison of different detection paradigms.

Model	mAP50	mAP50-95	F1	AP^s	AP^m	AP^l	AR^s	AR^m	AR^l	Param	GFLOPs
RetinaNet [51]	0.882	0.561	0.839	0.331	0.577	0.639	0.473	0.661	0.754	36.45M	178.18
ATSS [52]	0.900	0.595	0.856	0.391	0.604	0.690	0.492	0.679	0.780	51.12M	239.62
PAA [53]	0.905	0.603	0.869	0.410	0.618	0.750	0.579	0.747	0.857	32.13M	174.13
GFL [54]	0.897	0.605	0.870	0.387	0.606	0.717	0.477	0.676	0.794	32.27M	176.13
FCOS [55]	0.893	0.579	0.864	0.374	0.589	0.691	0.469	0.668	0.776	32.13M	169.99
TOOD [56]	0.913	0.621	0.879	0.396	0.611	0.736	0.497	0.683	0.799	32.03M	169.98
DINO [57]	0.929	0.609	0.899	0.416	0.618	0.756	0.561	0.713	0.835	47.55M	238.59
Faster-RCNN	0.908	0.616	0.866	0.376	0.596	0.689	0.460	0.656	0.766	41.38M	180.22
Mask-RCNN [58]	0.910	0.621	0.870	0.382	0.598	0.695	0.461	0.660	0.765	44.01M	233.47
Cascade-RCNN [59]	0.916	0.644	0.876	0.386	0.611	0.717	0.455	0.660	0.771	69.29M	208.89
Deformable-DETR [60]	0.925	0.583	0.899	0.388	0.588	0.713	0.506	0.684	0.799	40.1M	166.91
Conditional-DETR [61]	0.891	0.558	0.873	0.340	0.556	0.715	0.451	0.643	0.803	43.45M	84.84
DAB-DETR [62]	0.902	0.569	0.882	0.350	0.571	0.726	0.483	0.666	0.813	43.71M	81.15
DDQ-DETR [63]	0.928	0.613	0.894	0.410	0.618	0.750	0.579	0.747	0.828	42.34M	180.93
VFNET [64]	0.900	0.614	0.869	0.406	0.617	0.727	0.491	0.686	0.798	32.72M	162.82
Our	0.971	0.804	0.945	0.553	0.761	0.806	0.633	0.805	0.846	3.71M	22.2

By contrast, the proposed model achieves an mAP50 of 0.971 and an mAP50-95 of 0.804 with only 3.71M parameters and 22.2 GFLOPs. Its mAP50-95 represents a relative improvement of approximately 24% over Cascade-RCNN, the second-best method in Table 8. This improvement can be attributed to the complementary effects of thermal-clutter suppression, weak-boundary preservation, cross-scale feature alignment, task-conditioned localization, and teacher-guided geometric consistency. To further analyze scale-sensitive detection performance, Table 8 reports the AP and AR results for small, medium, and large targets following the COCO scale criteria defined above. The proposed model achieves AP^s, AP^m, and AP^l values of 0.553, 0.761, and 0.806, respectively. Compared with the strongest competing result for each scale, these values represent absolute improvements of 0.137, 0.143, and 0.050. The proposed model also obtains the highest AR^s and AR^m values of 0.633 and 0.805, exceeding the corresponding second-best results by 0.054 and 0.058, respectively. Although its AR^l of 0.846 is slightly lower than the best value of 0.857, it remains competitive for large targets. The more pronounced improvements for small and medium targets suggest that frequency enhancement, cross-scale feature interaction, task-conditioned localization, and teacher-guided geometric supervision collectively improve feature representation, detection coverage, and localization for distant and weak infrared ship targets. As illustrated in Fig. 14, the proposed model consequently forms a more balanced profile across scale-sensitive accuracy, recall, and computational-efficiency indicators.

Despite the favorable overall and scale-specific results, the proposed model does not eliminate all detection errors under particularly difficult conditions. Fig. 15 presents four representative failure cases, with the first row showing the original infrared images and the second row showing the corresponding detection results. In Fig. 15a, the substantial scale difference between the large foreground vessel and the distant extremely small target results in the loss of the latter's limited spatial features. Fig. 15b shows a target located at the image boundary whose partially visible appearance is further disturbed by densely distributed vessel and mast structures. In Fig. 15c, dense mooring, image blur, and weak target boundaries result in an oversized and displaced prediction box that does not satisfy the matching criterion, thereby producing a localization failure. In Fig. 15d, densely distributed distant targets overlap with complex shoreline thermal responses, leading to simultaneous missed and false detections. These failure cases indicate that extremely small targets, partial target visibility, ambiguous boundaries, and strong structured background interference

remain challenging for the proposed model. Future work will investigate higher-resolution shallow-feature preservation, partial-object contextual modeling, and targeted hard-sample training to improve detection robustness under these difficult conditions.

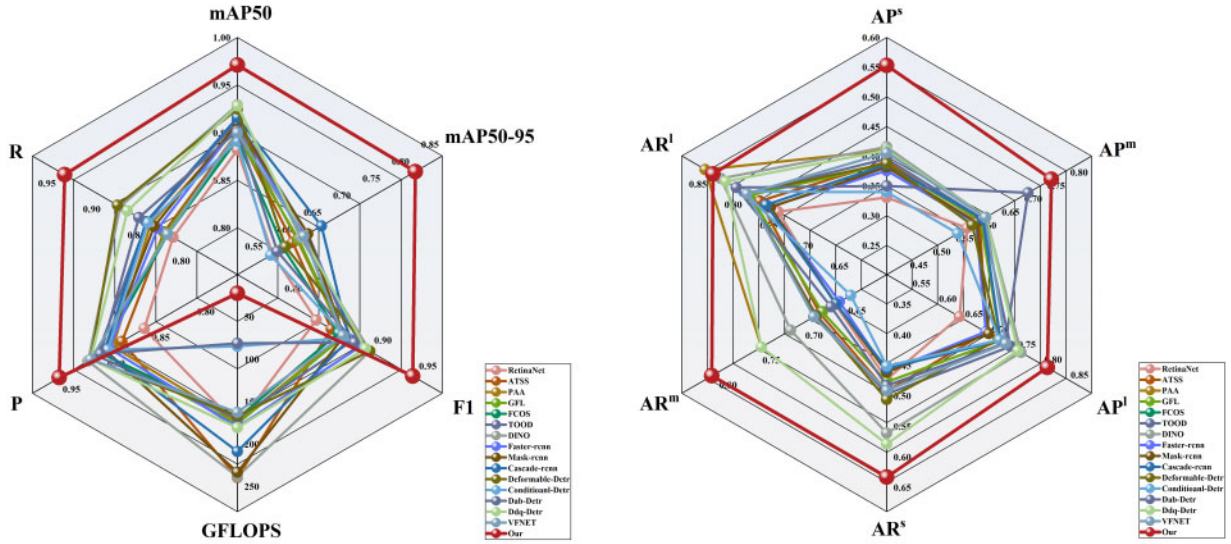

Figure 14: Comparison of different detection paradigms.

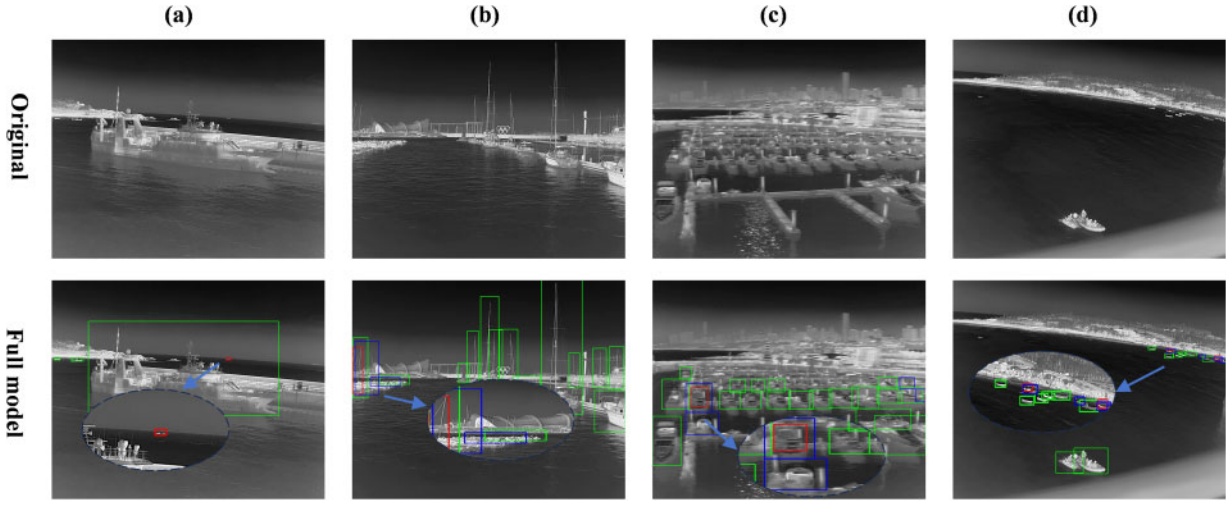

Figure 15: Representative failure cases of the proposed model: (**a**) extreme scale variation; (**b**) boundary truncation and structural interference; (**c**) image blur and weak boundaries; and (**d**) dense targets under shoreline thermal interference.

4.4 Performance and Cross-Scenario Adaptability Evaluation on Different Datasets

To further evaluate the adaptability of the proposed method to different maritime imaging conditions, extended experiments were conducted on ISDD and SMD, with the results summarized in Table 9. On the ISDD dataset, which is characterized by aerial top-down imaging, a high proportion of small targets, and scarce texture information, the proposed method improves mAP50-95 from 0.502 to 0.701, demonstrating strong adaptability in weak-small-target structure recovery and cross-scale representation. On the more challenging SMD dataset, which is characterized by complex shoreline interference, severe high-frequency

sea-surface clutter, and substantial multi-target interference, the proposed method achieves an mAP50 of 0.981 and an mAP50-95 of 0.835, demonstrating strong robustness to noise and stable localization performance under complex maritime background conditions. The PR and F1 curves shown in Fig. 16 indicate that it expands recall coverage while suppressing background false alarms, and preserves strong discriminative stability over a wide confidence-threshold range.

Table 9: Comparison of different datasets.

Datasets	Model	mAP50	mAP75	mAP50-95	P	R	F1
ISDD	Baseline	0.917	0.508	0.502	0.922	0.858	0.889
	Our	0.960	0.865	0.701	0.927	0.924	0.926
SMD	Baseline	0.927	0.733	0.683	0.924	0.872	0.897
	Our	0.981	0.933	0.835	0.976	0.969	0.972

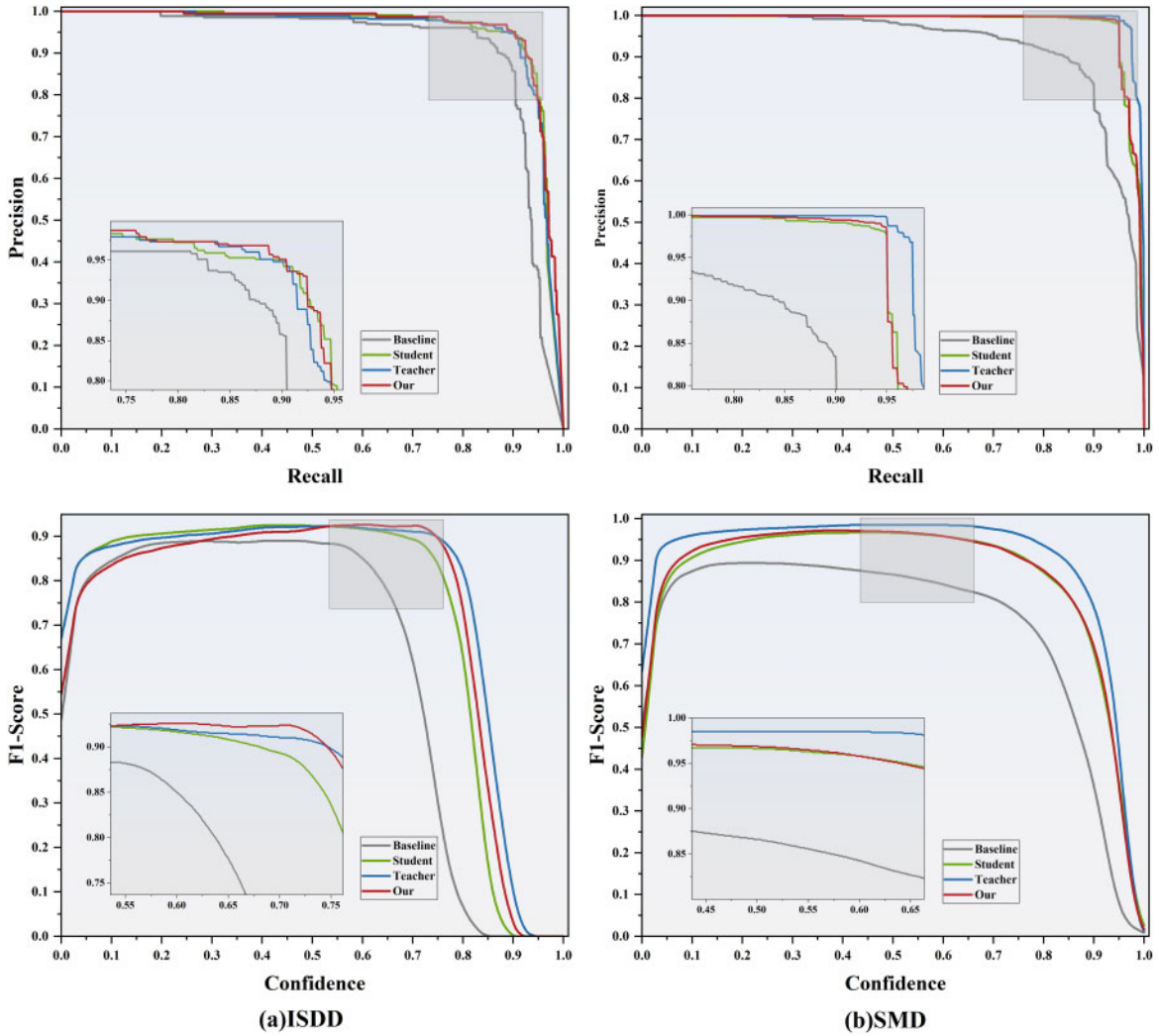

Figure 16: Comparison of PR and F1 curves of different datasets.

Fig. 17 provides intuitive visual evidence of missed detections and false positives. Compared with the baseline model, which is more susceptible to complex background interference and thus produces spurious responses (e.g., ISDD-Image4 and SMD-Image1) or fails to detect weak distant targets (e.g., ISDD-Image2 and SMD-Image6), the proposed model suppresses false alarms from sea-surface and shoreline backgrounds (e.g., ISDD-Image5 and SMD-Image4), recovers small and weak targets submerged by environmental interference (e.g., ISDD-Image2 and SMD-Image5), and achieves high-quality target separation. Taken together, these results indicate that the proposed architecture maintains consistent feature extraction capability and localization accuracy across heterogeneous maritime imaging platforms and complex environmental conditions.

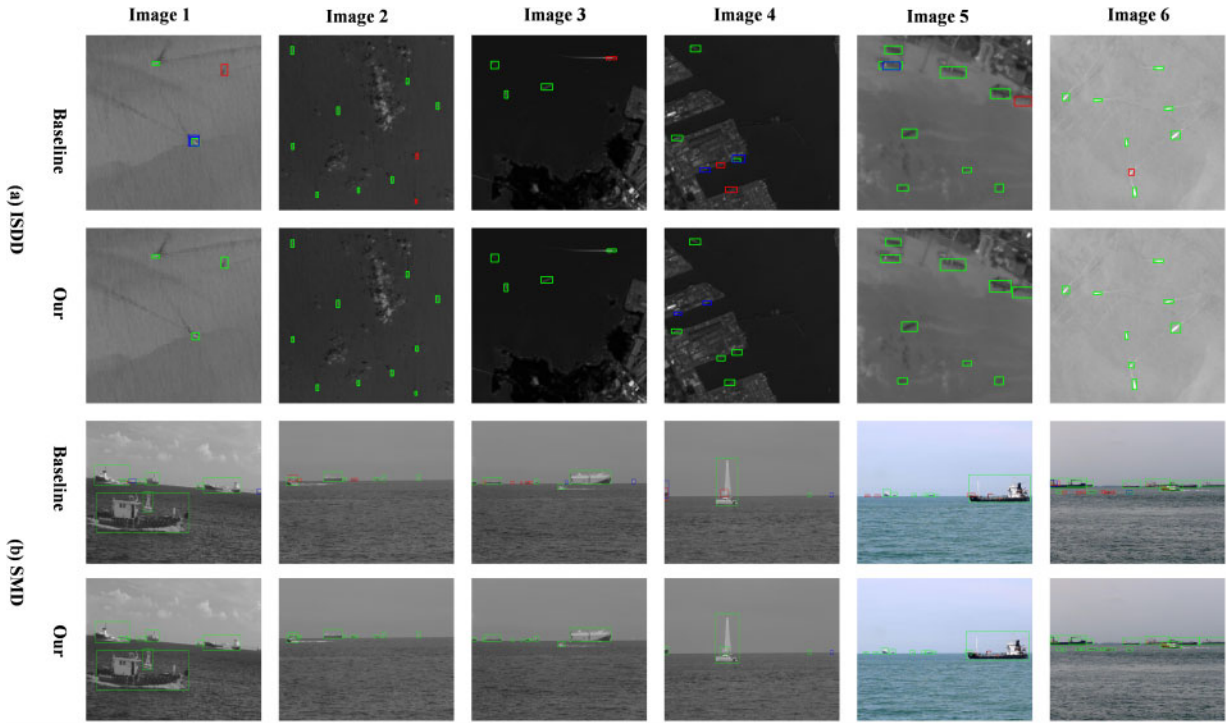

Figure 17: Visual comparison of different datasets.

To further evaluate the cross-scenario adaptability of the proposed architecture, experiments were extended to SSDD [65], HRSID [37], and the drone-view ship detection dataset constructed by Cheng et al. [66]. SSDD is a publicly available SAR ship-detection dataset containing 1160 images acquired from multiple SAR sensors and covering offshore and nearshore scenes. HRSID contains 5604 high-resolution SAR images and 16,951 ship instances collected under different spatial resolutions, polarizations, sea conditions, and coastal environments. The dataset constructed by Cheng et al. contains 3200 visible-light ship images captured by drones or from a drone-view perspective and was released through the authors' project repository. For each dataset, the baseline and proposed models were independently trained and evaluated using the same dataset-specific split and training configuration. Except for adapting the output-category dimension to the corresponding label space, the proposed architecture remained unchanged. The corresponding results are reported in Table 10.

Table 10: Adaptability evaluation of the proposed architecture across additional maritime datasets.

Datasets	Model	mAP50	mAP50-95	P	R	F1
SSDD	Baseline	0.986	0.720	0.967	0.951	0.959
	Our	0.989	0.878	0.967	0.977	0.972
HRSID	Baseline	0.891	0.646	0.915	0.789	0.847
	Our	0.944	0.795	0.935	0.873	0.903
Cheng et al.'s dataset	Baseline	0.624	0.356	0.732	0.550	0.628
	Our	0.771	0.566	0.834	0.703	0.763

As shown in Table 10, the proposed architecture consistently outperforms the baseline across all three additional datasets. In particular, mAP50-95 increases by 0.158, 0.149, and 0.210 on SSDD, HRSID, and Cheng et al.'s dataset, respectively, with corresponding improvements in recall and F1-score. These results demonstrate improved localization of densely distributed ships in SAR imagery and enhanced robustness to appearance and background variations in visible-light scenes. Together with the results on IR-Ship, ISDD, and SMD, the improvements obtained through independent training and evaluation on each dataset demonstrate that the proposed architecture remains effective across infrared, SAR, and visible-light maritime ship detection tasks, indicating its adaptability to different imaging modalities and scene characteristics.

4.5 Cross-Backbone Portability Evaluation

To evaluate whether the proposed architectural components are specifically dependent on YOLOv11, YOLOv12 was selected as an additional lightweight detector. The dataset split, input resolution, training schedule, optimizer, data augmentation, and evaluation protocol were kept identical across all configurations. WFEM and DMS-FFM were separately integrated into the corresponding feature-processing stages of YOLOv12, whereas TCUDH replaced its original detection head. Since TDSKD is a training-only optimization strategy and does not modify the inference-stage architectural interface, this portability experiment focuses on the three inference-stage components.

As shown in Table 11, the YOLOv12 baseline achieves 0.909 mAP50 and 0.636 mAP50-95. After introducing WFEM, mAP50-95 increases to 0.646, indicating that its frequency- and wavelet-domain enhancement remains effective after transfer to YOLOv12. DMS-FFM also increases mAP50-95 to 0.646, while improving mAP50 from 0.909 to 0.931 and recall from 0.858 to 0.890, indicating enhanced target coverage through cross-scale feature interaction. Both feature-enhancement modules introduce additional computational costs but retain inference speeds above 318 FPS. TCUDH produces the largest improvement, increasing mAP50 and mAP50-95 to 0.949 and 0.770, respectively, with absolute gains of 0.040 and 0.134. It also reduces the parameter count from 2.51M to 2.38M and the computational cost from 5.8 to 5.7 GFLOPs while retaining an inference speed of 560.5 FPS. Overall, the positive gains obtained by all three components demonstrate that their effectiveness is not strictly coupled to YOLOv11 and provide direct evidence of their portability to another lightweight detector.

Table 11: Cross-backbone portability evaluation of the proposed architectural components on YOLOv12.

Model	mAP50	mAP50-95	P	R	Params	GFLOPs	FPS
YOLOv12	0.909	0.636	0.904	0.858	2.51M	5.8	634.6
YOLOv12+WFEM	0.911	0.646	0.899	0.865	3.14M	12.4	318.5
YOLOv12+DMS-FFM	0.931	0.646	0.893	0.890	3.56M	13.5	329.7
YOLOv12+TCUDH	0.949	0.770	0.929	0.898	2.38M	5.7	560.5

4.6 Edge Deployment and Performance Analysis of Inference

Edge-deployment efficiency depends on the joint effects of model architecture, numerical precision, runtime optimization, and hardware configuration. Recent studies have evaluated lightweight YOLO variants on the Jetson Xavier NX by jointly considering accuracy, inference latency, GPU utilization, power consumption, and resource usage [67], while others have developed a lightweight multi-scale ship detector for wave gliders and validated its TensorRT FP16 deployment on the Jetson Orin Nano under a 15 W low-power mode, demonstrating the feasibility of real-time ship detection on resource-constrained maritime platforms [19]. These studies provide a hardware-aware context for evaluating the practical performance of object detectors on resource-constrained edge devices. Following this hardware-aware evaluation perspective, the runtime performance of the edge-side pipeline illustrated in Fig. 1 was evaluated in a real-world riverside infrared monitoring scenario. The inference application running on the Jetson Orin Nano 8GB was implemented in C++ using TensorRT FP16 inference and a Qt-based visualization interface. The TensorRT engine used an input resolution of 512 × 512 and CUDA Graph acceleration. Fig. 18 presents representative RTSP-based inference results, in which the original infrared stream frames and the corresponding detection outputs are displayed side by side. Runtime statistics collected during RTSP-based field operation show that the stream-capture rate ranged from 31.0 to 35.2 FPS, with an average of 32.55 FPS, while the effective inference throughput ranged from 27.3 to 28.9 FPS and averaged 28.23 FPS. The corresponding TensorRT inference latency ranged from 29.10 to 31.36 ms per frame, with an average of 30.34 ms, while the runtime jitter averaged 1.82 ms. The average processing interval corresponding to the effective inference throughput was approximately 35.42 ms per frame, compared with an average TensorRT inference latency of 30.34 ms. The difference of approximately 5.08 ms per frame provides an estimate of the combined overhead associated with RTSP stream decoding, image preprocessing, prediction postprocessing, thread scheduling, and interface rendering. As illustrated in Fig. 18, the proposed model maintains stable responses to key ship targets under low infrared contrast and thermal interference from shoreline buildings. Together with the measured latency and inference throughput, these results demonstrate the real-time operational feasibility of the complete infrared acquisition, inference, and visualization pipeline on the Jetson Orin Nano 8 GB. Because the available runtime statistics primarily characterize stream acquisition and inference efficiency, system-level indicators such as GPU utilization, memory consumption, and device-level power consumption are not included in the current evaluation. Comprehensive resource profiling under controlled power modes will be considered in future multi-platform deployment evaluations.

5 Discussion and Limitations

Although the experimental results demonstrate the effectiveness and deployment potential of the proposed method, its current validation scope and application boundaries should be considered. The following discussion analyzes its robustness under adverse infrared imaging conditions, the applicability,

parameter dependence, and limitations of the proposed components, and the adaptability of the deployment pipeline to other hardware platforms and infrared sensors.

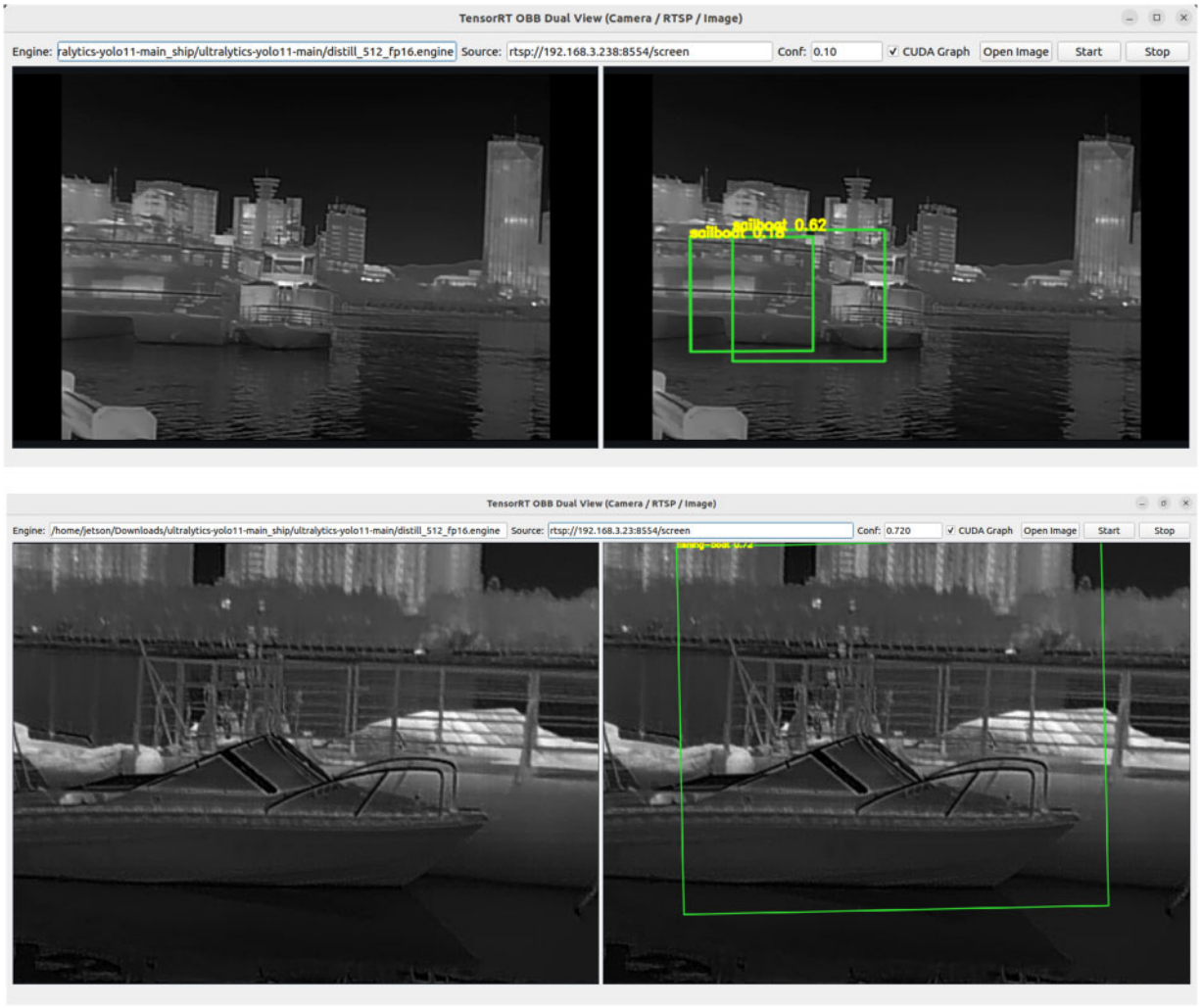

Figure 18: Edge deployment of real-time inference.

5.1 Robustness under Adverse Infrared Imaging Conditions

The current experiments mainly cover low target-background contrast, sea-surface thermal clutter, shoreline heat-source interference, dense target distributions, and substantial scale variations. More extreme imaging conditions may cause additional performance degradation. Severe fog and heavy rain can attenuate long-range thermal radiation and reduce the effective contrast between ships and their surroundings, while strong atmospheric attenuation and rain-induced interference may further weaken the already limited responses of distant targets. Thermal blooming or sensor saturation may broaden high-response regions, distort local intensity distributions, and obscure the boundaries of nearby targets. Under extremely low target-background contrast, the frequency components of weak ships and structured background interference may substantially overlap, making it difficult for frequency-domain enhancement to separate useful target responses from noise. Severe fog, heavy rain, thermal blooming, and sensor saturation are not

sufficiently represented in the current datasets. Therefore, the reported results should not be interpreted as comprehensive validation under all extreme weather and infrared imaging conditions.

5.2 Applicability, Parameter Considerations, and Limitations of the Proposed Components

The effectiveness of the proposed components depends on the characteristics of the target and background. WFEM is primarily designed to suppress structured thermal clutter while preserving the high-frequency details and boundaries of weak targets. Its benefit may therefore be limited when targets already exhibit high contrast and clear boundaries. Moreover, when weak target responses and structured background interference occupy overlapping frequency ranges, frequency enhancement may retain or amplify undesirable noise and may introduce local ringing artifacts. DMS-FFM mainly benefits scenes involving substantial scale variation and dense target distributions; its improvement may be smaller in sparse scenes dominated by a narrow target-scale range, while its cross-scale operations increase computational cost. TCUDH is designed for weak-boundary, elongated, and densely distributed targets. Its advantage may consequently be less pronounced for large, isolated, and high-contrast ships. When target boundaries or appearance cues are severely degraded or partially absent, task-conditioned modulation cannot fully recover the missing localization information. In addition, TDSKD depends on the quality and calibration of the teacher model, and inaccurate teacher predictions may transfer erroneous or biased supervision to the student model under substantial domain shift.

From the perspective of parameter dependence, the proposed framework contains both trainable variables and manually specified control parameters. The channel-wise enhancement factor (λ_E), the Fourier-wavelet balancing coefficient (γ), and the scale factors (s_l) are optimized jointly with the network rather than maintained as fixed manually selected weights. Among the principal fixed parameters, the cutoff frequency radius (D_0) controls the frequency range emphasized by WFEM. An excessively small (D_0) may broaden the enhanced frequency range and amplify background interference, whereas an excessively large value may weaken the recovery of target-boundary information. In TDSKD, a low distillation temperature (T) produces a sharper teacher distribution, while an excessively high value may over-smooth the localization responses. A low teacher-confidence threshold (τ) may introduce unreliable pseudo-targets, whereas a high threshold may remove weak or distant targets. The hard-sample coefficient (ρ) determines the strength of difficulty-aware weighting; an excessively small value may weaken the guidance for ambiguous targets, whereas an excessively large value may amplify unstable or noisy samples. The principal parameter settings and structural configurations were kept unchanged across all comparative and ablation experiments to isolate the contributions of the proposed components. An exhaustive joint search over every possible parameter and structural combination was not conducted; therefore, the reported conclusions should be understood under the adopted configuration. Together, these factors define the current applicability and parameter-sensitivity boundaries of the proposed framework.

5.3 Adaptability to Other Embedded Platforms and Infrared Sensors

The current deployment pipeline uses a Lenovo tablet as an RTSP relay because the Tianyan X3 thermal imager does not provide an independent SDK. This relay is an acquisition-interface solution rather than a requirement of the proposed detection model. Infrared sensors supporting RTSP, USB Video Class, GStreamer, or vendor-specific SDK interfaces can be connected directly to the preprocessing and inference pipeline. The trained network can be exported through ONNX and deployed using TensorRT on NVIDIA Jetson platforms. Adaptation to other embedded accelerators is also technically possible when the selected inference backend supports the required operators; however, model conversion, operator compatibility, memory allocation, and hardware-specific optimization must be reconsidered for each platform. Infrared

sensors with different spatial resolutions, bit depths, spectral ranges, thermal responses, or non-uniformity characteristics may require sensor-specific normalization, calibration, and, under substantial domain shift, additional model adaptation. Because the current real-world deployment evaluation is limited to one edge platform and one infrared acquisition pipeline, the reported inference latency, application-level throughput, and detection performance should not be directly generalized to other devices without platform- and sensor-specific validation.

6 Conclusion

This paper proposes an edge-oriented infrared ship detection method for deep learning-based pattern recognition in complex maritime scenes, aiming to improve vision-based perception reliability under low contrast, sea-surface thermal clutter, shoreline heat-source interference, dense target distribution, and large-scale variations in infrared imagery. The proposed method focuses on thermal clutter suppression, multi-scale feature enhancement, robust localization, and teacher-guided training optimization. WFEM suppresses thermal interference caused by sea-wave textures, shoreline heat sources, and background reflections, while preserving weak and small target structures through frequency-domain filtering and wavelet-based edge reconstruction. DMS-FFM further enhances multi-scale feature interaction and cross-scale semantic alignment, improving the representation of ships with large scale variations and dense distributions. Based on the enhanced features, TCUDH improves localization robustness for low-contrast and elongated infrared ship targets through task-conditioned modulation and shared prediction design. TDSKD improves the discriminative capability of the student model without increasing inference overhead. Extensive experiments on IR-Ship, ISDD, SMD, SSDD, HRSID, and Cheng et al.'s dataset demonstrate that the proposed method achieves consistent performance improvements and architectural adaptability across infrared, SAR, and visible-light maritime imaging scenarios. On the IR-Ship dataset, the final model attains 0.971 mAP50 and 0.804 mAP50-95 with 3.71M parameters and 22.2 GFLOPs, showing a favorable balance between accuracy and computational cost. Additional evaluations on ISDD and SMD verify its robustness under remote-sensing, shoreline-interference, and multi-target maritime conditions, while the consistent improvements on SSDD, HRSID, and Cheng et al.'s dataset further demonstrate its adaptability to SAR and visible-light imagery. Moreover, real-world deployment in a riverside monitoring scenario achieves an average effective inference throughput of 28.23 FPS on the Jetson Orin Nano 8 GB platform, demonstrating its engineering applicability for real-time edge-side visual perception and automated maritime target recognition. Although the proposed method demonstrates promising performance, several limitations remain. Quantitative validation of the unified head on segmentation tasks is still insufficient, and direct cross-dataset transfer without dataset-specific retraining requires further investigation. Future work will strengthen the quantitative evaluation of multi-task learning, explore more compact feature modeling and distillation strategies, and extend the framework toward broader maritime perception tasks, such as joint detection, segmentation, tracking, and open-set target recognition in more challenging real-world environments.

Acknowledgement: The authors acknowledge Yantai Raytron Technology Co., Ltd. for publicly releasing the Infrared Ship Database used in this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: data curation and conceptualization, Hongliang Tian; methodology, experimental implementation, and writing—original draft preparation, Chenying Pei; results analysis and interpretation, Jin Lei; manuscript review and formatting, Xiaoke Liu; validation and supervision, Xin Ma. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The Infrared Ship Database used in this study is publicly released by Yantai Raytron Technology Co., Ltd. and is available free of charge upon application and approval through its Infrared Open Source Platform (https://openai.raytrontek.com/apply/E_Sea_shipping.html/). The other public datasets used in this study are available from the respective sources cited in the manuscript. The experimental results generated during this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024.
2. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
3. Bakirci M. Advanced ship detection and ocean monitoring with satellite imagery and deep learning for marine science applications. Reg Stud Mar Sci. 2025;81:103975. doi:10.1016/j.rsma.2024.103975.
4. Jiang Z. DCB-YOLO: an infrared ship detection model with deformable convolution and bi-level routing attention. In: Proceedings of the 2025 7th International Conference on Information Science, Electrical and Automation Engineering (ISEAE); 2025 Apr 18–20; Harbin, China. New York, NY, USA: IEEE; 2025. p. 1072–8. doi:10.1109/ISEAE64934.2025.11042022.
5. Wang S, Feng Y, Jin W, Liu L, Zhou C, Tao H, et al. DSEE-YOLO: a dynamic edge-enhanced lightweight model for infrared ship detection in complex maritime environments. Remote Sens. 2025;17(19):3325. doi:10.3390/rs17193325.
6. Wu CM, Lei J, Li ZQ, Ren ML. Ship_YOLO: general ship detection based on mixed distillation and dynamic task-aligned detection head. Ocean Eng. 2025;323(2):120616. doi:10.1016/j.oceaneng.2025.120616.
7. Wang S, Feng Y, Tao H, Chen J, Jin W, Liu L, et al. EISD-YOLO: efficient infrared ship detection with param-reduced PEMA block and dynamic task alignment. Photonics. 2025;12(11):1044. doi:10.3390/photonics12111044.
8. Sun Y, Lian J. IRSD-net: an adaptive infrared ship detection network for small targets in complex maritime environments. Remote Sens. 2025;17(15):2643. doi:10.3390/rs17152643.
9. Man F, Li C, Guan T. Infrared remote sensing small ship target detection method based on spatial-semantic enhancement and feature reconstruction neck. J Vis Commun Image Represent. 2026;115:104683. doi:10.1016/j.jvcir.2025.104683.
10. Hu C, Dong X, Huang Y, Wang L, Xu L, Pu T, et al. SMPISD-MTPNet: scene semantic prior-assisted infrared ship detection using multitask perception networks. IEEE Trans Geosci Remote Sens. 2025;63:5000814. doi:10.1109/TGRS.2024.3516879.
11. Wang J, Su N, Zhao C, Xu C, Feng Q, Zhang C, et al. CIFDet: robust correlation-guided fusion and contrast-driven attention for misaligned multi-modal object detection. Inf Fusion. 2026;127:103833. doi:10.1016/j.inffus.2025.103833.
12. Lin Y, Peng D, Wang L, Jiang L, Nam H. MSCK-net: multiscale Chinese knot convolutional network for dim and small infrared ship detection. IEEE Trans Geosci Remote Sens. 2026;64:5602218. doi:10.1109/TGRS.2025.3649839.
13. Dong K, Liu T, Zheng Y, Shi Z, Du H, Wang X. Visual detection algorithm for enhanced environmental perception of unmanned surface vehicles in complex marine environments. J Intell Rob Syst. 2023;110(1):1. doi:10.1007/s10846-023-02020-z.
14. Mela JL, Sánchez CG. Yolo-based power-efficient object detection on edge devices for USVs. J Real Time Image Process. 2025;22(3):108. doi:10.1007/s11554-025-01682-2.
15. Zhang J, Jin J, Ma Y, Ren P. Lightweight object detection algorithm based on YOLOv5 for unmanned surface vehicles. Front Mar Sci. 2023;9:1058401. doi:10.3389/fmars.2022.1058401.
16. Sun H, Wang R, Li Y, Yang L, Lin S, Cao X, et al. SET: spectral enhancement for tiny object detection. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. New York, NY, USA: IEEE; 2025. p. 4713–23. doi:10.1109/CVPR52734.2025.00444.

17. Zhu H, Dong W, Yang L, Li H, Yang Y, Ren Y, et al. WaveMamba: wavelet-driven mamba fusion for RGB-infrared object detection. In: Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–25; Honolulu, HI, USA. New York, NY, USA: IEEE; 2026. p. 11219–29. doi:10.1109/ICCV51701.2025.01044.
18. Chen Y, Ren J, Li J, Shi Y. Enhanced adaptive detection of nearby and distant ships in fog: a real-time multi-scale target detection strategy. Digit Signal Process. 2025;158:104961. doi:10.1016/j.dsp.2024.104961.
19. Sang H, Lu Q, Sun X, Zhang S, Liu F. A lightweight multi-scale ship detection framework for wave gliders with spatial-channel attention fusion. Measurement. 2026;258:119280. doi:10.1016/j.measurement.2025.119280.
20. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022.
21. Dai X, Chen Y, Xiao B, Chen D, Liu M, Yuan L, et al. Dynamic head: unifying object detection heads with attentions. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. New York, NY, USA: IEEE; 2021. p. 7369–78. doi:10.1109/CVPR46437.2021.00729.
22. Yu Z, Huang H, Chen W, Su Y, Liu Y, Wang X. YOLO-FaceV2: a scale and occlusion aware face detector. Pattern Recognit. 2024;155:110714. doi:10.1016/j.patcog.2024.110714.
23. Wang J, Chen Y, Zheng Z, Li X, Cheng MM, Hou Q. CrossKD: cross-head knowledge distillation for object detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. New York, NY, USA: IEEE; 2024. p. 16520–30. doi:10.1109/CVPR52733.2024.01563.
24. Cao X, Hu Y, Zhang H. LKD-YOLOv8: a lightweight knowledge distillation-based method for infrared object detection. Sensors. 2025;25(13):4054. doi:10.3390/s25134054.
25. Li M, Liang X, Hu Q, Lin YE, Xia C. Multi-scale feature fusion with knowledge distillation for object detection in aerial imagery. Eng Appl Artif Intell. 2025;158(2):111518. doi:10.1016/j.engappai.2025.111518.
26. Zhou W, Ben T. IRMultiFuseNet: ghost hunter for infrared ship detection. Displays. 2024;81:102606. doi:10.1016/j.displa.2023.102606.
27. Han Y, Liao J, Lu T, Pu T, Peng Z. KCPNet: knowledge-driven context perception networks for ship detection in infrared imagery. IEEE Trans Geosci Remote Sens. 2023;61:5000219. doi:10.1109/TGRS.2022.3233401.
28. Prasad DK, Rajan D, Rachmawati L, Rajabally E, Quek C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: a survey. IEEE Trans Intell Transp Syst. 2017;18(8):1993–2016. doi:10.1109/TITS.2016.2634580.
29. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. New York, NY, USA: IEEE; 2020. p. 10778–87. doi:10.1109/cvpr42600.2020.01079.
30. Yang G, Lei J, Zhu Z, Cheng S, Feng Z, Liang R. AFPN: asymptotic feature pyramid network for object detection. arXiv:2306.15988. 2023.
31. Li H, Li J, Wei H, Liu Z, Zhan Z, Ren Q. Slim-neck by GSConv: a lightweight-design for real-time detector architectures. J Real Time Image Process. 2024;21(3):62. doi:10.1007/s11554-024-01436-6.
32. Huang Z, Shang W. DSP-YOLO: a SAR ship detection algorithm for multiscale sequence fusion based on fusion attention. In: Proceedings of the 2024 7th International Conference on Computational Intelligence and Intelligent Systems; 2024 Nov 22–24; Nagoya, Japan. p. 44–52. doi:10.1145/3708778.3708785.
33. Xu X, Jiang Y, Chen W, Huang Y, Zhang Y, Sun X. DAMO-YOLO: a report on real-time object detection design. arXiv:2211.15444. 2022.
34. Shan W, Yue Y. Apple defect detection in complex environments. Electronics. 2024;13(23):4844. doi:10.3390/electronics13234844.
35. Wang CY, Yeh IH, Liao HYM. YOLOv9: learning what you want to learn using programmable gradient information. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editor. Computer vision—ECCV 2024. Lecture notes in computer science. vol. 15089. Cham, Switzerland: Springer; 2025. p. 1–21. doi:10.1007/978-3-031-72751-1_1.

36. Shao Z, Wu W, Wang Z, Du W, Li C. SeaShips: a large-scale precisely annotated dataset for ship detection. IEEE Trans Multimed. 2018;20(10):2593–604. doi:10.1109/TMM.2018.2865686.
37. Wei S, Zeng X, Qu Q, Wang M, Su H, Shi J. HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation. IEEE Access. 2020;8:120234–54. doi:10.1109/ACCESS.2020.3005861.
38. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. In: Belgrave D, Zhang C, Lin H, Pascanu R, Koniusz P, Ghassemi M, et al., editors. Advances in neural information processing systems. vol. 38. Red Hook, NY, USA: Curran Associates, Inc.; 2025. p. 78433–57. doi:10.52202/085713-2627.
39. Lei M, Li S, Wu Y, Hu H, Zhou Y, Zheng X, et al. YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv:2506.17733. 2025.
40. Sapkota R, Cheppally RH, Sharda A, Karkee M. YOLO26: key architectural enhancements and performance benchmarking for real-time object detection. arXiv:2509.25164. 2025.
41. Wang C, He W, Nie Y, Guo J, Liu C, Wang Y, et al. Gold-YOLO: efficient object detector via gather-and-distribute mechanism. In: Advances in neural information processing systems. vol. 36. Red Hook, NY, USA: Curran Associates, Inc.; 2023. p. 51094–112. doi:10.52202/075280-2224.
42. Zhan W, Zhang C, Guo S, Guo J, Shi M. EGISD-YOLO: edge guidance network for infrared ship target detection. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:10097–107. doi:10.1109/JSTARS.2024.3389958.
43. Wang Z, Li C, Xu H, Zhu X, Li H. Mamba YOLO: a simple baseline for object detection with state space model. Proc AAAI Conf Artif Intell. 2025;39(8):8205–13. doi:10.1609/aaai.v39i8.32885.
44. Ye J, Yuan Z, Qian C, Li X. CAA-YOLO: combined-attention-augmented YOLO for infrared ocean ships detection. Sensors. 2022;22(10):3782. doi:10.3390/s22103782.
45. Xiao Y, Xu T, Xin Y, Li J. FBRT-YOLO: faster and better for real-time aerial image detection. Proc AAAI Conf Artif Intell. 2025;39(8):8673–81. doi:10.1609/aaai.v39i8.32937.
46. Wu CM, Lei J, Liu WK, Ren ML, Ran LL. Unmanned ship identification based on improved YOLOv8s algorithm. Comput Mater Contin. 2024;78(3):3071–88. doi:10.32604/cmc.2023.047062.
47. Guo L, Wang Y, Guo M, Zhou X. YOLO-IRS: infrared ship detection algorithm based on self-attention mechanism and KAN in complex marine background. Remote Sens. 2025;17(1):20. doi:10.3390/rs17010020.
48. Xue Y, Ju Z, Li Y, Zhang W. MAF-YOLO: multi-modal attention fusion based YOLO for pedestrian detection. Infrared Phys Technol. 2021;118:103906. doi:10.1016/j.infrared.2021.103906.
49. Wang Y, Wang B, Huo L, Fan Y. GT-YOLO: nearshore infrared ship detection based on infrared images. J Mar Sci Eng. 2024;12(2):213. doi:10.3390/jmse12020213.
50. Wang P, Zhao J. SOD-YOLO: enhancing YOLO-based detection of small objects in UAV imagery. arXiv:2507.12727. 2025.
51. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Itally. New York, NY, USA: IEEE; 2017. p. 2999–3007. doi:10.1109/iccv.2017.324.
52. Zhang S, Chi C, Yao Y, Lei Z, Li SZ. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. New York, NY, USA: IEEE; 2020. p. 9756–65. doi:10.1109/cvpr42600.2020.00978.
53. Kim K, Lee HS. Probabilistic anchor assignment with IoU prediction for object detection. In: Computer vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 355–71. doi:10.1007/978-3-030-58595-2_22.
54. Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editor. Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2020. p. 21002–12.
55. Tian Z, Shen C, Chen H, He T. FCOS: a simple and strong anchor-free object detector. IEEE Trans Pattern Anal Mach Intell. 2022;44(4):1922–33. doi:10.1109/TPAMI.2020.3032166.

56. Feng C, Zhong Y, Gao Y, Scott MR, Huang W. TOOD: task-aligned one-stage object detection. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. New York, NY, USA: IEEE; 2021. p. 3490–9. doi:10.1109/iccv48922.2021.00349.
57. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, et al. *DINO*: DETR with improved DeNoising anchor boxes for end-to-end object detection. arXiv:2203.03605. 2022.
58. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. New York, NY, USA: IEEE; 2017. p. 2980–8. doi:10.1109/ICCV.2017.322.
59. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. New York, NY, USA: IEEE; 2018. p. 6154–62. doi:10.1109/CVPR.2018.00644.
60. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020.
61. Meng D, Chen X, Fan Z, Zeng G, Li H, Yuan Y, et al. Conditional DETR for fast training convergence. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. New York, NY, USA: IEEE; 2022. p. 3631–40. doi:10.1109/ICCV48922.2021.00363.
62. Liu S, Li F, Zhang H, Yang X, Qi X, Su H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR. arXiv:2201.12329. 2022.
63. Zhang S, Wang X, Wang J, Pang J, Lyu C, Zhang W, et al. Dense distinct query for end-to-end object detection. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. New York, NY, USA: IEEE; 2023. p. 7329–38. doi:10.1109/CVPR52729.2023.00708.
64. Zhang H, Wang Y, Dayoub F, Sunderhauf N. VarifocalNet: an IoU-aware dense object detector. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. New York, NY, USA: IEEE; 2021. p. 8510–9. doi:10.1109/cvpr46437.2021.00841.
65. Zhang T, Zhang X, Li J, Xu X, Wang B, Zhan X, et al. SAR ship detection dataset (SSDD): official release and comprehensive data analysis. Remote Sens. 2021;13(18):3690. doi:10.3390/rs13183690.
66. Cheng S, Zhu Y, Wu S. Deep learning based efficient ship detection from drone-captured images for maritime surveillance. Ocean Eng. 2023;285:115440. doi:10.1016/j.oceaneng.2023.115440.
67. Bakirci M. Performance evaluation of low-power and lightweight object detectors for real-time monitoring in resource-constrained drone systems. Eng Appl Artif Intell. 2025;159:111775. doi:10.1016/j.engappai.2025.111775.