remote sensing

MDPI

Article

IRSD-Net: An Adaptive Infrared Ship Detection Network for Small Targets in Complex Maritime Environments

Yitong Sun and Jie Lian *

College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China; 1000530508@smail.shnu.edu.cn
* Correspondence: lianjie@shnu.edu.cn

Abstract

Infrared ship detection plays a vital role in maritime surveillance systems. As a critical remote sensing application, it enables maritime surveillance across diverse geographic scales and operational conditions while offering robust all-weather operation and resilience to environmental interference. However, infrared imagery in complex maritime environments presents significant challenges, including low contrast, background clutter, and difficulties in detecting small-scale or distant targets. To address these issues, we propose an Infrared Ship Detection Network (IRSD-Net), a lightweight and efficient detection network built upon the YOLOv11n framework and specially designed for infrared maritime imagery. IRSD-Net incorporates a Hierarchical Multi-Kernel Convolution Network (HMKCNet), which employs parallel multi-kernel convolutions and channel division to enhance multi-scale feature extraction while reducing redundancy and memory usage. To further improve cross-scale fusion, we design the Dynamic Cross-Scale Feature Pyramid Network (DCSFPN), a bidirectional architecture that combines up- and downsampling to integrate low-level detail with high-level semantics. Additionally, we introduce Wise-PIoU, a novel loss function that improves bounding box regression by enforcing geometric alignment and adaptively weighting gradients based on alignment quality. Experimental results demonstrate that IRSD-Net achieves 92.5% mAP_{50} on the ISDD dataset, outperforming YOLOv6n and YOLOv11n by 3.2% and 1.7%, respectively. With a throughput of 714.3 FPS, IRSD-Net delivers high-accuracy, real-time performance suitable for practical maritime monitoring systems.

Keywords: infrared ship detection; complex environmental background; multi-kernel convolution; cross-scale feature extraction; Wise-PIoU

Academic Editor: Hossein M. Rizeei

Received: 10 June 2025
Revised: 25 July 2025
Accepted: 28 July 2025
Published: 30 July 2025

Citation: Sun, Y.; Lian, J. IRSD-Net: An Adaptive Infrared Ship Detection Network for Small Targets in Complex Maritime Environments. *Remote Sens.* **2025**, *17*, 2643. https://doi.org/10.3390/rs17152643

1. Introduction

Ship target detection is a vital real-time monitoring task with significant applications in areas including port surveillance, maritime search and rescue, navigation safety, fisheries management, marine environmental protection, and modern naval warfare [1–3]. With advances in remote sensing imaging technologies enabling surveillance across extensive maritime areas, thermal infrared imaging technology has gained widespread use in ship detection owing to its excellent concealment, minimal power consumption, all-weather operation, and robust remote monitoring capabilities [4,5]. It has thus become one of the principal approaches for real-time maritime monitoring [6–8].

Conventional infrared small-target detection algorithms primarily encompass several main categories. These include background suppression techniques utilizing spatial domain

filtering [9], such as morphological Top-hat transformation [10]; saliency detection models that enhance local contrast [11] and extract directional gradient features [12]; and target extraction methods based on manually designed features, such as histograms [13] and edge response templates [14]. However, these methods rely on manually preset parameters and struggle to adapt to multi-scale targets and dynamic environments. Moreover, their high computational complexity makes them unsuitable for real-time detection in complex maritime environments [15,16].

With the emergence of deep learning, data-driven methods have demonstrated remarkable performance in complex visual tasks by learning hierarchical representations from raw data. For object detection, frameworks are broadly divided into two categories. Two-stage algorithms, such as region-based convolutional neural networks (R-CNNs) [17] and Faster R-CNN [18], generate and classify candidate regions in two steps. Single-stage approaches, such as You Only Look Once (YOLO) [19] and single-shot multi-box detector (SSD) [20], directly determine object locations and categories. While effective for general-purpose detection, these architectures often struggle with the small, low-contrast targets prevalent in infrared maritime imagery.

Recognizing these limitations, recent efforts have investigated multi-scale and lightweight designs tailored for ship detection. Chen et al. [21] introduced the lightweight Tiny YOLO-Lite model for synthetic-aperture radar (SAR) ship detection, which decreases model parameters and computational complexity through network pruning and knowledge distillation. Nevertheless, this approach suffers from performance degradation and requires sophisticated compensation strategies. Zhan et al. [22] developed EGISD-YOLO, employing Dense Cross Stage Partial (DCSP) modules and Deconvolution Channel Attention (DCA) modules to enhance feature reuse and semantic fusion for weak target detection, yet the substantial model size renders it inappropriate for deployment on resource-constrained platforms. Building upon YOLOv7, Deng et al. [23] introduced the efficient channel and spatial excitation attention module (ECSE) with dilated convolution-based weighted feature pyramid networks (DWFPN) for rotational ship detection, effectively enhancing small-target detection in foggy maritime environments. However, the multiple enhancement modules resulted in increased model parameters and computational overhead. Ge et al. [24] proposed RD-YOLO for infrared remote sensing ship detection by incorporating Receptive Field Convolution (RFConv) modules and Deep Convergence Networks (DCNnet) into YOLOv8s, effectively integrating high-dimensional information with detailed shallow features. Despite improved detection accuracy, the model exhibits higher sensitivity to infrared image degradation and complex maritime environmental interference. Sun et al. [25] presented a bidirectional feature fusion and angular classification (BiFA)-YOLO model, featuring Bidirectional Deep Feature Fusion Modules (Bi-DFFM) and Random Rotation Mosaic (RR-Mosaic) data augmentation. It effectively captures accurate ship orientation information while distinguishing closely docked vessels in complex port scenarios, but demonstrates insufficient sensitivity to extremely small targets.

Despite these advancements, detecting ships in infrared maritime imagery continues to pose significant challenges. Thermal images inherently lack color and texture, yielding low target–background contrast, especially for distant or small vessels. Environmental elements such as waves, clouds, and coastlines often exhibit thermal characteristics similar to ships, increasing the risk of false alarms and localization drift. Moreover, fixed feature fusion strategies inadequately balance information across scales, and geometric misalignment under long-range or low-visibility conditions further exacerbates detection errors. The blurred contours and simple structures of ship targets complicate robust feature learning [26–28].

To address these challenges, we introduce IRSD-Net, based on the YOLOv11n architecture, a lightweight deep network specifically designed for infrared ship detection

in low-contrast maritime scenes. The key contributions of this work are summarized as follows:

1. We design IRSD-Net, an efficient architecture that balances multi-scale feature extraction and cross-layer semantic fusion, achieving high accuracy with low computational overhead.
2. We develop a novel backbone named HMKCNet that leverages parallel convolutions with diverse kernel sizes for simultaneous local detail and contextual feature capture, enhanced by channel partitioning for reduced redundancy.
3. The DCSFPN is introduced to precisely integrate low-level detail information with high-level semantic information, which addresses information loss and the unidirectional flow of traditional methods.
4. Experiments on ISDD and IRSDSS datasets validate that IRSD-Net can achieve an mAP_{50} of 92.5% and 92.9%, respectively, outperforming existing methods with only 6.5 GFLOPs and 2.1M parameters, thus supporting real-time deployment in constrained environments.

2. Related Work

2.1. Traditional Infrared Ship Detection Methods

Initial infrared ship studies centered on image enhancement and boosting target saliency. Researchers introduced various detection techniques utilizing local contrast analysis and morphological operations. The Top-hat transformation, for instance, serves to reduce background influence and accentuate local bright spots, making it suitable for uniform background conditions [29]. Using sliding windows, Local Contrast Measure (LCM) amplifies local target signals, proving especially useful for identifying small ships at significant ranges [30]. To improve robustness, researchers developed advanced strategies including the Multi-Level Local Contrast Measure (MLLCM) [31], Multiscale Patch-Based Contrast Measure (MPCM) [32], and Relative Local Contrast Measure (RLCM) [33]. Background modeling offers another key approach. For example, Zhou et al. [34] proposed a Fourier domain method that creates separate models for fluctuating and stable sea regions, which can effectively isolate targets from sea clutter.

Machine learning approaches have enhanced discrimination capabilities in complex environments. Such techniques derive manually engineered features, including brightness gradients and regional entropy, and leverage Support Vector Machine (SVM) or Random Forest algorithms for binary classification tasks. For example, Lin et al. [35] combined Bayesian inference with SVM classification, using prior probability modeling to screen infrared ships and suppress false detections. Li et al. [36] introduced a detection technique utilizing hyperspectral imagery which integrates spectral and texture features with a Random Forest model for stable target recognition against complex backgrounds. These approaches established foundational feature-classifier frameworks exhibiting preliminary characteristics of end-to-end learning paradigms. Nevertheless, their dependence on handcrafted features limits their efficacy when confronted with complex maritime environments and multi-scale target variability.

2.2. Deep Learning-Based Infrared Ship Detection Methods

Deep learning has demonstrated superior capabilities in tasks related to infrared ship detection. Researchers have adapted various detection frameworks specifically for infrared scenarios, addressing difficulties including low contrast and structural deterioration in infrared images. Chen et al. [37] proposed a Combined Attention-Augmented (CAA)-YOLO model that incorporates a fine-scale shallow feature layer (P2) to preserve positional information, utilizing temporal attention modules and contextual attention mechanisms.

However, its complex architecture requires optimization to enhance inference efficiency. Li et al. [38] introduced a comprehensive YOLO-based vessel detection approach (CYSDM) to detect maritime targets in all weather conditions using thermal infrared remote sensing. They employed bicubic interpolation upsampling and gray value stretching techniques to enhance characteristics, though this method relies heavily on preprocessing and manual intervention for dataset creation. Transformer architectures introduced new methods for modeling remote sensing images. The Multi-Level TransUNet (MTU-Net) [39] combines CNNs with Vision Transformers to enable multi-scale feature fusion, allowing precise identification of small ship targets. Wu et al. [40] proposed a Swin Transformer combined with multi-scale atrous spatial pyramid pooling (STASPPNet) to improve feature representation and scale adaptability, although performance validation was largely limited to single datasets. Beyond architectural innovations, some researchers have focused on optimizing supervision strategies. The Local Patch Network (LPNet) [41] uses supervised attention modules with local patch networks, guiding the model to focus on sparse target regions. While effective for extreme class imbalance and pseudo-target interference, its patch-based partitioning strategy shows limitations in modeling continuous structures. Although deep learning-based approaches have achieved substantial advancements in infrared ship detection, particularly in feature modeling, multi-scale perception, and suppressing complex backgrounds, challenges remain in recognizing low-resolution small targets and designing lightweight models. Further optimization of architectural design and training strategies is crucial to improve their practicality and generalization in real-world applications.

2.3. Small-Target Detection Methods Against Complex Backgrounds

In infrared remote sensing imagery, the detection of small targets is significantly challenged by complex backgrounds. These challenges arise primarily from three factors: the high similarity between background texture patterns and target characteristics, limited thermal radiation contrast, and indistinct target morphological properties. Early research endeavors concentrated on developing techniques to enhance multi-scale feature representation. Liu et al. [42] proposed Multi-Branch Parallel Feature Pyramid Networks (MPFPN) that employ parallel branch structures for deep feature restoration and incorporate a Supervised Spatial Attention Module (SSAM) to suppress background interference, thereby enhancing the representation of small target features. Yue et al. [43] presented an infrared small-target detection approach termed the YOLO-MST framework, which combines super-resolution enhancement techniques with multi-scale dynamic detection heads and incorporates a Multi-Scale Feature Augmentation (MSFA) module, demonstrating enhanced detection performance in densely populated target environments. Researchers have addressed semantic relationship enhancement by developing contextual modeling mechanisms. For example, Zhang et al. introduced an attention-guided pyramid context network (AGPCNet) [44] that utilizes an Attention-Guided Context Block (AGCB) and Context Pyramid Module (CPM) to model local and multi-scale semantic relationships, enabling effective extraction of small targets from complex background environments. Background suppression methodologies typically establish robust mechanisms for differentiation between foreground and background regions based on image contrast properties, structural characteristics, or statistical features. Zhou et al. [34] developed a dual-mode sea background model to address oceanic background variations. This approach implements scene classification based on global contrast parameters and utilizes background block filtering combined with region attribute analysis and local correlation features to achieve accurate target isolation. To address edge interference phenomena, Xin et al. [45] proposed a Superpixel Patch Image (SPI) model that generates background-edge-aligned superpixel patches for foreground–background differentiation and employs adaptive threshold mecha-

nisms for target identification. Feature extraction optimization approaches have focused on architectural improvements. Bao et al. [46] developed an improved dense nested attention network (IDNANet) that integrates Swin Transformer and ACmix attention architectures and implements a weighted dice loss function to address foreground–background sample imbalance issues. Despite the significant advancements achieved through these various enhancement strategies, existing infrared ship detection methods continue to face critical limitations, including insufficient integration of multi-scale features, computational inefficiency, and limited adaptability to diverse maritime conditions. These persistent challenges emphasize the necessity of developing more robust detection frameworks capable of effectively addressing complex maritime surveillance requirements.

3. Method

To address the core challenges of low contrast, background clutter, and the detection of small objects in infrared maritime imagery, we propose IRSD-Net. Built upon the YOLOv11n baseline, IRSD-Net is a lightweight and robust detection framework, as shown in Figure 1. Given an input infrared image, IRSD-Net first extracts hierarchical features using the HMKCNet. At the core of the HMKCNet is the Hierarchical Multi-Kernel Convolution (HMKConv), which performs parallel convolutions with multiple kernel sizes at each level. By integrating channel splitting and group convolution strategies, HMKConv effectively captures both local details and global context across scales while ensuring computational efficiency. The extracted multi-scale features are then passed to the DCSFPN. This network begins by aligning channel dimensions using 1×1 convolutions, producing consistent representations across different feature levels. Feature fusion is performed bidirectionally using BiFPN-inspired fusion modules. To further enhance scale adaptability, the DCSFPN incorporates a Multi-Scale Convolution Block (MSCBlock [47]) for kernel-wise adaptive processing, and an Efficient Up-Convolution Block (EUCBlock [47]) that employs depthwise convolutions and channel shuffle operations for lightweight upsampling. Finally, the fused feature maps are processed by multi-scale detection heads that predict object classes and bounding boxes. For precise localization, we introduce Wise-PIoU, a novel loss function that combines a geometric alignment penalty with dynamic attention-based weighting to refine regression gradients. This design significantly improves localization accuracy in challenging maritime scenes characterized by low contrast and small, ambiguous targets.

3.1. Problem Formulation

Detecting ships in infrared marine environments faces significant challenges, such as low contrast, small target size, and complex backgrounds. This study models infrared ship detection as a multi-scale feature extraction problem, enabling accurate localization of small targets. The input infrared image is represented by $X_{\text{input}} \in \mathbb{R}^{B \times C \times H \times W}$, where B is the batch size, C is the number of channels, and H and W denote the height and width of the image. The intensity values of the pixels fall within the range $[0, 1]$. Infrared images undergo progressive feature extraction through the HMKCNet, generating multi-scale feature representations across multiple stages: $F_P = \{F_{P1}, F_{P2}^E, F_{P3}^E, F_{P4}^E, F_{P5}^H\}$. In this context, F_{P1} denotes the output of the basic feature extraction, F_{Pi}^E represents the enhanced features, and F_{P5}^H indicates the features processed by the Multi-Kernel Convolutional Block (HMKCBlock). Our objective is to optimize small-target detection by extracting and enhancing multi-scale features. This joint optimization problem can be expressed as follows:

$$F_{P\_\text{fused}} = \Phi_{\text{DCSFPN}}\left(F_{P_3}^E \oplus F_{P_4}^E \oplus F_{P_5}^H\right) \tag{1}$$

Here, F_{P_fused} represents the final fused multi-scale feature maps obtained after processing through the DCSFPN, which combines enhanced features from different scales to form a comprehensive representation for accurate target detection. Φ_{DCSFPN} serves as the DCSFPN processing function, while $\oplus$ symbolizes the adaptive weighted fusion operation.

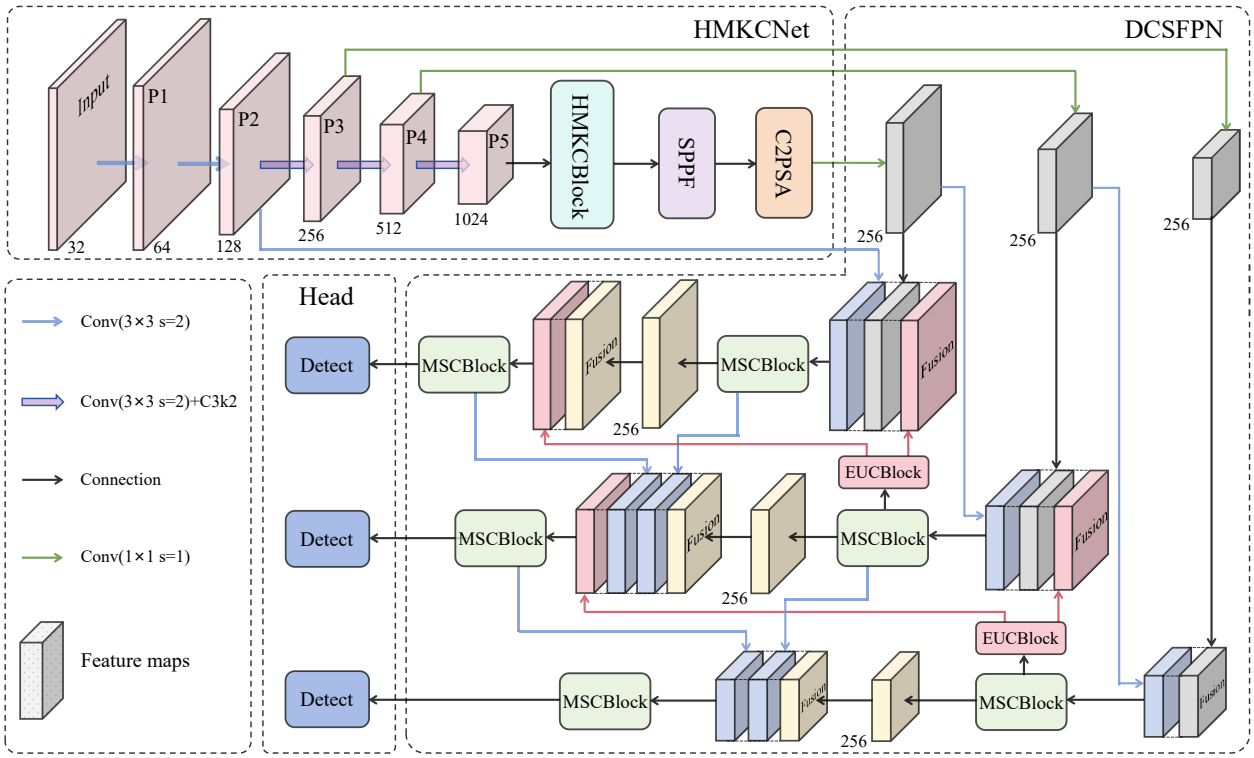

Figure 1. Network structure of IRSD-Net. Numbers represent the channel dimensions of feature maps at different network stages; C3k2 is the Cross Stage Partial bottleneck with kernel size 2 × 2 for efficient feature extraction; SPPF is the Spatial Pyramid Pooling Fast module for multi-scale feature aggregation; C2PSA is the Cross Stage Partial block with Position Self-Attention for enhanced feature representation.

3.2. HMKCNet

Traditional object detection models rely on stacked convolutional layers for feature extraction. These architectures suffer from a substantial computational burden and extensive parameter requirements, which severely limit processing speed and impede real-time capabilities. Moreover, such designs fail to effectively utilize the multi-scale information inherent in convolution operations, resulting in limited detection capabilities for objects of varying dimensions [48]. These limitations become particularly evident in infrared ship detection scenarios, where traditional methods struggle with scale variations and complex backgrounds [49]. To mitigate these challenges, we present the HMKCNet, to process richer hierarchical features with greater efficiency, as shown in Figure 2.

Inspired by the factorized convolutions [50] in Inception-v3 and the large kernel parameterization technique [51] from RepLKNet, we develop the HMKCBlock to enhance feature representation capabilities in this network. As the core component, HMKConv employs grouped convolutions to adapt to perception requirements for targets of various scales. It operates through a hierarchical kernel progression [1, 3, 5, 7] that captures complementary information across different scales. Unlike existing multi-kernel approaches that apply uniform kernel combinations regardless of target characteristics, the hierarchical kernel collaboration operates through strategic kernel distribution: the 1 × 1 kernel preserves fine-grained boundary details essential for precise small-target localization, the

3 × 3 and 5 × 5 kernels capture contextual features at medium scales that correspond to typical ship hull characteristics, and the 7 × 7 kernels integrate broader spatial context for robust target–background separation without amplifying noise interference.

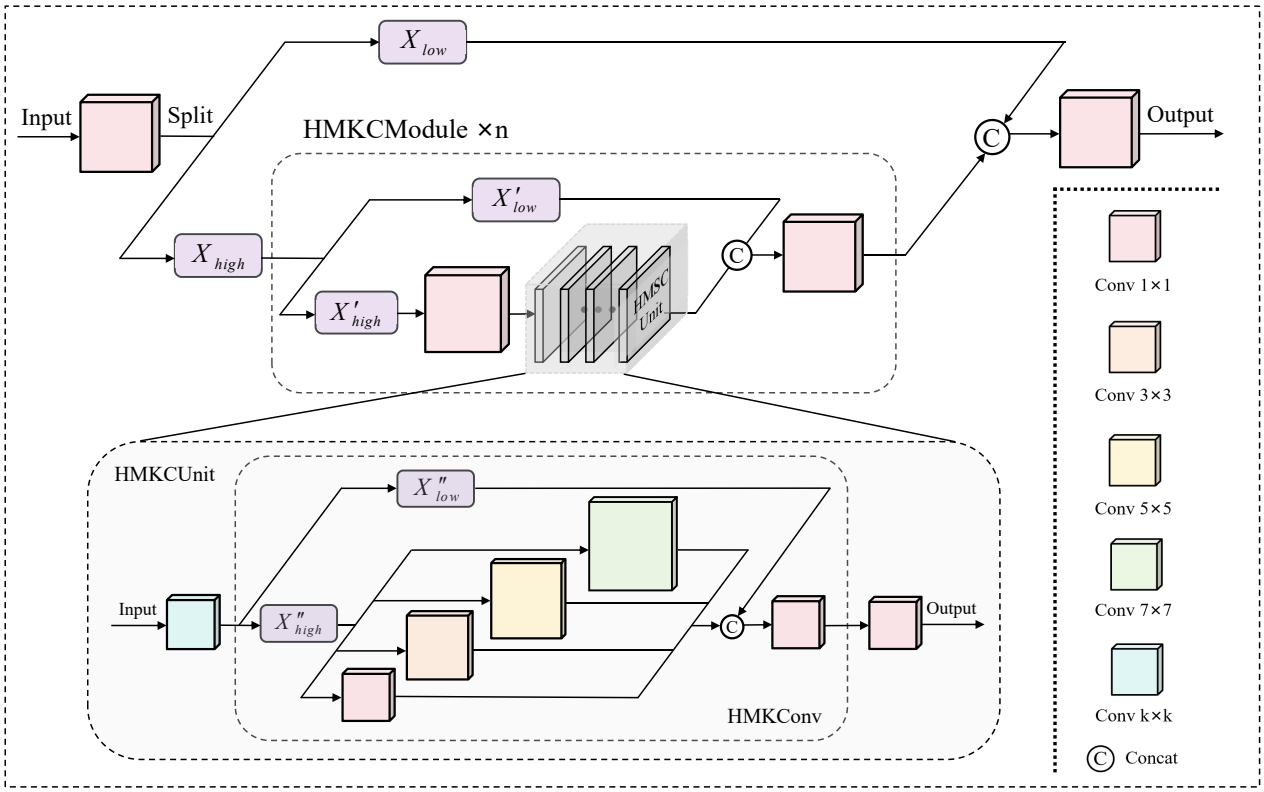

Figure 2. Structure of the HMKCBlock.

The HMKCBlock achieves robust cross-scale feature representation through the systematic integration of channel-partitioning mechanisms, multi-kernel collaboration strategies that establish diverse receptive fields, and hierarchical feature fusion operations for enhanced representational capacity. The core design adopts a proportional channel-splitting strategy that implements a dual-branch structure to balance computational efficiency and feature representation capability. Input feature channels are divided into two distinct pathways, namely, a lightweight pathway and a multi-kernel perception pathway. The lightweight pathway directly preserves original features to minimize computational overhead, while the multi-kernel perception pathway employs grouped spatial operations with varying kernel dimensions to capture multi-scale characteristics through the hierarchical kernel progression. This strategy enhances model generalizability and reduces computational requirements across diverse detection scenarios. To effectively integrate information from both branches, feature fusion is achieved through concatenation operations followed by 1 × 1 convolution for dimensional alignment. Additionally, residual connections are employed to combine the original input with processed results, ensuring the preservation of low-level detail information.

Specifically, this module is constructed upon a Cross Stage Partial Networks (CSPNet) [52] framework, which functions as the overall structural design for efficient feature processing across different stages. Input feature $F_{P5} \in R^{B\times C\times H\times W}$ is represented as X in Equation (2), with B, C, H, and W denoting the batch size, channel count, height, and width of the image, respectively. Following channel reduction and reorganization via

a 1×1 convolution, input features are divided according to the channel dimension into two sub-feature maps.

$$X_{low}, X_{high} = split(Conv_{1\times1}(X), [C/2, C/2], \text{dim} = 1) \tag{2}$$

where the $split(\cdot)$ operation partitions the feature tensor across the channel dimension into two equivalent parts, with dim $= 1$ indicating the second dimension of the tensor corresponding to the channel axis. Each part contains $C/2$ channels, which represents half of the original channel number C. The sub-feature map $X_{low} \in R^{B\times C/2\times H\times W}$ is utilized for feature concatenation to preserve the spatial and semantic information from the original input, minimizing feature loss. Meanwhile, the remaining sub-feature map $X_{high} \in R^{B\times C/2\times H\times W}$ undergoes multi-scale convolution processing through the HMKCModule.

In the dynamic path's feature processing, we further perform secondary channel-level splitting and processing on the input features below.

$$X'_{low}, X'_{high} = split(Conv_{1\times1}(X_{high}), [C'/2, C'/2], \text{dim} = 1) \tag{3}$$

where $C' = C/2$ represents the channel number of X_{high} from the first splitting operation. Then, the higher-level feature map X'_{high} from the dynamic processing branch enters the HMKCUnit after convolution processing. Each unit contains an embedded HMKConv, where input features undergo a custom $k \times k$ convolution. Within the HMKConv unit, similar channel-splitting and concatenation operations are performed. The feature map X'_{high} is further divided into X''_{high} and X''_{low}. Then, the $rearrange(\cdot)$ operation reorganizes the mixed $(G \cdot C_g)$ channels into G independent groups to meet the requirements of grouped convolutions, where $G = |K| = 4$ represents the number of kernel groups and $C_g = \frac{C''/2}{|K|}$ denotes the number of channels allocated to each scale group, with $C'' = C'/2 = C/4$. This results in a complete set of rearranged features X'''_{high}, which represents the rearranged feature map organized into G independent groups for grouped convolutions, as shown below:

$$X''_{low}, X''_{high} = split(Conv_{k\times k}(X'_{high}), [C''/2, C''/2], \text{dim} = 1) \tag{4}$$

$$X'''_{high} = rearrange(X''_{high}, B(G \cdot C_g)HW \rightarrow BC_gHWG, G = |K|), C_g = \frac{C''/2}{|K|} \tag{5}$$

Each subgroup is then processed through predefined convolution kernels $K = \{1, 3, 5, 7\}$. Subsequently, the rearranged features ${X'''_{high}}^{(k)}$ of each scale branch undergo independent convolution operations, where W_k represents the weights of the k-th convolutional kernel, producing outputs Y_k at different scales, as shown in the following equation:

$$\forall k \in K, \quad Y_k = \text{Conv}_{k\times k}({X'''_{high}}^{(k)}; W_k), \quad W_k \in R^{C_g\times C_g\times k\times k} \tag{6}$$

To effectively integrate multi-scale features, the output features $Y_k = [Y_1, Y_2, Y_3, Y_4]$ from different scale branches are first stacked along a new dimension and then rearranged through dimensional transformation. $stack(\cdot)$ denotes the operation that stacks tensors along a new dimension. This process converts the grouped tensor format $[G, B, C_g, H, W]$ into a channel-concatenated format $[B, G \cdot C_g, H, W]$, where G represents the number of kernel groups. The rearrangement operation forms a unified multi-scale feature representation Y_{high} that encapsulates contextual information from various scales.

$$Y_{high} = rearrange(stack([Y_k]), GBC_gHW \rightarrow B(G \cdot C_g)HW) \tag{7}$$

Then, Y_{high} and the identity path X''_{low} are concatenated along the channel dimension, where $concat(\cdot)$ represents the concatenation operation along the channel dimension. To

fuse these channel features, we employ pointwise convolution to integrate the various feature information. Specifically, a 1×1 convolution is utilized to exchange features extracted from each convolution, obtaining the final output F_{HConv} of the HMKConv. After another 1×1 convolution, F_{HUnit} is derived, which represents the output of the HMKCUnit, as shown below:

$$F_{HConv} = Conv_{1\times1}(concat(X''_{low}, Y_{high})) \tag{8}$$

$$F_{HUnit} = Conv_{1\times1}(F_{HConv}) \tag{9}$$

The dual-path structure subsequently concatenates and fuses the processed features F_{HUnit} with the identity path X'_{low}, forming $F_{HModule}$:

$$F_{HModule} = Conv_{1\times1}(concat(X'_{low}, Stack_n[F_{HUnit}])) \tag{10}$$

Through the cascading of multiple such modules, a deeper feature extraction network can be further formed. The block output F_{HBlock} is as follows:

$$F_{HBlock} = Conv_{1\times1}(concat(X_{low}, Stack_n[F_{HModule}])) \tag{11}$$

Complex backgrounds in infrared images, such as sea surfaces and sky regions, frequently interfere with accurate ship target recognition. The HMKCNet addresses this challenge by simultaneously extracting multi-scale features through parallel convolutions with varying kernel sizes, effectively capturing different receptive fields and enhancing discrimination between targets and complex backgrounds. The equal channel division strategy preserves robust feature representation while eliminating the computational redundancy and excessive overhead of typical deep convolutional networks. The HMKCNet maintains computational efficiency, satisfying real-time detection requirements without compromising detection accuracy for small ship targets in challenging environments.

3.3. DCSFPN

Our proposed DCSFPN draws inspiration from the design principles of the Bidirectional Feature Pyramid Network (BiFPN) [53]. The primary advantage of the BiFPN lies in its bidirectional information flow mechanism, which enables multi-level fusion across feature representations via both downward and upward pathways. This approach promotes efficient data propagation across multi-scale feature representations and effectively avoids the typical problems, including information loss and unidirectional flow, found in traditional feature fusion methods. As a result, it significantly enhances the ability to perceive global context [54]. In infrared ship detection, background noise can hinder detection performance, particularly under conditions where target–background contrast remains low [55,56]. Figure 3 demonstrates how the DCSFPN effectively combines low-level detail with high-level semantic information during both upsampling and downsampling processes. It also enhances the resolution of feature maps, enabling the network to capture global information with greater efficiency. This architecture shows improved performance in detecting targets that are low in contrast or located at a distance.

In detail, the input feature maps undergo initial processing through downsampling and three specialized convolutional layers. These layers systematically extract features from multi-scale feature maps and transform them into uniform-dimensional representations. Each layer's output subsequently undergoes downsampling operations, creating varied receptive fields that efficiently capture hierarchical image information across multiple scales. Specifically, F^E_{P2} from the shallowest layer P2 is directly downsampled, while F^E_{P3}, F^E_{P4}, and F^H_{P5} from layers P3, P4, and P5 undergo dedicated convolutional processing, which is also shown in Figure 1. The feature map P3, generated by shallower convolutional

layers, encapsulates larger-scale contextual information essential for scene understanding. The intermediate layer P4 captures medium-scale features that balance local and global information. The deeper convolutional layer P5 preserves fine-grained details critical for small-scale target detection. Each convolutional layer implements precisely calibrated kernel sizes and strides optimized for their respective scales. This methodical approach enables the network to extract scale-appropriate features and contextual information from feature maps across varying resolutions.

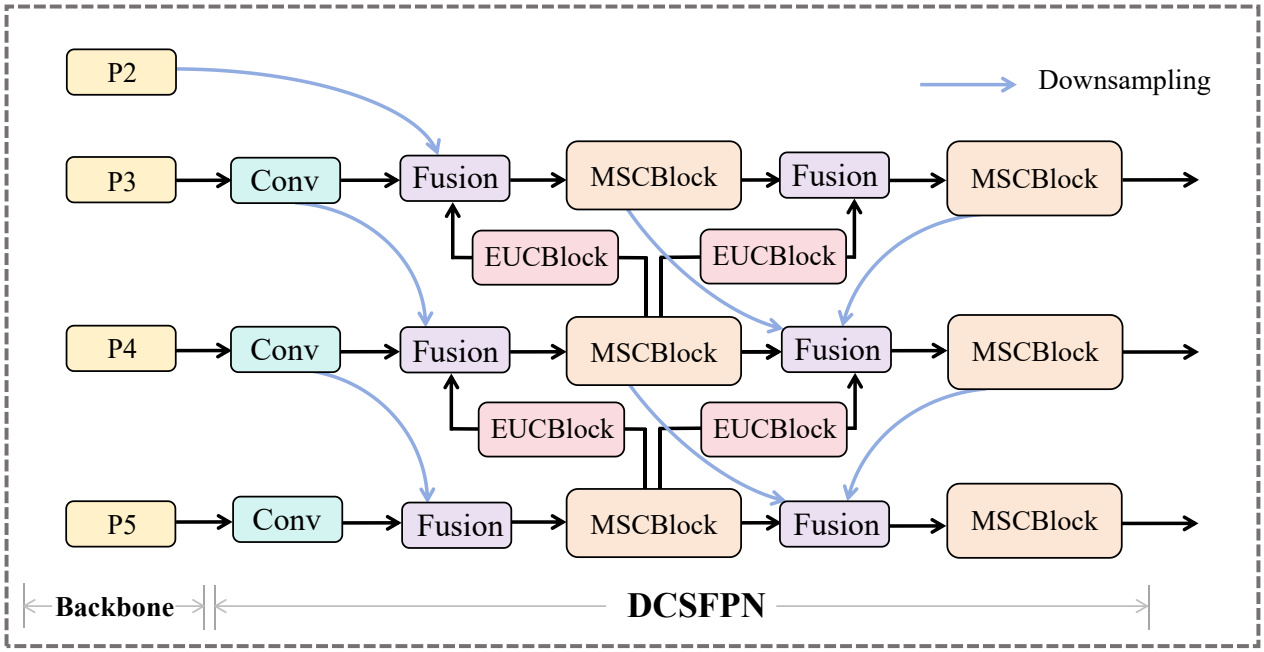

Figure 3. Structure of the DCSFPN.

To effectively integrate multi-scale feature information, we design a BiFPN-based feature fusion method called Fusion. This method can adaptively learn the importance of different feature maps, enabling more efficient feature aggregation. Traditional feature fusion methods usually employ simple element-wise addition or channel-wise concatenation, which cannot distinguish the importance of different feature maps, whereas the proposed fusion method employs an adaptive fusion strategy based on learnable weights.

At each fusion node in the DCSFPN, we collect feature maps from multiple sources: the current-level feature, downsampled features from lower levels, and features processed by the EUCBlock from higher levels. Given a set of input feature maps $\{F_1, F_2, \ldots, F_n\}$ at each fusion node, we assign a learnable weight parameter w_i to each feature map for adaptive aggregation. These parameters are automatically optimized during the network training process and are initialized as follows:

$$W = w_1, w_2, \ldots, w_n, \quad w_i = 1.0 \quad \forall i \tag{12}$$

where W represents the weight vector containing all learnable parameters, and w_i denotes the weight parameter associated with the i-th feature map, with each weight initially set to 1.0.

During training, these weights may become negative due to gradient updates. To ensure the stability of the fusion process, we further process these weights in the following steps. Initially, a Rectified Linear Unit (ReLU) [57] activation function is applied to the weights to prevent mutual cancellation between feature maps and to facilitate the preserva-

tion of information from each feature map. Subsequently, the weights are normalized to ensure that their sum equals one, as shown in the following equations:

$$w_i^+ = ReLU(w_i) = \max(0, w_i) \tag{13}$$

$$\hat{w}_i = \frac{w_i^+}{\sum_{j=1}^{n} w_j^+ + \varepsilon} \tag{14}$$

where w_i^+ represents the non-negative weight after ReLU activation, ensuring that all weights remain positive to prevent feature cancellation, $\hat{w}_i$ is the final normalized weight for the i-th feature map, $\sum_{j=1}^{n} w_j^+$ represents the sum of all non-negative weights, and ε is a small constant introduced to prevent division by zero and to ensure numerical stability. Finally, the fused feature map F_{fused} is obtained by performing a weighted summation:

$$F_{\text{fused}} = \sum_{i=1}^{n} \hat{w}_i \cdot F_i \tag{15}$$

This fusion method can adaptively adjust the contributions of different feature maps, which is particularly important for addressing common issues in infrared ship images, such as low contrast and complex backgrounds. The learnable weights enable the network to automatically emphasize the most informative features while suppressing less relevant information, improving target–background discrimination.

Subsequently, the MSCBlock further optimizes the feature fusion process through an effective residual connection mechanism. Within this module, parallel multi-scale convolution kernels extract diverse feature information simultaneously and significantly expand the network's receptive field capacity. Different-scale kernels focus on extracting local details and global contextual information. Through systematic layer-by-layer fusion of these features, the MSCBlock ensures complementarity between multi-scale features and effective information integration.

To complement the MSCBlock's multi-scale processing, the EUCBlock operates in the bottom-up pathway to enhance spatial resolution recovery. The EUCBlock collects gradient information from high-level feature maps via dense connections, enriching low-level representations with contextual information. This upsampling mechanism enhances inter-layer information flow and proves particularly effective for complex scenes, enabling precise target–background separation when handling multi-scale objects. Together, the MSCBlock and EUCBlock form a comprehensive bidirectional feature enhancement system within the DCSFPN.

3.3.1. Design of the MSCBlock

The MSCBlock serves as a pivotal part in the DCSFPN. It addresses the inherent limitations of traditional CNNs in multi-scale detection scenarios. Conventional networks with fixed-size kernels exhibit restricted perceptual capacity across varying object scales. For small-object detection, standard CNNs struggle to isolate small target features from surrounding contexts, which significantly compromises detection efficacy [58]. The MSCBlock resolves this constraint through its parallel multi-scale kernel implementation, effectively equipping the network with a larger receptive field capability, as shown in Figure 4a.

Specifically, the MSCBlock processes input feature maps through a systematic multi-scale feature extraction process, shown in Figure 4a. The module first expands the channel number of input feature maps F_{fused} through point-wise convolution (PWC_1, expansion factor = 2), which is followed by batch normalization (BN) [59] and a ReLU6 [60] activation layer in order to obtain enhanced feature representations. The resulting feature maps then undergo multi-scale depthwise convolution (MSDConv) [47] operations through

parallel processing branches where each branch utilizes depthwise convolutions with diverse kernel sizes to capture different receptive fields. The MSDConv operation can be expressed as follows:

$$MSDConv(X) = \text{concat}[G_p(X), G_q(X), G_s(X)] \tag{16}$$

$$G_k(X) = \text{ReLU6}(\text{BN}(\text{DWC}_k(X))) \tag{17}$$

where $X = \text{ReLU6}(\text{BN}(\text{PWC}_1(F_{\text{fused}})))$ represents the channel-expanded feature maps, and $G_k(\cdot)$ denotes a composite operator consisting of depthwise convolution (DWC), batch normalization, and ReLU6 activation. The kernel size k is selected from the predefined set $K_S = \{p, q, s\}$, with different combinations used at various network stages: $p = 1$, $q = 3$, $s = 5$ for lower stages to capture fine local features, $p = 3, q = 5, s = 7$ for intermediate layers to balance local and global information, and $p = 5, q = 7, s = 9$ for higher layers to expand the receptive field.

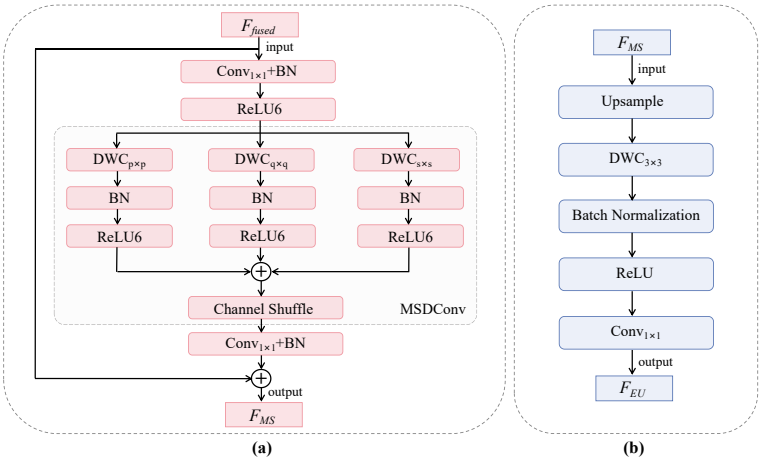

Figure 4. (**a**) MSCBlock, (**b**) EUCBlock.

However, parallel multi-scale convolution operations alone cannot fully prevent information isolation between different channels. To enhance inter-channel connectivity, the MSCBlock incorporates channel shuffle [61] operations to rearrange and group channels, enabling cross-channel feature interaction and improving the multi-scale feature integration ability. The feature map processed through the MSCBlock is as follows:

$$F_{MS} = \text{MSCBlock}(F_{fused}) = \text{BN}(\text{PWC}_2(\text{CS}(\text{MSDConv}(F_{fused})))) \tag{18}$$

where CS represents the channel shuffle operation, and PWC_2 adjusts the output to the required channel dimensions for downstream processing.

3.3.2. Design of the EUCBlock

In small-object detection, especially for infrared images, feature degradation during upsampling significantly limits detection performance. Standard upsampling methods increase the resolution but fail to preserve structural details, particularly for tiny objects with blurred edges, leading to feature attenuation and reduced model response. To address this, we develop the EUCBlock within the DCSFPN, which restores spatial resolution while minimizing information loss, as shown in Figure 4b.

The EUCBlock initially upsamples low-resolution feature maps to higher spatial resolutions, thereby enhancing the pixel-level representation for small objects. Afterwards, depthwise convolution (DWC) extracts local features independently per channel. This approach prevents inter-channel interference and enhances feature granularity. Additionally,

DWC reduces computational parameters, significantly improving target discriminability while maintaining efficiency.

Furthermore, to enhance feature map expressiveness, the EUCBlock implements channel shuffle technology similar to that implemented in the MSCBlock. This technique rearranges channel groups and enables cross-propagation of features between groups. The process strengthens cross-channel information flow and fusion while maintaining channel independence.

Finally, the EUCBlock adjusts the feature map channel count by a 1×1 convolution layer to maintain consistency with the subsequent network layers. Overall, the EUCBlock effectively addresses the limitations of traditional upsampling techniques in detail recovery and semantic preservation while remaining lightweight. This design significantly enhances the small-object modeling capability and detection accuracy. The output feature map is calculated as follows, with input F_{MS} from the MSCBlock:

$$F_{\mathrm{EU}} = \mathrm{EUCBlock}(F_{\mathrm{MS}}) = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{DWC}(\mathrm{Up}(F_{\mathrm{MS}}))))) \tag{19}$$

3.4. Wise-PIoU

In object detection, bounding box regression loss functions critically affect localization precision and convergence. Traditional IoU losses inadequately address geometric misalignment and dimensional variations, especially for small or irregular objects [62]. Therefore, we propose Wise-PIoU, which combines the geometric alignment penalty of Powerful-IoU (PIoU) [63] and the dynamic focusing mechanism of Wise-IoU [64], significantly enhancing regression performance.

PIoU enhances IoU by adding a geometric penalty term P that measures shape and position differences between predicted and target boxes, also addressing the boundary alignment insensitivity of the IoU. The penalty factor P depends solely on target box dimensions. The equation is as follows:

$$P = \frac{1}{4}\left(\frac{d_{w1}}{w_{gt}} + \frac{d_{w2}}{w_{gt}} + \frac{d_{h1}}{h_{gt}} + \frac{d_{h2}}{h_{gt}}\right) \tag{20}$$

where d_{w1}, d_{w2}, d_{h1}, and d_{h2} measure the boundary distance differences between predicted and target boxes in the horizontal and vertical directions. w_{gt} and h_{gt} denote the ground truth box width and height, respectively. This formulation enables balanced evaluation across targets of different sizes, avoiding the adverse effects of dimensional variations on model training.

Then, PIoU combines standard IoU loss with an additional geometric alignment quality penalty, expressed as follows:

$$f(P) = 1 - e^{-P^2} \tag{21}$$

$$PIoU = IoU - f(P) \tag{22}$$

$$L_{PIoU} = 1 - PIoU = L_{IoU} + f(P) \tag{23}$$

where L_{IoU} denotes the basic Intersection Over Union loss, assessing the overlap between predicted and target boxes, and $f(P)$ is a penalty function. The $1 - e^{-P^2}$ term enhances sensitivity to boundary alignment errors, notably when alignment quality is poor.

To enhance performance, PIoU v2 introduces a non-monotonic function $u(\cdot)$ that dynamically adjusts focus across anchor boxes of varying quality. By computing geometric penalty P, it defines an adjustment factor $q = e^{-P}$, reflecting the target prediction alignment quality. Smaller differences yield larger q values, prioritizing high-quality anchors, while larger differences produce smaller q values, reducing emphasis on low-quality anchors.

λ is a hyperparameter controlling the geometric alignment loss intensity, set to 1.3 in our implementation. Through this strategic allocation of training resources toward moderately aligned anchors, the regression process achieves more efficient convergence. The PIoU v2 loss function is defined as follows:

$$u(\lambda q) = 3 \cdot (\lambda q) \cdot e^{-(\lambda q)^2} \tag{24}$$

$$L_{PIoU\_v2} = u(\lambda q) \cdot L_{PIoU} \tag{25}$$

PIoU v2 effectively suppresses negative impacts from low-quality anchor boxes while accelerating optimization of medium-quality ones. Although PIoU v2 implements non-monotonic attention for handling anchor box quality variations, it demonstrates insufficient capability in dynamically modulating gradient gains for low-quality anchors. Wise-IoU v3 addresses this shortcoming through a dynamic non-monotonic focusing mechanism that evaluates anchor box outlier degrees. By dynamically adjusting gradient contributions based on these evaluations, Wise-IoU v3 successfully results in optimization processes with greater robustness.

The Wise-IoU loss function utilizes an outlier degree β to quantify sample quality by measuring the deviation between individual sample loss and global average loss values:

$$\beta = \frac{L^*_{IoU}}{L_{IoU}} \tag{26}$$

where L^*_{IoU} represents the IoU loss value that does not participate in backpropagation. L_{IoU} is the dynamically updated moving average value. According to this definition, when a predicted sample's error is significantly higher than the average value, its outlier degree β increases, indicating lower anchor box quality. The model will allocate higher attention to these samples.

To extend this foundation, a focus factor r is introduced to adaptively adjust the loss weight according to the value of β:

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}} \tag{27}$$

In this equation, α and δ serve as hyperparameters that regulate the focus factor, whereby the magnitude of loss weighting is modulated, with α = 1.7 and δ = 2.7 adopted in our method.

Wise-IoU incorporates distance attention using distance metrics, resulting in Wise-IoU v1 incorporating a dual-layer attention mechanism. R_{WIoU} is used to enhance IoU loss in average-quality anchor boxes, while L_{IoU} substantially reduces R_{WIoU} for high-quality anchor boxes. The mathematical expression is given as follows:

$$L_{WIoUv1} = R_{WIoU} \cdot L_{IoU} \tag{28}$$

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{29}$$

Here, the center coordinates of predicted and target bounding boxes are denoted by (x, y) and (x_{gt}, y_{gt}), respectively. W_g and H_g denote the width and height of the target's smallest bounding box, with these parameters being separated from the computational graph (*) to prevent R_{WIoU} from generating convergence-impeding gradients. By quantifying geometric disparities between predicted and target boxes, this formula enables dynamic loss adjustment, whereby anchor boxes with inferior alignment quality receive heightened attention during model training.

An outlier degree is subsequently employed to construct a non-monotonic focus coefficient, from which Wise-IoU v3 is derived:

$$L_{WIoUv3} = r \cdot L_{WIoUv1} = r \cdot (R_{WIoU} \cdot L_{IoU}) \tag{30}$$

Combining geometric penalties with a dynamic focusing mechanism, the final loss function for Wise-PIoU is as follows:

$$\begin{aligned} L_{Wise-PIoU} &= r \cdot (R_{WIoU} \cdot L_{PIoU\_v2}) \\ &= r \cdot \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \cdot \left(3 \cdot (\lambda q) \cdot e^{-(\lambda q)^2} \cdot L_{PIoU}\right) \end{aligned} \tag{31}$$

With this loss function, the model adaptively adjusts its focus on different samples, thereby enhancing localization precision, especially when target box position deviations are significant. Environmental factors such as background noise lead to geometric misalignment and localization errors. With Wise-PIoU, the model adaptively adjusts its focus on different samples, thereby enhancing localization precision, especially when target box position deviations are significant. This approach enhances the model's sensitivity to boundary alignment and enables more precise box adjustments. Additionally, the dynamic weighting mechanism reduces regression interference from inaccurate anchors while accelerating optimization for accurate ones. As a result, Wise-PIoU significantly improves detection precision and demonstrates enhanced robustness under complex environmental conditions.

4. Datasets And Experimental Setup

4.1. Datasets

This study utilized two high-quality public datasets, ISDD [65] and IRSDSS [66]. Constructed from Landsat 8 satellite remote sensing imagery, these datasets feature authentic data acquisition, high environmental complexity, and strong diversity. Representative examples are presented in Figure 5, with blue boxes indicating ships. Such characteristics effectively support performance evaluation for small-target detection under challenging conditions.

- ISDD: This collection comprises 1284 shortwave infrared images of 500 × 500 pixel resolution and 3061 annotated ship instances covering multiple regions. ISDD features small target scales, with an average target area occupying only 0.18% of the image area, diverse near-shore and open-ocean scenarios, and complex weather conditions including wind waves, thin clouds, and thick clouds. These characteristics impose higher requirements on detection networks for fine-grained modeling and robustness against complex backgrounds.
- IRSDSS: This collection includes 1491 infrared images of 640 × 640 pixel resolution, encompassing 4062 annotated ship targets. Ship lengths range from 3.32 to 85.88 m, with dimensions spanning from 16.28 to 4780.24 square meters and aspect ratios varying between 0.25 and 4.20, comprehensively covering ship instances of diverse scales and morphologies. Additionally, IRSDSS encompasses rich sea–land scenarios, varied wake features, and multiple kinds of weather interference like thin clouds, thick clouds, and wind waves, significantly increasing detection challenges.

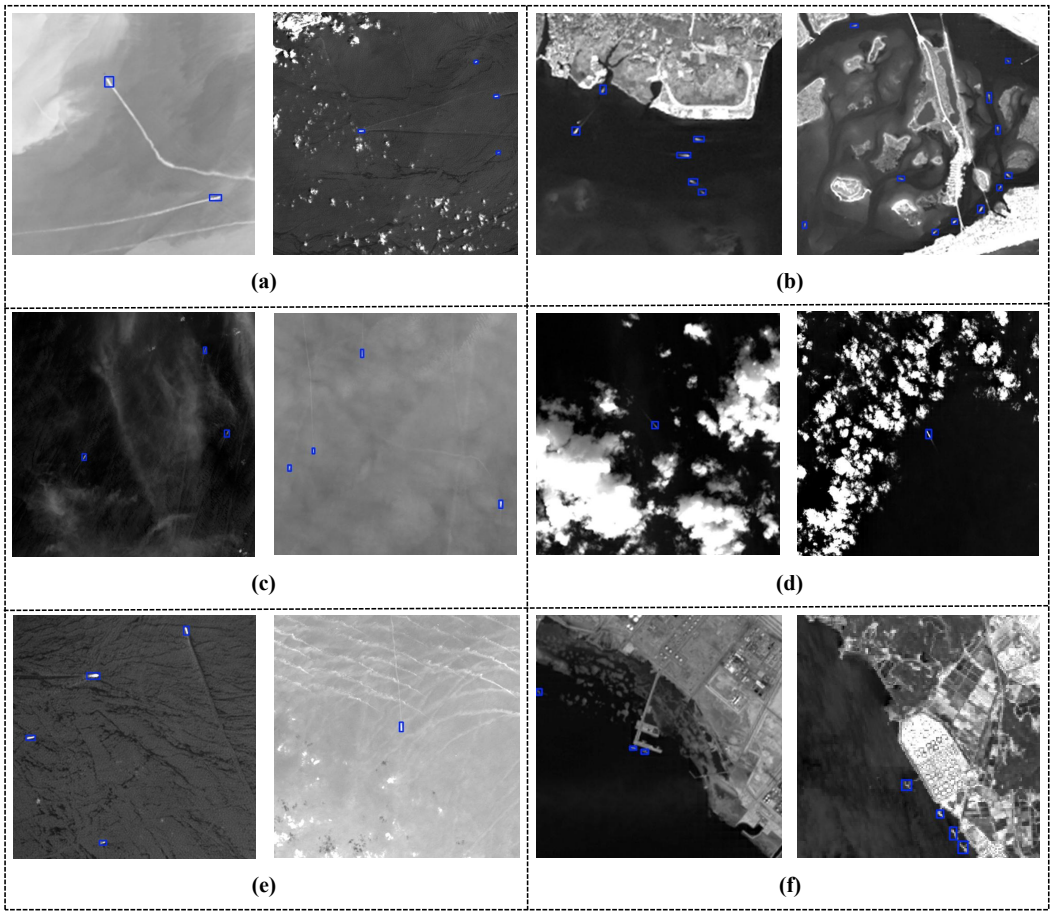

Figure 5. Scene diversity and weather condition diversity: (**a**) ships with trails; (**b**) inshore scenes; (**c**) thin clouds scenes; (**d**) thick clouds scenes; (**e**) sea wave scenes; and (**f**) berthing scenes. The blue box represents the ship.

4.2. Parameters and Settings

The experimental setup utilized a computational platform equipped with an NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) providing 24 GB of VRAM, alongside an Intel Xeon Platinum 8352V processor (Intel Corporation, Santa Clara, CA, USA) operating at 2.10 GHz across 16 cores, supported by 120 GB system memory. The computational environment was established on Ubuntu 20.04, implementing the software stack through Python 3.8 and PyTorch 1.10.0, with GPU computational acceleration facilitated by CUDA 11.3. Following the official dataset configurations, both datasets had a 7:3 split ratio for training and validation. Training configurations maintained consistent input image dimensions of 640 × 640 pixels. The experimental protocol employed a batch size of 32 samples across 150 training epochs. Optimization was achieved through Stochastic Gradient Descent (SGD), initialized with a learning rate of 0.01, a momentum parameter set to 0.937, and a weight decay coefficient of 0.0005. All additional hyperparameters retained their default configurations.

4.3. Evaluation Metrics

For comprehensive evaluation of the proposed detection approach, we utilized several standard assessment metrics: precision, recall, Giga Floating-Point Operations (GFLOPs), parameter count, and frames per second (FPS). These indicators provide a scientific analysis of model effectiveness and deployment feasibility across various aspects, encompassing detection precision, recall performance, computational overhead, and processing speed.

Specifically, the definitions and calculation formulas are as follows:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{32}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{33}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\mathrm{dR} \tag{34}$$

$$\mathrm{mAP} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathrm{AP}_i \tag{35}$$

Here, TP denotes correctly identified positive instances, FP denotes negative instances misclassified as positive, and FN denotes positive instances misclassified as negative by the model. Parameter N indicates the total class count, and AP_i represents the average precision for class i. We utilized mAP_{50} as our principal evaluation standard, calculated using an IoU threshold of 0.5. Additionally, following the COCO evaluation standard, we adopted scale-specific metrics including $\mathrm{AP_S}$@50 and $\mathrm{AP_M}$@50 to comprehensively assess detection performance across different target scales. $\mathrm{AP_S}$@50 represents the average precision for small targets (area < 32^2 pixels), while $\mathrm{AP_M}$@50 represents the average precision for medium targets ($32^2 \leq$ area < 96^2 pixels). It is worth noting that, due to the absence of ships in both the ISDD and IRSDSS datasets that conform to large-target standards, as stipulated by COCO, we have excluded the results related to $\mathrm{AP_L}$@50. GFLOPs quantifies the floating-point computations needed for single-pass inference, reflecting computational demands. FPS measures the processing rate in frames per second, where elevated values indicate superior real-time performance capabilities.

5. Results

5.1. Comparative Experiments

For evaluation of our proposed method's detection effectiveness, we conducted comprehensive benchmarking against multiple state-of-the-art real-time target detection algorithms. The comparative analysis utilized both ISDD and IRSDSS datasets, ensuring all approaches underwent evaluation with consistent training parameters to maintain experimental validity. Table 1 summarizes these comparative results.

As shown in Table 1, the proposed IRSD-Net model shows superior performance across various critical metrics compared to the representative baseline models, including SSD, Faster R-CNN, YOLOv3, YOLOv5n, YOLOv6n, YOLOv7-tiny, YOLOv8n, RT-DETR-l, YOLOv11n, YOLO-FIRI, Hyper-YOLO, DNA-Net, and ALCNet. On the ISDD dataset, it achieved substantial precision enhancements over conventional detectors: a 2.9% improvement relative to SSD, a 5.5% improvement over Faster R-CNN, and a 7.9% improvement over RT-DETR-l. In comparison to the YOLO series models, IRSD-Net showed a 2.5% improvement over YOLOv6n, a 2.3% improvement over YOLO-FIRI, and a 2.2% enhancement over YOLOv11n while maintaining competitive advantages against other variants. In comparison to non-YOLO models such as DNA-Net and ALCNet, the accuracy was enhanced by 1.6% and 1.9%, respectively. Its recall performance exhibited substantial advantages, particularly demonstrating a 9.5% improvement over RT-DETR-l and a 2.7% enhancement over YOLO-FIRI. Regarding the mAP_{50} metric, IRSD-Net surpassed all detector categories, with an 8.2% improvement over RT-DETR-l, a 3.2% improvement over YOLOv6n, and a 1.7% improvement over YOLOv11n. The comprehensive outperformance in terms mAP_{50-95} further confirms the model's capability in small-object detection and complex scene analysis.

Table 1. Overall comparison of experimental results for different models.

Dataset	Model	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)
ISDD	SSD [20]	89.3	80.4	88.4	41.5	81.1	88.5	87.5	24.4
	Faster R-CNN [18]	86.7	83.8	91.8	39.7	84.9	90.0	134	41.3
	YOLOv3 [67]	90.7	83.6	89.6	45.1	86.3	89.4	282.2	103.7
	YOLOv5n [68]	92.1	85.2	91.5	45.6	89.8	94.6	7.1	2.5
	YOLOv6n [69]	89.7	83.9	89.3	44.9	88.2	95.1	11.8	4.2
	YOLOv7-tiny [70]	91.8	81.1	90.0	42.3	87.9	93.1	13.2	6.0
	YOLOv8n [71]	92.1	84.8	91.9	45.2	90.5	94.7	8.1	3.0
	RT-DETR-l [72]	84.3	76.1	84.3	39.2	87.7	91.2	100.6	28.4
	YOLOv11n [73]	90.0	85.0	90.8	44.6	86.3	92.4	**6.3**	2.6
	YOLO-FIRI [74]	89.9	82.9	88.7	43.4	85.3	93.2	6.8	4.7
	Hyper-YOLO [75]	91.8	85.3	90.9	44.7	90.0	94.8	12.5	5.3
	DNA-Net [76]	90.6	84.6	91.7	42.5	90.1	92.7	44.0	10.5
	ALCNet [77]	90.3	81.7	89.1	40.9	89.7	94.9	68	14.7
	IRSD-Net	**92.2**	**85.6**	**92.5**	**45.7**	**91.2**	**96.8**	6.5	**2.1**
IRSDSS	SSD [20]	84.9	85.5	88.3	40.6	90.7	92.6	87.5	24.4
	Faster R-CNN [18]	88.9	87.3	91.8	39.7	89.1	93.2	134	41.3
	YOLOv3 [67]	92.0	88.1	92.2	43.5	91.0	92.0	282.2	103.7
	YOLOv5n [68]	91.9	**89.2**	92.8	42.8	92.8	95.9	7.1	2.5
	YOLOv6n [69]	91.7	86.9	91.8	43.3	91.9	96.4	11.8	4.2
	YOLOv7-tiny [70]	87.2	84.7	90.5	41.3	91.4	94.7	13.2	6.0
	YOLOv8n [71]	91.5	89.1	92.8	42.7	93.2	95.1	8.1	3.0
	RT-DETR-l [72]	83.8	83.5	86.3	38.0	92.7	94.8	100.6	28.4
	YOLOv11n [73]	91.3	87.9	91.8	42.6	92.9	94.5	**6.3**	2.6
	YOLO-FIRI [74]	90.8	88.0	91.9	41.1	93.8	95.7	6.8	4.7
	Hyper-YOLO [75]	91.9	86.7	92.1	40.9	94.2	96.1	12.5	5.3
	DNA-Net [76]	89.6	88.1	92.0	40.4	92.7	95.9	44.0	10.5
	ALCNet [77]	91.2	87.1	90.7	42.9	93.4	95.8	13.2	6.0
	IRSD-Net	**92.1**	88.7	**92.9**	**43.7**	**94.9**	**97.8**	6.5	**2.1**

Note: The best results are displayed in bold, and the second-best results are underlined.

Notably, the scale-specific performance analysis reveals IRSD-Net's exceptional capability in small-target detection. On the ISDD dataset, IRSD-Net achieved 91.2% AP_S@50, surpassing SSD, Faster R-CNN, YOLOv3, YOLOv5n, YOLOv6n, YOLOv7-tiny, YOLOv8n, RT-DETR-l, YOLOv11n, YOLO-FIRI, Hyper-YOLO, DNA-Net, and ALCNet by 10.1%, 6.3%, 4.9%, 1.4%, 3.0%, 3.3%, 0.7%, 3.5%, 4.9%, 5.9%, 1.2%, 1.1%, and 1.5%, respectively. For medium-scale targets, IRSD-Net attained 96.8% AP_M@50, demonstrating superior performance across all comparative methods with improvements of 8.3% over SSD and 6.8% over Faster R-CNN. These results validate the effectiveness of our multi-scale feature extraction approach in addressing the core challenge of small-target detection in infrared maritime imagery.

Validation on the IRSDSS dataset further substantiates IRSD-Net performance. The model achieved an optimal precision of 92.1% and recall of 88.7%, demonstrating clear advantages over the majority of comparative models. Although marginal differences existed in its recall performance compared to some YOLO variants, these variations are negligible. Its mAP_{50} reached 92.9% in the evaluation, achieving significant improvements of 6.6% over RT-DETR-l, 2.4% over YOLOv7-tiny, and 1.1% over YOLOv11n. The scale-specific analysis on the IRSDSS dataset shows that IRSD-Net achieved 94.9% AP_S@50, outperforming all comparative methods, and 97.8% AP_M@50, establishing a new state-of-the-art performance for both small- and medium-target detection. IRSD-Net attained 43.7% for the mAP_{50-95} metric across different scales. This establishes its optimal balance between detection accuracy and computational efficiency, rendering it particularly suitable for complex maritime surveillance applications.

The computational efficiency analysis on the ISDD dataset reveals the significant lightweight characteristics of IRSD-Net among contemporary detection methods. IRSD-Net required only 6.5 GFLOPs and 2.1 million parameters while achieving the highest mAP_{50}

of 92.5% and mAP_{50-95} of 45.7%. The HMKCNet reduces computational burden through hierarchical feature processing, where each HMKCUnit performs secondary channel-level splitting and reorganizes mixed channels into G independent groups, enabling efficient grouped convolutions that scale linearly. The lightweight design is further enhanced by the DCSFPN, which implements bidirectional feature fusion without traditional FPNs' parameter-heavy upsampling layers. In terms of computational complexity, IRSD-Net significantly outperforms existing methods. YOLOv3 necessitates 282.2 GFLOPs, RT-DETR-l requires 100.6 GFLOPs, and YOLOv7-tiny consumes 13.2 GFLOPs, all substantially exceeding IRSD-Net's requirements. Regarding parameter efficiency, IRSD-Net exhibits remarkable advantages with only 2.1 million parameters compared to the 103.7 million parameters of YOLOv3, the 28.4 million parameters of RT-DETR-l, and the 6.0 million parameters of YOLOv7-tiny. Even against the most recent YOLOv11n, with 6.3 GFLOPs and 2.6 million parameters, IRSD-Net achieves superior accuracy with reduced computational overhead.

On the IRSDSS dataset, IRSD-Net maintained consistent computational requirements of 6.5 GFLOPs and 2.1 million parameters while achieving 92.9% mAP_{50} and 43.7% mAP_{50-95}. Our method delivers superior computational efficiency with substantially reduced parameters compared to contemporary YOLO-based approaches. The computational analysis reveals that YOLOv5n, YOLOv8n, and YOLOv6n required 7.1 GFLOPs with 2.5 million parameters, 8.1 GFLOPs with 3.0 million parameters, and 11.8 GFLOPs with 4.2 million parameters, respectively, all exceeding the computational requirements of IRSD-Net. Enhanced variants including YOLO-FIRI and Hyper-YOLO exhibited even higher computational demands at 6.8 GFLOPs with 4.7 million parameters and 12.5 GFLOPs with 5.3 million parameters, respectively. In terms of CNN-based approaches, ALCNet required 13.2 GFLOPs and 6.0 million parameters, demonstrating that IRSD-Net achieves approximately half the computational cost. These results confirm that IRSD-Net achieves optimal computational efficiency while preserving detection accuracy, making it well-suited for practical maritime surveillance systems with limited computational resources.

5.2. Performance Comparison of Different Necks

For evaluating the effectiveness of the DCSFPN, we performed extensive benchmarking studies with several advanced neck architectures, including YOLOv11-Neck, BIFPN, GFPN, MAFPN, and HS-FPN. Consistent training procedures were applied across all models, where only the neck architecture differed while other components remained uniform. The experiments employed YOLO as the backbone network to ensure methodological rigor and experimental fairness. Systematic performance assessment was carried out using the ISDD and IRSDSS datasets. Table 2 displays the numerical outcomes from this comparative study.

Table 2. Overall comparison of experimental results for different necks.

Dataset	Model	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)
ISDD	YOLOv11-Neck [73]	91.3	85.3	91.2	45.6	90.4	94.2	6.3	2.5
	BIFPN [53]	90.6	85.8	91.4	45.1	90.9	94.5	6.2	1.9
	GFPN [78]	90.3	86.3	91.4	45.3	89.3	95.7	8.5	3.2
	MAFPN [79]	91.3	**86.8**	91.8	45.1	90.6	96.0	7.0	2.6
	HS-FPN [80]	91.9	85.5	91.6	45.6	90.2	95.9	**5.5**	**1.8**
	DCSFPN	**92.2**	85.6	**92.5**	**45.7**	**91.2**	**96.8**	6.5	2.1
IRSDSS	YOLOv11-Neck [73]	91.7	88.6	92.6	**43.7**	94.5	96.3	6.3	2.5
	BIFPN [53]	91.5	88.2	92.4	43.6	93.6	96.6	6.2	1.9
	GFPN [78]	91.6	**89.0**	92.8	43.3	93.9	95.8	8.5	3.2
	MAFPN [79]	92.0	88.6	92.3	43.6	94.3	97.2	7.0	2.6
	HS-FPN [80]	91.4	88.4	92.3	43.5	92.8	97.7	**5.5**	**1.8**
	DCSFPN	**92.1**	88.7	**92.9**	**43.7**	**94.9**	**97.8**	6.5	2.1

Note: The best results are displayed in bold, and the second-best results are underlined.

On the ISDD dataset, the DCSFPN demonstrated superior performance across multiple evaluation metrics, particularly excelling in precision, mAP_{50}, and mAP_{50-95} compared to other network architectures. Specifically, the DCSFPN achieved a precision of 92.2%, surpassing YOLOv11-Neck, BIFPN, GFPN, MAFPN, and HS-FPN by 0.9%, 1.6%, 1.9%, 0.9%, and 0.3%, respectively. For the mAP_{50-95}, the DCSFPN attained 45.7%, exhibiting notable improvements over comparative models. For different target scales, the DCSFPN demonstrated superior performance with 91.2% AP_S@50 for small targets and 96.8% AP_M@50 for medium targets, outperforming all comparative neck architectures. Notably, the DCSFPN surpassed the best-performing alternative, GFPN, by 1.9% in AP_S@50 and YOLOv11-Neck by 2.6% in AP_M@50. These results indicate that the DCSFPN possesses significant advantages in terms of precision and overall detection capability, particularly when processing complex infrared ship detection tasks, providing more reliable detection outcomes.

On the IRSDSS dataset, the DCSFPN similarly exhibited superior performance characteristics, achieving a precision of 92.1%. Regarding the mAP_{50} metric, the DCSFPN attained 92.9%, surpassing YOLOv11-Neck, BIFPN, GFPN, MAFPN, and HS-FPN by 0.3%, 0.5%, 0.1%, 0.6%, and 0.6%, respectively. In terms of target scale performance, the DCSFPN attained 94.9% AP_S@50 for small targets and 97.8% AP_M@50 for medium targets, demonstrating consistent superiority across all target scales with 0.4% and 0.6% improvements over the second-best method, respectively. The comparative analysis indicates that the DCSFPN establishes a more optimal equilibrium between precision and recall metrics relative to alternative neck architectures, thus demonstrating its exceptional comprehensive performance capabilities in complex detection scenarios.

In conclusion, the DCSFPN demonstrated statistically significant performance advantages across both the ISDD and IRSDSS datasets. These experimental results validate the efficacy of the DCSFPN in optimizing the neck structure and substantiate its superiority and feasibility for practical infrared ship detection applications.

5.3. Performance Comparison of Different Loss Functions

To analyze the performance of different loss functions in object detection, we conducted benchmarking analyses using CIoU, PIOU v2, Wise-IoU, MPDIoU, Inner-CIoU, and Wise-PIOU. Consistent training protocols were employed across all models, differing solely in the loss function implementation, using both the ISDD and IRSDSS datasets for assessment. Table 3 presents the comprehensive results.

Table 3. Overall comparison of experimental results for different loss functions.

Dataset	Loss Function	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)
ISDD	CIoU [81]	91.2	85.5	<u>91.1</u>	<u>45.1</u>	90.0	95.1	6.5	2.1
	PIOU v2 [63]	<u>91.3</u>	84.6	<u>91.1</u>	44.8	<u>90.1</u>	95.7	6.5	2.1
	Wise-IoU [64]	91.1	84.7	90.7	44.6	89.5	<u>96.3</u>	6.5	2.1
	MPDIoU [82]	88.2	**85.9**	90.6	44.9	90.0	95.7	6.5	2.1
	Inner-CIoU [83]	90.6	84.2	90.4	44.8	89.2	93.5	6.5	2.1
	Wise-PIoU	**92.2**	<u>85.6</u>	**92.5**	**45.7**	**91.2**	**96.8**	6.5	2.1
IRSDSS	CIoU [81]	91.5	87.2	<u>92.8</u>	42.3	93.2	95.8	6.5	2.1
	PIOU v2 [63]	91.3	87.7	91.9	43.5	<u>94.4</u>	95.9	6.5	2.1
	Wise-IoU [64]	91.0	87.8	92.0	<u>43.6</u>	94.1	96.2	6.5	2.1
	MPDIoU [82]	<u>92.0</u>	87.3	91.5	43.0	93.7	**97.9**	6.5	2.1
	Inner-CIoU [83]	91.9	<u>88.1</u>	92.3	43.1	92.9	96.6	6.5	2.1
	Wise-PIoU	**92.1**	**88.7**	**92.9**	**43.7**	**94.9**	<u>97.8</u>	6.5	2.1

Note: The best results are displayed in bold, and the second-best results are underlined.

On the ISDD dataset, the Wise-PIoU loss function exhibited excellent performance across various assessment metrics, especially precision and mAP_{50}, achieving 92.2% and 45.7%, respectively. Compared to other loss functions, Wise-PIoU improved precision by

1.0%, 0.9%, 1.1%, 4.0%, and 1.6% over CIoU, PIOU v2, Wise-IoU, MPDIoU, and Inner-CIoU, respectively. When examining different target scales, Wise-PIoU achieved superior results with 91.2% AP_S@50 for small targets and 96.8% AP_M@50 for medium targets. Notably, Wise-PIoU outperformed all comparative loss functions in small-target detection, achieving a 1.7% improvement over Wise-IoU and 2.0% improvement over Inner-CIoU in AP_S@50. The findings indicate that the Wise-PIoU loss function enhances detection precision while simultaneously improving overall detection capability, especially in complex infrared ship detection tasks.

On the IRSDSS dataset, the Wise-PIoU loss function similarly demonstrated outstanding performance, showing significant improvements in overall evaluation metrics compared to other loss functions. In terms of the mAP_{50-95} metric, Wise-PIoU achieved 43.7%, representing an improvement of 1.4%, 0.2%, 0.1%, 0.7%, and 0.6% over the comparative loss functions, respectively. Regarding target scale metrics, Wise-PIoU attained exceptional performance with 94.9% AP_S@50 for small targets and 97.8% AP_M@50 for medium targets. For small-target detection, Wise-PIoU surpassed the second-best method by 0.5%, while, for medium targets, it achieved comparable performance to MPDIoU but with superior overall metrics. These results highlight its advantages in low-contrast conditions and its detail extraction capabilities.

From a comprehensive perspective, the Wise-PIoU loss function significantly enhances detection precision, training convergence rate, and stability in infrared ship detection tasks through its optimization of target localization accuracy and improved small-object detection capabilities.

5.4. Ablation Experiments

5.4.1. Ablation Experiments on the HMKConv Kernel Configuration

To validate the effectiveness of different kernel combinations in the HMKCNet module, we conducted comprehensive ablation experiments with HMKConv on both the ISDD and IRSDSS datasets. The experiments evaluated various kernel configurations to determine the optimal hierarchical arrangement for infrared ship detection tasks, as shown in Table 4.

Table 4. Ablation experiment results for different kernel configurations in HMKConv.

Dataset	Kernel Configuration	mAP_{50} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)
ISDD	3 × 3	91.5	<u>91.1</u>	91.6	2.10	6.7
	[1, 3]	90.6	<u>91.1</u>	95.3	<u>2.09</u>	6.7
	[3, 5]	91.3	91.0	94.7	**2.07**	6.7
	[5, 7]	89.5	88.3	96.2	<u>2.09</u>	6.6
	[1, 3, 5]	90.2	90.8	92.1	**2.07**	6.6
	[3, 5, 7]	90.7	89.6	95.9	<u>2.09</u>	6.5
	[1, 3, 5, 9]	<u>92.1</u>	91.0	96.5	2.10	6.7
	[3, 5, 7, 9]	91.5	90.2	**97.8**	2.12	6.8
	[1, 3, 5, 7] (ours)	**92.5**	**91.2**	<u>96.8</u>	**2.07**	6.5
IRSDSS	3 × 3	89.4	90.8	89.1	2.10	6.7
	[1, 3]	89.0	92.0	91.5	<u>2.09</u>	6.7
	[3, 5]	90.7	91.4	92.3	**2.07**	6.7
	[5, 7]	90.5	91.6	96.6	<u>2.09</u>	6.6
	[1, 3, 5]	<u>91.7</u>	**92.2**	94.7	**2.07**	6.6
	[3, 5, 7]	91.2	88.7	95.2	<u>2.09</u>	6.5
	[1, 3, 5, 9]	91.5	91.8	96.9	2.10	6.7
	[3, 5, 7, 9]	91.3	90.1	<u>97.2</u>	2.12	6.7
	[1, 3, 5, 7] (ours)	**92.9**	<u>92.1</u>	**97.8**	**2.07**	6.5

Note: The best results are displayed in bold, and the second-best results are underlined.

The HMKConv module uses a hierarchical kernel arrangement [1, 3, 5, 7] to balance detailed target extraction and contextual understanding in infrared ship detection. The 1 × 1 kernel preserves pixel-level thermal gradients for distant vessels, the 3 × 3 kernel

captures spatial relationships for thermal boundaries, and the 5 × 5 and 7 × 7 kernels handle extended thermal signatures from engine heat and wake patterns. This progressive expansion provides necessary context while maintaining sensitivity to small targets.

The hierarchical integration uses channel splitting to distribute computational resources across different scales, preventing feature dilution of small targets. The experimental results show that the [1, 3, 5, 7] configuration achieved the optimal performance, with an mAP_{50} of 92.5% on the ISDD dataset and 92.9% on the IRSDSS dataset. Removing the 1 × 1 kernel degrades small-target detection, while kernels larger than 7 × 7 introduce background noise, as shown by the performance drops in the [1, 3, 5, 9] and [3, 5, 7, 9] configurations. The consistent performance across target scales confirms that the hierarchical design effectively captures infrared ship characteristics.

5.4.2. Hyperparameter Sensitivity Analysis of Wise-PIoU

To address concerns about hyperparameter selection in the Wise-PIoU loss function and demonstrate the robustness of our approach, we conducted a comprehensive sensitivity analysis on the key parameters that control the geometric alignment and dynamic focusing mechanisms. The Wise-PIoU loss function incorporates three critical hyperparameters: λ, controlling the geometric alignment loss intensity from the PIoU, and α and δ, regulating the dynamic non-monotonic focusing mechanism from the Wise-IoU.

As shown in Table 5, we systematically tested $\lambda \in [1.1, 1.3, 1.5]$ based on the PIoU's established optimal range [63]. For the focusing mechanism, $\alpha \in [1.5, 1.7, 1.9]$ and $\delta \in [2.5, 2.7, 3.0]$ were evaluated to cover conservative-to-aggressive focusing strategies, including the Wise-IoU optimal configuration [64]. The sensitivity analysis reveals that $\lambda = 1.3$, $\alpha = 1.7$, and $\delta = 2.7$ achieved superior performance with an mAP_{50} of 92.5% on the ISDD dataset and 92.9% on the IRSDSS dataset. This configuration demonstrates that moderate geometric alignment strength combined with balanced focusing parameters effectively allocates attention between high-quality and medium-quality anchor boxes without over-suppressing valuable training samples.

Table 5. Hyperparameter Sensitivity Analysis of Wise-PIoU.

Dataset	λ	α	δ	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)
ISDD	1.1	1.5	2.5	91.6	<u>85.4</u>	92.0	<u>45.5</u>
	1.1	1.7	2.7	91.5	85.3	<u>92.3</u>	44.7
	1.3	1.5	2.7	91.7	84.3	89.9	45.1
	1.3	1.7	2.5	<u>92.0</u>	85.1	92.1	44.8
	1.3	1.7	3.0	91.7	85.2	92.0	45.2
	1.3	1.9	3.0	90.8	84.7	91.4	45.0
	1.5	1.7	2.7	91.9	85.3	92.0	45.2
	1.3	1.9	3.0	90.2	84.2	91.7	44.3
	1.3 (ours)	1.7 (ours)	2.7 (ours)	**92.2**	**85.6**	**92.5**	**45.7**
IRSDSS	1.1	1.5	2.5	90.5	88.6	92.1	42.1
	1.1	1.7	2.7	91.7	88.2	91.8	<u>43.6</u>
	1.3	1.5	2.7	90.8	**89.2**	92.2	42.0
	1.3	1.7	2.5	91.4	87.3	92.5	43.5
	1.3	1.7	3.0	91.6	88.4	92.0	42.6
	1.3	1.9	3.0	<u>91.9</u>	87.9	91.4	42.9
	1.5	1.7	2.7	91.4	88.5	<u>92.7</u>	41.8
	1.3	1.9	3.0	91.1	87.4	91.2	41.4
	1.3 (ours)	1.7 (ours)	2.7 (ours)	**92.1**	<u>88.7</u>	**92.9**	**43.7**

Note: The best results are displayed in bold, and the second-best results are underlined.

5.4.3. Ablation Experiments on the Overall Model

To examine how the HMKCNet, DCSFPN components, and the Wise-PIoU loss function contribute to the enhancement of the baseline model, we performed comprehensive ablation studies. These investigations measured the effect of individual components and

their combinations across essential evaluation criteria, including precision, recall, mAP_{50}, mAP_{50-95}, AP_S@50, AP_M@50, GFLOPs parameters, and FPS through progressive incorporation. Table 6 presents the findings from the ISDD dataset analysis.

Table 6. Overall ablation experiment results for the ISDD dataset.

Original	HMKCNet	DCSFPN	Wise-PIoU	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)	FPS
✓				90.0	85.0	90.8	44.6	86.3	92.4	6.3	2.6	**909.1**
✓	✓			91.6	<u>85.9</u>	<u>92.3</u>	45.1	88.5	93.9	6.3	2.5	<u>833.3</u>
✓		✓		<u>92.1</u>	85.2	91.3	45.0	90.1	93.1	6.5	2.1	769.2
✓			✓	90.4	**86.1**	91.2	45.5	89.4	95.3	6.3	2.6	<u>833.3</u>
✓	✓	✓		91.2	85.5	91.1	45.1	90.0	95.1	6.5	2.1	625.0
✓	✓		✓	91.3	85.3	91.2	<u>45.6</u>	<u>90.9</u>	94.2	6.3	2.5	**909.1**
✓		✓	✓	91.5	85.2	91.7	45.3	90.5	<u>95.6</u>	6.5	2.1	714.3
✓	✓	✓	✓	**92.2**	85.6	**92.5**	**45.7**	**91.2**	**96.8**	6.5	2.1	714.3

Note: The best results are displayed in bold, and the second-best results are underlined. Checkmarks (✓) indicate which modules were enabled in each experiment.

Upon integration of the HMKCNet alone, the model exhibited substantial performance improvements across multiple metrics. HMKCNet's parallel multi-kernel architecture effectively captures multi-scale features through diverse receptive fields, enhancing small-target discrimination on complex backgrounds. Precision increased from 90.0% to 91.6%, recall improved to 85.9%, and mAP_{50} reached 92.3%, representing a 1.5 percentage point enhancement compared to the baseline, while mAP_{50-95} advanced to 45.1%. Concerning target scale performance, the HMKCNet demonstrated significant improvements, with AP_S@50 increasing from 86.3% to 88.5% for small targets and AP_M@50 improving from 92.4% to 93.9% for medium targets. These improvements demonstrate the HMKCNet's efficacy in enhancing identification precision, especially for small objects and complex entities in challenging scenarios. The channel-splitting strategy optimizes computational efficiency while preserving critical feature information. Computational overhead remained modest, with GFLOPs maintained at 6.3, the parameter count slightly reduced to 2.5 M, and inference speed decreased to 833.3 FPS, establishing a favorable trade-off between computational cost and performance enhancement.

Isolated introduction of the DCSFPN further enhanced performance. Precision increased to 92.1%, showing a 2.1 percentage point enhancement compared to the baseline, while recall reached 85.2%, mAP_{50} reached 91.3%, and mAP_{50-95} remained at 45.0%. The bidirectional fusion mechanism of the DCSFPN addresses traditional FPN limitations, enabling superior cross-scale information integration that particularly excels in small-target detection. Across target size categories, the DCSFPN achieved 90.1% AP_S@50 for small targets while maintaining strong medium-target performance at 93.1% AP_M@50, which can be attributed to its adaptive weighted fusion which automatically learns feature importance for effective multi-scale aggregation. Although GFLOPs slightly increased to 6.5, inference speed remained at 769.2 FPS, indicating real-time performance preservation.

Following Wise-PIoU integration, bounding box regression accuracy improved. Precision showed a slight enhancement, while recall increased to 86.1%, representing a 1.1 percentage point improvement compared to the baseline. Additionally, mAP_{50} reached 91.2%, and mAP_{50-95} improved to 45.5%. When analyzing performance across different target scales, Wise-PIoU achieved 89.4% AP_S@50 for small targets and 95.3% AP_M@50 for medium targets, demonstrating balanced improvements through its outlier degree weighting mechanism that suppresses low-quality anchors while accelerating high-quality sample optimization.

The experimental results demonstrate that combining the HMKCNet, DCSFPN, and Wise-PIoU significantly enhances overall model performance. The integration of the HMKCNet with the DCSFPN achieved 91.2% precision, 85.5% recall, and 91.1% mAP_{50}, and improved mAP_{50-95} from 44.6% to 45.1%. This improvement highlights how the HMKCNet's enhanced deep feature extraction, when combined with the DCSFPN's multi-scale feature fusion, substantially improves detection capabilities for small objects and complex backgrounds. The combination of the HMKCNet and Wise-PIoU further optimized performance, yielding 91.3% precision and 91.7% mAP_{50}. When pairing the DCSFPN with Wise-PIoU, the model achieved 91.5% precision, 85.2% recall, 91.7% mAP_{50}, and 45.3% mAP_{50-95}. This configuration excels in small-object detection and complex background processing. The complementary effects enhance detection capabilities across objects of varying scales. Despite increased computational complexity with 6.5 GFLOPs and 2.1 M parameters, the model maintained high FPS. This demonstrates its excellent real-time performance alongside improved accuracy. These pairwise integrations achieved substantial improvements in both small- and medium-target detection, with small-target performance reaching 90.0%, 90.9%, and 90.5%, respectively, and medium-target performance reaching 95.1%, 94.2%, and 95.6%, respectively, demonstrating that each pairing exhibits distinct strengths in addressing different aspects of infrared ship detection challenges.

When combining thr HMKCNet, DCSFPN, and Wise-PIoU, the model achieved optimal performance. Precision increased to 92.2%, recall reached 85.6%, mAP_{50} rose to 92.5%, and mAP_{50-95} reached 45.7%. The comprehensive integration achieved outstanding scale-specific performance with 91.2% AP_S@50 for small targets and 96.8% AP_M@50 for medium targets, representing the highest performance across all target scales. Despite a slight increase in computational overhead with 6.5 GFLOPs, 2.1 M parameters, and an FPS slightly decreasing to 714.3, the model maintained excellent real-time performance while improving detection precision and recall. The synergistic effect of these components enhances the model's accuracy and adaptability across various detection tasks, validating their crucial role in complex scenarios.

In Table 7, the results on the IRSDSS dataset demonstrate improvements. The HMKCNet optimizes deep feature extraction, delivering enhanced detection performance. The model demonstrated strong capabilities, with precision and recall reaching 91.4% and 88.5%, respectively, while mAP_{50} and mAP_{50-95} reached 92.4% and 42.7%. Compared to the baseline, this represents a notable 0.6% improvement in both recall and mAP_{50}. For small-target detection, the HMKCNet achieved 91.4% AP_S@50.

Table 7. Overall ablation experiment results for the IRSDSS dataset.

Original	HMKCNet	DCSFPN	Wise-PIoU	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)	FPS
✓				91.3	87.9	91.8	42.6	91.3	94.5	6.3	2.6	**1111.1**
✓	✓			91.4	88.5	92.4	42.7	91.4	95.4	6.3	2.6	833.3
✓		✓		91.9	87.4	92.3	43.2	91.9	94.7	6.5	2.1	666.7
✓			✓	91.2	**89.1**	92.1	<u>43.5</u>	91.2	95.9	6.3	2.6	<u>909.1</u>
✓	✓	✓		91.5	87.2	<u>92.8</u>	42.3	91.5	95.8	6.5	2.1	625.0
✓	✓		✓	91.7	88.6	92.6	**43.7**	91.7	96.3	6.3	2.5	<u>909.1</u>
✓		✓	✓	<u>92.0</u>	88.3	92.4	42.9	<u>92.0</u>	<u>96.7</u>	6.5	2.1	769.2
✓	✓	✓	✓	**92.1**	<u>88.7</u>	**92.9**	**43.7**	**92.1**	**97.8**	6.5	2.1	769.2

Note: The best results are displayed in bold, and the second-best results are underlined. Checkmarks (✓) indicate which modules were enabled in each experiment.

The introduction of the DCSFPN brings further enhancements through multi-scale feature fusion. Precision increased to 91.9% alongside an 87.4% recall, while mAP_{50} reached 92.3% and mAP_{50-95} rose to 43.2%. Medium-target performance demonstrated particular strength with 94.7% AP_M@50, validating the DCSFPN in terms of its effectiveness in

multi-scale feature processing. Despite the computational cost increasing to 6.5 GFLOPs, the DCSFPN successfully enhanced both localization precision and detection accuracy. Meanwhile, Wise-PIoU integration elevated model performance by optimizing bounding box regression. This resulted in balanced metrics of 91.2% precision and 89.1% recall, with mAP_{50} and mAP_{50-95} reaching 92.1% and 43.5%, respectively. Wise-PIoU attained an AP_M@50 of 95.9% for medium objects, exceeding the baseline by 1.4% and thereby demonstrating improved localization accuracy.

When combining the HMKCNet with the DCSFPN, precision reached 91.5% and recall 87.2%. The mAP_{50} increased to 92.8% while mAP_{50-95} rose to 42.3%. This combination leverages advantages from both deep feature extraction and multi-scale information fusion, performing exceptionally well for small objects and complex backgrounds. Despite a computational cost of 6.5 GFLOPs and 2.1M parameters, real-time performance remained excellent, with a frame rate of 625.0. By integrating the HMKCNet with Wise-PIoU, the model achieved better performance in precision and recall, 91.7% and 88.6%, respectively. The mAP_{50} improved from 91.8% to 92.6%, and mAP_{50-95} increased from 42.6% to 43.7%. This pairing not only optimizes feature extraction but also enhances localization precision, showing significant effectiveness in improving recall and handling complex targets. With 6.3 GFLOPs computational cost and 2.5M parameters, real-time performance remained at a high level. After combining the DCSFPN with Wise-PIoU, the model exhibited a higher precision of 92.0% and a recall of 88.3%, with an mAP_{50} at 92.4% and an mAP_{50-95} reaching 42.9%. These combinations consistently achieved superior performance compared to the baseline across both small- and medium-target detection metrics.

When jointly using the three components, the model achieved the optimal comprehensive performance. Precision reached 92.1% and recall 88.7%, while mAP_{50} and mAP_{50-95} reached 92.9% and 43.7%, respectively. The complete integration achieved the highest performance, with AP_S@50 reaching 92.1% for small targets and AP_M@50 reaching 97.8% for medium targets across all target scales. Although the computational cost was 6.5 GFLOPs with 2.1 M parameters and the frame rate slightly decreased to 769.2, the model demonstrated exceptional advantages in complex object detection tasks with significant performance improvements.

5.5. Visual Presentation

To facilitate better assessment of the detection performance of the model, we performed a comprehensive visual analysis of the target objects.

In Figure 6, the infrared ship detection performance across diverse maritime environments achieved using RT-DETR-l, YOLOv8n, YOLOv11n, YOLO-FIRI, DNA-Net, and our proposed model is illustrated. Our approach exhibited excellent capabilities relative to all other comparative methods, particularly excelling against complex maritime backgrounds. The comparative analysis shows that existing models face challenges when operating under adverse imaging. These models exhibit significantly reduced confidence scores and inferior localization precision in complex or wave-dominated scenes, leading to incorrect identifications and detection failures under high-interference conditions.

Conversely, our proposed model enhances detection performance by integrating the HMKCNet and DCSFPN to optimize the receptive fields and multi-scale feature integration. Incorporating Wise-PIoU further enhances target localization stability while strengthening cross-scale information exchange. The visual comparison reveals that our model effectively addresses diverse, challenging scenarios with enhanced feature extraction, resulting in significantly improved ship localization precision. Our heterogeneous convolution kernel-based approach mitigates interference effects and yields higher confidence detections. The model achieves precise localization due to a reinforced feature hierarchy information flow

and effective background interference suppression, thus enabling superior small-target detection under adverse conditions.

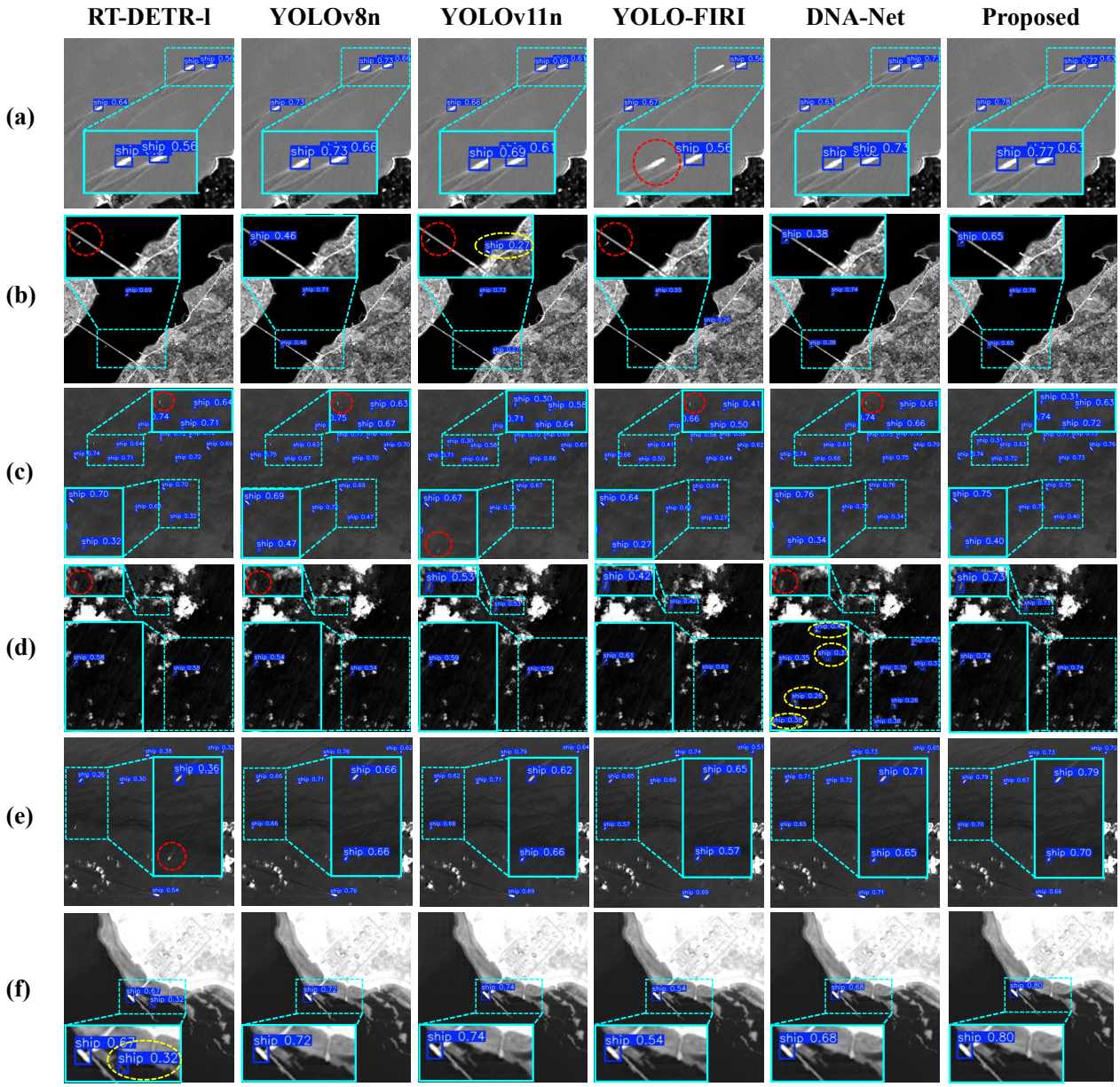

Figure 6. Visual comparison of detection results across different models. (**a**) Ships with trails; (**b**) inshore scene; (**c**) thin clouds scene; (**d**) thick clouds scene; (**e**) sea wave scene; and (**f**) berthing scene. The blue boxes are magnified views of specific areas in the original image. False positives and false negatives are marked with yellow and red circles, respectively.

Figure 7 illustrates performance differences among various models for small ship target detection under different background conditions. The heatmap colors represent model confidence levels. Warmer colors like red and yellow signify elevated confidence, while cooler colors like blue and green correspond to reduced confidence or background regions. As depicted in Figure 7, YOLOv11n exhibits limitations when processing complex backgrounds, presenting scattered response regions and relatively low confidence scores across diverse scenarios. These shortcomings are particularly evident in low-contrast areas, resulting in inaccurate localization and target detection failures. To mitigate these issues, the HMKCNet was incorporated. The HMKCNet generates expanded response regions through enhanced feature extraction capabilities. Although the response areas become

broader, this modification improves the model's sensitivity to ship targets, effectively capturing more comprehensive target features despite the increased spatial coverage. Subsequent integration with the DCSFPN leads to more compact and precise response regions. This component facilitates cross-scale information transfer and refines the features expanded by the HMKCNet, thereby optimizing detection accuracy across various scales. As a result, background interference is effectively diminished in the heatmap, showcasing the enhanced adaptability in complex environments. The proposed IRSD-Net demonstrates clear superiority across all scenarios, characterized by highly concentrated response regions and consistently elevated confidence scores. This integrated approach effectively manages complex backgrounds, low-contrast conditions, and small targets, delivering superior localization accuracy and diminished false detection rates.

5.6. Generalization Analysis

5.6.1. Cross-Model Validation

To validate the robustness and generalizability of our proposed approach across different network architectures, we conducted comprehensive experiments by applying IRSD-Net to larger YOLO variants, specifically, YOLOv11s. Significant architectural differences exist between YOLOv11n and YOLOv11s: YOLOv11n, as the Nano version, features 2.6 M parameters and 6.3 GFLOPs, with a focus on extreme lightweight design, whereas YOLOv11s, as the Small version, scales up to 9.4 M parameters and 21.5 GFLOPs, incorporating deeper network structures and enhanced feature representation capabilities. This cross-architectural migration experiment aimed to demonstrate that the performance improvements achieved by IRSD-Net represent fundamental enhancements rather than architecture-specific optimizations. The comprehensive results are presented in Table 8.

The experimental results demonstrate consistent performance improvements across both backbone network variants. On the ISDD dataset, with the progressive integration of the HMKCNet, DCSFPN, and Wise-PIoU modules, the YOLOv11s-based model performance exhibited a clear incremental trend, advancing from the baseline mAP_{50} of 91.8% to 92.3%, 92.7%, and, ultimately, 93.1%. Correspondingly, the parameter count stabilized at 7.5 M from the baseline 7.9 M, while GFLOPs increased marginally from 22.0 to 22.3. Similar improvement patterns were observed for the IRSDSS dataset, with mAP_{50} progressively advancing from the baseline of 92.9% to a final value of 93.7%. Although the improvement margins were slightly reduced in larger network variants, attributable to the diminishing returns phenomenon inherent in more complex baseline models, the persistent positive improvements across all configurations substantiate that IRSD-Net addresses the intrinsic challenges in maritime infrared target detection rather than merely compensating for lightweight architecture limitations, thereby validating the robustness and practical applicability of our approach.

Table 8. Generalization analysis of YOLOv11s.

Dataset	Model Configuration	Precision (%)	Recall (%)	mAP_{50} (%)	mAP_{50-95} (%)	AP_S@50 (%)	AP_M@50 (%)	GFLOPs	Parameters (M)
ISDD	Baseline	91.5	87.4	91.8	45.4	90.2	91.8	22.0	7.9
	+ HMKCNet	92.1	<u>88.2</u>	92.3	45.3	90.9	92.8	22.0	7.9
	+ HMKCNet + DCSFPN	<u>92.7</u>	88.1	<u>92.7</u>	<u>45.8</u>	<u>91.4</u>	**93.9**	22.3	7.5
	+ HMKCNet + DCSFPN + Wise-PIoU	**93.3**	**88.4**	**93.1**	**46.1**	**92.3**	<u>93.6</u>	22.3	7.5
IRSDSS	Baseline	91.8	88.3	92.9	44.8	92.3	95.6	22.0	7.9
	+ HMKCNet	92.0	88.9	93.2	<u>45.7</u>	92.5	97.3	22.0	7.9
	+ HMKCNet + DCSFPN	<u>92.6</u>	<u>89.2</u>	<u>93.4</u>	44.9	<u>92.7</u>	<u>97.6</u>	22.3	7.5
	+ HMKCNet + DCSFPN + Wise-PIoU	**92.9**	**89.6**	**93.7**	**45.9**	**92.9**	**97.9**	22.3	7.5

Note: The best results are displayed in bold, and the second-best results are underlined.

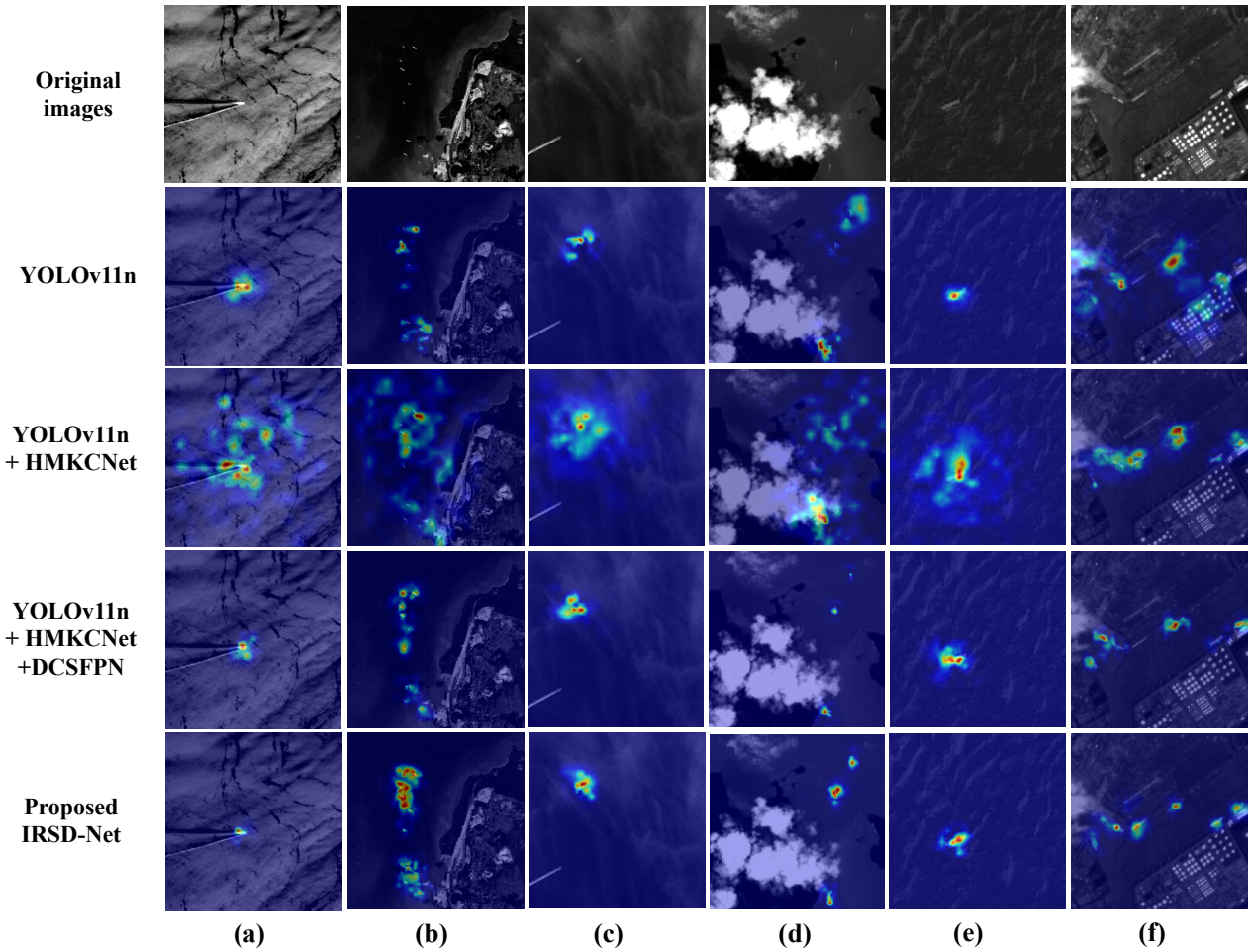

Figure 7. Heatmap comparison of detection accuracy across four networks. (**a**) Ships with trails; (**b**) inshore scene; (**c**) thin clouds scene; (**d**) thick clouds scene; (**e**) sea wave scene; and (**f**) berthing scene.

5.6.2. Cross-Platform Performance Analysis

To assess the practical performance of IRSD-Net across diverse computational environments, we conducted comprehensive benchmarking experiments on three representative computing platforms. The evaluation encompassed a high-performance RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), mid-range RTX 3060 GPU (NVIDIA Corporation, Santa Clara, CA, USA), and Intel i7-12700K CPU (Intel Corporation, Santa Clara, CA, USA), representing basic computational environments. All evaluations utilized a 640 × 640 input resolution, with memory consumption measured during actual inference operations. The experimental outcomes demonstrate robust cross-platform adaptability, as detailed in Table 9.

The experimental outcomes demonstrate robust cross-platform adaptability. On the RTX 4090 GPU, IRSD-Net achieved 714.3 FPS and 769.2 FPS on the ISDD and IRSDSS datasets, respectively, delivering accuracy improvements of 1.7 and 1.1 percentage points over YOLOv11n. The RTX 3060 GPU testing maintained an efficient performance with 242.5 FPS and 238.2 FPS, preserving accuracy advantages despite modest speed decreases compared to YOLOv11n's 260.5 FPS and 270.3 FPS. The CPU evaluation using Intel i7-12700K yielded 6.09 FPS and 5.93 FPS, meeting maritime surveillance real-time requirements of 1 to 5 FPS while maintaining consistent detection accuracy across all platforms. Memory consumption ranged from 50.2 to 78.0 MB for the GPU and from 374 to 378 MB for the CPU, demonstrating comparable resource efficiency to YOLOv11n. With 2.1 M parameters versus 2.6 M for YOLOv11n, IRSD-Net achieves superior accuracy and parameter efficiency through strategic speed trade-offs, making it particularly suitable for

Table 9. Performance comparison of IRSD-Net and YOLOv11n across different computing platforms.

Dataset	Platform	Model	FPS	Latency (ms)	mAP_{50} (%)	GFLOPs	Parameters (M)	Memory (MB)
ISDD	RTX4090	YOLOv11n	909.1	1.1	90.8	6.3	2.6	73.7
		IRSD-Net	714.3	1.4	92.5 (+1.7)	6.5	2.1	78.0
	RTX3060	YOLOv11n	260.5	3.8	90.5	6.3	2.6	50.2
		IRSD-Net	242.5	4.1	91.6 (+1.1)	6.5	2.1	53.2
	CPU	YOLOv11n	13.07	76.5	90.5	6.3	2.6	378
		IRSD-Net	6.09	164.3	92.1 (+1.6)	6.5	2.1	374
IRSDSS	RTX4090	YOLOv11n	1111.1	0.9	91.8	6.3	2.6	73.7
		IRSD-Net	769.2	1.3	92.9 (+1.1)	6.5	2.1	78.0
	RTX3060	YOLOv11n	270.3	3.7	91.7	6.3	2.6	50.2
		IRSD-Net	238.2	4.2	92.8 (+1.1)	6.5	2.1	53.2
	CPU	YOLOv11n	12.85	77.8	91.4	6.3	2.6	378
		IRSD-Net	5.93	168.6	91.5 (0.1)	6.5	2.1	374

accuracy-prioritized maritime surveillance applications. Although IRSD-Net slightly lags behind YOLOv11n in FPS, this difference is primarily due to IRSD-Net's adoption of a more complex network structure and more meticulous inference strategies to ensure higher detection accuracy across diverse maritime monitoring scenarios. Furthermore, IRSD-Net focuses more on precision and reliability, ensuring effectiveness and robustness in critical applications even at the expense of some processing speed. Nonetheless, IRSD-Net shows excellent performance in terms of memory consumption and parameter efficiency, proving its suitability for environments with limited resources.

Although technical constraints prevented actual deployment testing on NVIDIA Jetson series devices, IRSD-Net demonstrated exceptional edge deployment potential based on computational complexity analysis. Our model requires only 6.5 GFLOPs and 2.1 M parameters, representing a 19.2% parameter reduction compared to the YOLOv11n baseline. Recent investigations have established strong correlations between theoretical computational metrics and actual edge device performance. The LEAF-YOLO-N model [84] with 5.6 GFLOPs and 1.2 M parameters achieved real-time inference exceeding 30 FPS on the NVIDIA Jetson AGX Xavier (NVIDIA Corporation, Santa Clara, CA, USA), exhibiting computational complexity highly similar to our IRSD-Net architecture. Studies have validated that lightweight detection algorithms with computational loads below 10 G FLOPs and under 5 M parameters demonstrate stable operation across various edge platforms [85,86]. The HMKCNet backbone employs grouped convolution and channel-partitioning mechanisms, while the DCSFPN neck reduces information transfer overhead through bidirectional feature fusion, adhering to resource-constrained design principles. Based on theoretical analysis and comparable model cases, IRSD-Net is expected to achieve 45–55 FPS on the Jetson Xavier NX and 40–50 FPS on the Jetson Orin Nano (NVIDIA Corporation, Santa Clara, CA, USA), meeting real-time maritime target monitoring requirements. Future research will focus on comprehensive deployment verification on actual edge hardware.

6. Discussion

To provide a comprehensive evaluation of IRSD-Net's performance characteristics, we selected several challenging scenarios where our model exhibited false positives and false negatives to analyze the underlying causes and limitations.

Although IRSD-Net demonstrated robust performance across diverse scenarios, analysis of the detection errors revealed several limitations, as shown in Figure 8. The method experiences performance degradation under extreme weather conditions such as heavy fog or severe storms, primarily when ships exhibit minimal thermal radiation differences compared to surrounding seawater with a less than 3% contrast ratio, causing detection accuracy to decrease by 10–25%. Another source of errors arises from sea surface reflections

or cloud boundaries with ship-like thermal signatures being misidentified as targets, leading to an mAP_{50} reduction of approximately 8%. Additionally, complex port infrastructure creates false detections, where docked vessels or maritime equipment are erroneously classified as active targets. Despite multi-scale optimization by the DCSFPN, occasional missed detections occur for extremely small targets smaller than 10×10 pixels, and the HMKConv module may lack effective adaptation when significant environmental differences exist between testing and training scenarios.

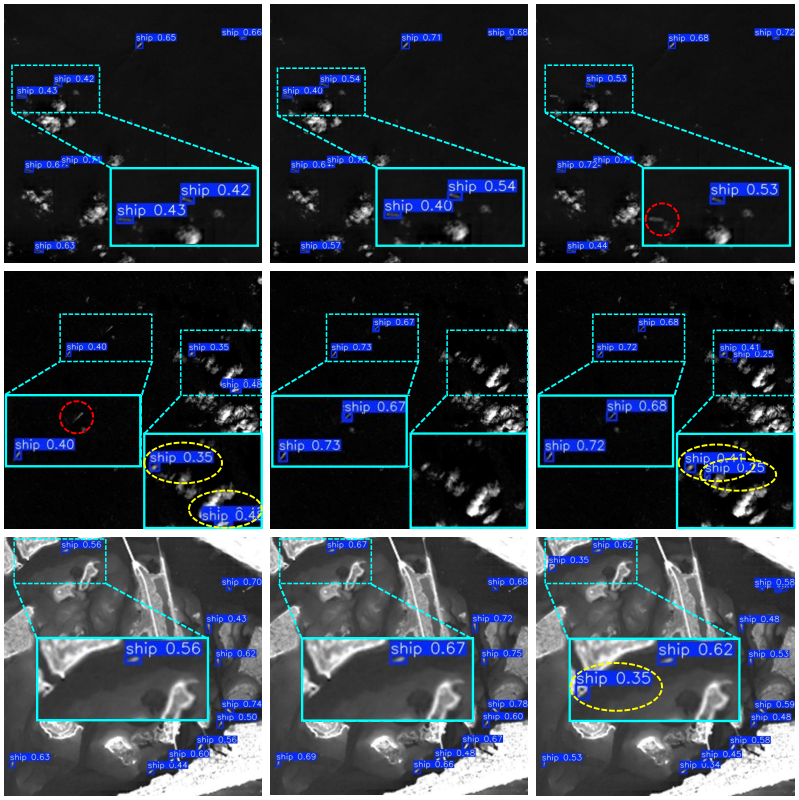

Figure 8. Visual analysis of false positive and false negative cases across different models. The blue boxes are magnified views of specific areas in the original image. False positives and false negatives are marked with yellow and red circles, respectively.

7. Conclusions

This paper presents IRSD-Net, which addresses multiple challenges in infrared ship target detection and achieves significant experimental results. The following are the three main innovations and their analysis:

- In this work, we introduce the HMKCNet into the backbone architecture to enhance feature extraction capabilities. This novel component employs parallel convolutions with varying kernels and receptive field regulation, thereby optimizing multi-scale target processing. As a result, the HMKCNet demonstrates particularly significant performance improvements when detecting small targets and processing low-contrast images.
- In terms of network architecture, we propose the DCSFPN. This innovation optimizes the fusion and transmission mechanisms of multi-scale features, thus enhancing global contextual perception capabilities. To achieve this, we incorporate two key components: the MSCBlock and EUCBlock. The components enable efficient integration and transmission of cross-scale semantic information.

- We implement Wise-PIoU, combining geometric alignment penalties with dynamic focusing. This addresses the performance degradation caused by numerous low-quality examples in detection tasks. Unlike traditional IoU-based loss functions, Wise-PIoU handles bounding box misalignments more accurately, enhancing regression precision while preventing domination by extreme samples. This approach optimizes small-ship detection in dynamic scenarios and complex environments, significantly improving both precision and recall metrics.

Despite robust performance, our model's computational complexity limits real-time applications on resource-constrained devices. Future research directions include implementing knowledge distillation and pruning techniques to reduce computational overhead for maritime device deployment. Meanwhile, integrating infrared imaging with visible light and radar data could enhance detection robustness under extreme weather conditions. Beyond performance optimization, extending the framework to detect multiple maritime objects such as floating debris and rescue equipment would enable comprehensive surveillance capabilities.

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

Funding: This work was funded by the National Natural Science Foundation of China (no. U2142206) and the National Key Research and Development Program of China (2022YFB4501704).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Zhang, X.; Wang, A.; Zheng, Y.; Mazhar, S.; Chang, Y. A detection method with antiinterference for infrared maritime small target. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3999–4014. [CrossRef]
2. Yang, P.; Dong, L.; Xu, H.; Dai, H.; Xu, W. Robust infrared maritime target detection via anti-jitter spatial–temporal trajectory consistency. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
3. Dong, L.; Wang, B.; Zhao, M.; Xu, W. Robust infrared maritime target detection based on visual attention and spatiotemporal filtering. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3037–3050. [CrossRef]
4. Gao, Y.; Wu, C.; Ren, M.; Feng, Y. Refined anchor-free model with feature enhancement mechanism for ship detection in infrared images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 12946–12960. [CrossRef]
5. Yao, T.; Hu, J.; Zhang, B.; Gao, Y.; Li, P.; Hu, Q. Scale and appearance variation enhanced siamese network for thermal infrared target tracking. *Infrared Phys. Technol.* **2021**, *117*, 103825. [CrossRef]
6. 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.* **2024**, *17*, 20. [CrossRef]
7. Cao, Z.; Kong, X.; Zhu, Q.; Cao, S.; Peng, Z. Infrared dim target detection via mode-k1k2 extension tensor tubal rank under complex ocean environment. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 167–190. [CrossRef]
8. Liu, Y.; Li, C.; Fu, G. PJ-YOLO: Prior-Knowledge and Joint-Feature-Extraction Based YOLO for Infrared Ship Detection. *J. Mar. Sci. Eng.* **2025**, *13*, 226. [CrossRef]
9. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-mean and max-median filters for detection of small targets. In Proceedings of the Signal and Data Processing of Small Targets, Denver, CO, USA, 19–23 June 1999; SPIE: Washington, DC, USA, 1999; Volume 3809, pp. 74–83.
10. Li, Y.; Li, Z.; Li, J.; Yang, J.; Siddique, A. Robust small infrared target detection using weighted adaptive ring top-hat transformation. *Signal Process.* **2024**, *217*, 109339. [CrossRef]
11. Lin, F.; Bao, K.; Li, Y.; Zeng, D.; Ge, S. Learning contrast-enhanced shape-biased representations for infrared small target detection. *IEEE Trans. Image Process.* **2024**, *33*, 3047–3058. [CrossRef]

12. Hao, C.; Li, Z.; Zhang, Y.; Chen, W.; Zou, Y. Infrared Small Target Detection Based on Adaptive Size Estimation by Multi-directional Gradient Filter. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5007915. [CrossRef]
13. Yang, X.; Li, Y.; Li, D.; Wang, S.; Yang, Z. Siam-AUnet: An end-to-end infrared and visible image fusion network based on gray histogram. *Infrared Phys. Technol.* **2024**, *141*, 105488. [CrossRef]
14. Wu, J.; He, Y.; Zhao, J. An infrared target images recognition and processing method based on the fuzzy comprehensive evaluation. *IEEE Access* **2024**, *12*, 12126–12137. [CrossRef]
15. Gan, C.; Li, C.; Zhang, G.; Fu, G. DBNDiff: Dual-branch network-based diffusion model for infrared ship image super-resolution. *Displays* **2025**, *88*, 103005. [CrossRef]
16. Wang, Y.; Wang, B.; Fan, Y. PPGS-YOLO: A lightweight algorithms for offshore dense obstruction infrared ship detection. *Infrared Phys. Technol.* **2025**, *145*, 105736. [CrossRef]
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
21. Chen, S.; Zhan, R.; Wang, W.; Zhang, J. Learning slimming SAR ship object detector through network pruning and knowledge distillation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1267–1282. [CrossRef]
22. 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*, 10097–10107. [CrossRef]
23. Deng, H.; Zhang, Y. FMR-YOLO: Infrared ship rotating target detection based on synthetic fog and multiscale weighted feature fusion. *IEEE Trans. Instrum. Meas.* **2023**, *73*, 1–17. [CrossRef]
24. Ge, Y.; Ji, H.; Liu, X. Infrared remote sensing ship image object detection model based on YOLO In multiple environments. *Signal Image Video Process.* **2025**, *19*, 1–12. [CrossRef]
25. Sun, Z.; Leng, X.; Lei, Y.; Xiong, B.; Ji, K.; Kuang, G. BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images. *Remote Sens.* **2021**, *13*, 4209. [CrossRef]
26. Wang, W.; Li, Z.; Siddique, A. Infrared maritime small-target detection based on fusion gray gradient clutter suppression. *Remote Sens.* **2024**, *16*, 1255. [CrossRef]
27. Guo, F.; Ma, H.; Li, L.; Lv, M.; Jia, Z. FCNet: Flexible convolution network for infrared small ship detection. *Remote Sens.* **2024**, *16*, 2218. [CrossRef]
28. Zhou, A.; Xie, W.; Pei, J. Background modeling combined with multiple features in the Fourier domain for maritime infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
29. Xie, B.; Hu, L.; Mu, W. Background suppression based on improved top-hat and saliency map filtering for infrared ship detection. In Proceedings of the 2017 International Conference on Computing Intelligence and Information System (CIIS), Nanjing, China, 21–23 April 2017; IEEE: New York, NY, USA, 2017; pp. 298–301.
30. Cui, Z.; Yang, J.; Li, J.; Jiang, S. An infrared small target detection framework based on local contrast method. *Measurement* **2016**, *91*, 405–413. [CrossRef]
31. Sun, H.; Jin, Q.; Xu, J.; Tang, L. Infrared small-target detection based on multi-level local contrast measure. *Procedia Comput. Sci.* **2023**, *221*, 549–556. [CrossRef]
32. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]
33. Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [CrossRef]
34. Zhou, A.; Xie, W.; Pei, J. Maritime infrared target detection using a dual-mode background model. *Remote Sens.* **2023**, *15*, 2354. [CrossRef]
35. Lin, J.; Yu, Q.; Chen, G. Infrared ship target detection based on the combination of Bayesian theory and SVM. In Proceedings of the MIPPR 2019: Automatic Target Recognition and Navigation, Wuhan, China, 2–3 November 2019; SPIE: Washington, DC, USA, 2020; Volume 11429, pp. 244–251.
36. Li, N.; Ding, L.; Zhao, H.; Shi, J.; Wang, D.; Gong, X. Ship Detection Based on Multiple Features in Random Forest Model for Hyperspectral Images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 891–895. [CrossRef]

37. Ye, J.; Yuan, Z.; Qian, C.; Li, X. Caa-yolo: Combined-attention-augmented yolo for infrared ocean ships detection. *Sensors* **2022**, *22*, 3782. [CrossRef]
38. Li, L.; Jiang, L.; Zhang, J.; Wang, S.; Chen, F. A complete YOLO-based ship detection method for thermal infrared remote sensing images under complex backgrounds. *Remote Sens.* **2022**, *14*, 1534. [CrossRef]
39. Wu, T.; Li, B.; Luo, Y.; Wang, Y.; Xiao, C.; Liu, T.; Yang, J.; An, W.; Guo, Y. MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
40. Wu, H.; Huang, X.; He, C.; Xiao, H.; Luo, S. Infrared small target detection with Swin Transformer-based multi-scale atrous spatial pyramid pooling network. *IEEE Trans. Instrum. Meas.* **2024**, *74*, 5003914.
41. Chen, F.; Gao, C.; Liu, F.; Zhao, Y.; Zhou, Y.; Meng, D.; Zuo, W. Local patch network with global attention for infrared small target detection. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 3979–3991. [CrossRef]
42. Liu, Y.; Yang, F.; Hu, P. Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks. *IEEE Access* **2020**, *8*, 145740–145750. [CrossRef]
43. Yue, T.; Lu, X.; Cai, J.; Chen, Y.; Chu, S. YOLO-MST: Multiscale deep learning method for infrared small target detection based on super-resolution and YOLO. *Opt. Laser Technol.* **2025**, *187*, 112835. [CrossRef]
44. Zhang, T.; Li, L.; Cao, S.; Pu, T.; Peng, Z. Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [CrossRef]
45. Xin, J.; Luo, M.; Cao, X.; Liu, T.; Yuan, J.; Liu, R.; Xin, Y. Infrared superpixel patch-image model for small target detection under complex background. *Infrared Phys. Technol.* **2024**, *142*, 105490. [CrossRef]
46. Bao, C.; Cao, J.; Ning, Y.; Zhao, T.; Li, Z.; Wang, Z.; Zhang, L.; Hao, Q. Improved dense nested attention network based on transformer for infrared small target detection. *arXiv* **2023**, arXiv:2311.08747.
47. Rahman, M.M.; Munir, M.; Marculescu, R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 11769–11779.
48. Raza, A.; Liu, J.; Liu, Y.; Liu, J.; Li, Z.; Chen, X.; Huo, H.; Fang, T. IR-MSDNet: Infrared and visible image fusion based on infrared features and multiscale dense network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3426–3437. [CrossRef]
49. Chen, X.; Qiu, C.; Zhang, Z. A multiscale method for infrared ship detection based on morphological reconstruction and two-branch compensation strategy. *Sensors* **2023**, *23*, 7309. [CrossRef]
50. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
51. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31 $\times$ 31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
52. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
53. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
54. Wang, S.; Wang, Y.; Chang, Y.; Zhao, R.; She, Y. EBSE-YOLO: High precision recognition algorithm for small target foreign object detection. *IEEE Access* **2023**, *11*, 57951–57964. [CrossRef]
55. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust infrared small target detection network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
56. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 877–886.
57. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Lauderdale, FL, USA, 11–13 April 2011; JMLR Workshop and Conference Proceedings; pp. 315–323.
58. Zhang, Y.; Jiu, B.; Wang, P.; Liu, H.; Liang, S. An end-to-end anti-jamming target detection method based on CNN. *IEEE Sen. J.* **2021**, *21*, 21817–21828. [CrossRef]
59. Santurkar, S.; Tsipras, D.; Ilyas, A.; Madry, A. How does batch normalization help optimization? *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 2488–2498.
60. Krizhevsky, A.; Hinton, G. Convolutional deep belief networks on cifar-10. *Unpubl. Manuscr.* **2010**, *40*, 1–9.
61. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.

62. Zhao, L.; Fu, L.; Jia, X.; Cui, B.; Zhu, X.; Jin, J. YOLO-BOS: An Emerging Approach for Vehicle Detection with a Novel BRSA Mechanism. *Sensors* **2024**, *24*, 8126. [CrossRef]
63. Liu, C.; Wang, K.; Li, Q.; Zhao, F.; Zhao, K.; Ma, H. Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism. *Neural Net.* **2024**, *170*, 276–284. [CrossRef]
64. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
65. 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.* **2022**, *61*, 1–19. [CrossRef]
66. Hu, C.; Dong, X.; Huang, Y.; Wang, L.; Xu, L.; Pu, T.; Peng, Z. SMPISD-MTPNet: Scene Semantic Prior-Assisted Infrared Ship Detection Using Multi-Task Perception Networks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5000814. [CrossRef]
67. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
68. Fei, X.; Guo, M.; Li, Y.; Yu, R.; Sun, L. ACDF-YOLO: Attentive and Cross-Differential Fusion Network for Multimodal Remote Sensing Object Detection. *Remote Sens.* **2024**, *16*, 3532. [CrossRef]
69. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [CrossRef]
70. Zhu, R.; Jin, H.; Han, Y.; He, Q.; Mu, H. Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv7-Tiny Network. *IEEE Access* **2025**, *13*, 48904–48922. [CrossRef]
71. Hussain, M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines* **2023**, *11*, 677. [CrossRef]
72. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
73. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
74. Li, S.; Li, Y.; Li, Y.; Li, M.; Xu, X. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE Access* **2021**, *9*, 141861–141875. [CrossRef]
75. Feng, Y.; Huang, J.; Du, S.; Ying, S.; Yong, J.H.; Li, Y.; Ding, G.; Ji, R.; Gao, Y. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 2388–2401. [CrossRef]
76. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2022**, *32*, 1745–1758. [CrossRef] [PubMed]
77. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
78. Jiang, Y.; Tan, Z.; Wang, J.; Sun, X.; Lin, M.; Li, H. GiraffeDet: A heavy-neck paradigm for object detection. *arXiv* **2022**, arXiv:2202.04256.
79. Yang, Z.; Guan, Q.; Yu, Z.; Xu, X.; Long, H.; Lian, S.; Hu, H.; Tang, Y. MHAF-YOLO: Multi-Branch Heterogeneous Auxiliary Fusion YOLO for accurate object detection. *arXiv* **2025**, arXiv:2502.04656.
80. Chen, Y.; Zhang, C.; Chen, B.; Huang, Y.; Sun, Y.; Wang, C.; Fu, X.; Dai, Y.; Qin, F.; Peng, Y.; et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Comput. Biol. Med.* **2024**, *170*, 107917. [CrossRef]
81. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34; pp. 12993–13000.
82. Ma, S.; Xu, Y. Mpdiou: A loss for efficient and accurate bounding box regression. *arXiv* **2023**, arXiv:2307.07662. [CrossRef]
83. Zhang, H.; Xu, C.; Zhang, S. Inner-iou: More effective intersection over union loss with auxiliary bounding box. *arXiv* **2023**, arXiv:2311.02877.
84. Nghiem, V.Q.; Nguyen, H.H.; Hoang, M.S. LEAF-YOLO: An Edge-Real-Time and Lightweight YOLO for Small Object Detection. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4868866 (accessed on 10 May 2025).
85. Alqahtani, D.K.; Cheema, M.A.; Toosi, A.N. Benchmarking deep learning models for object detection on edge computing devices. In Proceedings of the International Conference on Service-Oriented Computing, Tunis, Tunisia, 3–6 December 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 142–150.
86. Mittal, P. A comprehensive survey of deep learning-based lightweight object detection models for edge devices. *Artif. Intell. Rev.* **2024**, *57*, 242. [CrossRef]