Computer Science and Information Systems 23(1):133–164 https://doi.org/10.2298/CSIS250613005L

ADN-YOLO: An Improved Ship Detection Model Based on YOLOv11

Tao Li[1], Dezhi Han[1], Songyang Wu[2⋆], Xiang Shen[1,3], Liqi Zhu[1], and Wenqi Sun[2]

[1] Shanghai maritime university, 1550 Pudong Ave
201306 Shanghai, China
202430310151@stu.shmtu.edu.cn
dezhi@shmtu.edu.cn
shenxiang1107@163.com
zhuliqi0309@163.com
[2] Network Security Center, The Third Research Institute of the Ministry of Public Security
200031 Shanghai, China
wusongyang@stars.org.cn
sunwenqi@gass.ac.cn
[3] School of Computer Science, The University of Sydney, Sydney, NSW2006.Australia
shenxiang1107@163.com

Abstract. Existing infrared imaging techniques have garnered considerable attention and have achieved notable progress in all weather ship target detection tasks, owing to their robustness against varying ambient lighting conditions. However, due to the inherent limitations of infrared images,such as low spatial resolution and insufficient texture information the performance of multi-scale ship target detection remains suboptimal. These challenges significantly hinder the overall improvement of detection accuracy. To address this issue and enhance the detection performance of multi-scale ship targets, particularly small ones, in infrared imagery,this paper proposes an improved You Only Look Once (YOLO) based detection model named ADN-YOLO. The model first introduces a Dynamic Upsampler (Dysample) module, which more effectively integrates semantic information across different layers. This integration balances low level detailed features with high level semantic representations, thereby enhancing the model's ability to perceive target edges and structural characteristics. Second, a lightweight downsampling module (ADown) is incorporated to reduce the parameter count while improving both the efficiency and representational capacity of feature extraction. Additionally, to address the issue of small targets being highly sensitive to localization errors, a new loss function is designed based on the Wasserstein distance. This function combines the Normalized Wasserstein Distance (NWD) with the Complete Intersection over Union (CIoU), thereby enhancing the model's ability to accurately localize small targets. Comprehensive experimental validation is conducted on a marine infrared target detection dataset. Compared to the standard YOLOv11 model, the proposed ADN-YOLO reduces the number of parameters by 20.3%, achieves a 1.9%increase in mAP, a 1.9% boost in Recall, and lowers FLOPs by 1.1G, demonstrating its effectiveness and practicality for infrared image target detection tasks.

Keywords: Target detection; Infrared images; Multi-scale objects; Deep learning;

⋆ Corresponding author

1. Introduction

With the deepening development of the modern technological revolution, the development of marine resources has risen to become a strategic priority for major coastal nations worldwide[1]. In this context, intelligent target recognition technology for maritime applications has emerged as a core enabler for ensuring marine safety. Particularly in nearshore waters characterized by high vessel traffic density and a mix of heterogeneous vessels, the ability to accurately identify and dynamically monitor diverse targets such as cargo ships, fishing vessels, and passenger ferries has become a critical challenge for enhancing maritime governance efficiency and risk management capabilities [2]. The existing ship target detection system primarily relies on three technological pathways: visible light imaging, synthetic aperture radar (SAR), and infrared thermal imaging[3][4]. While visible light imaging offers significant advantages such as high spatial resolution, rich texture features, and a wide detection range, its target recognition accuracy in low-light reflective environments is inherently limited. SAR technology, while enabling all-weather and long-range detection capabilities, is subject to multiple constraints in marine applications, including electromagnetic interference, land-based obstruction effects, port clutter, and atmospheric attenuation. In contrast, infrared imaging technology demonstrates unique advantages. Its passive radiation reception mechanism not only ensures the stealthiness of the detection process[5] but also effectively overcomes the environmental sensitivity of visible light imaging through features such as smoke interference resistance and independence from lighting conditions.

Traditional infrared ship detection methods primarily employ spatial filtering and feature modeling techniques. Filters based on morphological operations suppress uniform background noise using predefined structural templates. However, in non-uniform interference scenarios such as cloud reflections and solar flares, their linear filtering mechanism struggles to distinguish between the statistical similarity of targets and complex backgrounds, leading to a significant increase in false detection rates. To overcome this limitation, a spectral residual model based on the human visual system (HVS)[6][7][8] attempts to extract salient features from the frequency domain perspective. However, this method lacks sensitivity to dynamic sea surface clutter and struggles to effectively suppress high-frequency interference such as wave breaking. Further developed local contrast methods can enhance edge responses of high-contrast targets but fail to adapt to the thermal radiation attenuation characteristics of targets and backgrounds under low-visibility conditions like sea fog or rain/snow due to their reliance on fixed thresholds. The recently emerging low-rank representation method [9][10] achieves sparse separation of background and targets through matrix decomposition, demonstrating potential in low signal-to-noise ratio scenarios. However, it lacks the ability to model the morphological diversity of small ship targets, especially when targets and wave interference exhibit motion coupling, leading to a sharp increase in separation error. The common shortcomings of the above methods reveal the inherent limitations of traditional algorithms in complex sea conditions. Their modeling paradigm based on artificial feature engineering is unable to effectively characterize the nonlinear coupling relationship between infrared ship targets and dynamic sea backgrounds. This bottleneck has prompted researchers to turn to data-driven deep learning frameworks to break through the representation capabilities of traditional methods through end-to-end feature learning.

Current deep learning-based ship detection algorithms primarily follow two technical approaches: two-stage and one-stage, which continue to evolve. Two-stage detectors utilize a cascaded architecture composed of candidate region generation and fine-grained regression, demonstrating significant advantages in rotating ship detection tasks. For example, Gu et al.[11] developed an improved Faster R-CNN model that constructs multi-scale feature maps through multi-layer convolutional concatenation, thereby obtaining feature vectors with richer semantic information and enhancing model performance. Pang et al.[12] proposed the rotating Libra R-CNN, which introduces a three-dimensional balanced mechanism between the feature layer, sample layer, and target layer, effectively controlling the prediction error of ship azimuth and further improving the detection accuracy of rotating targets. However, the multi-stage processing workflow of such methods introduces significant computational redundancy, severely limiting their real-time performance. In contrast, one-stage detectors (such as the YOLO series) adopt a parallel architecture combining "anchor box prediction and classification regression," significantly improving detection speed while maintaining detection accuracy, better aligning with the real-time response requirements of ship detection tasks. For example, Zheng et al. [13][proposed two novel gated fusion units that effectively fuse feature maps generated by the intermediate layers of SSD, constructing various stacked and gated hybrid fusion versions; Li et al. [14] proposed YOLO-FIRI, which enhances the feature representation capability for small infrared targets through shallow network expansion and multi-scale attention mechanisms; and Luo et al. [15] developed YOLO-IR, which significantly improves target recall rates in low-contrast scenes using a focus loss function. Si et al. [16] proposed an improved YOLO-RSSD algorithm, where an enhanced bidirectional feature pyramid network structure is embedded into the feature fusion component. This enables cross-layer multi-scale weighted feature fusion and introduces a channel attention mechanism in the convolutional unit to further enhance the detection performance of ship targets in infrared images.

Although infrared image detection algorithms that incorporate deep learning have achieved significant performance improvements, their high model complexity and computational costs still pose constraints on practical deployment. As a result, lightweight design has become a research hotspot in recent years. For example, Xing et al. [17] proposed a lightweight detection algorithm named YOLOv8-FAS, which incorporates efficient FasterNet components and attention mechanisms to enhance the feature extraction capabilities of the backbone network, achieving a good balance between performance and efficiency. These methods demonstrate excellent performance in ship image detection under visible light conditions, but they still face numerous challenges in ship detection tasks targeting infrared images. Due to the complex thermal radiation distribution characteristics of infrared images, their physical imaging mechanisms significantly differ from those of visible light images. For example, the thermal wake of a ship is prone to frequency domain aliasing with background thermal radiation, leading to feature confusion; thermal diffusion effects between adjacent targets can cause boundary blurring; and under low-contrast conditions, the saliency information of targets often severely degrades, thereby affecting the visibility and accuracy of detection results. As shown in Figures 1(a) and (b), the day-night temperature difference causes the ocean and land areas to exhibit obvious gray-scale bipolarity [18][19], with high variability in contrast between ships and their backgrounds. Figure 1(c) shows that ship targets in infrared images are typically

small in scale, especially small vessels anchored near the coast, which are easily obscured by complex backgrounds. Additionally, due to the lack of obvious semantic structure in infrared images, existing methods are limited in terms of robustness and generalization capabilities.

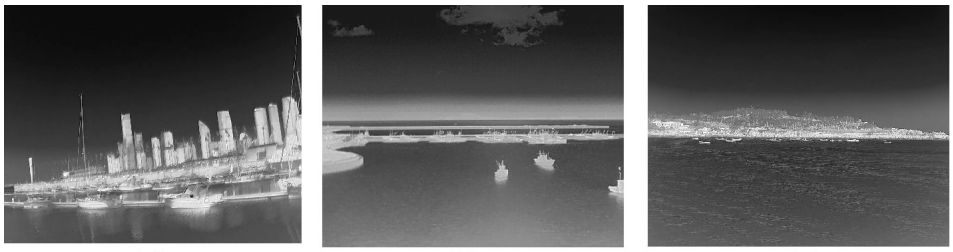

Fig. 1. (a) Targets under close imaging, (b) Targets near a harbor, (c) Multiple targets near a harbor with a complex low-contrast background

In summary, current infrared ship target detection methods continue to face significant challenges, including the difficulty of balancing detection accuracy with real time performance and the limited capability of feature extraction. At the core of these issues is the fact that existing approaches have yet to effectively incorporate the thermophysical priors inherent to infrared imaging.To overcome these limitations, there is an urgent need to develop a hybrid intelligent detection framework that seamlessly integrates thermodynamic a priori knowledge with data driven learning mechanisms. Such a framework would be better equipped to tackle the fundamental challenges of infrared ship detection, including complex background interference, low contrast imaging, and target blurring, ultimately advancing both the robustness and generalization capabilities of detection systems in real world maritime environments.

To address the aforementioned challenges, this paper proposes an infrared ship detection method based on an improved YOLOv11 architecture, named ADN-YOLO (Attention-guided Downsample and Normalized Distance enhanced YOLO). The proposed method significantly enhances the detection accuracy and feature discrimination for small- and medium-scale targets in infrared images, while preserving the lightweight characteristics essential for real-time deployment. Specifically, the backbone of YOLOv11 is modified by introducing an improved downsampling module, ADown, to replace parts of the original CBS (Conv-BN-SiLU) structure. ADown integrates an attention mechanism with a dynamic convolution strategy, effectively reducing both the parameter count and computational complexity. This design notably boosts the expressive power of the feature extractor, demonstrating superior robustness in typical infrared scenarios characterized by low contrast and complex backgrounds. To further address the issue of small targets being easily obscured by background information, a point-based dynamic upsampling module, Dysample, is introduced. Dysample adaptively enhances the spatial detail representation of low-resolution feature maps, effectively preserving shallow semantic and positional information. This enables the model to better perceive and detect multiscale targets, par-

ticularly weak infrared signals.In addition, this paper proposes a novel loss function that combines normalized Wasserstein distance with traditional bounding box regression. By modeling both the predicted and ground truth boxes as two-dimensional Gaussian distributions,and computing their normalized Wasserstein distance, this metric captures the distributional differences between predicted and actual targets. This approach supplements the geometric constraints of CIoU by incorporating statistical distribution information, thereby enhancing the accuracy of bounding box regression, particularly for small and irregularly shaped targets, while also improving the model's generalization ability and robustness.The main contributions of the proposed ADN-YOLO model are as follows:

(1) An improved detection model, ADN-YOLO, based on YOLOv11 is proposed specifically for infrared ship detection scenarios. By incorporating convolutional enhancement and a dynamic downsampling mechanism, the model effectively improves the detection capability of weak and low-contrast targets.

(2) A novel ADown downsampling module is designed to replace the original CBS structure, reducing model complexity while simultaneously enhancing the expressive power of feature extraction.

(3) The Dysample module is introduced to enhance the model's multi-scale perception of small targets, effectively addressing the detection challenges posed by blurred and low-contrast ships in infrared images.

(4) The NWD-CIoU loss function is proposed, which fuses the normalized Wasserstein distance with CIoU to address the robustness limitations of traditional IoU in small target detection, thereby improving both localization accuracy and overall detection performance.

The remainder of this paper is organized as follows: Section 2 reviews key recent research in the field of target detection. Section 3 provides a detailed description of the YOLOv11 model and the improvements introduced in our proposed method. Section 4 presents the experimental results of our model and compares its performance with several other target detection approaches. Finally, Section 5 concludes the paper by summarizing the findings and discussing the limitations of the proposed model.

2. Related work

In recent years, researchers have proposed a variety of strategies and algorithms to cope with the problems of large computational overhead, significant scale changes, and difficulty in effectively extracting target features in complex backgrounds faced by infrared images during processing. On the basis of reviewing the traditional target detection methods for infrared images, this paper focuses on the target detection technology based on deep learning, systematically combs through the main improvement directions of single-phase and two-phase methods, and discusses in depth the limitations and challenges of each.

2.1. Traditional infrared image detection

Traditional infrared target detection methods typically model infrared images as consisting of three components: the target, the background, and noise. The core idea is to

enhance the saliency of the target region by effectively suppressing background interference and image noise, thereby achieving more accurate target detection. These methods primarily rely on classical image processing techniques or manually designed feature extraction strategies and do not incorporate large-scale data-driven learning mechanisms, resulting in certain limitations under specific conditions. Among numerous traditional methods, detection algorithms based on the human visual system (HVS) are representative. For example, the Three-Layer Local Contrast Measurement (TLLCM) method enhances the saliency of the target region by simulating the human eye's sensitivity to local contrast, thereby improving target detectability. Although these methods have improved the response capability to target regions to some extent, they rely on fixed contrast enhancement mechanisms and struggle to adapt to the varying features of targets in complex backgrounds. To further enhance target detection performance, Zhao and Kong[20] proposed a spatial filtering-based infrared image target detection method. This method designs filters based on the differences in grayscale distribution between targets and backgrounds to selectively suppress background regions and highlight target regions. However, this method performs poorly when handling isolated small noise points, often leading to misclassification or omission of target regions, thereby limiting overall detection accuracy. To address these issues, Anju and Raj [21] proposed a frequency-domain-based target detection method inspired by frequency-domain analysis. This method analyzes the differences in frequency distribution between the target and background in an image, decomposing the image into high-frequency components (corresponding to target information) and low-frequency components (corresponding to background information), thereby achieving more effective target extraction. Experiments show that compared to spatial filtering methods, frequency-domain filtering performs better in suppressing background interference and improves detection accuracy, but its computational complexity significantly increases, making it unsuitable for real-time application deployment. Additionally, Jiao[22] proposed modeling infrared image target detection as a problem of separating low-rank matrices and sparse matrices, and introduced sparse representation methods to achieve precise extraction of target regions in images. This method outperforms traditional filtering methods in terms of signal-to-clutter ratio (BCR) and background suppression factor (BSF), demonstrating excellent target detection capabilities. However, due to the non-local autocorrelation characteristics of infrared image backgrounds, the distinction between targets and backgrounds is often blurred, leading to limitations in the method's suppression capabilities under complex backgrounds. In summary, while traditional infrared image detection methods still hold practical value in specific scenarios, the increasing miniaturization of targets, growing complexity of backgrounds, and continuous decline in BCR make the trade-off between detection accuracy and processing efficiency increasingly challenging. Additionally, their high computational resource consumption makes it challenging to meet the dual requirements of speed and energy efficiency for real-time detection systems. Therefore, there is an urgent need for more efficient and robust detection methods to address target recognition tasks in complex infrared scenarios.

2.2. Deep learning Based InfiRay image detection

Deep learning, a key branch of computer vision, has achieved significant advancements in tasks such as target detection. By building deep neural network models with

multilayer convolutional architectures, these methods automatically learn high level semantic features from images, thereby substantially improving target detection accuracy in complex scenes.

Two-stage object detectors In 2014, R. Girshick et al. [23] proposed the groundbreaking two-stage object detector R-CNN, whose core idea is to divide object detection into two steps: generating proposals and classification regression prediction. Although this method achieved certain breakthroughs in detection accuracy, its high computational complexity, lack of weight sharing, and slow inference speed limited its practical application. To enhance detection efficiency and accuracy, Fast R-CNN [24] and Faster R-CNN [25] were subsequently proposed, introducing ROI (Region of Interest) pooling operations and the Region Proposal Network (RPN), respectively, significantly improving overall performance. In the field of infrared image object detection, Ghose et al. [26] proposed an improved method based on the classic Faster R-CNN, incorporating saliency maps to enhance key regions in infrared images. However, the saliency network in this method was not trained in an end-to-end, multi-task joint training manner with Faster R-CNN, resulting in a time-consuming training process. Devaguptapu et al. [27] developed a multimodal Faster R-CNN that utilizes RGB channels to extract high-level infrared features, further improving detection performance. However, this multimodal fusion strategy significantly increases training computational overhead. Overall, although two-stage detectors excel in terms of accuracy, their high computational complexity and large inference latency make them unsuitable for real-time object detection requirements, particularly in applications with stringent timeliness demands such as ship monitoring. Additionally, their detection capabilities for small objects are also suboptimal. Therefore, when selecting ship target detection algorithms, it is essential to fully consider actual application requirements and scene characteristics, balancing detection accuracy with computational efficiency. Chen et al. [28] proposed the CAAN model, which designs a new method for calculating absolute positions based on the coordinates of each image region and the actual size of the image.

Single-stage object detectors To address the limitations of two-stage detectors in terms of computational efficiency and real-time performance, researchers have proposed single-stage object detection algorithms, with the YOLO (You Only Look Once) series being a representative example. This series of algorithms divides the input image into multiple grid units and detects each unit as a potential target candidate region. YOLO leverages a fully convolutional neural network architecture to integrate object classification and bounding box regression into a unified end-to-end processing workflow. Compared to traditional two-stage methods (such as Faster R-CNN), YOLO significantly improves computational efficiency and inference speed while maintaining detection accuracy. Therefore, it demonstrates excellent application potential and broad prospects in scenarios with high requirements for real-time performance and low latency, such as video surveillance and autonomous driving. YOLO's efficient detection capabilities have inspired a large number of related studies. For example, Li et al. [29] proposed a cross-layer attention network to enhance the feature expression capabilities of small objects in visible light images. In recent years, positive progress has also been made in the detection of weak small objects in infrared images. Guo et al. [30] designed a bidirectional attention feature pyramid network (BAFPN) for near-shore vessel detection, significantly improving detection performance for miniature vessels. Chen et al. [31] utilized optimized global context information to dynamically modulate and filter image features, combined with

spatial location information, to further enhance the modeling of image label dependencies and the model's inference capabilities. Shen et al. [32] proposed a sparse attention network (GFSNet) that overcomes the limitations of traditional spatial domain methods, enhancing the network's ability to perceive multi-scale features and extract significant regions. Although deep learning-based infrared image ship detection technology has made significant progress, it still faces numerous challenges that require further research and breakthroughs. First, infrared images themselves have low resolution, are often blurry, and frequently contain significant noise, severely affecting detection accuracy. Second, infrared images are mostly acquired from coastal viewpoints, which, while offering a broader field of view, also introduce more complex background interference. There are still notable shortcomings in terms of small target recognition accuracy and feature expression stability. To address these challenges, this paper proposes an improved object detection model based on the YOLOv11 framework—ADN-YOLO. The detailed structure and implementation of this method will be discussed in the next section.

3. Methodology

Addressing the challenges of marine vessel detection, such as significant multi-scale variation and the weakening of small targets in infrared images, this paper proposes an improved target detection algorithm based on the YOLOv11 framework, named ADN-YOLO. This chapter systematically details the innovative architectural design and key technological enhancements of the method. First, the ADown module is introduced to replace part of the convolutional structure in the YOLOv11 backbone. Its multi-branch design enhances feature fusion and information interaction, thereby improving the detection accuracy of distant small targets. Second, to further strengthen small target perception, the Dysample module is adopted to dynamically upsample low-resolution feature maps, effectively preserving fine target details. Finally, a novel loss function optimization strategy is proposed, combining Normalized Wasserstein Distance (NWD) with Complete IoU (CIoU) to replace the traditional IoU metric, further enhancing target localization accuracy. Specifically, Section 3.1 introduces the fundamental characteristics of the YOLOv11 model. Section 3.2 elaborates on the structure and improvements of the ADN-YOLO algorithm.

3.1. YOLOv11 Modeling

YOLOv11 is the latest iteration in the YOLO series, optimized from YOLOv8 with significant improvements. It features a redesigned backbone and neck architecture, enabling it to outperform YOLOv8 in multi-target detection tasks. Efficient feature extraction is crucial for accurate target localization and classification, and YOLOv11 greatly enhances both detection sensitivity and accuracy. Infrared images lack visible light color information, resulting in weaker texture and boundary features. Although YOLOv8 supports multi-scale detection, it is prone to missed or false detections in infrared scenarios due to the low signal-to-noise ratio.YOLOv11 also introduces a more efficient framework and training procedure. Its detection head incorporates two Depthwise Separable Convolution (DWConv) modules, which substantially reduce the model's parameters and computational load while maintaining accuracy and accelerating processing speed. Similar to

YOLOv8, YOLOv11 uses convolutional layers to downsample the input image, reducing spatial dimensions while increasing channel depth. However, in the backbone, YOLOv11 replaces YOLOv8's C2f module with a new C3k2 module. Additionally, following the SPPF module, the C2PSA module is introduced to enhance spatial attention within the feature maps. The backbone serves as the foundational component of the YOLO network, responsible for extracting high level semantic features from the input image through a series of convolutional and pooling layers. These extracted feature maps are then passed to subsequent network components for target detection. The neck module of YOLOv11 further processes these feature maps to prepare them for detection. Its main function is to extract and fuse multi-scale features, which are then fed to the head network for classification and bounding box regression. Compared to YOLOv8, YOLOv11's use of the C3k2 module in the neck allows for more detailed multi-scale feature extraction. The C3k2 module offers flexibility through parameterization: when set to false, it behaves like the C2f module with a standard bottleneck; when set to true, the C3k2 module replaces the bottleneck with a C3 module, as illustrated in Figure 2.

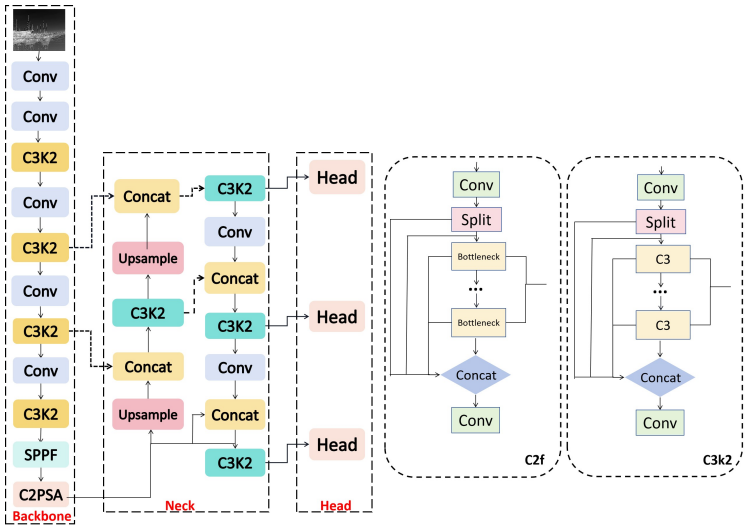

Fig. 2. Comparison of the original YOLOv11 models C2f and C3K2

3.2. Proposed methodology

In this paper, we propose an efficient and lightweight network, ADN-YOLO, designed to address the challenges of missing texture details in small infrared targets caused by low scale and limited spatial resolution, as well as the problem of highly variable multi-scale ship distributions. Instead of conventional convolutions in the feature extraction network, the ADown module is employed to capture locally significant features while preserving global information. It combines average and max pooling to achieve multi-scale feature extraction. The innovative ADown downsampling component plays a crucial role in preserving original high-frequency details, minimizing information loss, and enhancing

computational efficiency. This results in a significant improvement in small-target detection performance while reducing model parameters and increasing detection speed. To further boost feature extraction for small targets, the upsampling process is dynamically adjusted using DySample, which requires fewer parameters and computations than traditional upsampling methods, effectively reducing model complexity and computational resource consumption. Given the common issues of positional errors and shape distortions in small target detection, we combine the Normalized Wasserstein Distance (NWD) loss with the Complete IoU (CIoU) loss. The NWD loss optimizes localization accuracy, while the CIoU loss refines both the shape and position of the target bounding boxes at a finer granularity. The overall structure of the ADN-YOLO network is illustrated in Figure 3.

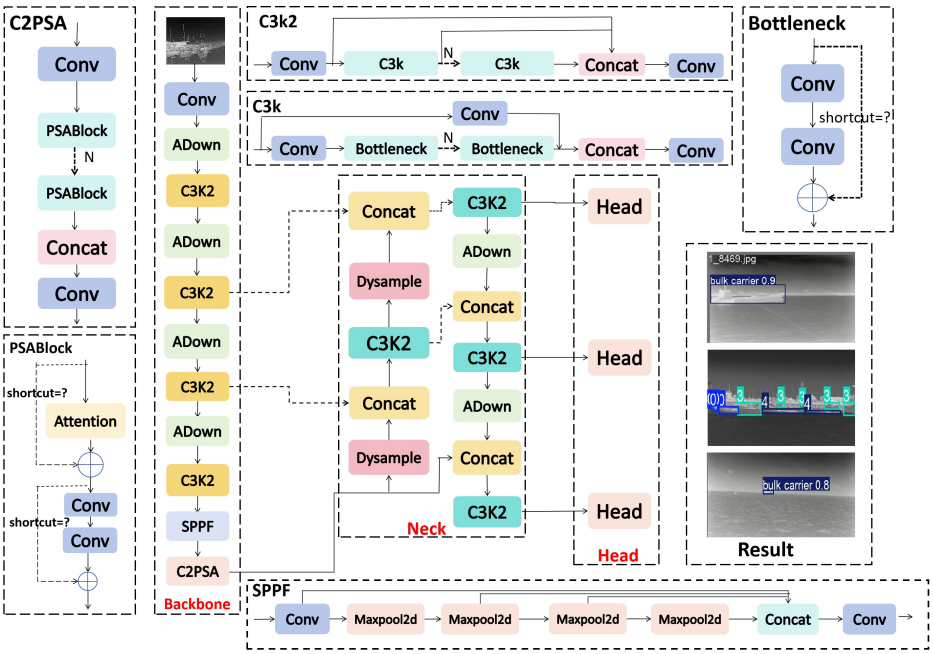

Fig. 3. Structure of ADN-YOLO

ADown Since the resolution of infrared images is lower than that of visible light images, they often appear more blurred, resulting in limited information extraction from small target vessels and the loss of critical details, which negatively impacts detection performance. Additionally, infrared images frequently contain significant noise and numerous distant objects. As the feature map size decreases through the network, fine features of distant targets may be lost due to downsampling, hindering the effective capture of small target details this poses particular challenges for densely packed targets in infrared images. The ADown module addresses this by combining average pooling, which preserves global information, with max pooling, which captures important local features. This dual-pooling strategy enables feature extraction at multiple scales. Therefore, this study in-

tegrates the ADown module into the backbone network of YOLOv11, replacing several conventional convolutional downsampling operations to enhance the model's ability to capture image features more effectively.

As shown in Figure 4., in the ADown downsampling module, the input feature map X first undergoes an average pooling operation. The feature map is sliced to halve its size and divided into two parts along the channel dimension, as shown in Eqs (1)(2). The feature map is then divided into two sub-feature maps along the channel dimension and . for, a 3×3 convolutional layer Conv1 operation with a step size of 2 and padding of 1 is used. And is first subjected to a maximum pooling operation with the aim of retaining important features. Then, a 1×1 convolutional layer Conv2 with step size 1 and padding 0 is used to capture the features, as shown in Eqs(3)-(5).

$$X' = AvgPool(\mathrm{X}, 2, 1, 0)\,. \tag{1}$$

$$\mathrm{X}_1, \mathrm{X}_2 = chunk\,(X', 2, 1)\ . \tag{2}$$

$$\mathrm{X}_1' = Conv\,(\mathrm{X}_1, kernel = 3, stride = 2, padding = 1) \tag{3}$$

$$\mathrm{X}_2' = MaxPool\,(\mathrm{X}_2, kernel = 3, stride = 2, padding = 1) \tag{4}$$

$$\mathrm{X}_2'' = Conv\,(\mathrm{X}_2', kernel = 1, stride = 1, padding = 0) \tag{5}$$

Finally, the two sub feature maps, X_1 and X_2, are concatenated to produce the output feature map Y of the ADown module.

$$Output = Concat\,(\mathrm{X}_1, \mathrm{X}_2, dim = 1) \tag{6}$$

ADown significantly enhances target detection accuracy while reducing false positives and missed detections by effectively suppressing background noise and redundant information. Small targets often suffer from poor localization accuracy and are prone to being overlooked; the ADown module addresses this by extracting rich feature information and fusing multi-scale features to produce more representative feature maps. The introduction of the ADown's multi-branch structure enables greater feature combination and information interaction, preserving more contextual details and preventing the loss of small target features. Whether dealing with images of varying resolutions or complex, diverse backgrounds, ADown allows the network to flexibly adjust its focus, thereby improving overall detection performance and boosting model robustness.

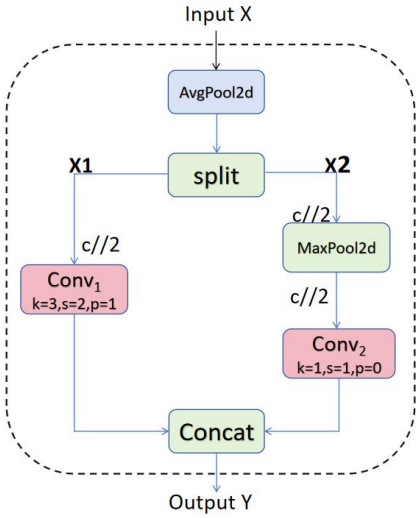

Fig. 4. ADown structure

Dysample DySample introduces an innovative approach that replaces the traditional dynamic convolution based upsampling with a point sampling technique, significantly improving resource efficiency. In the original YOLOv11 model, the UpSample operation imposes substantial computational demands and a large number of parameters, limiting the model's lightweight performance in ship detection tasks. In practical infrared ship monitoring scenarios, infrared images often suffer from pixel distortion, resulting in the loss of fine details that hinder effective feature learning. To address this, DySample was developed as a lightweight and efficient dynamic upsampling module designed to replace the conventional UpSample method. This point-based sampling technique generates content-aware upsampled features concisely and efficiently, without requiring additional high-resolution inputs.DySample reduces model complexity and computational overhead while maintaining high accuracy. Through an appropriate loss function and forward propagation training, it adaptively adjusts sampling point positions and weights, significantly enhancing upsampling accuracy and efficiency. This makes DySample particularly well-suited for the complex target detection tasks addressed in this paper. The flowchart of DySample upsampling is shown in Figure 5. Given an input feature map of size C × H × W and a sampling coordinate set of size 2 × H × W (where the first two dimensions correspond to x and y coordinates), the feature map is resampled using the $grid\_sample$ function. Bilinear interpolation is applied to generate a new feature map of size $C \times H_1 \times W_1$, as mathematically expressed in (7).

$$X = grid_sample(X, \delta) \tag{7}$$

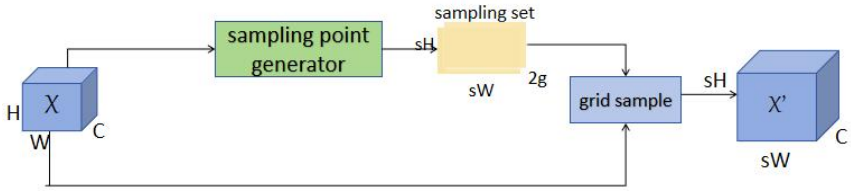

Fig. 5. Sampling process on DySample

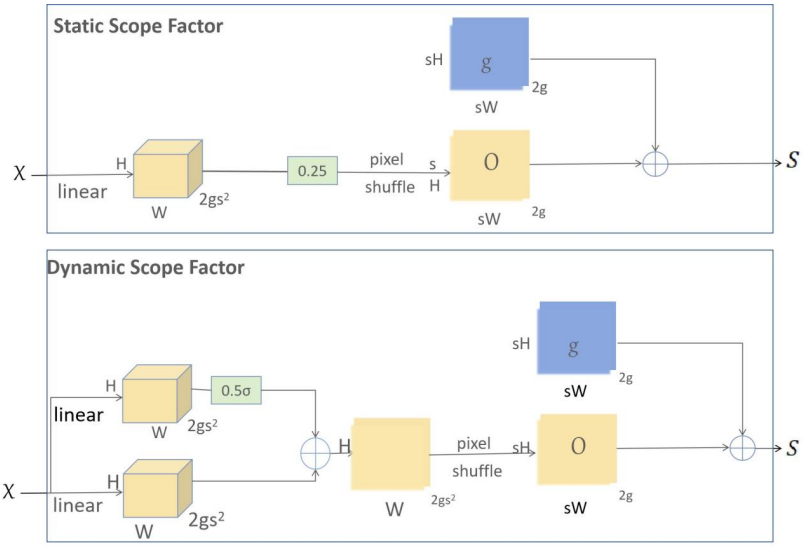

Fig. 6. Point sampling set generation process

As shown in Figure 6, given an upsampling scaling factor s and a feature map X of size $C \times H \times W$, a linear layer (with the number of input and output channels C and $2s^2 \times H \times W$, respectively) is used to generate an offset O of size $2 \times sH \times sW$, which is then reshaped to 2 × sH × sW by Pixel Shuffling [33] . through Pixel Shuffling, each pixel of the original feature map is mapped to a higher resolution and the pixels are aligned by offsets. The sampling set δ is the sum of the offset O and the original sampling grid G. The sampling set in this case represents the "offset" operation on the original grid, which results in a more accurate position in high resolution space. G is a grid, usually consisting of pixel coordinates or positions. O By learning the offset of each pixel, the position of each pixel can be adjusted so that features can be more accurately localized at higher resolutions. That is, the reshaping operation is omitted. Finally, an upsampled feature map X' of size $C \times sH \times sW$ is generated and produced by the sampling set function, which is shown in (8)-(9) as follows.

$$O = linear(X), \tag{8}$$

$$\delta = G + O. \tag{9}$$

The core function of the point sampling process is to accurately upsample feature maps through offset generation, pixel shuffling, and sampling set adjustment. This enhances the model's ability to capture fine details and improves overall accuracy while maintaining computational efficiency. By increasing the resolution of feature maps, this process boosts the network's performance, particularly in preserving image details and detecting small objects. As a result, the model's perceptual capability is strengthened, enabling more precise localization and identification of true vibration sites.

NWD-Ciou In YOLOv11, the CIoU metric is used to evaluate the overlap between predicted and ground truth bounding boxes. Compared to the traditional IoU (Intersection over Union) metric, CIoU considers not only the overlap area but also the normalized distances between box centers, as well as differences in width and height, making it more robust to variations in bounding box location and size[34][35] .In ship detection, targets vary greatly in size, posing specific challenges. For small ships, the bounding boxes are usually very small, so even minor positional deviations can cause significant fluctuations in IoU values. While the CIoU loss improves IoU stability by incorporating center distance and aspect ratio, it remains highly sensitive to slight bounding box deviations for small targets. This sensitivity can lead to unstable training and ultimately degrade model performance, as the model becomes overly sensitive to the exact position of small targets.Furthermore, CIoU's sensitivity to size changes is especially pronounced for small targets, which may cause the model to overemphasize size adjustments at the expense of other important features like appearance and contextual information.To address these limitations, NWD (Normalized Wasserstein Distance) has been introduced as a novel approach for small target detection[36][37] . Instead of relying on IoU, NWD models each bounding box as a two-dimensional Gaussian distribution and uses the Wasserstein distance to measure the similarity between these distributions. Unlike IoU, Wasserstein distance can effectively measure similarity even when there is little or no overlap. This method is inherently more adaptable to multi-scale objects and is particularly well-suited for measuring similarity among small targets.

In this chapter, NWD is combined with the CIoU method and integrated into the regression loss function to fully utilize the advantages of both in ship detection. CIoU excels in measuring spatial location and overlap, and is particularly suitable for medium- to large-scale ship detection, where the alignment accuracy of the bounding box is critical. On the other hand, NWD exhibits remarkable robustness in dealing with small-scale ships, capturing the distributional features of targets effectively and being insensitive to scale variations. The advantage of NWD lies in its ability to recognize small distributional differences of small targets, details that are often overlooked by conventional methods. Therefore, by combining CIoU's precision in spatial alignment with NWD's sensitivity to distributional features, the robustness and accuracy of the model at different scales are significantly improved. This fusion captures the location and features of pedestrians in a more comprehensive and detailed way, significantly improving the detection performance, especially in scenes with large variations in target size. The formulation of this regression loss function is shown below:

$$L_{reg} = (1 - \gamma)L_{NWD} + \gamma L_{CIoU}, \tag{10}$$

where is the regression loss function, is a weighting factor to balance the contribution of NWD and CIoU in the regression loss. The is usually between 0 and 1. It has been verified that optimal results are obtained when = 0.7. The CIoU loss function is defined as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v, \tag{11}$$

Wherein, IoU is used to assess the degree of overlap between the predicted bounding box and the real bounding box . by calculating the Euclidean distance between the center point of the predicted box and the center point of the real box. denotes the diagonal length of the smallest closed region that can enclose the predicted and true boxes. is the trade-off parameter used to apply the importance of the center distance in the loss function. is a function of the quantized aspect ratio metric.

$$v = \frac{4}{\pi^2}\left(\arctan\frac{wgt}{hgt} - \arctan\frac{w}{h}\right)^2 \tag{12}$$

In(12) , w and h denote the height and width of the predicted frame, respectively, and denote the height w_{gt} and h_{gt} width of the true frame, respectively. The NWD loss function is defined as follows:

$$L_{NWD} = 1 - NWD(N_a, N_b), \tag{13}$$

$$NWD(N_a, N_b) = \exp\left(-\frac{\sqrt{w_2^2(N_a, N_b)}}{C}\right), \tag{14}$$

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right]^T \right) \right\|_2^2 \tag{15}$$

c is a constant with strong correlation to the dataset and $W_2^2(N_a N_b)$ is a distance metric. $\mathbb{N}_a$ and $\mathbb{N}_b$ are Gaussian distributions modeled by the bounding boxes $A(cx_a, cy_a, w_a, \hat{h}_a)$ and $B(c_{X_b}, c_{Y_b}, w_b, \hbar_b)$ their centers.

4. Experiments and discussions

This section primarily describes the composition of the dataset used in the experiments, along with relevant details and evaluation metrics. To thoroughly validate the effectiveness of the proposed improved algorithm, a series of comparative experiments have been designed and conducted. To ensure fairness and reproducibility, all experiments are performed without using pre-trained weights, and consistent parameter settings and training strategies are strictly maintained throughout.

4.1. Introduction to the data set

The dataset used in this study is primarily sourced from the InfiRay infrared open platform , which provides real-scene infrared ship image data designed to support the

validation and evaluation of infrared target detection algorithms. The dataset is captured using infrared imaging devices with varying resolutions and focal lengths, covering a wide range of typical ship types. It contains a total of 8,402 infrared images labeled with seven categories of ship targets: liner, bulk carrier, warship, sailboat, canoe, container ship, and fishing boat. The image resolutions vary widely, with some images at 384×288 and 1280×1024 pixels, while the remainder are below 640×640 pixels. During model training, all images are uniformly resized to 640×640 pixels. The dataset is split into training, validation, and test sets at an approximately 8:1:1 ratio, with 6,048 images for training, 1,277 for validation, and 1,077 for testing. The images are mainly captured in coastal ports and active ship areas, where sailboats and canoes constitute a relatively high proportion, and fishing boats are commonly distributed in coastal port zones. The presence of numerous small targets in the dataset adds to the challenge of infrared ship detection and realistically reflects the difficulties encountered in maritime environments. To further validate the generalization ability and robustness of the proposed method, the high-resolution satellite image dataset HRSID[38] is also introduced in this paper. This dataset is widely used in remote sensing image processing and computer vision research, and contains 99 images from Sentinel-1B, 36 images from TerraSAR-X, and 1 image from TanDEM-X, with the image resolution ranging from 1 m to 15 m, and unified in 800×800 image format. A total of 16951 vessels are labeled in the whole dataset, distributed in 5604 sliced SAR images, which provides sufficient samples for the robust training and performance evaluation of the model. In the experiments, the HRSID dataset is used as 70% for training, 20% for testing, and 10% for validation in order to comprehensively measure the detection effect and adaptability of the model in different types of images.

The detailed parameters of the two datasets are presented in Table 1., which outlines the dataset sources, sizes, the number of images in the training, validation, and test sets, as well as the total number of vessels included in each dataset. Figure 7. showcases sample images from both datasets. The InfiRay dataset contains infrared images captured from various angles, featuring different vessel types and diverse backgrounds. In contrast, the HRSID dataset primarily consists of small vessel targets, making it particularly suitable for evaluating the model's performance in small object detection.

Table 1. Parameters of the HRSID dataset and InfiRay dataset

Parameters	HRSID	InfiRay
Data source	Sentinel-1B,TerraSAR-X,TanDem-X	InfiRay
Image size/pixels	800 × 800	384x2881280x1024640x640
Number of training set images	3922	6048
Number of test images	1121	1077
Number of val images	561	1277
Total number of ship target	16951	31307

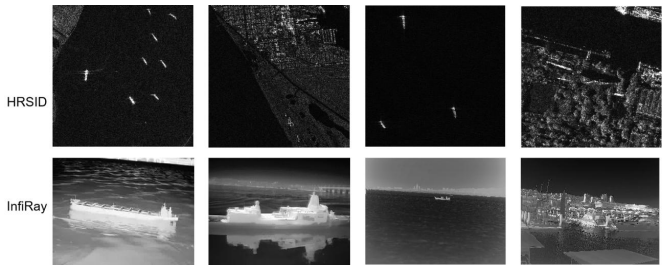

Fig. 7. Example of HRSID training set and InfiRay training set

4.2. Experimental environment

The experiments in this study were conducted on a computer running Windows 10. The hardware configuration includes a 12th-generation Intel® Core™ i7-4214R processor with a base clock speed of 2.40 GHz, paired with an NVIDIA GeForce RTX 3080 Ti graphics card. The software environment was built using Python 3.8 and PyTorch 2.2.1. Detailed specifications of the experimental platform are provided in Table2.

Table 2. Configuration information for the experimental platform

Configuration	Versions
Operating system	Windows 10
CPU	12 vCPU Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz
GPU	NVIDIA GeForce RTX 3080Ti
RAM	32GB
Toolkit	CUDA 12.1
Compiler	Python 3.8
Framework	PyTorch 2.2.1

The experimental parameters for training the model are shown in Table3. The experiments were conducted using an input image size of 640 × 640 pixels over 150 training cycles. The initial and final learning rates were set at 0.01.

Table 3. Experimental parameters of the training model

Name	Configuration
Momentum	0.937
Data enhancement	Mosaic
Epochs	150
Batch size	16
Workers	16
Image size	640×640
Initial learning rate	0.01
Final learning rate	0.01

4.3. Performance evaluation

In this study, to comprehensively evaluate the model's performance, we employ the COCO evaluation metrics, including Precision (P), Recall (R), mean Average Precision (mAP), GFLOPS, and the number of parameters.Precision (P) represents the probability that a detected target is correctly identified and is defined as follows.

$$P = \frac{TP}{TP + FP} \times 100\% \tag{16}$$

In the target detection task, TP (True Positive) denotes the number of correctly identified positive samples, while FP (False Positive) refers to the number of incorrectly detected negative samples.Recall (R) represents the probability of correctly identifying positive samples among all actual positive samples and is defined as follows:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{17}$$

Here, TP (True Positive) refers to the number of correctly identified positive samples, while FN (False Negative) represents the number of missed positive samples. Average Precision (AP) indicates the average detection accuracy for a single category and is defined as follows.

$$AP = \int_0^1 P \, dR \tag{18}$$

mAP (mean Average Precision) is used as the core index to evaluate the performance of the target detection model, and its higher value indicates the better detection accuracy of the model, which is defined as follows.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}. \tag{19}$$

GFLOPS (Giga Floating-point Operations Per Second) quantifies the computational capacity of the model, representing the number of billions of floating-point operations the model can perform per second. Meanwhile, model parameters—comprising the network's weights and biases—directly impact the model's computational complexity.

4.4. Ablation experiments

The proposed ADN-YOLO model demonstrates enhanced detection performance across all categories in the infrared ship dataset (infiRay). It significantly improves the detection of small targets in multi-scale scenarios and performs robustly under conditions of dense occlusion. This study introduces three innovative modules built upon the YOLOv11 framework: ADown, DySample, and NWD-CIoU. As shown in Table 4.,Table 5.. eight ablation experiments were conducted to rigorously assess the contribution of each module to the overall detection accuracy. The ADown module combines average pooling and max pooling to enable multi-scale feature extraction in infrared ship images. This approach led to a 0.9% increase in Recall, a 0.5% improvement in mAP@0.5, a reduction of 0.46M in parameters, and a decrease of 1.2 GFLOPs. DySample replaces traditional

dynamic convolution with a point-based sampling technique, yielding a 0.7% increase in mAP@0.5 and a 0.8% improvement in Recall. The NWD-CIoU module enhances the model's sensitivity to subtle distributional differences in small targets by integrating Normalized Wasserstein Distance (NWD) with Complete IoU (CIoU). This integration resulted in a 1.2% gain in Recall and a 0.8% increase in mAP@0.5.Collectively, these enhancements significantly boost the model's performance across all key evaluation metrics.

Table 4. Comparison tests on the InfiRay dataset

ADown	Dysample	NWD-CIoU	Precision (%)	Recall (%)	mAP@0.5 (%)
			90.1	83.1	88.7
✓			89.7	84.0	89.2
	✓		90.0	83.9	89.4
		✓	89.5	84.3	89.5
✓	✓		90.8	84.3	89.8
✓		✓	90.4	85.1	90.0
	✓	✓	90.7	84.4	89.9
✓	✓	✓	91.1	85.0	90.6

Table 5. Comparison tests on the InfiRay dataset

ADown	Dysample	NWD-CIoU	mAP@0.5:0.95 (%)	Params (M)	FLOPs (G)
			62.1	2.56	6.3
✓			61.8	2.10	5.1
	✓		62.4	2.59	6.3
		✓	62.0	2.56	6.3
✓	✓		62.5	2.14	5.2
✓		✓	61.9	2.10	5.3
	✓	✓	62.3	2.62	6.5
✓	✓	✓	62.8	2.14	5.2

PR curves are generated by computing precision and recall at various threshold levels. The horizontal axis represents recall, indicating the proportion of actual positive cases that the model correctly identifies. The vertical axis represents precision, indicating the proportion of predicted positive cases that are truly positive. PR curves provide a comprehensive view of model performance across different thresholds. Typically, curves that are closer to the upper right corner reflect a model's ability to achieve both high precision and high recall. A larger area under the curve (AUC) signifies better overall performance, with the blue curve representing the average performance across all categories.Figure 8 illustrates the PR curves for both ADN-YOLO and YOLOv11. In this figure, the proximity of a curve to the upper right corner indicates superior detection performance. Figure 8.(a) and Figure 8.(b) show the mAP@0.5 values for various categories in the dataset at the lower left. On the infrared ship dataset, the results in Figure 8.(b) demonstrate notable improvements across all categories compared to Figure 8.(a), with particularly significant

gains for the 'liner' and 'fishing boat' categories. This highlights the enhanced ability of ADN-YOLO to capture target features in complex maritime environments.Overall, the PR curves confirm that ADN-YOLO achieves a well-balanced trade-off between precision and recall.

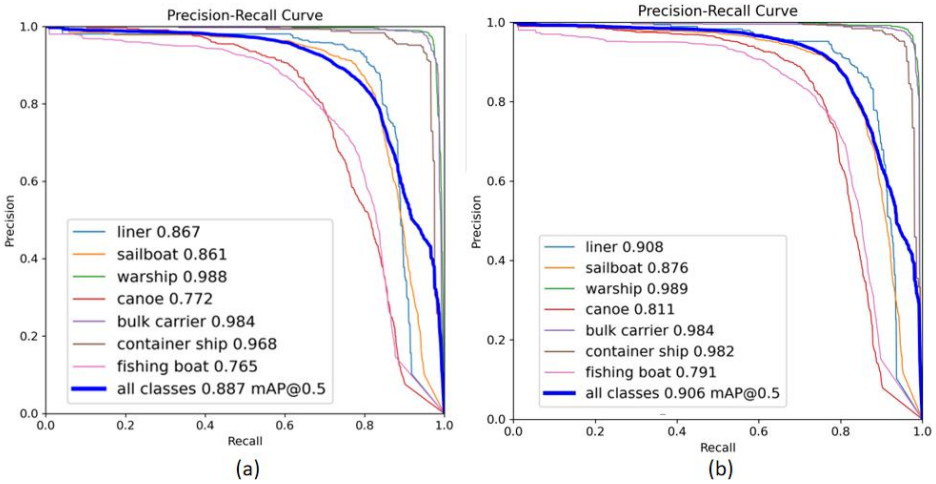

Fig. 8. Comparison of PR curves for the test set models: the PR curves of (a) YOLOv11 and (b) ADN-YOLOv11

A confusion matrix is a visualization tool that presents the performance of a classification algorithm in a matrix format, offering a clear view of how the model performs across different categories. Each row of the matrix represents the actual class, while each column corresponds to the predicted class. In this context, the horizontal axis denotes the true labels, and the vertical axis denotes the predicted precision. Each horizontal coordinate of the matrix represents a category in the dataset, including seven types of ship targets: liner, bulk carrier, warship, sailboat, canoe, container ship, and fishing boat.The elements along the main diagonal represent the number of correctly classified samples. The lower left triangular region indicates missed detections (false negatives), whereas the upper right triangular region reflects false detections (false positives).When values are high in the lower-left region, it suggests that the model has missed many detections. Conversely, high values in the upper-right region indicate a greater number of false detections. As shown in Figure 9., the confusion matrix of ADN-YOLO (after improvements to YOLOv11) reveals a notable enhancement in detection accuracy. Compared to Figure 9.(a), the reduced values in the upper-right region of Figure 9.(b) indicate a significant decrease in false detections. Similarly, the lower values in the lower-left region of Figure 9.(b) demonstrate a reduction in missed detections.These results confirm that the improved model is more effective in detecting targets within complex maritime environments, with marked improvements in both false and missed detections.

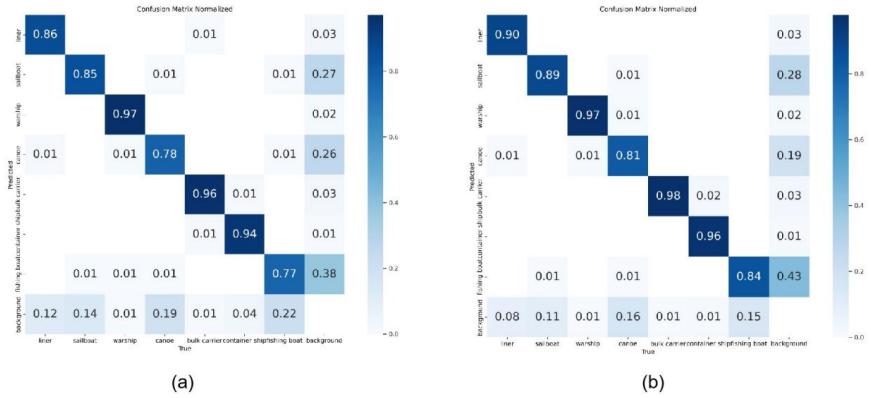

Fig. 9. Comparison of the confusion matrices of the test set models: This figure presents a visual comparison between the confusion matrices of two models evaluated on the test set: (a) the baseline YOLOv11 and (b) the improved ADN-YOLOv11. The comparison highlights differences in classification performance, including true positives, false positives, and false negatives across different categories

Figure 10. presents the visual results of ship detection using HiResCAM for both YOLOv11 and ADN-YOLO. HiResCAM is a model interpretability technique that visualizes attention by capturing gradient information to generate heatmaps. Compared to Grad-CAM—which tends to spread attention and may be better suited for segmentation tasks—HiResCAM offers a more faithful and class-specific interpretation of model decisions [39]. This method provides an intuitive means of visualizing how and where the model focuses during inference, helping to explain its decision-making process. In Figure 10.(a), the detection region of the ship in the original image is annotated to clarify the ship's position. This baseline aids in comparing the attention maps produced by the two models. As seen in Figure 10.(c), ADN-YOLO generates a heatmap with concentrated, darker regions, indicating a strong and focused response to the actual target. This suggests that the model effectively concentrates on critical features related to the ship, improving its precision in infrared ship detection.In contrast, Figure 10.(b) shows that YOLOv11 produces a more dispersed heatmap, with attention distributed over a broader area, including potentially irrelevant background regions. This contrast highlights the superior feature extraction and localization capabilities of ADN-YOLO, particularly in detecting occluded or small ship targets.Overall, compared to the original algorithm, ADN-YOLO demonstrates enhanced focus on ship features and better suppression of background noise. The highlighted areas in the heatmaps more closely align with the true locations of ships in the original image, underscoring the model's improved accuracy, robustness, and interpretability in complex maritime environments.

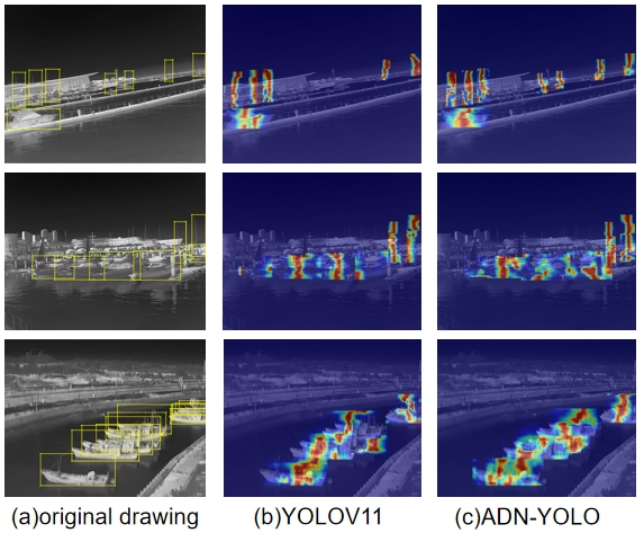

Fig. 10. Visualization of the infiRay dataset.(a) shows the original annotation boxes of the ships, (b) represents the heatmap of YOLOv11, and (c) represents the heatmap of ADN-YOLO

In complex maritime environments, the wide variation in ship scales, combined with the challenges of infrared imaging,such as low-light conditions and target occlusion, makes accurate detection particularly difficult [40,41]. Small targets, in particular, often exhibit weak texture features, and the limited feature extraction capability of YOLOv11 may result in issues such as false detections or missed targets. As illustrated in Figure 11.(a) comparison of the prediction results between YOLOv11 and ADN-YOLO reveals that small targets like canoes are not detected by YOLOv11 in Figure 11.(a), largely due to the reduced visibility inherent in infrared imaging. In densely occluded scenes, or when the target appears at a small scale, YOLOv11 is prone to interference, which negatively impacts detection and localization accuracy. In contrast, Figure 11.(b) demonstrates the results obtained using the proposed ADN-YOLO model. Even under low-visibility conditions, ADN-YOLO is capable of accurately detecting small targets such as canoes. Compared to YOLOv11, ADN-YOLO exhibits significantly enhanced performance, particularly in scenes with heavy occlusion. It achieves superior results in terms of detection accuracy, robustness, and reliability. These improvements make ADN-YOLO better suited for real-world applications in ship detection, especially in challenging infrared imaging scenarios.

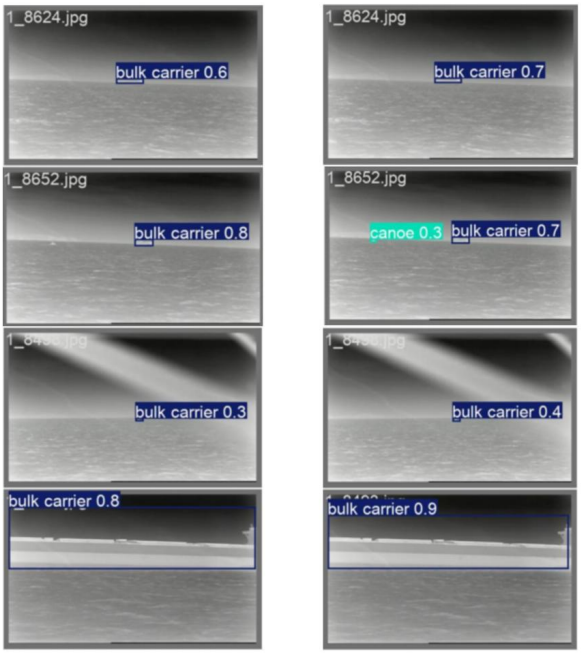

Fig. 11. Comparison of infrared ship detection.(a) shows the ship detection results of YOLO11, and (b) shows the ship detection results of ADN-YOLO

4.5. Comparisons with other methods

In this section, to qualitatively evaluate the detection performance of the proposed ADN-YOLO model on the InfiRay dataset, we compare it against several mainstream YOLO models, including YOLOv3-tiny[42] , YOLOv5 [43], YOLOv6[44] , YOLOv8[45] , YOLOv9-tiny[46] , YOLOv10[47] , YOLOv11[48] , and YOLOv12[49] , using the test set for ship image detection. The comparison results are presented in Table 6.Among the models, YOLOv8 achieves the highest average accuracy at 89.9%. However, its large number of parameters and high FLOPs make it less suitable for real-time applications. In contrast, the proposed ADN-YOLO significantly outperforms the baseline YOLOv11 model, achieving improvements of 1.0%, 1.9%, and 1.9% in precision, recall, and mAP@0.5, respectively reaching 91.1%, 85.0%, and 90.6%.Furthermore, ADN-YOLO reduces the number of parameters by 0.4 million compared to YOLOv11, and by 0.9 million when compared to YOLOv8, demonstrating its superior efficiency. These results confirm that ADN-YOLO is more suitable for real-time ship detection in infrared images, offering a favorable balance between detection accuracy and computational efficiency. It achieves strong detection performance while maintaining a lightweight model architecture, making it highly effective for practical deployment in complex maritime scenarios.

Table 6. Comparison tests on the InfiRay dataset

Model	Precision/%	Recall/%	Map@0.5/%	map@0.5:0.95/%	Params/M	FLOPs/G
Yolov3-tiny	89.8	79.1	85.2	59.4	12.1	18.9
Yolov5	88.8	82.7	88.9	61.3	2.5	7.1
Yolov6	89.8	82.2	88.2	61.5	4.2	11.8
Yolov8	90.1	85.3	89.9	62.5	3.0	8.1
Yolov9-tiny	90.6	84.3	89.8	62.7	1.9	7.6
Yolov10	86.6	81.1	88.8	61.2	2.7	8.2
Yolov11	90.1	83.1	88.7	62.1	2.6	6.3
Yolov12	90.3	84.4	89.0	61.4	2.55	6.3
ADN-YOLO	91.1	85	90.6	62.8	2.1	5.2

presents the evaluation results of the proposed model in terms of Figure 12mAP@0.5 and Recall after 150 training epochs. Due to challenges such as low resolution, insufficient illumination, and noise interference inherent in infrared imagery, vessel targets are often missed during detection. As shown in Figure 12(b), the proposed ADN-YOLO model significantly outperforms the comparative models in Recall, demonstrating a clear advantage in reducing missed detections. This indicates the model's effectiveness in mitigating the adverse effects of complex backgrounds, poor lighting conditions, and other interference factors common in infrared images.The mAP@0.5 metric is a critical indicator of object detection performance—higher values signify better model capability in distinguishing targets from background noise and reducing both missed and false detections. As illustrated in Figure 12.(a), the mAP@0.5 curve of the proposed model is consistently higher than those of other models, underscoring its superior detection accuracy. This suggests that ADN-YOLO is more effective in recognizing vessel targets within infrared imagery and is capable of detecting a greater number of valid targets.In summary, ADN-YOLO demonstrates higher precision and recall in infrared ship detection tasks, reflecting its robustness and overall performance advantages in complex maritime environments.

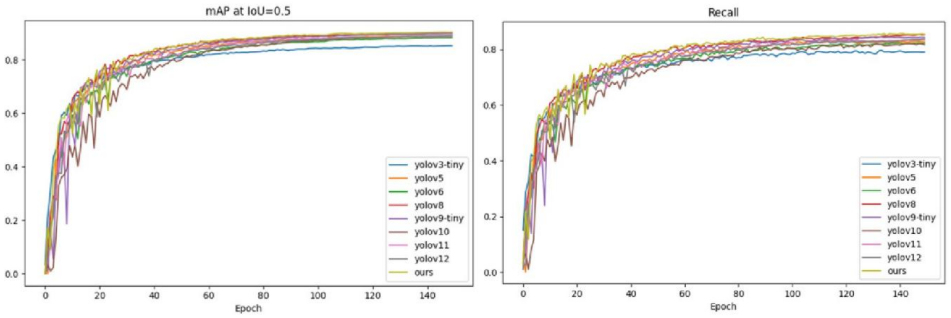

Fig. 12. mAP for each model in the InfiRay dataset and Recall.(a) shows the curves of each model with respect to the mAP@0.5 metric, and (b) shows the curves of each model in terms of Recall

In this study, comparative experiments were conducted on the HRSID dataset to further validate the stability and generalization ability of the proposed model. The performance of ADN-YOLO was compared with several mainstream detection algorithms, including YOLOv3-tiny [42], YOLOv5 [43], YOLOv6 [44], YOLOv8 [45], YOLOv9-tiny [46], YOLOv10 [47], YOLOv11 [48], and the latest YOLOv12 [49]. The comparison focuses on the models' performance in ship image detection on the validation set. The specific experimental results are presented in Table 7.As shown in the table, YOLOv9-tiny has the smallest number of parameters. While it offers certain advantages in terms of lightweight design, its mAP@0.5 and Recall scores are 0.7% and 2.0% lower, respectively, than those of the proposed ADN-YOLO, indicating some limitations in detection performance. Compared with the baseline model YOLOv11, ADN-YOLO achieves improvements of 1.9%, 2.1%, and 2.9% in Precision, Recall, and mAP@0.5:0.95, respectively, significantly enhancing overall detection capability.This performance improvement is particularly meaningful in practical applications especially in complex maritime scenarios where small vessels are more prone to being missed or misidentified. Such advancements contribute to enhanced system robustness and safety. In summary, ADN-YOLO not only demonstrates superior detection performance for ship targets in noisy, low-light infrared images but also shows strong adaptability in small-object detection tasks, highlighting its broad application potential.

Table 7. Comparison tests for the HRSID dataset

Model	Precision/%	Recall/%	Map@0.5/%	map@0.50.95/%	Params/M	FLOPs/G
Yolov3-tiny	90.4	70.4	81.6	58.2	12.1	18.9
Yolov5	89.2	82.3	88.3	63.1	2.50	7.1
Yolov6	88.1	81.0	86.7	60.7	4.23	11.8
Yolov8	89.8	83.6	88.5	66.5	3.01	8.1
Yolov9-tiny	92.3	81.5	90.8	66.4	1.97	7.6
Yolov10	91.6	80.6	90.3	66.0	2.69	8.2
Yolov11	89.7	81.4	91.0	63.3	2.58	6.3
Yolov12	89.4	79.9	88.4	63.1	2.56	6.3
ADN-YOLO	91.6	83.5	91.5	66.2	2.25	5.2

Figure 13 presents various evaluation metrics after 150 training epochs on the HRSID dataset. The vertical position of each circle represents the mAP@0.5 higher value placement indicates better detection performance, especially for small vessel targets at sea. The size of each circle intuitively reflects the number of model parameters, which corresponds to the model's complexity and learning capacity.From the figure, it is evident that our model achieves superior performance in terms of parameter efficiency compared to other models. Horizontally, the position of each circle indicates the GFLOPs, with a greater distance to the right representing a heavier computational load. As seen in the figure, YOLOv3-tiny lies farthest to the right, indicating a higher computational burden. In contrast, our model is positioned closest to the left edge of the horizontal axis, reflecting its low computational cost and faster inference speed—making it well suited for real-time detection of small vessel targets.In summary, the proposed model not only offers faster inference speed than competing models but also achieves a higher mAP@0.5 score. It ef-

fectively detects small vessel targets while preserving detailed features, even under rapid inference conditions. Additionally, it significantly reduces both false positives and missed detections, highlighting its practical advantage in real-world applications.

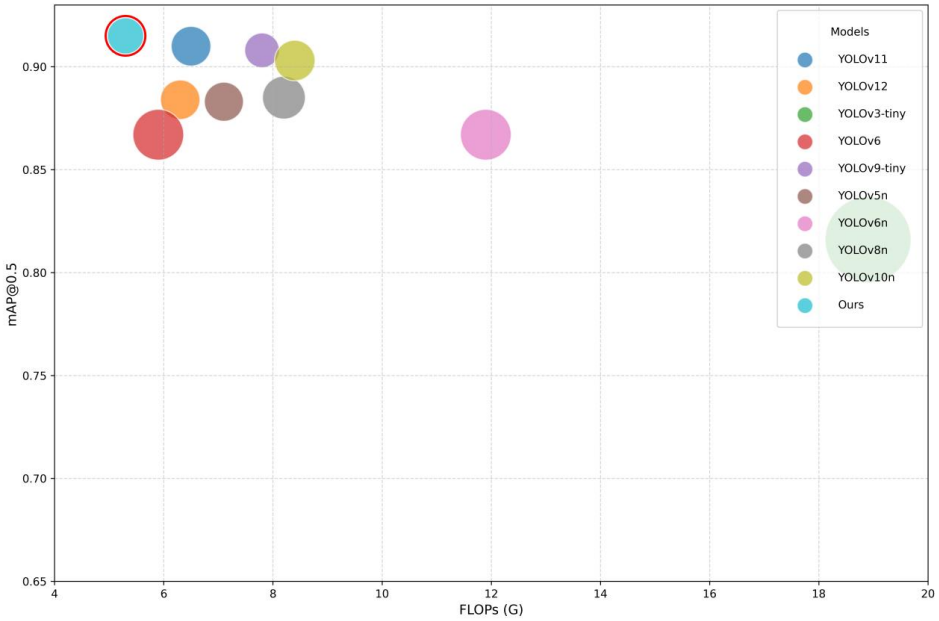

Fig. 13. Bubble plots for each model in the HRSID dataset

In complex marine environments, noise interference is common due to the structural similarities among vessels. Detecting small targets often encounters challenges such as overlapping objects and weak responses on heat maps. To better illustrate these issues, we visualized detection results on the HRSID dataset[50][51][52], enabling direct observation of missed and misdetected vessels. To compare our improved model with the baseline, we selected three representative images from the HRSID dataset depicting complex sea scenes, as shown in Figure 14 These images contain numerous small vessel targets. In our visualization, different colored bounding boxes are used to distinguish detection outcomes: yellow boxes indicate ground-truth annotations from the original dataset, blue boxes represent correctly identified vessels by the model, orange boxes mark vessels missed by the model, and purple boxes denote falsely detected vessels. This color scheme facilitates a clearer understanding and evaluation of the model's accuracy and errors. As shown in Figure 14.(a), small vessels are densely clustered, making detection heavily influenced by the surrounding environment. In Figure 14.(b), the baseline model exhibits both false positives and missed detections, with environmental interference exacerbating these errors. In contrast, Figure 14.(c) displays results from our improved model, demonstrating significant reductions in missed detections of small vessels and accurate recognition even in complex scenarios.

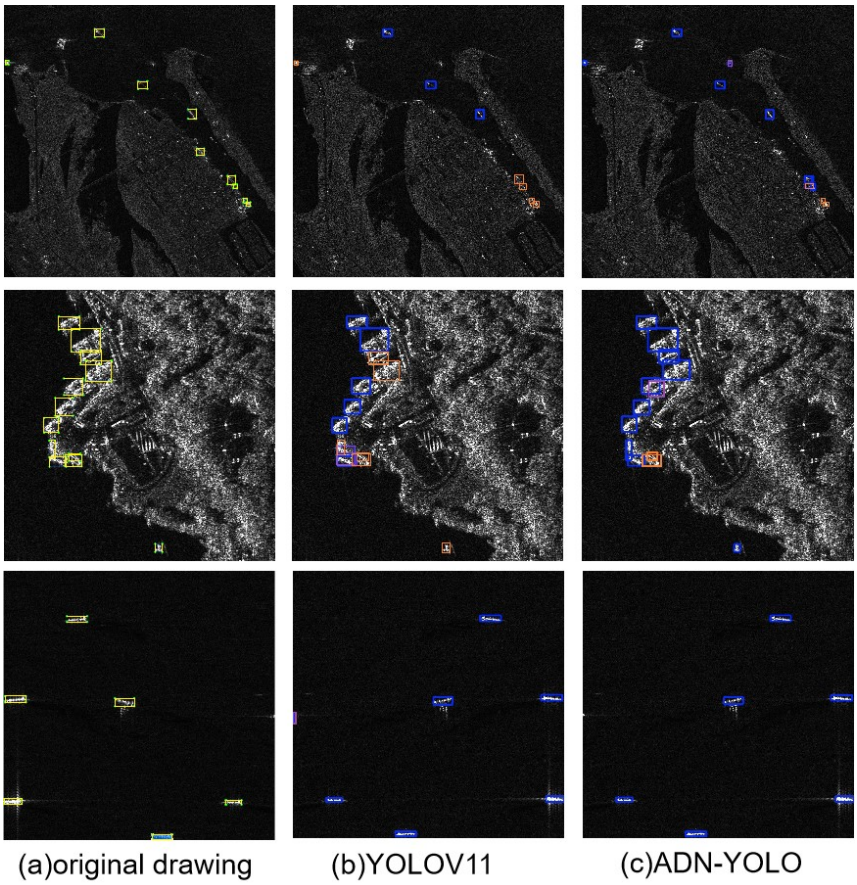

Fig. 14. Heatmaps on the HRSID dataset (a) Vessel distribution in the original image, with yellow boxes representing the ground-truth annotations (b) Detection results from YOLOv11: blue boxes indicate correctly detected vessels, orange boxes indicate missed detections, and purple boxes indicate false detections. (c) Detection results from ADN-YOLO: blue boxes denote correctly detected vessels, and orange boxes denote missed detections

5. Conclusion

In this study, we propose an improved infrared ship detection model, ADN-YOLO, designed to address the challenges inherent to infrared scenes, such as the wide distribution of small target ships, ambiguous features, and susceptibility to background interference. The model is structurally optimized based on the YOLOv11 framework and incorporates dynamic sampling, attention mechanisms, and a novel loss function to comprehensively enhance detection accuracy and robustness for infrared small targets.Specifically, ADN-YOLO introduces a lightweight and efficient ADown downsampling module into

the backbone network, replacing the traditional CBS structure. This module enhances the extraction of salient features for small targets while reducing the model's parameter count and computational complexity. Additionally, the Dysample dynamic upsampling module reconstructs the upsampling process from a point-sampling perspective, effectively preserving shallow features and spatial location information. This further improves the model's accuracy and robustness in detecting small infrared targets, especially under complex backgrounds and weak target conditions.Moreover, to tackle the instability of bounding box regression using the traditional IoU loss function in small target detection, this study proposes the NWD-CIoU metric. By modeling the predicted and ground-truth boxes as two-dimensional Gaussian distributions and employing the normalized Wasserstein distance to measure their differences, NWD-CIoU significantly enhances the model's generalization across targets of varying scales and shapes.Comprehensive experimental results demonstrate that ADN-YOLO achieves superior detection accuracy and recall in infrared image target detection, particularly for small targets, validating the effectiveness and practical value of the proposed method.

However, the method proposed in this paper still has certain limitations in practical applications. While the model has achieved significant improvements in detection accuracy, there remains room for optimization in inference efficiency, as the current running speed and resource consumption do not fully meet the demands of extreme real-time scenarios. Additionally, the model has yet to be thoroughly validated across a broader range of infrared detection tasks, and its adaptability to complex environments and cross-scene generalization requires further investigation. In future work, we plan to extend the model to other infrared target detection applications such as night surveillance and border defense warning to further verify its generality and stability.

Acknowledgments. This work was Supported by Key Lab of Information Network Security, Ministry of Public Security,the National Natural Science Foundation of China under Grants 61672338, the Natural Science Foundation of Shanghai under Grant 21ZR1426500.

References

1. Shuhong Wang, Weiyao Li, and Lu Xing. A review on marine economics and management: How to exploit the ocean well. *Water*, 14(17):2626, 2022.
2. Mingming Cui, Dezhi Han, Han Liu, Kuan-Ching Li, Mingdong Tang, Chin-Chen Chang, Ferheen Ayaz, Zhengguo Sheng, and Yong Liang Guan. Secure data sharing for consortium blockchainenabled vehicular social networks. *IEEE Transactions on Vehicular Technology*, 2024.
3. Liyuan Li, Linyi Jiang, Jingwen Zhang, Siqi Wang, and Fansheng Chen. A complete yolo-based ship detection method for thermal infrared remote sensing images under complex backgrounds. *Remote Sensing*, 14(7):1534, 2022.
4. Chongqing Chen, Dezhi Han, and Xiang Shen. Clvin: Complete language-vision interaction network for visual question answering. *Knowledge-Based Systems*, 275:110706, 2023.
5. Renke Kou, Chunping Wang, Zhenming Peng, Zhihe Zhao, Yaohong Chen, Jinhui Han, Fuyu Huang, Ying Yu, and Qiang Fu. Infrared small target segmentation networks: A survey. *Pattern recognition*, 143:109788, 2023.
6. Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007.

7. Chongqing Chen, Dezhi Han, and Xiang Shen. Clvin: Complete language-vision interaction network for visual question answering. *Knowledge-Based Systems*, 275:110706, 2023.
8. Chongqing Chen, Dezhi Han, Zihan Guo, and Chin-Chen Chang. Towards bias-aware visual question answering: Rectifying and mitigating comprehension biases. *Expert Systems with Applications*, 264:125817, 2025.
9. Landan Zhang and Zhenming Peng. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sensing*, 11(4):382, 2019.
10. Dezhi Han, Nannan Pan, and Kuan-Ching Li. A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection. *IEEE Transactions on Dependable and Secure Computing*, 19(1):316–327, 2020.
11. GU Jiaojiao, LI Bingzhen, LIU Ke, and JIANG Wenzhi. Infrared ship target detection algorithm based on improved faster r-cnn. *Infrared Technology*, 43(2):170–178, 2021.
12. Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 821–830, 2019.
13. Yang Zheng, Izzat H Izzat, and Shahrzad Ziaee. Gfd-ssd: Gated fusion double ssd for multispectral pedestrian detection. *arXiv preprint arXiv:1903.06999*, 2019.
14. Shasha Li, Yongjun Li, Yao Li, Mengjun Li, and Xiaorong Xu. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE access*, 9:141861–141875, 2021.
15. Xiao Luo, Hao Zhu, and Zhenli Zhang. Ir-yolo: Real-time infrared vehicle and pedestrian detection. *Computers, Materials & Continua*, 78(2), 2024.
16. Jihao Si, Binbin Song, Jixuan Wu, Wei Lin, Wei Huang, and Shengyong Chen. Maritime ship detection method for satellite images based on multiscale feature fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:6642–6655, 2023.
17. Bowen Xing, Wei Wang, Jingyi Qian, Chengwu Pan, and Qibo Le. A lightweight model for real-time monitoring of ships. *Electronics*, 12(18):3804, 2023.
18. Nan Wang, Bo Li, Xingxing Wei, Yonghua Wang, and Huanqian Yan. Ship detection in spaceborne infrared image based on lightweight cnn and multisource feature cascade decision. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4324–4339, 2020.
19. Jiatao Li, Dezhi Han, Tien-Hsiung Weng, Huafeng Wu, Kuan-Ching Li, and Arcangelo Castiglione. A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. *Computer Standards & Interfaces*, 91:103887, 2025.
20. K Zhao and X Kong. Background noise suppression in small targets infrared images and its method discussion. *Optics and Optoelectronic Technology*, 2(2):9–12, 2004.
21. TS Anju and NR Nelwin Raj. Shearlet transform based image denoising using histogram thresholding. In *2016 International Conference on Communication Systems and Networks (ComNet)*, pages 162–166. IEEE, 2016.
22. P Jiao. Research on image classification and retrieval method based on deep learning and sparse representation. *Xi'an University of Technology: Xi'an, China*, 2019.
23. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
24. Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
25. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
26. Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. Pedestrian detection in thermal images using saliency maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

27. Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
28. Chongqing Chen, Dezhi Han, and Chin-Chen Chang. Caan: Context-aware attention network for visual question answering. *Pattern Recognition*, 132:108980, 2022.
29. Yangyang Li, Qin Huang, Xuan Pei, Yanqiao Chen, Licheng Jiao, and Ronghua Shang. Cross-layer attention network for small object detection in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2148–2161, 2020.
30. Hao Guo and Dongbing Gu. Closely arranged inshore ship detection using a bi-directional attention feature pyramid network. *International Journal of Remote Sensing*, 44(22):7106–7125, 2023.
31. Chongqing Chen, Dezhi Han, and Chin-Chen Chang. Mpcct: Multimodal vision-language learning paradigm with context-based compact transformer. *Pattern recognition*, 147:110084, 2024.
32. Xiang Shen, Dezhi Han, Chin-Chen Chang, Ammar Oad, and Huafeng Wu. Gfsnet: Gaussian fourier with sparse attention network for visual question answering. *Artificial Intelligence Review*, 58(6):1–30, 2025.
33. Wenji Yang and Xiaoying Qiu. A lightweight and efficient model for grape bunch detection and biophysical anomaly assessment in complex environments based on yolov8s. *Frontiers in Plant Science*, 15:1395796, 2024.
34. Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.
35. Yufei Zhao, Yong Liang Guan, Dong Chen, Afkar Mohamed Ismail, Xiaoyan Ma, Xiaobei Liu, and Chau Yuen. Exploring rcs diversity with novel oam beams without energy void: An experimental study. *IEEE Transactions on Vehicular Technology*, 2024.
36. Liliang Zhou, Huaming Ran, Rongling Xiong, and Ruijie Tan. Nwd-yolov5: A yolov5 model for small target detection based on nwd loss. In *2024 6th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pages 542–546. IEEE, 2024.
37. Yufei Zhao, Ziyang Wang, Yilong Lu, and Yong Liang Guan. Multimode oam convergent transmission with co-divergent angle tailored by airy wavefront. *IEEE Transactions on Antennas and Propagation*, 71(6):5256–5265, 2023.
38. Xiao Tang, Jiufeng Zhang, Yunzhi Xia, and Huanlin Xiao. Dbw-yolo: A high-precision sar ship detection method for complex environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
39. Tobias Clement, Truong Thanh Hung Nguyen, Mohamed Abdelaal, and Hung Cao. Xai-enhanced semantic segmentation models for visual quality inspection. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–4. IEEE, 2024.
40. Dun Li, Dezhi Han, Noel Crespi, Roberto Minerva, Syed Mohsan Raza, Reza Farahbakhsh, Wei Liang, and Zibin Zheng. Blockchain in the digital twin context: A comprehensive survey. *ACM Computing Surveys*, 2025.
41. Rui Jiang, Hang Shi, Jiahong Ni, Jiatao Li, Yi Feng, Xinqiang Chen, and Yinlin Li. Lsdformer: Lightweight sar ship detection enhanced with efficient multi-attention and structural reparameterization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–20, 2025.
42. Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
43. Bharat Mahaur and KK Mishra. Small-object detection based on yolov5 in autonomous driving systems. *Pattern Recognition Letters*, 168:115–122, 2023.

44. Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.
45. Muhammad Hussain. Yolov1 to v8: Unveiling each variant–a comprehensive review of yolo. *IEEE access*, 12:42816–42833, 2024.
46. Jialin Zou and Hongcheng Wang. Steel surface defect detection method based on improved yolov9 network. *IEEE Access*, 2024.
47. Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.
48. Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.
49. Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.
50. Jiatao Li, Dezhi Han, Shuxin Shi, Xiaoqi Xin, Kuan-Ching Li, and Chin-Chen Chang. An active client selection scheme based on blockchain for federated learning in shipping. *IEEE Trans. Intell. Transp. Syst.*, 26(11):20669–20684, Nov. 2025.
51. Rui Jiang, Ruixiang Zhu, Hu Su, Yinlin Li, Yuan Xie, and Wei Zou. Deep learning-based moving object segmentation: Recent progress and research prospects. *Int. J. Automat. Comput.*, 20(3):335–369, Jun 2023.
52. Jiatao Li, Dezhi Han, Tien-Hsiung Weng, Huafeng Wu, Kuan-Ching Li, and Arcangelo Castiglione. A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. *Comput. Stand. Interfaces*, 91, Jan. 2025. Art. no. 103887.

Tao Li is currently pursuing a master's degree in the School of Information Engineering at the China University of Shipping in Pudong. His current research interests are in network security.

Dezhi Han received his Bachelor of Science in applied physics from Hefei University of Technology in 990, and his Master of Science and Doctor of Philosophy in computer science from Wuhan University of Science and Technology in China in 2001 and 205, respectively. He is currently a professor in the Department of Computer Science at the China University of Shipping in Shanghai Pudong, where he has been employed since 200. His current research interests include cloud computing and security, blockchain, wireless communication security, network, and information security.

Songyang Wu is a researcher and director of the Security Center at the Third Research Institute of Ministry of Public Security, and also serves as the deputy director of the National Engineering Research Center for Network Security Grade and Security Technology and the deputy director of the Key Laboratory of Information Network Security of Ministry of Public Security. He received his Bachelor of Science in computer science and technology from Tongji University in 200 and his Doctor of Philosophy in computer application from Tongji University in 2011. The same year, he joined the Ministry of Public Security Center. He received his Doctor Philosophy in computer application from Tongji University in 2011. The same year, he joined the Network Security Center at the Third Research Institute of Ministry of Public Security. research interests include cybercrime investigation, electronic data forensics, big data security, and artificial intelligence security.

Xiang Shen received his Master of Science in computer science technology from the Shanghai Maritime University in Shanghai Pudong, China, in 2022 and is currently pursuing a doctoral degree in the School of Information Engineering. current research interests are in network security.

Liqi Zhu received his Master of Science in computer science and technology from the Shanghai Maritime University in Shanghai Pudong,, in 2025 and is currently pursuing a doctoral degree in the School of Information Engineering. His current research interests are in blockchain federated learning.

Wenqi Sun is an associate researcher and research engineer at the Network Security Center at the Third Research Institute of Ministry of Public Security; she received her computer science and technology degree from Northeast University 2010 and her doctorate in computer science and technology from Tsinghua University in 2016. In 2018, she joined the Network Center at the Third Research Institute of Ministry of Public Security; her current research interests are in network investigations.

Received: June 13, 2025; Accepted: December 14, 2025.