MHOE-DETR: A Ship Detection Method for Small and Fuzzy Targets Based on Satellite Remote Sensing Image Data

Zhuhua Hu, *Senior Member, IEEE*, Xiyu Fan, Yaochi Zhao, Wei Wu, and Jie Liu

***Abstract*—Pinpointing elusive and minor target vessels from satellite-based images is recognized as a considerable obstacle in the specialized areas of computer vision and the examination of remote sensing imagery. The majority of existing methods are based on the YOLO architecture, which relies on manually designed anchor points and nonmaximum suppression (NMS) postprocessing. The detection of small targets in a single scene, the phenomenon of "catastrophic forgetting" due to the streaming of data, and the issue of an "information bottleneck" present significant challenges in this field. In order to address these issues, we propose the following solutions. A hybrid explicit spatial prior MH-Net network based on Manhattan distance is designed. By decomposing the self-attention matrix and the spatial attenuation matrix, the spatial correlations of different directions and positions are captured, thus effectively alleviating the problem of catastrophic forgetting. We propose an online convolutional reparameterization efficient layer aggregation networks cross-stage fusion network. Through equivalent transformations, the complex network architecture is compressed into a single linear layer. The network groups and processes input features in parallel, integrating both low-dimensional and high-dimensional features to alleviate the information bottleneck problem. The prediction head of the model uses the DINO decoder and applies contrastive denoising to remove useless prediction boxes. This allows the proposed MHOE-DETR model to avoid thresholding and NMS), reducing the model's computational complexity. The experimental results demonstrate that the MHOE-DETR algorithm, designed for this purpose, markedly enhances the detection performance of small and indistinct targets in private remote sensing datasets. The average accuracy, recall, and AP50 reached 96.3%, 91.4%, and 95.4%, respectively, while maintaining a low GFLOPS value (54.4 G) and parametric count (77.3 M). These findings offer substantial technical justification for the implementation of sea area management and maritime safety monitoring strategies.**

***Index Terms*—Fuzzy targets, information bottlenecks, model catastrophic forget, remote sensing data, small targets.**

Received 3 March 2025; revised 25 April 2025 and 8 July 2025; accepted 14 July 2025. Date of publication 25 July 2025; date of current version 22 August 2025. This work was supported in part by the Key Research and Development Project of Hainan Province under Grant ZDYF2022GXJS348 and Grant ZDYF2024GXJS021, in part by the National Natural Science Foundation of China under Grant 62361024 and Grant 62161010, and in part by the Hainan Seed Industry Laboratory under Grant B23H10004. *(Corresponding author: Yaochi Zhao.)*

Zhuhua Hu, Xiyu Fan, Wei Wu, and Jie Liu are with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China (e-mail: eagler_hu@hainanu.edu.cn; 22210810000042@hainanu.edu.cn; wuweido@126.com; 20233006003@hainanu.edu.cn).

Yaochi Zhao is with the School of Cyberspace Security, Hainan University, Haikou 570228, China (e-mail: zhyc@hainanu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2025.3592850

I. Introduction

MARINE remote sensing is a critical field within the broader domain of remote sensing target detection and identification technology, with ships being the primary mode of marine transportation [1]. Current methods for detecting marine vessels generally derive from synthetic aperture radar, infrared imaging technology, visible light remote sensing, and so on. Among these, visible light remote sensing technology, supported by optical satellites, has become a significant technology for marine remote sensing vessel detection due to its lower data transmission costs, a certain resistance to harsh weather interference, and the clarity of the remote sensing images it captures. However, due to the serious imbalance between the remote sensing image foreground and the ship target background and the lack of ship information in bad weather, it is very difficult to detect small target ships and fuzzy target ships, so it is of great significance to realize small target and fuzzy target detection in remote sensing images.

With the deep-learning-based computer vision field receiving considerable attention internationally, it has also achieved tremendous success in the field of marine remote sensing target detection [2]. In contrast to conventional algorithms, deep learning methodologies are characterized by their utilization of multilayered architectures, exemplified by neural networks, for the purpose of feature extraction across diverse datasets. In the early development of deep learning technology, the application of deep-learning-based methods in ship detection was significantly constrained by the scarcity of high-quality annotated remote sensing datasets. With the efforts of many researchers in the field of remote sensing, remote sensing datasets have been introduced, such as the common SSDD [3] AIR-SAR Ship, HRSID, LS-SSDDv1.0, and so on. More and more researchers are adopting deep learning methods in this field to realize the tasks.

Currently, the mainstream deep-learning-based methods for remote sensing vessel detection primarily involve predicting a set of detection boxes and category labels for vessels [4], and common object detectors solve the remote sensing ship prediction task with indirect methods by defining regression and classification on a large number of a priori proposals [5], [6], anchor frames, and sliding windows. However, these methods face the following issues when applied to the task of vessel detection in remote sensing imagery.

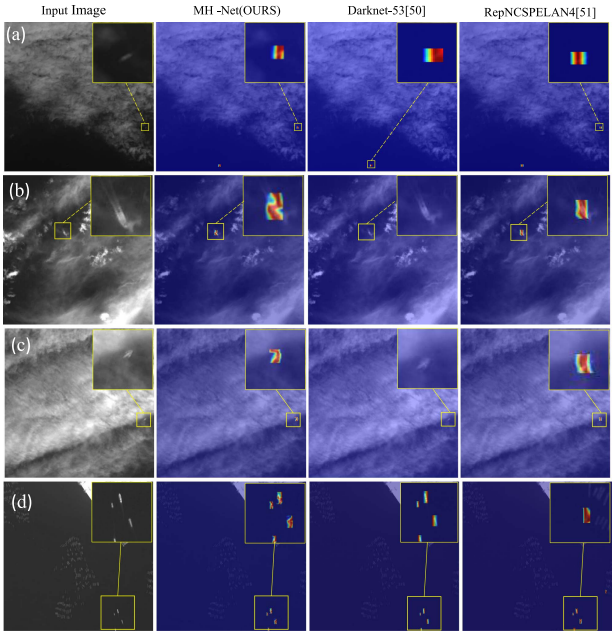


Fig. 1. Visualization of MH-Net network under cloud occlusion and underwater aquaculture cage.

1) *Issue of model computational complexity:* Mainstream detectors are generally based on preselected anchors and postprocessing problems. Therefore, it is necessary to reduce postprocessing operations in the model to lower computational complexity while maintaining detection accuracy.
2) *Issue of model catastrophic forgetting:* During neural network training, data are gradually fed into the model as a stream for learning. However, in remote sensing images, ships often appear in complex environments with large variations within the same class. This causes significant differences between batches of data, which can lead to catastrophic forgetting, where the model forgets what it has learned before when learning new data.
3) *Information Bottleneck Problem:* Compared with conventional ship datasets, remote sensing images contain a larger number of blurry target ships with fewer pixels. During the layer-by-layer feature extraction and spatial transformation process, this often leads to the emergence of information bottlenecks.

To solve the aforementioned problem, inspired by DETR. In this article, we propose an efficient end-to-end detector-based framework, which we call MHOE-DETR. In the backbone network part, we design an MH-NET network. A spatial prior of self-attention is introduced through the spatial decay of the Manhattan distance, thus extending the temporal decay mechanism of the model to the spatial domain. We decompose the self-attention and spatial decay matrices along the horizontal and vertical directions of the remote sensing imagery to capture the spatial correlation of different directions and locations, and thus, open up the catastrophic forgetting problem of the link model. Specifically, our feature maps for different salient orientations according to the MH-NET module are shown in Fig. 1. To address the confusion between small and blurry ship targets in remote sensing images, we propose a cross-stage fusion network named online convolutional reparameterization efficient layer aggregation networks (OREPAELAN). This module incorporates two online convolutional reparameterization new CSPNet (OREPANCSP) blocks and introduces linear feature scaling in the channel dimension as a substitute for conventional convolutional normalization operations. This linear scaling enables multipath reparameterization in different directions, thereby enhancing the interaction between high-dimensional semantic features and spatial localization features. To alleviate the computational burden of the model, we further adopt a system-level optimization strategy based on the DINO framework. Specifically, a contrastive denoising training approach is employed to discard redundant prediction boxes, and deformable attention with linear complexity is integrated to reduce the overall computational cost. Under the condition of maintaining a high detection accuracy of 96.3% on the private dataset, the proposed model achieves a computational complexity of only 54.5 GFLOPs, demonstrating excellent tradeoffs between performance and efficiency.

In summary, the main contributions of this article are as follows.

1) This study designs a novel network for the detection of small and fuzzy target vessels in remote sensing imagery. This network is an end-to-end system that does not require postprocessing such as nonmaximum suppression (NMS) and can still perform vessel detection tasks in environments with significant intraclass variations, featuring a lightweight structure. Extensive experiments on remote sensing imagery datasets demonstrate that the designed MHOE-DETR outperforms state-of-the-art methods.
2) In response to the phenomenon of catastrophic forgetting in model training, this investigation has crafted a novel backbone network, referred to as MH-Net, which efficiently harnesses the explicit spatial domain. In particular, it introduces a spatial prior of self-attention based on the spatial decay of the Manhattan distance. As the model transmits data, the spatial decay of the distance (network layer) provides the model with a substantial amount of spatial prior knowledge. Furthermore, the issue of model forgetfulness is mitigated by decomposing the self-attention and spatial decay matrices to capture the spatial correlation of disparate directions and locations.
3) The objective of this research is to tackle the challenge of discerning minute and indistinct targets within remote sensing datasets by leveraging the concept of information bottleneck during sequential feature extraction and spatial transformation. In pursuit of this goal, we have engineered an OREPAELAN symplectic fusion network, which condenses the intricate multilayered architecture into an integrated linear layer via a series of equivalent transformations.

II. Related Work

Historically, traditional detection algorithms in remote sensing imagery have relied heavily on extensive prior knowledge,

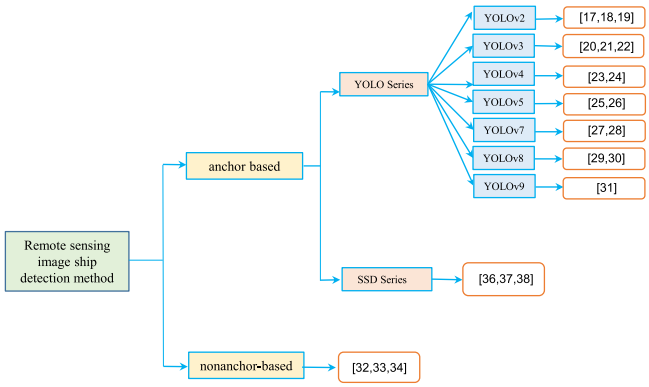


Fig. 2. Ship detection methods in remote sensing images.

including manually selected features and limited shallow learning representations [7]. The primary steps encompass preprocessing, candidate region extraction, and detector discrimination. Among these, the constant false alarm rate is a commonly employed method for extracting candidate regions [8], fundamentally an algorithm based on segmentation concepts. It categorizes pixels into ship and background classes based on the grayscale values of the remote sensing imagery, amalgamating ship pixels into ship regions. This approach is largely contingent upon the statistical modeling of ocean clutter and the likelihood estimation of the model, which is ill-equipped to handle complex scenarios.

Nowadays, to achieve the detection task of remote sensing vessels, most methods are based on deep learning. Common methods are mainly anchor-based [9] and nonanchor-based [10] approaches. Anchor-based methods typically involve a two-stage process: generating proposals, and then, classifying and localizing the targets. At the proposal generation phase, the system identifies a set of candidate regions potentially housing target objects. Following this, the classifier evaluates these regions to confirm the presence of targets and to pinpoint their exact positions. Therefore, proposal-based methods can also be categorized into single-stage and two-stage detectors, and the common two-stage detectors are Faster R-CNN [7], R-FCN [11], Mask-RCNN [13], and so on. Among them, the Faster RCNN is the basic work, and most of the subsequent remote sensing ship detection methods are improved. The main parts of the improvements include the Backbone network, anchor box prediction methods, Region Proposal Network, loss functions, and NMS. Our approach to the detection of ships from remotely sensed data is summarized in Fig. 2.

The task of two-stage detectors to first generate candidate frames for remote sensing images, and subsequently, to recognize and regress the candidate frames takes too much time and has too much computational complexity [55]. Therefore, single-stage detectors are subsequently used for target detection in remote sensing. The classical single-stage detectors are SSD [14], YOLO [15], and RetinaNet [16]. In particular, the YOLO series of object detection algorithms has been widely favored by many researchers. The authors in [14] and [19] used YOLOv2 to detect ships in remote sensing images. Deng et al. [18] improved YOLOv2 by constructing a reduced module, which enhanced the detection accuracy of the model for remote sensing vessels. Chaudhary et al. [20] and Jiang et al. [24] used YOLOv3 to directly detect ships in remote sensing images. In a study by Zhang et al. [21], the DarkNet-19 network was employed as the foundation for enhancements to the YOLOv3 system, which accelerated the convergence speed of the model and made it more lightweight. Hong et al. [22] improved the K-means algorithm, changing the way YOLOv3 generates anchor boxes, and improved the detection performance of YOLOv3 for multiscale ships. Jiang et al. [23] improved YOLOv4 and proposed YOLOv4-light, which reduced the number of model parameters, accelerated the inference speed, and also reduced memory consumption. Sun et al. [24] proposed a three-stage network based on YOLOv4, which considered global features when extracting ship features. In [25], inspired by YOLOv5, the model's architecture has been optimized by integrating a convolutional attention mechanism into the backbone and by employing a compact GSConv structure within the Neck, thereby significantly boosting its detection capabilities for maritime targets in remote sensing imagery. Zhang et al. [26] introduced an asymmetric pyramid-shaped nonlocal block and a sim attention mechanism into YOLOv5 to reduce the interference of coastal background for remote sensing vessel detection, and improved the feature representation capability of the model through cross-stage learning by improving the C3 module. Chen et al. [27] introduced an enhanced module termed SAS-FPN, which integrates spatial pyramid pooling and attention randomization techniques to refine the performance of YOLOv7. This approach enables the model to prioritize salient features and filter out nonessential details, thereby mitigating the feature loss associated with detecting small targets in remote sensing imagery. Liu et al. [28] proposed a top-down bounding box remote sensing vessel detection model (YOLOv7oSAR) for YOLOv7, optimizing the network's architectural depth and width to render the model more streamlined and efficient. Nie et al. [29] addressed the issue of low resolution and complex backgrounds of small target vessels in remote sensing imagery by proposing the HPANet network to improve YOLOv8n, enhancing the model's capability to swiftly and precisely detect minute target objects against intricate background settings. Zhao et al. [30] improved YOLOv8 to address the low accuracy and slow speed of traditional methods for detecting small target vessels in remote sensing images, adopting a dual-path attention module with a ViT architecture to improve the precision (P) of small target vessel detection. Yang et al. [31] proposed a metalearning balanced small shot target detector (B-FSDet) based on the GELAN version of YOLOv9, solving the problem of incomplete annotation objects in remote sensing data that could disrupt the balance of the few-shot principle, and also proposed a steady-state feature extraction module and corresponding steady-state rapid prediction method, forming a steady-state metalearning pattern to provide a method for detecting small target vessels in remote sensing images. Anchor-based methods require the setting of many hyperparameters, such as scale and aspect ratio, and this approach can produce many redundant boxes, increasing computation and memory consumption. Nonanchor-based methods, on the other hand,

only require regression tasks for the target center points and width and height on feature maps of different scales, reducing computational power and time consumption. Therefore, many scholars use nonanchor-based methods to complete the detection task of vessels in remote sensing images. Wang et al. [32] addressed the issue of small vessel scale and susceptibility to background interference in remote sensing imagery, proposing a feature enhancement and fusion target detection algorithm based on the CenterNet network. Sun et al. [33] proposed a point-based dual-domain alignment-guided RepPoints remote sensing target detection method to alleviate potential misalignment issues between the spatial domain and feature domain during the training process of remote sensing vessel data. Bai et al. [34] proposed an FBUA-NET network, a new type of anchor-free detector, and a globally contextual feature pyramid was introduced for balanced feature extraction. Focusing on the geometric characteristics of vessels in the spatial and scale domains reduces the interference of surrounding noise points and improves the accuracy of vessel detection on the SSDD and HRSID [35] remote sensing datasets.

III. Designed Network Framework

A. Overall Network Framework of MHOE-DETR

Compared to the general target size of remote sensing ship target detection tasks, most of the remote sensing ship small and fuzzy targets we study have a size of only 8–24 pixels or even smaller [38]. Nowadays, the mainstream deep learning methods mainly focus on finding the optimal objective function to make the predicted values closer to the real values. They also design a good architecture to extract enough information from the dataset to perform the prediction task [39]. However, data passing through "A deep learning network with an excessive number of layers tends to undergo multiple feature extraction and spatial transformations. Therefore, a lot of information is lost in the high-level feature maps [40]. The environment in which the ship is located in the remote sensing data is complex, and the data are input to the neural network in the form of data streams, and the problem of catastrophic model forgetting will also occur. Therefore, we designed an MH-Net network as a backbone network to extract shallow features to alleviate the problem of catastrophic model forgetting." In the initial phase, the final trio of stages within the MH-Net backbone, namely P3, P4, and P5, serve as the initial inputs for the encoder. This encoder then converts the multiscale features into a coherent stream of image features, facilitated by the intrascale feature interaction module, AIFI, in conjunction with the feature extraction network, OREAPN. AIFI functions to enhance feature interaction within the same scale. Meanwhile, an OREPAELAN module is designed based on the enlightenment of online convolutional parameterization. This module is a cross-scale feature fusion network, which mainly transforms the multiscale feature output from AIFI into image feature sequences. IoU-aware query selection is used to select a fixed number of small and fuzzy target features as initial query objects for the decoder. Finally, DION-head, a contrast denoising auxiliary head, is used as a decoder to reject useless anchor frames while iterating the optimized query objects to generate prediction frames and confidence scores. The network is an end-to-end network that does not require the postprocessing of iterative prediction. We call it the MHOE-DETR network. The overall network architecture is shown in Fig. 3.

B. Hybrid Explicit Spatial Prior MH-Net Network

Target detection of ships from remote sensing images is regarded as one of the most important technological directions in the field of remote sensing intelligence, which is of great importance for naval planning, military reconnaissance, and other fields [41]. However, the target detection model based on deep neural networks must be trained with all the training data at once. As data are incrementally introduced to the model in a sequential data stream format, the model can lead to the problem of catastrophic forgetting. This means that the predictive performance of the model for previous data decreases significantly due to parameter updates. For complex scenarios such as small target ships, ships in bad weather, ships docking, ports, etc., the inconsistency in the distribution of labels in the dataset leads to the catastrophic forgetting problem.

To solve the aforementioned problems, inspired by the Manhattan self-attention mechanism, this study proposes a backbone network MH-Net with fully exploited explicit space, in particular, we introduce a spatial prior of self-attention based on the spatial decay of the Manhattan distance [43]. Extending the temporal decay mechanism of the model to the spatial domain alleviates the problem of catastrophic forgetting of traditional models. In the spatial decay matrix we have designed, the attenuation of a target ship's attention score increases with the surrounding distance from the target ship. The target token is allowed to perceive a range of global information while allocating different degrees of attention [42] depending on the distance from the target ship. Using this spatial decay matrix, an explicit spatial prior is always introduced into the visual backbone network. Specifically, we transform the observed notion of unidirectional, 1-D temporal decay by upscaling and applying it to bidirectional, 2-D spatial decay. This transformation allows us to introduce an explicit spatial prior based on the Manhattan Distance into the backbone of visual processing, which in turn bolsters the model's capacity to discern spatial correlations within imagery. Furthermore, we develop an efficient and straightforward method that simultaneously decomposes the self-attention and spatial decay matrices along the two main axes of the image (horizontal and vertical directions). This decomposition strategy not only improves computational efficiency but also enables the model to more accurately capture spatial correlations in different directions and locations in the image.

The Manhattan self-attention mechanism uses a retention mechanism for sequence modeling that adds a temporal decay parameter to the neural network model, a factor not present in conventional models. Its data retention feature considers sequence modeling in a recursive manner and can be written as

$$\text{Export}_n = \sum_{m=1}^{n} \gamma^{n-m} \left(Q_n e^{in\theta}\right) \left(K_m e^{im\theta}\right)^{\dagger} v_m. \quad (1)$$

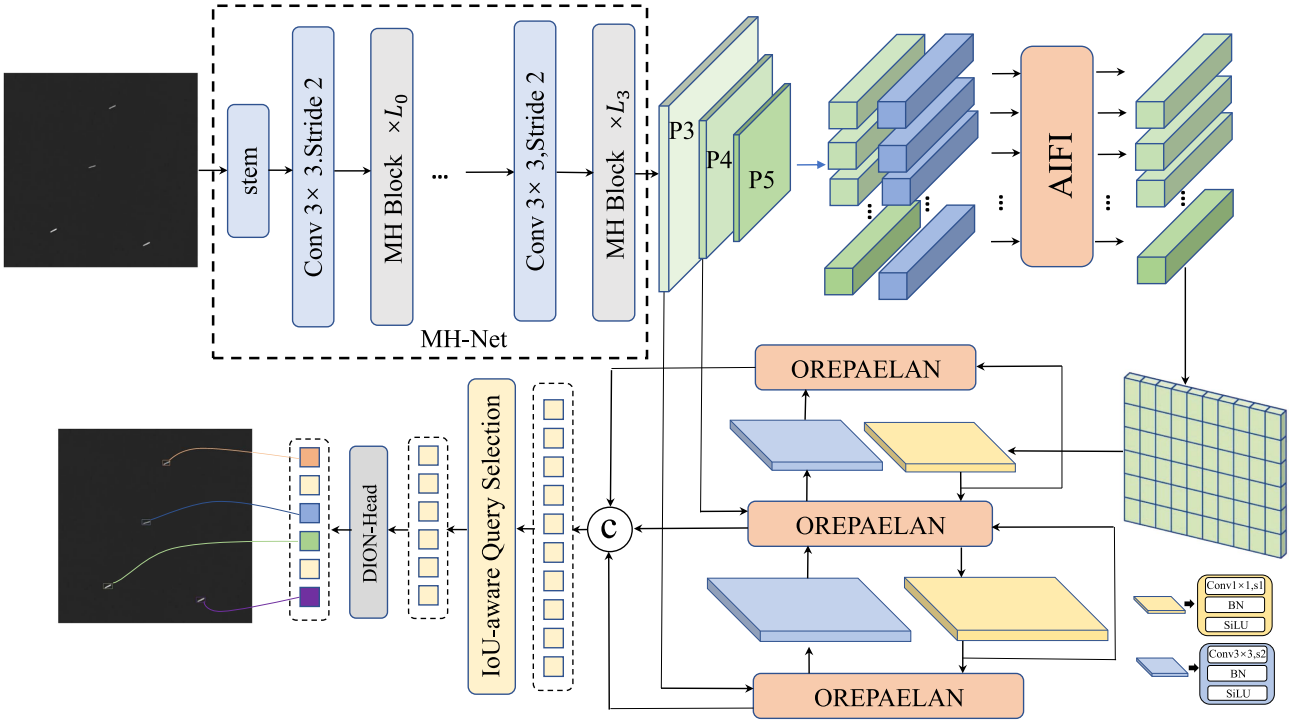


Fig. 3. MHOE-DETR overall network architecture for detecting remote sensing small target and fuzzy target ships is proposed.

And the aforementioned equation during training can be expressed as

$$Q = XW_Q \odot \theta, K = XW_{\mathrm{k}} \odot \bar{\theta}, V = XW_V,$$

$$\theta_n = e^{in\theta}, D_{nm} = \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases}$$

$$\mathrm{Reserve}X = (QK^T \odot D)V \tag{2}$$

where $\bar{\theta}$ is the conjugate of θ, and the decay matrix of D $\in \mathbb{R}^{|x|\times|x|}$ contains exponential decay with causal masking. Represents relative distances in a single 1-D sequence, providing an explicit temporal prior for the data stream. In addition, K_m represents the *m*th position of the Key vector in the sequence and Q_n represents the *n*th position of the Query vector in the sequence.

In traditional transformer-based deep learning networks, the model retains the data after one-way decay due to the causal nature of the data stream. In the traditional model, in terms of the dimension of time, each token only focuses on the tokens after the features are extracted earlier, and no connection is established for the tokens after the features are extracted from the later data streams. Specifically, the way it retains the features is considered in a recursive form of modeling, which can be represented as follows for bidirectional modeling:

$$\mathrm{BiReserve}X = (QK^T \odot D_{\mathrm{Bi}})V, D^{\mathrm{Bi}}_{nm} = \gamma^{|n-m|}. \tag{3}$$

As for the 2-D remote sensing images, although bidirectional modeling is supported, the modeling capability is still maintained at the 1-D level, so it is necessary to extend the 1-D ReserverX to both vertical and horizontal dimensions. Assuming that each token is uniquely positioned within the 2-D coordinates, the *n*th token can be written as (x_n, y_n). This represents the Manhattan distance according to the 2-D coordinates of each token. So, the matrix D can be written as

$$D^{\mathrm{Bi}}_{nm} = \gamma^{|x_n - x_m| + |y_n - y_m|}. \tag{4}$$

And Manhattan self-attention can be expressed as

$$\mathrm{ManHa}(X) = (\mathrm{Softmax}(QK^T) \odot D^{2d})V$$

$$D^{2d}_{nm} = \gamma^{|x_n - x_m| + |y_n - y_m|}. \tag{5}$$

High token counts escalate the computational cost for self-attention mechanisms in global data stream modeling. Therefore, for the Manhattan attention mechanism, we compute the horizontal and vertical attention scores in the input remote-sensing images. Subsequently, a 1-D decay matrix is applied to these attentional weights, and following (4) we obtain the horizontal and vertical Manhattan distances between each tokens as $D^W_{nm} = \gamma^{|x_n - x_m|}, D^H_{nm} = \gamma^{|y_n - y_m|}$. Thus, the decomposed Manhattan attention mechanism can be expressed as

$$\mathrm{Attention}_W = \mathrm{Softmax}(Q_W K_W^T) \odot D^W$$

$$\mathrm{Attention}_H = \mathrm{Softmax}(Q_H K_H^T) \odot D^H$$

$$\mathrm{ManHa}(X) = \mathrm{Attention}_H(\mathrm{Attention}_W V)^T. \tag{6}$$

For the decomposition of the Manhattan attention mechanism, the visualization of the feeling field of each token is shown in Fig. 4. After the derivation of the aforementioned formulas and the visualization of Fig. 4, it is shown that the decomposition method completely retains the explicit spatial a priori. The MH-Net network architecture we designed is a model based

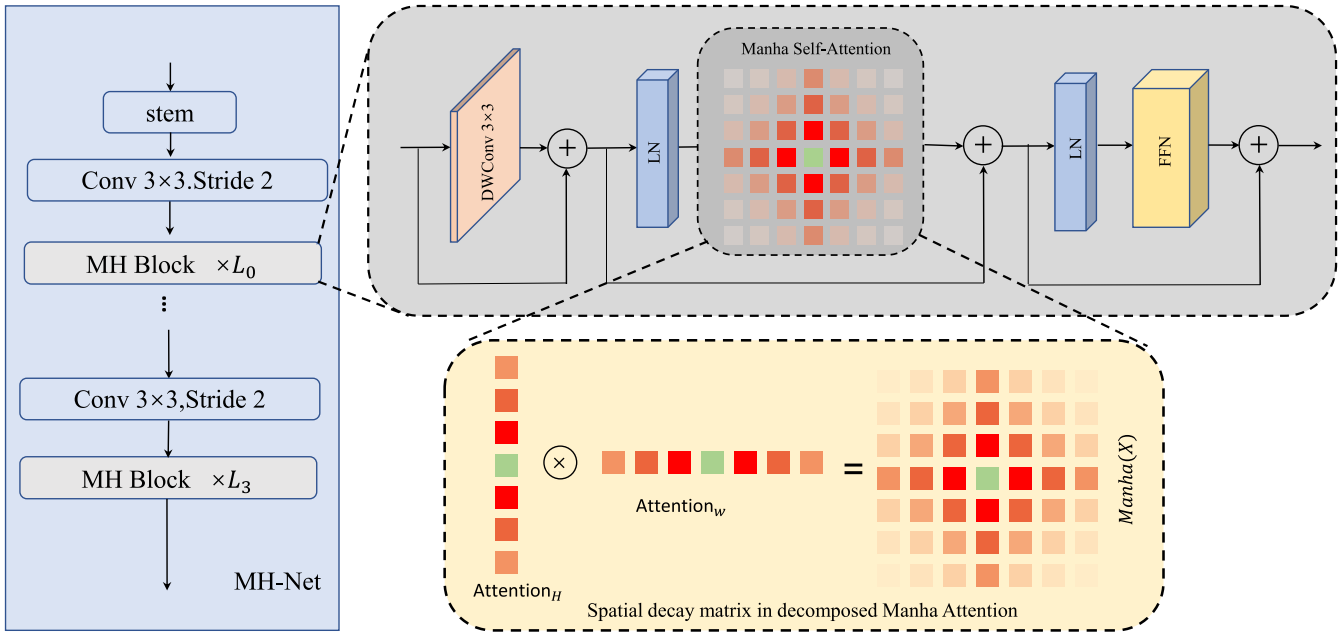


Fig. 4. Network structure of MH-Net.

on the Manhattan attention mechanism, which incorporates the decomposition of the Manhattan attention mechanism at each stage of feature extraction, retaining the explicit spatial a priori so that when the data enter the neural network as a data stream, the contextual information before and after it can still be retained, mitigating the catastrophic forgetting problem of the model. The overall network architecture of the MH-Net is shown in Fig. 4. We also introduced the contextual question enhancement module of DWConv to extract features further to improve the local representation of the Manhattan attention mechanism.

C. OREPAELAN Cross-Stage Feature Fusion Network

In this section, we present the structure of the OREPAELAN feature fusion network. The MH network designed in Section III-A introduces the spatial decay of the Manhattan distance and introduces the spatial prior of self-attention to improve the generalization of the model to complex scenes and mitigate the problem of catastrophic model forgetting. However, it contains less information for small and fuzzy targets in remote sensing image data. After spatial transformation and feature extraction layer by layer, a lot of information will be lost, i.e., the information bottleneck problem. Aiming at the aforementioned problem, inspired by online convolutional reparameterization (OREPA) [44], we designed an OREPAELAN network architecture, which is an extended high aggregation network based on OREPA. Its network structure is shown in Fig. 5.

A thorough discussion of approaches to efficient architecture design typically focuses on the core elements of parameter count, computational volume, and computational density. However, from the unique perspective of access overhead, Ding et al. [45] have further demonstrated the significant impact of input/output channel ratio, number of architectural branches, and element operations on network inference speed. Meanwhile, Ross and Dollár [46] additionally considered the dimension of activation in their study of model scaling strategies, i.e., they paid more attention to the number of elements in the output tensor of the convolutional layer in order to evaluate and optimize the model performance more comprehensively. When designing an efficient network, it is often the case that controlling the shortest gradient path allows deeper layers of the network to learn and converge efficiently. Inspired by the OREPA network in this study, an OREPANELAN network is designed. The network is designed to efficiently utilize information from small, indistinct targets in remote sensing imagery, maintaining model lightness to minimize information loss.

The number of stacked ELAN [48] network modules with gradient path lengths is already relatively stable in large networks, no matter how they are designed. Endless stacking of ELAN networks can be counterproductive since the parameter utilization of the whole network decreases and the learned data features decrease with the number of layers of the whole network. The OREPANELAN network we have designed uses the method of merging bases and expansion to achieve a method that does not destroy the gradient paths but still improves the learning ability of the network for small and fuzzy targets. The network mainly consists of OREPANCSP and OREP-Net. OREPANCSP mainly combines features across low and high dimensions through parallel learning of clustered inputs to preserve feature information. The OREP-Net network is mainly used to reduce the number of parameters in the model by transforming complex convolutional modules into simple single convolutional layers to reduce the computational complexity of the model. The structure is based on the effectiveness of the linear deflation layer as an alternative to the normalization structure of the general convolution, making the model more parametric. Instead of removing the normalization layer, a linear deflation of the feature channels is introduced to replace the normalization operation of the general

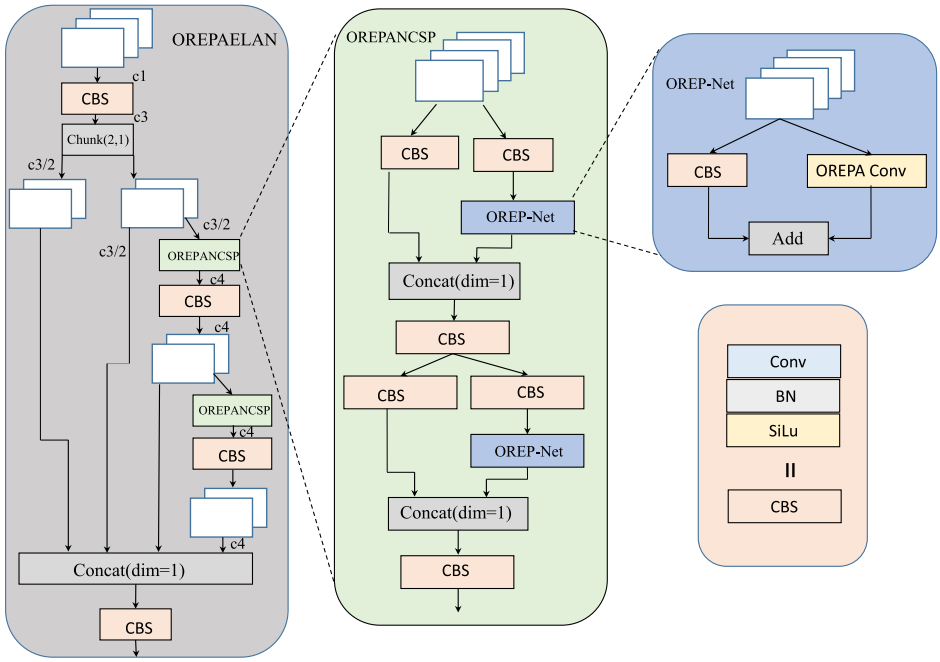


Fig. 5. Network structure of OREPAELAN.

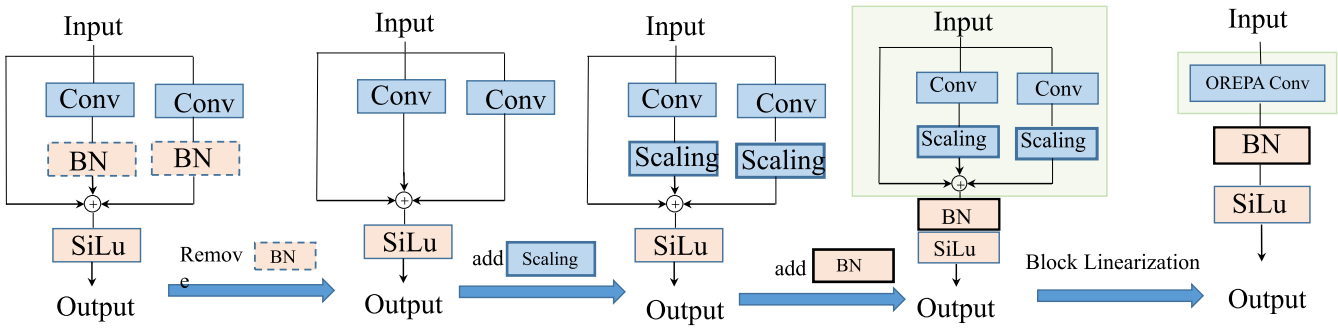


Fig. 6. Network structure of OREP-Net.

convolution. The linear deflation layer allows multiple paths to be optimized for reparameterization in different directions. After the linear deflation of the convolution, only the linear layer exists in the convolutional network, then all the convolution-based feature extraction networks can be merged and compressed into a single convolutional kernel for feature extraction. This approach diminishes the computational burden of the entire network while enhancing the efficiency of parameter usage. The linearization process for the OREPA convolutional structure in the OREP-Net network is shown in Fig. 5.

The process of convolution linearization is divided into three main steps. In the first step, we remove all nonlinear structures (BN) in the original network layer, and in the second step, we introduce a linear deflation layer instead of the BN structure, which better optimizes the network and reduces the computational complexity. In the third step, a linear compression operation is performed at the end to compress the linear deflation layer into the OREPA Cov. We ensure the efficiency of the model in extracting features while reducing the computational complexity.

The OREPANELAN network contains several OREP nets as described previously, and the realization process can be represented as as follows, let the input feature map be I, first go through the CV1 network layer, and perform the initial feature extraction on the feature map (Fig. 6):

$$Y_1 = \mathrm{BN}(\sigma\{\mathrm{Conv}_{c1\to c3}(I, k=1, s=1)\}) \tag{7}$$

where Y_1 denotes the feature map output from CV1, as the output of the first set of channels, which is directly connected to the final feature fusion layer. Specifically, the input I is first convolved by a 1×1 point-by-point convolution. This method maintains the spatial dimensions of the input feature map, decreases the model's parameter count, and preserves model expressiveness. It also facilitates feature integration across various channels, promoting information flow and enhancing model accuracy, while dynamically adjusting output channels to modulate feature map dimensionality. The Y_1' will pass through the OREPANCSP network and the final feature fusion layer, respectively, the OREPANCSP network that mainly contains the OREPANCSP

and the OREP-Net network as shown in Fig. 5. The implementation process can be expressed as follows:

$$Y_2 = \text{BN}(\sigma\{\text{Conv}_{c3/2\to c4}(\text{OREPANCSP}(Y_1'), k=3, s=1)\}) \tag{8}$$

$$Y_3 = \text{BN}(\sigma\{\text{Conv}_{c4\to c4}(\text{OREPANCSP}(Y_2'), k=3, s=1)\}) \tag{9}$$

where Y_2 is denoted as the feature map output of CV2 and the output of the first set of channels, which is also directly connected to the final feature fusion layer. The CV2 network layer first takes Y_1' from the second channel output of CV1 and passes it through the OREPANCSP network. The goal is to improve the representation of the model through efficient convolution and feature fusion while keeping the computational complexity and number of parameters low. And Y_3 is to perform further feature extraction and fusion operations on CV2. Finally, high-dimensional and high-state feature fusion is performed, and downscaled splicing is performed, which can be expressed as

$$\text{OREPANELAN} = \text{concat}_{c3+2*c4}(Y_1', Y_1, Y_2, Y_3)_{\text{dim}=1}. \tag{10}$$

The network fuses and splices shallow and deep features to form a richer feature representation, which is conducive to retaining small and fuzzy target ship information in remote sensing imagery, and alleviates the concern that the model loses data information as the number of network layers increases.

D. Loss Function and Label Matching

MHOE-DETR, whose detection head construction is heavily inspired by the DINO framework [47], uses contrast denoising to discard useless prediction frames without the steps of thresholding and NMS. Specifically, MHOE-DETR adopts the decoder structure of DINO and integrates a deformable attention mechanism with linear complexity. This choice not only ensures the efficiency of the model in processing complex remote sensing scenes but also significantly improves the detection accuracy. In the training phase, MHOE-DETR cleverly exploits the denoising idea of DINO to improve the quality of bilateral matching samples, thus accelerating the convergence speed of the training process. This strategy effectively solves the sample noise problem that traditional target detection algorithms may encounter during the training process and significantly improves the robustness and generalization ability of the model. The Hungarian matching algorithm is still employed for the assignment of labels during the training phase. Its implementation is based on DINO, which is a denoising function that considers the learning of negative samples.

In the task of detecting small and fuzzy targets in remote sensing imagery, there is a significant class imbalance issue stemming from the limited pixel information available for these targets. Therefore, the simple use of the binary cross-entropy loss function provides only a small loss for the backpropagation of the network, which makes it more difficult to update the gradient parameters of the neural network. In this study, we introduce GIoU loss and binary cross-entropy loss, and at the same time, we introduce "IoU soft labeling" on the category loss to solve the problem of category imbalance, and this method also becomes IoU-aware query selection. In the training process, the prediction frames obtained after more model prediction tasks, the IoU between them, and the real frames are used as the category prediction labels. By constraining the detector during training, a high classification score is generated for features with high IoU and a low classification score for features with low IoU. Specifically, the method introduces the IoU score into the objective function of the classification branch to achieve consistency constraints on the classification and localization of positive samples. The optimization objective can be expressed as

$$\begin{aligned}\mathcal{L}(\hat{y}, y) &= \mathcal{L}_{\text{box}}(\hat{b}, b) + \mathcal{L}_{\text{cls}}(\hat{c}, \hat{b}, y, b) \\ &= \mathcal{L}_{\text{box}}(\hat{b}, b) + \mathcal{L}_{\text{cls}}(\hat{c}, c, \text{IoU})\end{aligned} \tag{11}$$

where $\hat{y}$ represents the predicted value, y represents the true value, $\hat{y}{=}(\hat{c}, \hat{b}), y{=}(c, b)$, c denotes the category prediction, and d denotes the bounding box prediction. According to the traditional one-hot method, there may be a situation where the anchor box localization is not yet accurate enough, while the category loss has basically converged. Given that the class labels are binary, either 0 or 1. Therefore, if IoU acts as the category label, then category learning is modulated by regression, and the category prediction task only learns well when the regression task does well enough. Therefore, the training regimen tailored for remote sensing datasets is capable of significantly mitigating the issue of class imbalance.

IV. Experiments and Analyses

A. Datasets

The remote sensing data used in this study come from the Hainan 01 satellite in Hainan Province, which provides a lot of data support for ocean management, and the remote sensing image data transmitted to the ground include both offshore and offshore ships. However, the color depth of the satellite-transmitted remote sensing data is 16 bits, with color values ranging from 0 to 255. The data are single-channel, and the image size is 28000 × 28000 pixels. Only 10 bits of the available pixel value are utilized, making it impossible for the naked eye to discern the ship's specific location. Therefore, a linear transformation was applied to the data, converting it into a conventional 3-channel, 8-bit format. Subsequently, the data were cropped to a size of 1024× 1024 pixels for remote sensing images. Select small ships, inshore ports, ship clusters, ships remote sensing images of ship proximity, inclement weather, and other conditions. A total of 3827 remote sensing images are shown in Fig. 8. About 80% of these ships have a pixel size of only 8–24 pixel values. This makes the difficulty of small target and fuzzy target ship detection to the extreme.

B. Experimental Environment and Parameter Settings

We implemented MHOE-DETR on Pytorch 2.3.0 and leveraged a GeForce RTX 3090 GPU for both training and testing phases. The operating system used for training and testing was Ubuntu, and the specific details of the experimental environment are shown in Table I.

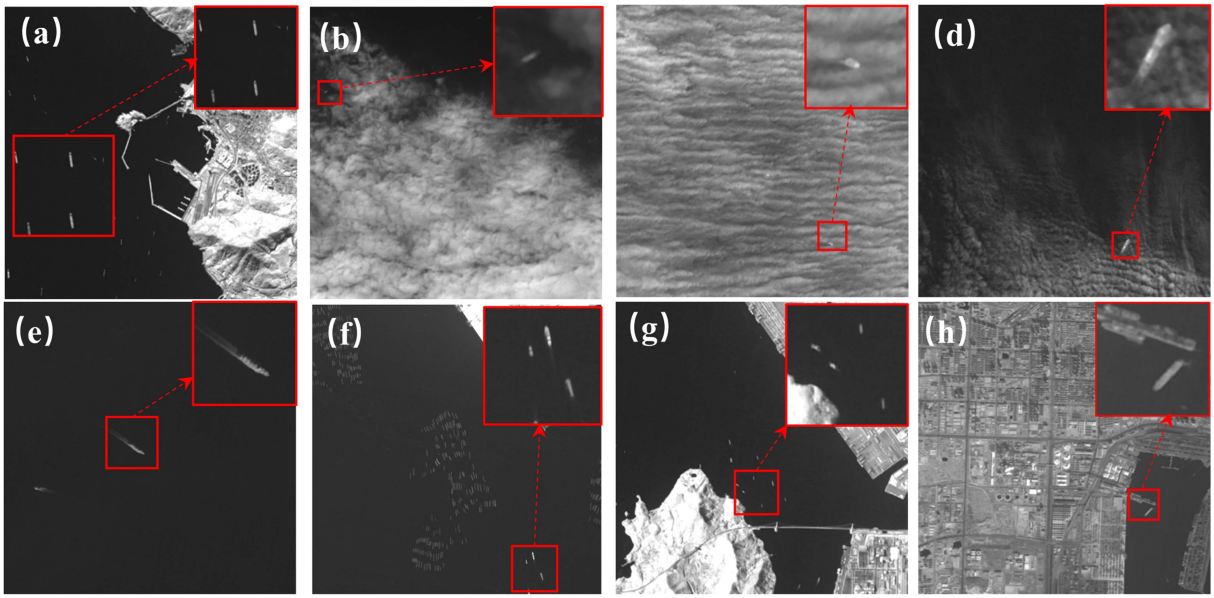


Fig. 7. Small target and fuzzy target ship in complex scene. (a) Harbor scene with ships. (b) Cloud-covered ship. (c) Ship in waves. (d) Fuzzy ship target. (e) Small target in dark water. (f) Ship formation. (g) Coastal ships. (h) Urban port ships.

TABLE I
EXPERIMENTAL SETTING

Item	Value
CPU	Intel(R) Xeon(R) Silver 4310 CPU at 2.10 GHz
GPU	GeForce RTX 3090
Cuda Version	12.0
Data Processing	Python3.9
Deep Learning Framework	Pytorch

In order to guarantee that the MHOE-DETR network can be adequately trained, the batch size of the neural network was set to 4. The experiment was conducted to evaluate the efficacy of MH-Net's architecture as a backbone network and to determine its potential to mitigate the issue of catastrophic forgetting within the model. We set the initial learning rate to 0.01 and configured the image patch size to 640 × 640. Employing stochastic gradient descent as the optimizer, we updated the gradients for the IoU-aware query selection method. Training was conducted for 300 epochs from scratch, without the use of a pretrained model.

C. Experimental Indicators

In order to demonstrate the advantages of MHOE-DETR, this study adopts P, recall (R), mean accuracy (mAP) AP_{50}, $AP_{50:95}$, number of model parameters, FPS, and GFLOPs as evaluation metrics to measure the algorithm (Fig. 7). It is assumed that the corresponding labels of remote sensing image types contain N+1 categories from T_0 to T_k. Y_{ij} denotes the number of targets labeled as class i but predicted to be class j. The number of targets labeled as class i but predicted to be class j is the number of targets labeled as class j. Y_{ii}, Y_{ji}, and Y_{ij} represent the counts of true positives, false positives, and false negatives, respectively. Precision and Recall are denoted as

$$\text{Precision} = \frac{Y_{ii}}{Y_{ii} + Y_{ij}} \tag{12}$$

$$\text{Recall} = \frac{Y_{ii}}{Y_{ii} + Y_{ji}}. \tag{13}$$

The average precision (AP) can be expressed as

$$\text{mAP}_{50} = \frac{\sum_{i=1}^{k} \text{AP}_i}{k}, \text{IoU} \in [0.5] \tag{14}$$

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^{n} \text{AP}_i. \tag{15}$$

D. Results and Analysis

Table II shows the comparison of our proposed MHOE-DETR with other state-of-the-art target detection algorithms. Specifically, we compare the proposed network with YOLOv7, YOLOv7X, YOLOv8-S, YOLOv9, YOLOv10, YOLOv12, and RT-DETR. The experimental data are shown in Table II.

Table II shows the comparison results of our proposed MHOE-DETR with other real-time target detection methods on our remote sensing dataset. For the detection of small and fuzzy targets, we can see that the current mainstream detectors do not do a good enough job. Overall, the network with the best accuracy (AP) of the existing methods for the detection task of small and fuzzy targets in this dataset is YOLOv8-X, which achieves an AP of 94.7% and an R of 90.3%. However, the computational complexity of its model reaches 259.9 G and the number of parameters reaches 258.1 M. Compared to our design MHOE-DETR, the computational complexity of MHOE-DETR is reduced by 79.0% and the number of parameters is reduced by 70.0%. In addition, the AP is improved by 1.7%, and the R value is improved by 1.2%. In terms of the lightweighting of the network, the better networks are YOLOv8-N, YOLOv10-N, and YOLOv12-N. The computational cost and parameters of YOLOv12-N are 6.3 G and 9.8 M, respectively, and these values are slightly lower than those of YOLOv8. This benefits from the

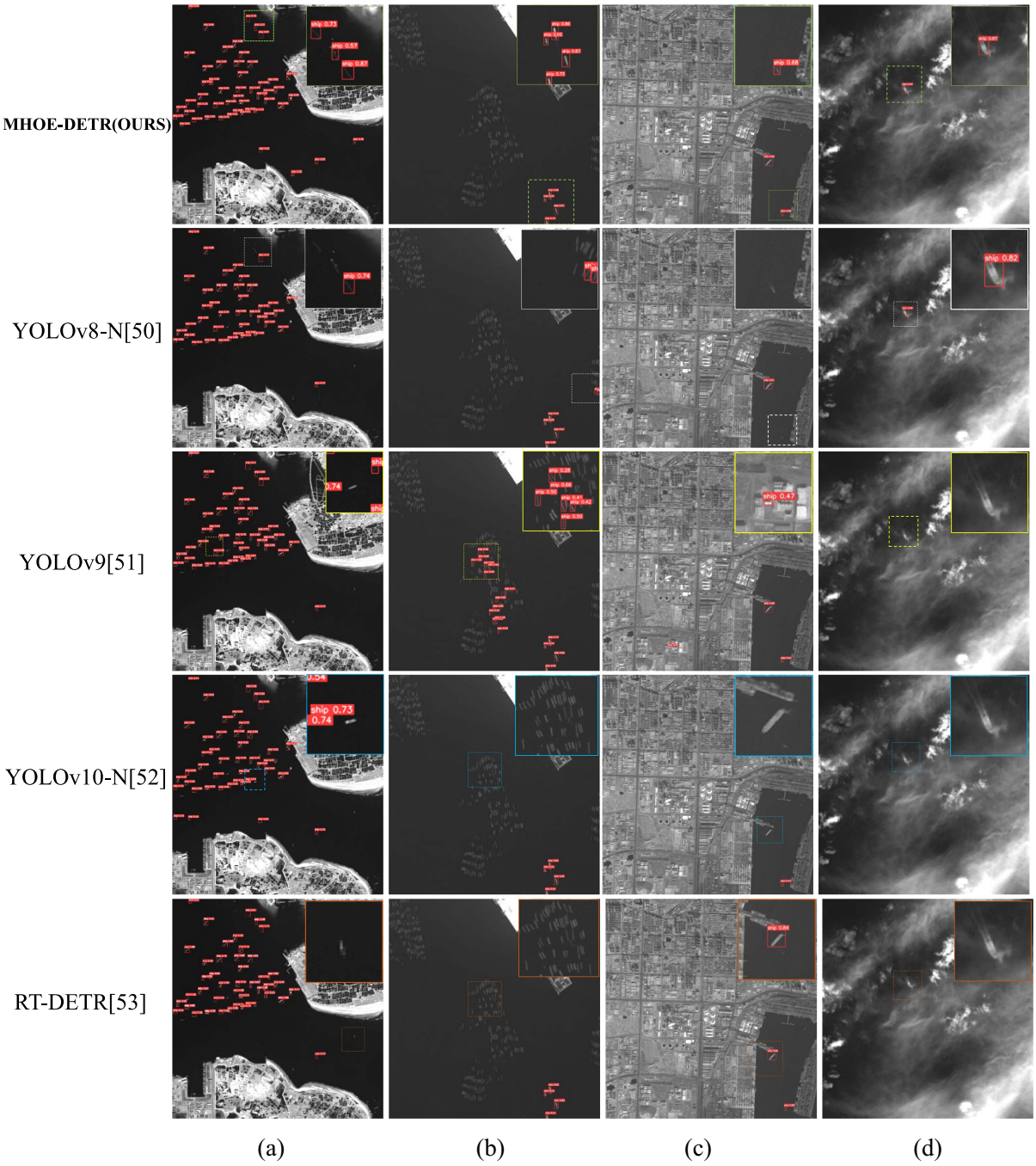


Fig. 8. Comparison of the effects of different detectors in complex scenes. (a) Dense ships. (b) Ship-net interaction. (c) Port interaction. (d) Cloud cover.

efficiency of the area attention module proposed by YOLOv12. Meanwhile, its AP and R values also remained at 86.9% and 86.2%, respectively. Compared with YOLOv8-N, YOLOv10-N, and YOLOv12-N, the AP values of MHOE-DETR increased by 8.9%, 8.0%, and 10.8%, respectively, while the R values also increased by 1.2%, 12.9%, and 6.0%, respectively.

For the large models YOLOv7-X, YOLOv9-E, and YOLOv10-X, these three large models stack many network layers and increase the number of model parameters. However, their detection tasks for small and fuzzy remote sensing targets in this dataset are not satisfactory. The number of model parameters of YOLOv7-X reaches 270.1 M, and the values of its detection P and R are unsatisfactory, with the P only reaching 84.8% and the R only reaching 78.2%. Compared to YOLOx7-X, MHOE-DETR reduces the number of parameters by 71.4%, reduces the amount of computation by 71.1%, and significantly improves the AP by 13.6% and the R value by 17.2%. YOLOv9-E has a model with 264.8 M parameters, but its accuracy (AP) is only 90.6% and its R is only 87.9%. Compared to YOLOv9-E, MHOE-DETR reduces the number of parameters by 70.8% and the amount of computation by 77.7%, and the AP is also improved by 6.3%. While YOLOv10-X has 120.8 M model

TABLE II
RESULTS OF DIFFERENT METHODS

Method	AP^{val} (%)	R^{val} (%)	AP^{val}_{50} (%)	$AP^{val}_{50:95}$(%)	FLOPS(G)	Param(M)
YOLOv7	83.0	78.0	85.3	40.4	141.9	141.8
YOLOv7-X	84.8	78.2	85.7	39.8	188.9	270.1
YOLOv7-AF	90.4	88.6	93.2	51.2	130.9	166.0
YOLOv8-N	88.4	90.3	93.3	52.0	8.2	11.4
YOLOv8-S	90.9	91.2	94.0	54.1	28.6	42.4
YOLOv8-L	91.1	91.5	95.1	62.0	165.4	166.4
YOLOv8-X	94.7	90.3	95.3	63.1	259.9	258.1
YOLOv9	88.1	89.5	93.1	51.0	266.1	231.9
YOLOv9-C	90.1	88.8	93.7	54.1	238.9	194.5
YOLOv9-E	90.6	87.9	93.6	53.1	244.8	264.8
YOLOv10-N	88.3	81.0	89.5	48.5	8.4	10.3
YOLOv10-M	90.1	88.3	93.2	53.1	64.0	62.9
YOLOv10-X	90.5	89.7	94.0	53.9	171.0	120.8
YOLOv12-N	86.9	86.2	90.4	48.0	6.3	9.8
YOLOv12-M	91.1	89.1	93.8	54.9	67.7	76.8
YOLOv12-X	91.7	90.0	94.7	56.6	199.8	255.5
RT-DETR-R18	92.4	87.0	90.9	49.2	57.3	76.1
MHOE-DETR(ours)	**96.3**	**91.4**	**95.4**	**54.4**	**54.5**	**77.3**

parameters, the detection accuracy (AP) of remote sensing small targets and fuzzy target ships only reaches 90.5%, and the R reaches 89.7%. Compared to YOLOv10-X, the AP of MHOE-DETR has been improved by 6.4%, the number of parameters has been significantly reduced by 36.0%, and the computational complexity has been reduced by 68.1%. The parameters of the YOLOv12-X model reached 255.5 M, but the AP for small target and fuzzy target ships reached 91.7%, and the R reached 90.0%. The detection effect was relatively good. However, compared with YOLOv12-X, the MHOE-DETR model has improvements of 5.0% and 1.6% in both AP and R values, while the number of parameters has also decreased by 69.7%. Although the area attention module proposed by YOLOv12 is simple and efficient, its feature extraction ability for small and fuzzy target ships in remote sensing images is not ideal.

For the generic models YOLOv8-L, YOLOv9-C, and RT-DETR, MHOE-DETR shows an increase in AP value by 5.7%, 6.9%, and 4.2%, respectively, and a significant decrease in computational time by 67.0%, 77.2%, and 4.8%, respectively. The number of parameters decreases by 53.5% and 60.3% compared to YOLOv8-L and YOLOv9-C, respectively, and increases slightly compared to RT-DETR. The aforementioned results show that our proposed MHOE-DETR model is improved in all aspects compared to the existing methods.

On the other hand, we also demonstrate the effectiveness of different current target detection networks for ship detection on remote sensing image datasets, as shown in Fig. 8. We have chosen to focus on a representative scene. Fig. 8(a) shows the scene with a dense distribution of ships near the harbor. It can be seen that YOLOv8-N fails to detect low-contrast ships in this scenario. YOLOv9 and YOLOv10-N do the same and also fail to detect densely distributed ships. Fig. 8(b) shows a scenario where a ship interacts with an underwater farmed net box. We can see that the contours of the farmed nets are very similar to those of the ship. YOLOv8-N and YOLOv9 both appear to recognize the farmed nets as ships. MHOE-DETR, on the other hand, accurately identifies the ship with no misses. Fig. 8(c) shows the scenario where the port interacts with the ship. At this point, the contrast between ships and land is similar, and it is difficult for general models to distinguish the features of ships and land. YOLOv8-N and YOLOv10-N both show missed detection of small target ships. YOLOv9, on the other hand, misclassified small buildings on land as ships. MHOE-DETR, on the other hand, can accurately detect two ships in port. Fig. 8(d) shows the blurred scene of the ship in a cloudy state. It can be seen that both YOLOv9 and YOLOv10-N are unable to identify the ships in the blurred target. However, MHOE-DETR can accurately identify the blurred ship under the cloud cover with a confidence level of 87%.

Finally, we show a comparison of convergence speed versus efficiency for training multiple target detectors as shown in Fig. 9. MHOE-DETR clearly has a faster convergence speed and higher performance. Specifically, in the training of 100-epoch, MHOE-DETR improved the AP by 2.7%, 5.2%, 4.5%, 0.6%, and 4.0%, respectively compared with YOLOv8-n, YOLOv9, YOLOv10-N, RT-DETR, and YOLOv12-X. In the training of 200 epoch, it improved by 3.6%, 6.8%, 4.7%, 4.1%, and 3.4%, respectively, compared with YOLOv8-n, YOLOv9, YOLOv10-N, RT-DETR, and YOLOv12-X.

E. Continual Learning Validation Experiments

To validate the antiforgetting capabilities of our proposed method, we conducted comprehensive continual learning experiments on temporal remote sensing datasets and compared with mainstream detection methods and dedicated continual learning approaches.

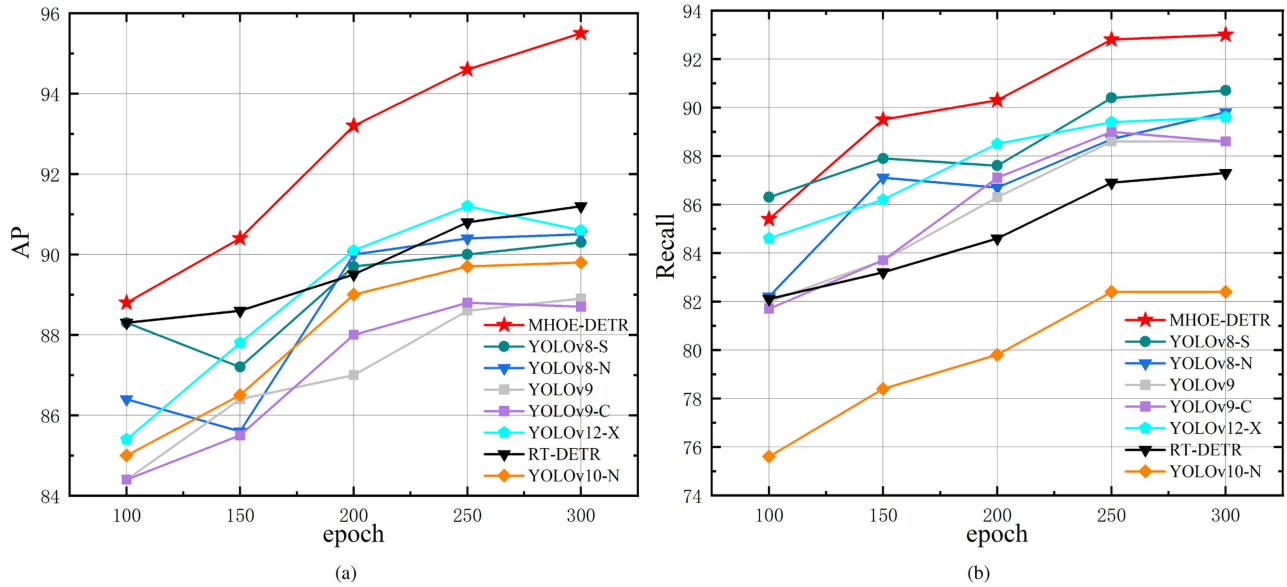


Fig. 9. Comparison of the convergence speeds of different models. (a) Average precision (AP) graphs for different models. (b) Recall (R) rate curves for different models.

1) Temporal Dataset Validation on xView2: We evaluated our model on the xView2 dataset, dividing it into two temporal phases: T1 (predisaster) and T2 (postdisaster) to simulate real-world data distribution changes.

Results demonstrate that our MHOE-DETR achieves negative forgetting rates (−0.2% and 0.0%), indicating performance improvement rather than degradation, while YOLOv8s and RT-DETR-R18 exhibit forgetting rates of 3.6%–8.1% and 9.5%–20.1%, respectively.

2) Sequential Task Learning With GDumb Comparison: We conducted sequential learning experiments using our Hainan 01 dataset, divided into two tasks: Task 1 (optimal conditions) and Task 2 (challenging conditions including inclement weather, low light, and complex port environments). We compared with GDumb, a strong continual learning baseline.

Our model demonstrates exceptional continual learning with negative forgetting rates (−2.08% and −8.98%), indicating positive knowledge transfer where challenging scenarios enhance understanding of ship characteristics. YOLOv8s with GDumb still exhibits positive forgetting rates despite memory mechanisms.

3) Key Advantages and Analysis: The superior antiforgetting performance stems from two architectural innovations: *Manhattan distance-based spatial prior* in MH-Net provides consistent spatial context across different imaging conditions, and *OREPAELAN feature preservation* maintains critical representations while adapting to new scenarios.

Compared to existing continual learning methods such iCaRL, our approach offers several advantages, which are as follows.

1) *No memory overhead:* Achieves antiforgetting through architectural design rather than exemplar storage.
2) *Positive transfer:* Enables performance improvement rather than mere preservation.
3) *Domain suitability:* Manhattan-distance-based modeling aligns with geometric properties of ship targets in satellite imagery (Table III).

Results demonstrate *zero forgetting* with negative forgetting rates across all metrics, making our method particularly suitable for real-world deployment in dynamic remote sensing environments where data distributions evolve over time.

Note: Forgetting rate is calculated as (Initial Performance—Subsequent Performance) / Initial Performance × 100% (Table IV). Negative values indicate performance improvement.

F. Ablation Studies

In order to verify the effectiveness of the designed MH-Net and OREPAELAN networks, this subsection performs ablation experiments on the self-generated remote sensing ship image dataset. To verify the adaptability of the MH-Net model to the complex environment of remote sensing ships and to mitigate the model forgetting problem. We set up Experiment 1: Introducing the CNN as the backbone of the whole network in the ablation experiment, while keeping the OREPAELAN network and comparing it with the MH-DETR network. To verify the effect of the OREPAELAN re-referencing network on the ablation of the network. We set up Experiment 2: We introduce the RepC3 network, leaving the rest of the network framework unchanged. The research objective of this ablation experiment is to verify that the proposed model does not forget the knowledge of different situations such as clouds, ports, and dense ships where the ship is located, and at the same time, ensure that there is no loss of accuracy to make the model lighter.

The results of the experiments are shown in Table V. The average accuracy of the MH-DETR model on the test set is 96.3%, the model comprises a mere 77.3 M parameters. The backbone network of the model is the spatial a priori network MH-Net,

TABLE III
CONTINUAL LEARNING RESULTS ON XVIEW2 DATASET

Model	Stage	Test Dataset	mAP50	mAP50-95	Forgetting Rate
3***YOLOv8s**	T1 Training	T1 Test	0.666	0.397	-
	T2 Training	T2 Test	0.571	0.305	-
	T2 Training	T1 Test	0.642	0.365	3.6% / 8.1%
3***RT-DETR-R18**	T1 Training	T1 Test	0.645	0.374	-
	T2 Training	T2 Test	0.488	0.229	-
	T2 Training	T1 Test	0.584	0.299	9.5% / 20.1%
3***MHOE-DETR (Ours)**	T1 Training	T1 Test	0.661	0.391	-
	T2 Training	T2 Test	0.625	0.350	-
	T2 Training	T1 Test	0.662	0.391	**-0.2% / 0.0%**

TABLE IV
SEQUENTIAL TASK LEARNING RESULTS WITH GDUMB COMPARISON

Model	Stage	mAP50	mAP50-95	Forgetting Rate
3***MHOE-DETR**	Stage 1 (Optimal Conditions)	0.913	0.635	-
	Stage 2 (Challenging Conditions)	0.841	0.570	-
	Test on Task 1	0.932	0.692	**-2.08% / -8.98%**
3* **YOLOv8s + GDumb**	Stage 1 (Optimal Conditions)	0.890	0.621	-
	Stage 2 (Challenging Conditions)	0.872	0.607	-
	Test on Task 1	0.850	0.601	4.49% / 3.22%

TABLE V
ABLATION STUDY FOR EACH COMPONENT

MH-Net	PResnet	OREPANELAN	RepC3	AP^{val}	AP^{val}_{50}	$AP^{val}_{50:95}$	Param.(M)
✓		✓		**96.3**	**95.4**	**54.4**	**77.3**
✓			✓	96.1	94.8	54.6	81.5
	✓		✓	92.7	91.9	48.6	75.8

which is based on the Manhattan attention mechanism, and the OREPANELAN network, which is improved by referencing the OREP network as a heavy-parameter convolutional module; and the mixing and interaction of features are carried out.

For Experiment 1, for the information bottleneck problem, if we continue to use MH-Net as the backbone network and OREPANELAN as the impulsive parameterized convolution module for the feature mixing part, its average accuracy is 96.1% and AP_{50} is 94.8%, but the number of parameters is 81.5M. Therefore, the OREPELAN network has a slight increase in P and R, while the number of parameters of the model has been reduced by 7.6%, which alleviates the information bottleneck problem of the model and makes the network lighter.

For Experiment 2, the model used is PResNet, which is commonly used for general detectors, as the backbone network, and RepC3 as the impulsive parameterized convolutional module. When tested on the remote sensing dataset, the average accuracy is only 92.7% and the AP_{50} is only 91.9%. The AP value of the MH-DETR model is improved by 3.9% and the AP_{50} by 3.8%. This indicates that the designed model mitigates the problem of the catastrophic model forgetting to some extent.

G. Theoretical Basis of the Manhattan Spatial Prior and the Verification of Its Compatibility With Ship Detection

The geometric superiority of the Manhattan distance spatial prior derives from its orthogonality decomposition property, which enables effective spatial relationship modeling through non-Euclidean geometric mapping in Cartesian coordinates. As illustrated by the diamond-shaped receptive field in Fig. 4, this representation exhibits distinct advantages in ship detection tasks, with the associated distance metric defined as

$$\begin{aligned} d_M &= |\Delta x| + |\Delta y| \\ &= |x_1 - x_2| + |y_1 - y_2|. \end{aligned} \tag{16}$$

Among them, d_M is the Manhattan distance, which is the shortest path distance of two points on a grid-like plane (such as a pixel grid), equal to the sum of the absolute values of the lateral distance and the vertical distance, and can better capture the axially aligned rectangular feature that ships usually have.

This geometric structure precisely matches the rectangular characteristics of vessels, as evidenced by remote sensing observations: 90% of ship principal axes exhibit deflection angles

TABLE VI
FEATURE COMPARISON BETWEEN MANHATTAN DISTANCE AND POLAR COORDINATES

Feature	Manhattan Distance	Polar Coordinates
Gradient Formula	$\frac{\partial D_M}{\partial \Delta x} = \ln\gamma \cdot \gamma^{\Delta x+\Delta y} \cdot (\text{sign}(\Delta x))$	$\frac{\partial D_P}{\partial \Delta\theta} = \gamma^{d_p} \ln\gamma \cdot \frac{r_{rT}\sin\Delta\theta}{d_p}$
Long-axis response ($p > 5$)	Attenuation gradient $\approx 1/p$	Isotropic attenuation
Ship adaptability	Maintains continuous hull attention	Keypoint feature dispersion

< 15 (per statistics from our proprietary dataset), with a median aspect ratio $p = 4.2$.

We analyzed the spatial attenuation characteristics of both polar coordinates and Manhattan distance through differential equations, and the comparison results are shown in Table VI. Here, γ represents the attenuation factor, $\Delta x + \Delta y$ represents the Manhattan distance, and $\text{sign}(\Delta x)$ will return 1 or -1 based on the positive or negative value of Δx, indicating the direction information. $\ln\gamma$ ensures that the gradient direction is opposite to the direction of distance increase. Furthermore, for the special target where the aspect ratio is greater than $p > 8$ (which accounts for approximately 30% of the dataset), we designed a parameterized attenuation function as follows:

$$D_{\text{adapt}} = \exp\left(-\beta\left(\frac{|\Delta x|}{\lambda_x} + \frac{|\Delta y|}{\lambda_y}\right)\right) \tag{17}$$

where λ_x said captain direction attenuation compensation, the $\lambda_x = p^{0.5}$. lambda$_y$ said the beam direction attenuation, the $\lambda_y = p^{-0.5}$. β is an attenuation coefficient, when the value is around 0.35, the model has the highest ship R rate on the private dataset.

H. Visualization and Analysis of MHOE-Net

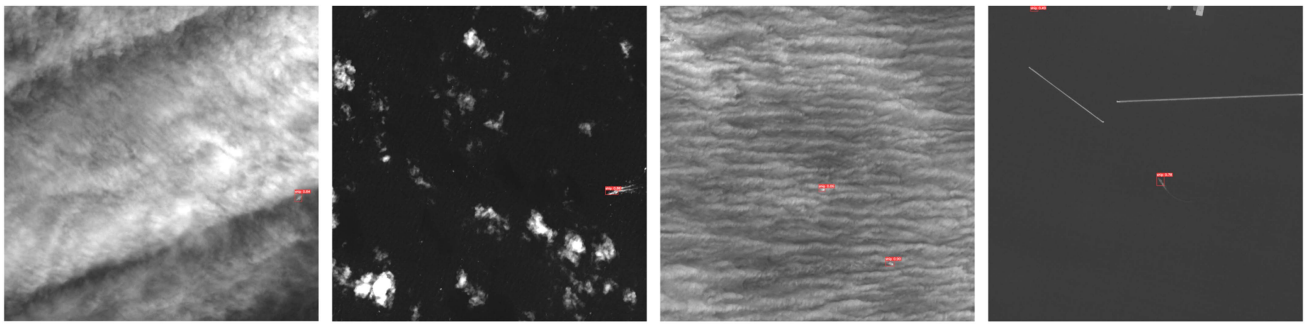
This section explores and visualizes the problem of modeling catastrophic pasts. To further understand how the MH-Net network, which introduces a self-attention prior, works, we perform a visualization of the MH-Net network on small and fuzzy target ship attention. For this visualization, we have extracted the attention scores from the MH block module of the fourth stage of the MH-Net, which is the main stage that consumes most of the computation in extracting features for small and fuzzy targets. When the remote sensing images are streamed into the network, there is a problem of model forgetting due to the large intraclass differences in the environment in which the ships in the remote sensing images are located (ocean nets, cloud occlusion, etc.). Therefore, this visualization provides a more intuitive view of the network's mitigation of the catastrophic model forgetting problem.

In Fig. 1, we show the results of the model's feature maps for different backbone network architectures in the case of large intraclass differences. Darknet-53 [49] represents the backbone network of YOLOv8-X, and RepNCSPELAN4 represents the network consisting of four GELANs. Fig. 1(a) and (b) both shows tiny ships under cloud cover, Fig. 1(c) is a scene of ocean waves, and Fig. 1(d) is a scene of underwater aquaculture nets whose contours are similar to those of ships. Despite the large differences within the scene classes, MH-Net can still capture long-range object relationships and accurately extract the features of small and fuzzy targets. The heat map shows that the MH-Net network extracts the features of small targets and the orientation of the ship, and does not pay much attention to the background around the ship. For the backbone network Darknet-53 [49] in Fig. 1(a)–(c) scenario, it is very difficult to see the position of the small target ship, then in the subsequent neck of the model, the network loses more features that can distinguish between the categories, leading to catastrophic model forgetting. For the RepNCSPELAN4 network, despite its combination of two neural network architectures, CSPNet [53] and ELAN [48], for gradient path planning, it can be seen from the underwater aquaculture netting scenario in Fig. 1(d) that the RepNCSPELAN4 network may not be able to discriminate between objects with similar contours (ships and underwater aquaculture nets). Therefore. For complex environments such as the ocean, data streams fed into neural networks are bound to have large intraclass variations, and both the mainstream Darknet-53 [49] backbone network and the backbone network composed of GELAN [50] are unsatisfactory. The MH-Net we design can effectively solve the problem of catastrophic model forgetting.

V. DISCUSSIONS

In the task of detecting small targets with fuzzy ships in remotely sensed image data, Table II by comparing the detection performance of RT-DETR with the algorithms YOLOv7-YOLOv9, which use a lot of postprocessing for target detection, and the end-to-end target detector YOLOv10. The end-to-end MHOE-DETR network proposed in this study outperforms other existing methods in terms of target detection performance. Although the number of parameters in the MHOE-DETR model is 65.9M and 34.9M more than that of YOLOv8-N and YOLOv8-S, respectively, the AP values of the MHOE-DETR model for the detection of small and fuzzy remote sensing targets and ships are improved by 8.9% and 5.9%, respectively, which reduces the missdetections and false detections of the ships, and the increase of the number of parameters is not too much in exchange of a better detection accuracy of the model. This is acceptable.

Furthermore, the detection results presented in Fig. 8 demonstrate that the MHOE-DETR algorithm continues to demonstrate accurate detection in scenarios involving dense ships, underwater aquaculture nets, harbor-land interaction, and cloud cover. In addition, it is evident that this algorithm outperforms other methods. As illustrated in Fig. 8(b), the remaining target detectors erroneously identify the nets as ships in the marine



(a) (b) (c) (d)

Fig. 10. Proposed model deals with three types of typical complex ship scenarios. (a) Small target ship detection scenario with obvious noise interference. (b) Scattered clouds cover the scene. (c) Thin clouds cover the scene densely. (d) The high-speed form of a ship forms a trailing track.

net box scenario. The limited number of marine net box scenes in the dataset may hinder the model's ability to adapt to scenes with a greater number of categories. In addition, the shape of the underwater net box bears resemblance to that of a ship, which may result in the model learning targets with similar shape contours and ships during the training process, potentially leading to the misidentification of the underwater net box as a ship.

In the scenario depicted in Fig. 8(a), the distribution of the small target ships is more intricate, resulting in the other detectors in the comparison failing to register a detection. It should be noted that the proposed MHOE-DETR does not include any leakage detection capabilities. However, it has been observed that the model incorrectly identifies two ships as a single entity in the context of a compact ship scenario. This is due to the fact that there are fewer scenarios in the dataset in which ship compactness occurs, which prevents the model from obtaining more features for the aforementioned case. In addition, the bow and stern of the two ships are facing the same direction and are in close proximity, which results in the shape profile of the two ships together being similar to the shape profile of one ship. Consequently, the model is unable to learn the features of the ships when they are in close proximity.

Finally, we conducted a comprehensive analysis of the proposed model's performance in addressing three typical complex maritime scenarios, specifically including: small target ship detection under significant noise interference, ship-environment feature similarity scenarios, and vessel trailing artifact conditions, as detailed in Fig. 10. Fig. 10(a) illustrates the challenging scenario of small vessel detection under pronounced speckle noise interference, a phenomenon typically observed in satellite remote sensing imagery captured during heavy sea wave conditions. The proposed model demonstrates robust performance in accurately identifying small maritime targets despite substantial noise contamination. Fig. 10(b) and (c) presentd two representative cases of feature confusion between vessels and environmental elements. Fig. 10(b) depicts a vessel situated within sparse cloud cover accompanied by faint trailing artifacts, while Fig. 10(c) shows a ship obscured by dense thin cloud cover with spectral characteristics analogous to cloud layers—both scenarios being prone to false positives or detection failures. The model maintains high detection accuracy in these challenging conditions, primarily attributed to the integration of the Manhattan-distance-based hybrid explicit spatial prior network (MH-Net) and the linear reparameterized cross-stage feature aggregation network (OREPAELAN). Fig. 10(d) demonstrates the model's performance in high-speed vessel detection with motion-induced trailing effects. Two high-speed vessels exhibiting pronounced linear blur artifacts were not successfully detected. This limitation stems from the following dual factors:

1) the high-speed motion artifacts manifest as bright linear streaks that obscure structural vessel features,
2) insufficient training samples containing such trailing characteristics in the dataset, which restricts the model's capacity to learn representative discriminative features.

These findings highlight potential directions for future dataset enhancement and model optimization.

VI. Conclusion

In this study, we use the ship dataset with small and fuzzy targets in multicategory scenarios from a previous work based on Hainan-1 satellite remote sensing data. This dataset contains 3800 remote sensing images of ships, and the scenarios include cloud cover, ocean nets, dense ships, and other complex situations. In order to be easily deployed on end devices and to eliminate the manual a priori anchor and NMS as postprocessing operations, we design an end-to-end network MHOE-DETR to solve the problem of difficulty in detecting small and fuzzy target ships in multicategory scenes in the field of ocean remote sensing.

The MH-Net network was designed with the Manhattan attention mechanism as its foundation. This approach offers an efficient and spatial prior that effectively addresses the catastrophic forgetting problem commonly encountered in machine learning models. In this study, the last stage of the MH-Net network output is visualized and analyzed, and the results show that the MH-Net still has a strong and stable performance in multicategory remote sensing image scenarios. To tackle the challenges in feature extraction for small and obscure ship targets, we designed the OREPAELAN feature amalgamation network, which simplifies complex normalization outcomes into a streamlined linear layer. This network facilitates parallel processing of input features, merging both low- and high-dimensional characteristics while minimizing information loss. The ablation experiments show that the introduction of OREPAELAN has improved both the

lightness and the accuracy of the model. The MHOE-DETR designed by combining MH-Net and OREPAELAN shows strong competitiveness, and this end-to-end network can be used to solve the problem of identifying small and fuzzy target ships in multiclass scenarios in the field of marine remote sensing. We conducted experiments on the proposed remote sensing ship dataset and compared it with mainstream advanced detection models (YOLOv7, YOLOv8 series, YOLOv9 series, YOLOv10 series,YOLOv12 series, and RT-DETR). The experimental results show that our proposed MHOE-DETR model improves the AP, R, and AP_{50} to 96.3%, 91.4%, and 95.4% for small and fuzzy target ship detection, respectively. Meanwhile, the AP_{50} of the model we designed has increased by 10.8%, 5.7%, and 5.0%, respectively compared with YOLOv12-N, YOLOv12-M, and YOLOv12-X. The R rates have increased by 6.0%, 2.5%, and 1.6%, respectively. While ensuring the detection accuracy of ships with small and fuzzy targets, the computational complexity and parameter count have decreased by 72.7% and 66.7%, respectively, compared with the YOLOv12-X model. This significantly reduces the computational complexity of the model.

REFERENCES

[1] R. Song, T. Li, and T. Li, "Ship detection in haze and low-light remote sensing images via colour balance and DCNN," *Appl. Ocean Res.*, vol. 139, 2023, Art. no. 103702.

[2] M. Er Joo et al., "Ship detection with deep learning: A survey," *Artif. Intell. Rev.*, vol. 56, no. 10, pp. 11825–11865, 2023.

[3] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models, Methods Appl.*, 2017, pp. 1–6.

[4] T. Bai et al., "Deep learning for change detection in remote sensing: A review," *Geo-Spatial Inf. Sci.*, vol. 26, no. 3, pp. 262–288, 2023.

[5] Z. Liu et al., "YOLO-extract: Improved YOLOv5 for aircraft object detection in remote sensing images," *IEEE Access*, vol. 11, 2023, pp. 1742–1751.

[6] X. Zhang, D. Zhu, and R. Wen, "SwinT-YOLO: Detection of densely distributed maize tassels in remote sensing images," *Comput. Electron. Agriculture*, vol. 210, 2023, Art. no. 107905.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[8] J. Ai et al., "AIS data aided Rayleigh CFAR ship detection algorithm of multiple-target environment in SAR images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 2, pp. 1266–1282, Apr. 2022.

[9] Q. Huang, H. Sun, Y. Wang, Y. Yuan, X. Guo, and Q. Gao, "Ship detection based on YOLO algorithm for visible images," *IET Image Process.*, vol. 18, no. 2, pp. 481–492, 2024.

[10] Y. Liang, J. Feng, X. Zhang, J. Zhang, and L. Jiao, "MidNet: An anchor-and-angle-free detector for oriented ship detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612113.

[11] S. Syed and K. Malathi, "Using modified R-FCN object-detection algorithm over fast R-CNN to increase classification accuracy for objects," *J. Surv. Fisheries Sci.*, vol. 10, no. 1S, pp. 2797–2806, 2023.

[12] S. Chen et al., "Info-FPN: An informative feature pyramid network for object detection in remote sensing images," *Expert Syst. Appl.*, vol. 214, 2023, Art. no. 119132.

[13] Y. Li et al., "Detection of the foreign object positions in agricultural soils using Mask-RCNN," *Int. J. Agricultural Biol. Eng.*, vol. 16, no. 1, 2023, pp. 220–231.

[14] G. Wen et al., "MS-SSD: Multi-scale single shot detector for ship detection in remote sensing images," *Appl. Intell.*, vol. 53, no. 2, pp. 1586–1604, 2023.

[15] X. Chen, M. Wang, J. Ling, H. Wu, B. Wu, and C. Li, "Ship imaging trajectory extraction via an aggregated you only look once (YOLO) model," *Eng. Appl. Artif. Intell.*, vol. 130, Art. no. 107742, 2024.

[16] T. Zhao et al., "Ship detection with deep learning in optical remote-sensing images: A survey of challenges and advances," *Remote Sens.*, vol. 16, no. 7, pp. 1145, 2024.

[17] Y. L. Chang et al., "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, 2019, vol. 11, no. 7, Art. no. 786.

[18] Z. Deng, H. Sun, S. Zhou, and J. Zhao, "Learning deep ship detector in SAR images from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, pp. 4021–4039, 2019.

[19] H. M. Khan and C. Yunze, "Ship detection in SAR image using YOLOv2," in *Proc. 37th Chin. Control Conf.*, 2018, 9495–9499.

[20] Y. Chaudhary, M. Mehta, N. Goel, P. Bhardwaj, D. Gupta, and A. Khanna, "YOLOv3 remote sensing SAR ship image detection," in *Data Analytics and Management*. Singapore: Springer, 2021, pp. 519–531.

[21] T. Zhang, X. Zhang, J. Shi, and S. Wei, "High-speed ship detection in SAR images by improved YOLOv3," in *Proc. 16th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process.*, Chengdu, China, Dec. 2019, pp. 149–152.

[22] Z. Hong et al., "Multi-scale ship detection from SAR and optical imagery via a more accurate YOLOv3," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6083–6101, 2021.

[23] J. H. Jiang et al., "High-speed lightweight ship detection algorithm based on YOLO-V4 for three-channels RGB SAR image," *Remote Sens.*, vol. 13, 2021, Art. no. 1909.

[24] B. Sun et al., "Automatic ship object detection model based on YOLOv4 with transformer mechanism in remote sensing images," *Appl. Sci.*, vol. 13, no. 4, 2023, Art. no. 2488.

[25] J. Jian et al., "Optical remote sensing ship classification and recognition based on improved YOLOv5," *Remote Sens.*, vol. 15, no. 17, 2023, Art. no. 4319, doi: 10.3390/rs15174319.

[26] Z. Zhang, R. Ouyang, and J. Xie, "Application of a remote-sensing ship dataset based on the YOLOv5 model," in *Proc. 16th Int. Conf. Mach. Learn. Comput.*, 2024, pp. 52–56.

[27] Z. Chen et al., "Multi-scale ship detection algorithm based on YOLOv7 for complex scene SAR images," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 2071.

[28] Y. Liu et al., "YOLOv7oSAR: A lightweight high-precision ship detection model for SAR images based on the YOLOv7 algorithm," *Remote Sens.*, vol. 16, no. 5, 2024, Art. no. 913.

[29] H. Nie et al., "A lightweight remote sensing small target image detection algorithm based on improved YOLOv8," *Sensors*, vol. 24, no. 9, Art. no. 2024, 2952.

[30] X. Zhao and Y. Song, "Improved ship detection with YOLOv8 enhanced with MobileViT and GSConv," *Electronics*, vol. 12, no. 22, 2023, Art. no. 4666.

[31] Z. Yang et al., "Few-shot object detection in remote sensing images via data clearing and stationary meta-learning," *Sensors*, vol. 24, no. 12, 2024, Art. no. 3882.

[32] S. Wang, Z. Cai, and J. Yuan, "Automatic SAR ship detection based on multi-feature fusion network in spatial and frequency domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4102111.

[33] Z. Sun et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, 2021.

[34] L. Bai, C. Yao, Z. Ye, X. Xue, X. Lin, and M. Hui, "A novel anchor-free detector using global context-guide feature balance pyramid and united attention for SAR ship detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 4003005.

[35] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020, doi: 10.1109/ACCESS.2020.3005861.

[36] J. Kolluri and R. Das, "Ship detection from satellite images with advanced deep learning model (single shot detector (SSD))," in *Proc. Int. Conf. Front. Intell. Comput., Theory Appl.*, Singapore, 2022, pp. 337–349.

[37] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion SSD for remote sensing object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5501309.

[38] W. Wu et al., "Ship detection and recognition based on improved YOLOv," *Comput., Mater. Continua*, vol. 76, no. 1, 2023, pp. 489–498, doi: 10.32604/cmc.2023.039929.

[39] Y. Chuang et al., "Infrared small target detection based on multiscale local contrast learning networks," *Infrared Phys. Technol.*, vol. 123, 2022, Art. no. 104107.

[40] C. Yu et al., "Pay attention to local contrast learning networks for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3512705.

[41] C. Yu et al., "Precise segmentation of remote sensing cage images based on SegNet and voting mechanism," *Appl. Eng. Agriculture*, 2022, vol. 38, no. 3, pp. 573–581.
[42] L. Xu, Z. Hu, C. Zhang, and W. Wu, "Remote sensing image segmentation of mariculture cage using ensemble learning strategy," *Appl. Sci.-Basel*, vol. 12, no. 16, 2022, Art. no. 8234.
[43] Q. Fan, H. Huang, H. Liu, and R. He, "RMT: Retentive networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 2024, pp. 5641–5651.
[44] M. Hu et al., "Online convolutional re-parameterization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 568–577.
[45] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style convnets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.* 2021, pp. 13733–13742.
[46] T. Y. Ross and G. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2980–2988.
[47] H. Zhang, "DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations* 2023.
[48] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2023, pp. 7464–7475.
[49] J. Glenn, "YOLOv8 release v8.1.0," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics/releases/tag/v8.1.0
[50] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.
[51] A. Wang et al., "YOLOv10: Real-time end-to-end object detection," in *Proc. NeurIPS*, 2024.
[52] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16965–16974.
[53] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*. 2020, pp. 390–391.
[54] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-centric real-time object detectors," 2025, *arXiv:2502.12524*.
[55] J. Chen et al., "Discrete edge feature guided rotation detection method for remote sensing ship wake," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, early access, Apr. 11, 2025, doi: 10.1109/JSTARS.2025.3560200.

Zhuhua Hu (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the Jilin University, Changchun, China, in 2002 and 2005, respectively, and the Ph.D. degree from the Hainan University, Haikou, China, in 2019.

He was a Software Engineer with Ningbo BIRD Research Institute of China from 2005 to 2006. He was a Software Engineer with Nanjing Research Institute of ZTE from 2006 to 2007. He was a Minister with the Software Department, Shanghai Aoxun Information Technology Company, Ltd., from 2007 to 2009. He has been a Professor and Doctorial Tutor with the School of Information and Communication Engineering, Hainan University, since 2020. He led the "Multimodal information intelligent processing and decision control" innovation team, and has authored and more than 130 academic papers in journals such as IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN, *Optics Express*, and *Computers and Electronics in Agriculture*, authorized 19 patents, and hosted more than ten large-scale commercial projects that have been successfully implemented. His current research interests include artificial intelligence and signal and information processing.

Dr. Hu is a Senior Member of the China Computer Federation. He is currently a high-level talent in Hainan Province. He was the Reviewer for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSIN, IEEE INTERNET OF THINGS JOURNAL, *Ocean Engineering*, Knowledge-Based Systems, Engineering Applications of Artificial Intelligence, Image and Vision Computing, IEEE International Conference on Acoustics, Speech and Signal Processing 2023 and 2024, and IEEE International Conference on Multimedia and Expo 2024.

Xiyu Fan received the M.Eng. degree majored in information and communication engineering from the School of Information and Communication Engineering, Hainan University, Haikou, China, in 2025.

His main research interests include image object detection and digital image processing.

Yaochi Zhao received the M.S. degree in pattern recognition and intelligent system from Central South University, Changsha, China, in 2005, and the Ph.D. degree in computer applied technology from the Tianjin University, Tianjin, China, in 2023.

She worked with Ningbo BIRD Research Institute and Shanghai Wingtech Communication Co., Ltd., for three years. Later, she was involved in teaching and research work with the College of Information Science and Technology, Hainan University, Haikou, China, where she is currently an Associate Professor with the School of Cyberspace Security. Her current research interests include deep learning, image processing, and computer vision.

Wei Wu received the B.Eng. degree majored in electronic information engineering from the Beijing University of Chemical Technology, Beijing, China, and the M.Eng. degree majored in software engineering from Tsinghua University, Beijing. He is currently working toward the Ph.D. degree in information and communication engineering with the School of Information and Communication Engineering, Hainan University, Haikou, China.

His main research interests include digital image processing and computer systems.

Jie Liu is currently working toward the B.Eng. degree in communication engineering with the School of Information and Communication Engineering, Hainan University, Haikou, China.

His research interests include object detection and digital image processing.