GLAF-DETR: Detection Transformer With Global–Local Adaptive Fusion Attention for Infrared Maritime Object Detection

Wenbo Zhang, Dongsheng Guo, *Member, IEEE*, Yilin Shang, Weidong Zhang, *Senior Member, IEEE*, and Zhuhua Hu, *Senior Member, IEEE*

***Abstract*—Infrared maritime object detection is a crucial technology for sea surface monitoring in low-light conditions within maritime Internet of Things (IoT) systems. In practical applications, this task faces significant challenges, including diverse target sizes and stringent real-time processing requirements. To address these challenges, a DEtection TRansformer (DETR) with global–local adaptive fusion (GLAF) attention for infrared maritime object detection (GLAF-DETR) is proposed. The GLAF attention mechanism is designed to capture both global contextual information and fine local details of objects. GLAF dynamically adjusts attention across regions by integrating long-range dependencies with short-range positional information, significantly enhancing detection performance for targets of varying sizes in complex maritime environments. In addition, the dynamic adaptive multiscale feature fusion (DAMFF) module is proposed to promote cross-channel interaction among multiscale features. Guided by GLAF, DAMFF dynamically fuses these features, further enhancing the accuracy of multiscale object detection. The lightweight HGNetv2-IRLight backbone is designed to minimize network complexity and ensure real-time performance by reducing redundant information while maintaining strong infrared feature extraction. Extensive experiments conducted on an infrared maritime object dataset show that GLAF-DETR surpasses state-of-the-art methods in both detection accuracy and inference speed. It demonstrates outstanding performance, particularly in detecting objects across different scales, offering enhanced accuracy and robustness in challenging maritime scenarios.**

***Index Terms*—DEtection TRansformer (DETR), global–local adaptive fusion (GLAF) attention, infrared maritime object detection, multiscale feature fusion.**

I. Introduction

With the rapid expansion of maritime activities and the integration of smart devices, the maritime Internet of Things (IoT) has become crucial in advancing the digitalization and intelligence of marine systems [1], [2]. However, the marine environment is dynamic and often unpredictable. Conditions, such as low light and adverse weather pose significant challenges to traditional visible light detection systems. These factors undermine the reliability of such systems for consistent, all-weather monitoring. To address these challenges, infrared maritime object detection has emerged as a vital real-time sensing technology within maritime IoT frameworks. Its enhanced performance under low-light and harsh weather conditions significantly improves the accuracy of maritime monitoring [3]. Consequently, infrared maritime object detection is increasingly employed in critical applications, such as coastal surveillance, maritime search and rescue, maritime safety, and intelligent unmanned marine platforms [4].

In the vast and complex maritime environment, infrared object detection faces several critical challenges. Firstly, significant physical size differences exist among various targets, such as cargo ships, speedboats, and buoys. Additionally, differences in the distance between the targets and the infrared camera lead to variations in their apparent size in infrared images. Together, these factors necessitate a detection system that can effectively adapt to multiple scales, posing stringent technical challenges [5]. Secondly, the inherent limitations of infrared imaging, including low resolution, limited feature information, and lack of texture, further complicate detection. These issues are exacerbated by the dynamic and unpredictable nature of the maritime environment, making accurate detection even more challenging. Lastly, maritime monitoring requires real-time performance, crucial for effective integration into maritime IoT systems to ensure continuous, reliable surveillance and prompt responses to changing conditions.

To address the challenges outlined above, traditional infrared maritime object detection methods often rely on manually crafted features. However, these methods lack the flexibility to adapt to complex and dynamic ocean environments. In recent years, deep learning techniques have made significant advances in the field of object detection, enabling automatic learning and feature extraction from large maritime datasets. Compared to conventional approaches,

Received 10 July 2025; accepted 4 August 2025. Date of publication 11 August 2025; date of current version 24 October 2025. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0119901; in part by the National Natural Science Foundation of China under Grant U2141234, Grant 62463004, and Grant U24A20260; in part by the Hainan Province Science and Technology Special Fund under Grant ZDYF2024GXJS003; and in part by the Scientific Research Fund of Hainan University under Grant KYQD(ZR)23025. *(Corresponding author: Dongsheng Guo.)*

Wenbo Zhang, Dongsheng Guo, Yilin Shang, and Zhuhua Hu are with the School of Information and Communication Engineering, Hainan University, Haikou 570288, China (e-mail: gdongsh2022@hainanu.edu.cn).

Weidong Zhang is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China.

Digital Object Identifier 10.1109/JIOT.2025.3597820

these deep learning-based techniques have demonstrated superior performance, particularly in complex maritime scenarios. Among them, Miao et al. [6] introduced an infrared ship detection algorithm based on multiscale feature extraction, enhancing detection accuracy by combining features from different candidate regions. While effective, this method's reliance on a two-stage detection framework limits its real-time applicability, making it less suitable for time-sensitive maritime surveillance tasks.

Compared to two-stage detection approaches, single-stage frameworks like the YOLO [7] series directly predict object categories and bounding boxes. The YOLO algorithms strike a balance between detection accuracy and inference speed, making them well-suited for real-time infrared maritime detection. Several improvements have been proposed to improve YOLO's performance in detecting infrared objects of varying scales within complex ocean environments. Ye et al. [8] and Wang et al. [9] introduced different attention-based feature fusion methods to improve YOLOv5, which effectively suppress background noise while improving multiscale feature fusion. Deng and Zhang [10] optimized YOLOv7 by proposing a weighted feature fusion approach based on dilated convolutions, incorporating channel and spatial excitation attention modules to improve the stability of multiscale ship target detection.

The YOLO series has made rapid advancements in infrared maritime object detection. It relies on non-maximum suppression (NMS) for post-processing, which reduces the stability and inference speed. To address these limitations, Carion et al. [11] proposed end-to-end object DEtection TRansformer (DETR). DETR eliminates the need for NMS, simplifying the detection pipeline. To improve DETR's real-time performance, Zhao et al. [12] proposed the real-time DETR (RT-DETR), which outperforms YOLO detectors of similar scale in both speed and accuracy. This makes RT-DETR better suited to meet the real-time infrared maritime detection requirements in maritime IoT systems. However, RT-DETR, optimized for RGB images, struggles with infrared maritime challenges. RT-DETR's generic convolutional backbone and attention mechanism fail to effectively balance global context with fine-grained local details in sparse infrared data, resulting in diminished accuracy for multiscale objects. Furthermore, its static multiscale fusion lacks dynamic adaptability, limiting robustness to the diverse scales of maritime targets in complex ocean environments.

To address these limitations, a DETR with global–local adaptive fusion (GLAF) attention for infrared maritime object detection (GLAF-DETR) is proposed. GLAF-DETR enhances RT-DETR for real-time maritime IoT systems. First, a GLAF attention mechanism is introduced, whereby global and local features are dynamically fused to improve detection of infrared targets across varying sizes. Then, a dynamic adaptive multiscale feature fusion (DAMFF) module is designed, whereby multiscale features are adaptively fused through attention-guided mechanisms to enhance robustness against scale variations. Finally, HGNetv2-IRLight, a lightweight backbone, is developed for sparse grayscale infrared data. It significantly improves computational efficiency while preserving feature representation for real-time performance. Extensive experiments conducted on multiple infrared maritime datasets demonstrate that GLAF-DETR effectively addresses the challenges of sparse, multiscale targets. It surpasses state-of-the-art methods and provides an efficient and robust solution for maritime IoT systems.

The main contributions of this article are as follows.

1) A lightweight GLAF attention mechanism is proposed to enhance feature extraction for infrared targets of various sizes. The GLAF mechanism dynamically fuses global context and local details, significantly strengthening the feature representation for sparse, small targets. By reducing computational cost, GLAF maintains real-time efficiency and achieves superior detection accuracy and speed compared to conventional attention mechanisms.
2) A DAMFF module is introduced, which adaptively fuses multiscale features through attention-guided mechanisms, thereby improving detection performance across a wide range of target scales. In contrast to the static fusion of FPN, DAMFF enhances scale adaptability, all while ensuring efficient real-time performance.
3) HGNetv2-IRLight, a lightweight backbone optimized for infrared maritime detection, is introduced. By reducing depth and channel dimensions to eliminate redundancy in grayscale data, it preserves robust feature representation and achieves high real-time performance.

II. Proposed Method

A. Model Architecture

The proposed GLAF-DETR architecture comprises three core components: a lightweight backbone network, a dynamic multiscale fusion encoder, and a transformer decoder. The GLAF-DETR architecture is presented in Fig.1. The backbone is based on the lightweight HGNetv2-IRLight, which efficiently extracts target features from infrared maritime images while ensuring real-time performance. The GLAF attention mechanism is integrated into the backbone to enhance the representation of sparse and low-contrast targets. GLAF allows the model to capture both global contextual information and fine-grained local details across scales. Multiscale features from the final three stages of the backbone $\{S_3, S_4, S_5\}$ are fed in parallel to the encoder. The DAMFF module within the encoder dynamically merges features at different scales, improving the model's robustness to scale variation in complex maritime environments. The fused features are then passed to the uncertainty-minimal query selection module and subsequently processed by the transformer decoder with auxiliary prediction heads to generate final bounding boxes and confidence scores. Overall, GLAF-DETR is specifically designed for infrared maritime object detection tasks. It effectively addresses challenges, such as target sparsity, scale variation, and environmental complexity, while meeting the real-time inference requirements of maritime IoT systems.

B. Global Local Adaptive Fusion Attention

The vast and complex maritime environment contributes to the diverse scales of targets in infrared maritime images,

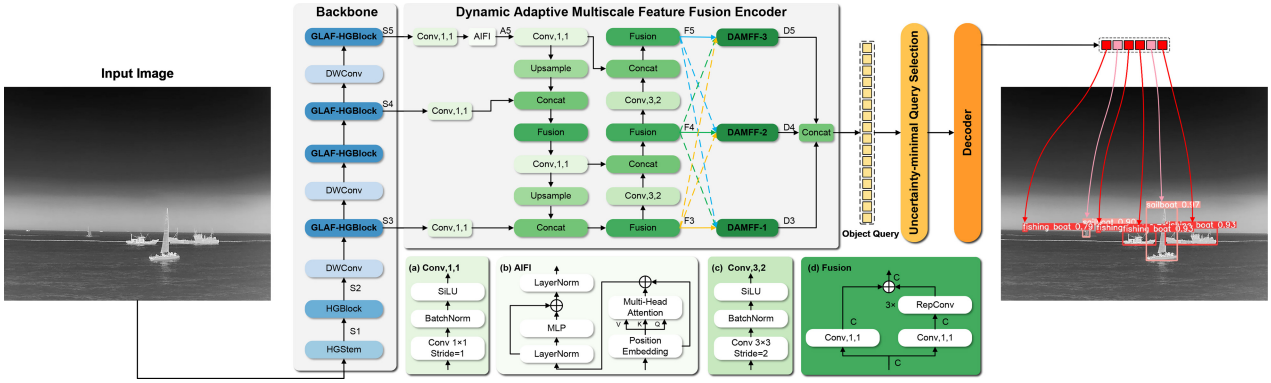


Fig. 1. Model architecture of GLAF-DETR.

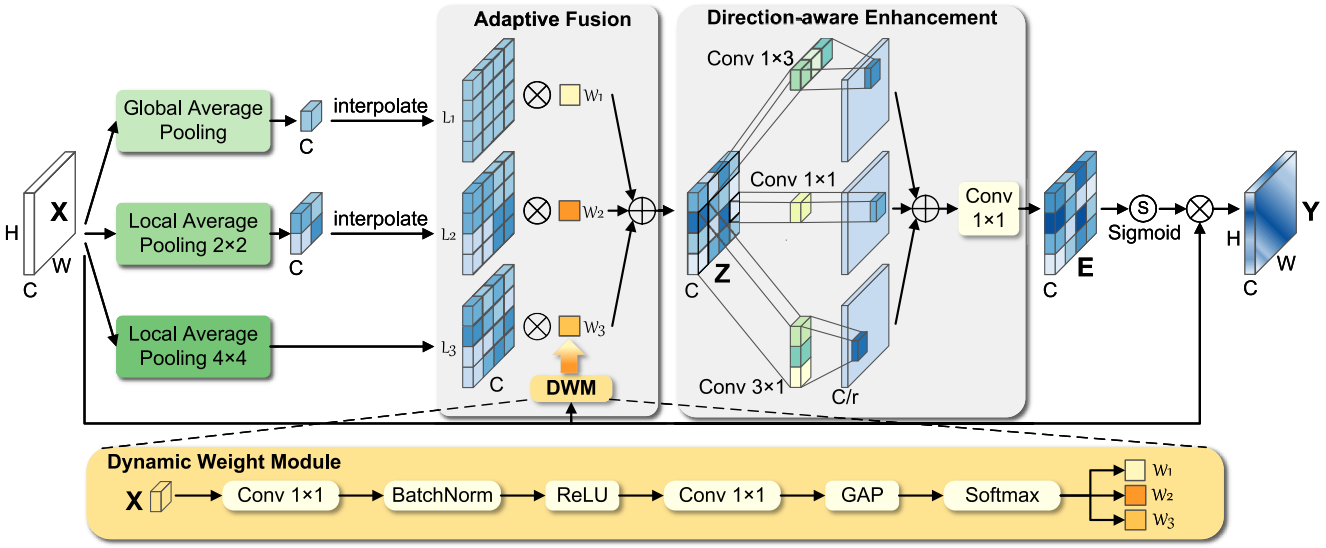


Fig. 2. Network architecture of the GLAF attention mechanism. The GLAF mechanism first applies global average pooling and multiscale pooling to capture both global contextual and local structural features, enhancing spatial structure representation across multiple scales. The DWM then computes adaptive fusion weights to balance the contributions of global and local information. Additionally, the DE module refines both spatial and directional consistency, improving the model's adaptability to targets of varying shapes and sizes. This process ultimately generates effective fusion weights, optimizing overall feature extraction performance.

posing significant detection challenges. Additionally, marine noise leads to sparse infrared target information with limited texture, further complicating detection. To address these issues, detection algorithms [6], [8], [9], [10] often use attention mechanisms to enhance target features and suppress noise, thereby improving detection accuracy. However, existing attention mechanisms, such as SE [13], ECA [14], and CBAM [15] primarily focus on either global or local features, often relying on single-scale pooling or limited spatial cues [16], [17]. This constrains their ability to model spatial relationships across targets of varying sizes and hinders effective extraction of critical features in complex maritime environments. To overcome these limitations, the GLAF attention mechanism is introduced. GLAF effectively integrates global and local information, dynamically adjusts weights, and optimizes feature fusion, significantly enhancing detection capabilities for infrared targets of varying sizes.

The GLAF mechanism encompasses global and local information extraction, adaptive fusion, direction-aware enhancement (DE), and attention generation, as depicted in Fig. 2. GLAF leverages global average pooling to capture global contextual information and establish long-range feature dependencies. At the same time, it employs multiscale average pooling to extract local features, highlighting high-resolution spatial details at shorter distances. This integration of global and local features enhances the ability to represent diverse feature characteristics and spatial structures across various scales, with the specific operation as follows:

$$L = [\text{Scale}(P(X, 1)), \text{Scale}(P(X, 2)), P(X, 4)]^T \quad (1)$$

where $P(\cdot, \cdot)$ denotes the adaptive average pooling operation, which transforms input feature maps of arbitrary sizes into fixed-resolution representations based on specified values. The values 1, 2, and 4 represent target dimensions, capturing spatial granularity levels from global semantics to local details, rather than specific target sizes. The input feature is denoted as $X \in \mathbb{R}^{C\times W\times H}$. Specifically, $P(X, 1)$ compresses the entire feature map into a single value per channel, capturing global information [13]. $P(X, 2)$ generates a 2×2 feature map, preserving intermediate spatial structures, while $P(X, 4)$

produces a 4×4 feature map, retaining finer local details. Compared to global average pooling in $P(X, 1)$, the higher spatial resolutions of $P(X, 2)$ and $P(X, 4)$ enhance feature representations, providing richer structural details for small to medium targets and improved contextual integration for large targets. The Scale$(\cdot)$ operation adjusts these outputs to a consistent dimension, improving the comprehensive representation of diverse features across multiple scales.

Simple weighted fusion of global and local information often proves insufficient for addressing the challenges posed by varying multiscale targets and complex marine environments. The GLAF mechanism adaptively adjusts fusion weights based on input features, achieving a more refined balance between global and local information. This adaptive fusion process enables GLAF to dynamically adjust focus on the most relevant regions based on changes in the input features. As a result, the detector becomes more flexible in responding to varying environmental conditions. A lightweight dynamic weight module (DWM) is introduced to generate adaptive weights, as shown in Fig. 2. The DWM automatically evaluates and adjusts the importance of global and local features through a learnable lightweight combination and global pooling. The specific operations are as follows:

$$W = \mathrm{DWM}(X) \tag{2}$$

$$Z = W^T L \tag{3}$$

where $W \in \mathbb{R}^3$ is a learnable parameter vector. The DWM calculates W from global features of original input X. These global features retain complete semantic context, enabling more precise guidance of weights for features across varying distances. Consequently, DWM endows GLAF with the ability to generalize across a continuous range of object scales.

In complex maritime environments, targets vary not only in scale but also in aspect ratio and shape. The GLAF introduces a DE strategy to enhance the model's adaptability to diverse target shapes and forms. The structure of the DE module is shown in Fig. 2. The specific operation is expressed as follows:

$$E = \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{1\times3}(Z) + \mathrm{Conv}_{1\times1}(Z) + \mathrm{Conv}_{3\times1}(Z)) \tag{4}$$

where $\mathrm{Conv}_{1\times1}(\cdot)$ preserves spatial details, while the asymmetric kernels $\mathrm{Conv}_{1\times3}(\cdot)$ and $\mathrm{Conv}_{3\times1}(\cdot)$ enhance edge responses along the horizontal and vertical directions, respectively. They enrich feature representations by emphasizing salient local patterns, thereby improving the model's ability to perceive structural cues from various orientations. The integration of these features strengthens the detection of infrared maritime objects with diverse shapes and poses.

The attention generation component employs a Sigmoid gating function to capture relationships among features, directing the model's focus to the most relevant regions. The processed features are aggregated to create the final weight map. By multiplying this weight map with the original features X, attention is distributed across various target regions in the infrared image. This adaptive mechanism enhances significant features while suppressing irrelevant information, resulting in the final output defined as

$$Y = \mathrm{Scale}(\sigma(E)) \otimes X \tag{5}$$

where $\sigma(\cdot)$ represents the Sigmoid gating function [14], Scale$(\cdot)$ adjusts the outputs to the shape of X, and $\otimes$ denotes channel-wise multiplication.

The GLAF attention mechanism adopts a lightweight design approach. The adaptive average pooling operation is non-learnable and incurs negligible computational cost. Furthermore, the DWM module and other components within GLAF utilize small-kernel convolutions, introducing only a minimal number of parameters and low computational complexity. As a result, this design allows GLAF to deliver robust detection performance and efficiency in complex maritime environments, while maintaining a low computational overhead.

GLAF is a plug-and-play universal module designed for seamless integration into object detection models. When integrated into the backbone network, it forms the GLAF-HGBlock, as illustrated in Fig. 3. This GLAF-HGBlock acts as the primary module in stages 2, 3, and 4 of the backbone network. The GLAF attention mechanism dynamically adjusts its focus based on the size of infrared maritime targets, thereby balancing global and local feature extraction. This capability enables the detector to effectively manage small objects with fine details while maintaining robustness for larger objects that require broader contextual information, enhancing adaptability in complex maritime environments.

C. Dynamic Adaptive Multiscale Feature Fusion Encoder

The multiscale feature fusion network integrates both shallow and deep feature information. Shallow features capture fine texture details and the distinctive characteristics of smaller objects, while deep features focus on high-level semantic and larger object attributes. This fusion network is widely utilized in infrared maritime target detection, enhancing feature representation. However, mainstream fusion networks, such as FPN [18] and PANet [19] employ static strategies like direct summation or concatenation, lacking scale sensitivity. Although BiFPN [20] and ASFF [21] introduce weight-based fusion, their fixed weighting schemes fail to adapt to the dynamic scale variations of targets in complex maritime environments. To overcome these limitations, we propose the DAMFF encoder. It combines intra-scale interactions with cross-scale fusion. The cross-scale fusion is achieved through the DAMFF module. Unlike static fusion, DAMFF utilizes an attention-guided mechanism that dynamically adjusts fusion weights based on the relative importance of features at different scales. This adaptive approach enables the model to prioritize relevant features and balance contributions of small and large objects, optimizing detection accuracy.

The encoder integrates the outputs from the final three stages of the backbone network $\{S_3, S_4, S_5\}$, as shown in Fig. 1. Initially, the attention-based intrascale feature interaction (AIFI) is applied to the S_5 features to facilitate intra-scale information exchange. The structure of AIFI is shown in Fig. 1(b). AIFI employs the transformer-based self-attention mechanism to enhance global information capture and strengthen long-range dependencies between targets. The multiscale features $\{S_3, S_4, A_5\}$ are initially fused through

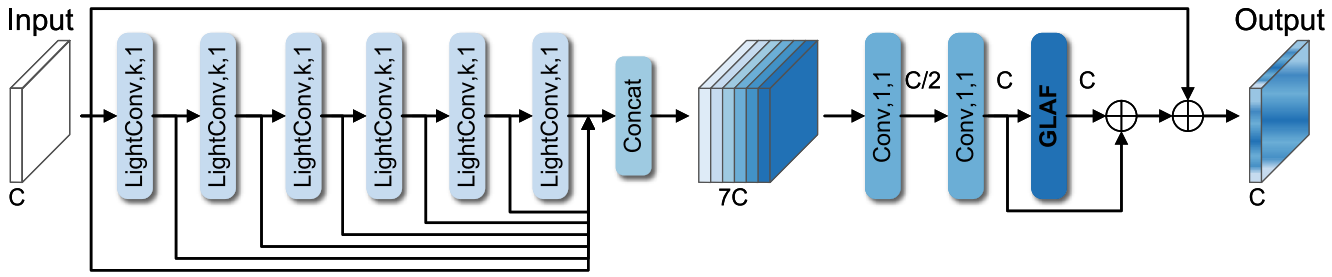


Fig. 3. Network architecture of GLAF-HGBlock. k represents the size of the convolutional kernel.

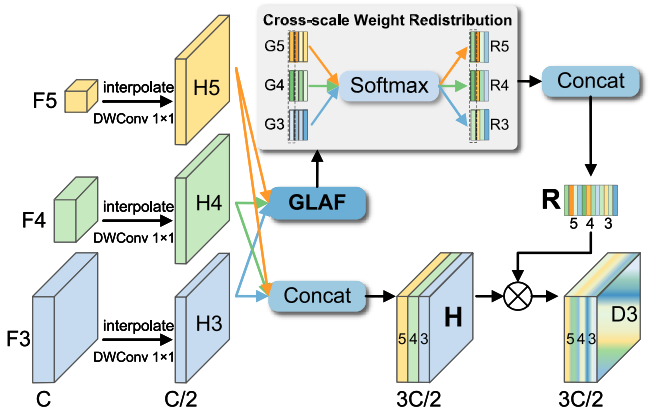


Fig. 4. Network architecture of DAMFF-1 module.

a path incorporating bottom-up, top-down, and lateral connections, yielding the fused multiscale features $\{F_3, F_4, F_5\}$. These features are then fed into the DAMFF module for further adaptive fusion, producing the final features $\{D_3, D_4, D_5\}$. The DAMFF module is structured into three variants: DAMFF-1, DAMFF-2, and DAMFF-3, based on output scales, with each variant following the same process, as illustrated in Fig. 1. The detailed structure of DAMFF-1 is shown in Fig. 4.

The DAMFF-1 module first scales the multiscale features $\{F_3, F_4, F_5\}$ to match the resolution of F_3, resulting in $\{H_3, H_4, H_5\}$. These scaled features are then concatenated along the channel dimension to form

$$H = \text{Cat}([H_3, H_4, H_5]). \tag{6}$$

The DAMFF module inputs each multiscale feature $\{H_3, H_4, H_5\}$ into the GLAF attention mechanism, generating attention weights for each feature map

$$G_i = \text{GLAF}(H_i) \tag{7}$$

where i represents the feature scales 3, 4, and 5. This operation assesses the importance of relevant regions within the features. To establish cross-level feature correlations, DAMFF redistributes the attention weights across scales, enabling global information exchange between channels. The module applies the Softmax normalization function to the attention weights of the three feature levels G_i, resulting in the final distributed weights for each feature channel in the fused representation, expressed as

$$R_i = \text{Softmax}(G_i) = \frac{\exp(G_i)}{\sum_{i=3}^{5} \exp(G_i)}. \tag{8}$$

The DAMFF module concatenates the attention-guided weights R_i in channel order, then applies these weights R to the feature map H, resulting in the final fused feature

$$D_3 = R \otimes H = \text{Cat}([R_3, R_4, R_5]) \otimes H. \tag{9}$$

The DAMFF module leverages GLAF attention along with cross-scale weight redistribution. This combination allows it to dynamically and adaptively adjust fusion weights across different scales based on the feature map content. DAMFF highlights the most relevant features of multiscale targets. This enhances the detector's adaptability and improves detection performance for targets of various scales in complex maritime environments. In the DAMFF module, small-kernel depthwise separable convolutions are used to reduce computational cost. Its cross-scale branches share the GLAF module, avoiding repeated attention blocks and further lowering complexity, thus meeting real-time requirements in maritime IoT scenarios.

D. Lightweight Backbone Network

In practical applications, real-time performance is critical for infrared maritime target detection. The RT-DETR algorithm surpasses YOLO detectors of similar scale in real-time capabilities. To preserve this advantage, the GLAF-DETR algorithm is built on the RT-DETR framework. HGNetv2-L and HGNetv2-X backbone networks from RT-DETR achieved strong feature extraction [12]. However, increasing the network depth by 61.4% in HGNetv2-X led to only marginal improvements in detection performance compared to HGNetv2-L, as shown in Table I. To resolve this, we performed an in-depth analysis of the HGNetv2 architecture to optimize it for infrared maritime object detection. Our goal was to reduce network complexity while maintaining effective feature extraction, ensuring the architecture remains efficient for real-time applications.

We conduct a detailed structural analysis and decomposition of the HGNetv2-X and HGNetv2-L architectures, with their configurations summarized in Table II. Both networks follow a modular stacking design, where parameters a and b control the number of channels, while c and d determine the number of repeated blocks in each stage. HGNetv2-X is configured

TABLE I
DETAILED COMPARISON OF NETWORK STRUCTURE PARAMETERS BETWEEN (LEFT) HGNETV2-L, (MIDDLE) HGNETV2-X, AND (RIGHT) HGNETV2-IRLIGHT

Stage	Output Size	HGNetv2-L	HGNetv2-X	HGNetv2-IRLight
Stem	160×160	$Conv, 3\times3, 32$,stride 2 $Conv, 2\times2, 16$ $Conv, 2\times2, 32$ $Conv, 3\times3, 32$,stride 2 $Conv, 1\times1, 48$	$Conv, 3\times3, 32$,stride 2 $Conv, 2\times2, 16$ $Conv, 2\times2, 32$ $Conv, 3\times3, 32$,stride 2 $Conv, 1\times1, 64$	$Conv, 3\times3, 16$,stride 2 $Conv, 2\times2, 8$ $Conv, 2\times2, 16$ $Conv, 3\times3, 16$,stride 2 $Conv, 1\times1, 16$
Stage1	160×160	$\begin{bmatrix}[Conv,3\times3,48]\times6\\ Conv,1\times1,64\\ Conv,1\times1,128\end{bmatrix}\times6$	$\begin{bmatrix}[Conv,3\times3,64]\times6\\ Conv,1\times1,64\\ Conv,1\times1,128\end{bmatrix}\times6$	$\begin{bmatrix}[LightConv,3\times3,16]\times6\\ Conv,1\times1,32\\ Conv,1\times1,64\end{bmatrix}\times3$
Stage2	80×80	$DWConv, 3\times3, 128$,stride 2 $\begin{bmatrix}[Conv,3\times3,96]\times6\\ Conv,1\times1,256\\ Conv,1\times1,512\end{bmatrix}\times6$	$DWConv, 3\times3, 128$,stride 2 $\begin{bmatrix}[Conv,3\times3,128]\times6\\ Conv,1\times1,256\\ Conv,1\times1,512\end{bmatrix}\times6\times2$	$DWConv, 3\times3, 64$,stride 2 $\begin{bmatrix}[LightConv,3\times3,32]\times6\\ Conv,1\times1,128\\ Conv,1\times1,256\end{bmatrix}\times3$
Stage3	40×40	$DWConv, 3\times3, 512$,stride 2 $\begin{bmatrix}\begin{bmatrix}[LightConv,5\times5,192]\times6\\ Conv,1\times1,512\\ Conv,1\times1,1024\end{bmatrix}\times6\end{bmatrix}\times3$	$DWConv, 3\times3, 512$,stride 2 $\begin{bmatrix}\begin{bmatrix}[LightConv,5\times5,256]\times6\\ Conv,1\times1,512\\ Conv,1\times1,1024\end{bmatrix}\times6\end{bmatrix}\times5$	$DWConv, 3\times3, 256$,stride 2 $\begin{bmatrix}\begin{bmatrix}[LightConv,5\times5,64]\times6\\ Conv,1\times1,256\\ Conv,1\times1,512\end{bmatrix}\times3\end{bmatrix}\times2$
Stage4	20×20	$DWConv, 3\times3, 1024$,stride 2 $\begin{bmatrix}[LightConv,5\times5,384]\times6\\ Conv,1\times1,1024\\ Conv,1\times1,2048\end{bmatrix}\times6$	$DWConv, 3\times3, 1024$,stride 2 $\begin{bmatrix}[LightConv,5\times5,512]\times6\\ Conv,1\times1,1024\\ Conv,1\times1,2048\end{bmatrix}\times6\times2$	$DWConv, 3\times3, 512$,stride 2 $\begin{bmatrix}[LightConv,5\times5,128]\times6\\ Conv,1\times1,512\\ Conv,1\times1,1024\end{bmatrix}\times3$
Number of layers		316	510	**170**
Params(M)		13.6	33.3	**1.9**
GFLOPs		44.6	115.1	**8.2**

with (a = 2, b = 4, c = 6, d = 5), resulting in a network depth of 510 layers and a larger channel capacity. In contrast, HGNetv2-L uses (2, 3, 6, 3), with a depth of 316 layers and a more lightweight design. Both architectures rely on stacking convolutional modules in later stages to enhance representational capacity. However, infrared images are typically grayscale and lack the rich color information present in visible-spectrum images. When processed as three-channel inputs, the channels remain identical, offering no additional color variation. Infrared maritime images also tend to exhibit sparse features and limited texture details. Overly deep networks often fail to capture meaningful infrared features. They tend to amplify noise and increase the risk of overfitting. HGNetv2-X and HGNetv2-L include higher channel widths and repeated modules. These designs cause computational and structural redundancy in infrared image processing.

To address this, we propose a lightweight redesign of HGNetv2-L by progressively reducing the number of channels and block repetitions. Specifically, we adjust the parameters a, b, c, and d to explore multiple architectural variants. Extensive experiments on multiple infrared maritime datasets and hyperparameter tuning identified an optimal configuration. We propose HGNetv2-IRLight, a lightweight model designed specifically for infrared object detection. The structure diagram of HGNetv2-IRLight is illustrated in Fig. 1, with detailed architecture provided in Table I. We reduced channel dimensions and module depth in the network. Standard convolution blocks were replaced with lightweight convolution (LightConv) blocks. This further compresses the network and lowers complexity. Compared with HGNetv2-L, the proposed HGNetv2-IRLight reduces network depth, parameter count, and GFLOPs by 46.2%, 86%, and 81.61%, respectively, significantly lowering computational complexity. When adopted as the backbone of RT-DETR and evaluated on an infrared maritime dataset, the detection performance is reported in Table IV. HGNetv2-IRLight serves as the backbone of GLAF-DETR. It reduces model size and computational cost. The network maintains efficient feature extraction and inference speed for infrared maritime object detection.

TABLE II
GENERALIZED ARCHITECTURE TEMPLATE OF HGNETV2 SERIES WITH PARAMETER DEFINITIONS

Stage	HGNetv2
Stem	$Conv, 3\times3, 16\times\mathbf{a}$, stride 2 $Conv, 2\times2, 8\times\mathbf{a}$ $Conv, 2\times2, 16\times\mathbf{a}$ $Conv, 3\times3, 16\times\mathbf{a}$, stride 2 $Conv, 1\times1, 16\times\mathbf{b}$
Stage1	$\begin{bmatrix}[Conv,3\times3,16\times\mathbf{b}]\times6\\ Conv,1\times1,32\times\mathbf{a}\\ Conv,1\times1,64\times\mathbf{a}\end{bmatrix}\times\mathbf{c}$
Stage2	$DWConv, 3\times3, 64\times\mathbf{a}$, stride 2 $\begin{bmatrix}[Conv,3\times3,32\times\mathbf{b}]\times6\\ Conv,1\times1,128\times\mathbf{a}\\ Conv,1\times1,256\times\mathbf{a}\end{bmatrix}\times\mathbf{c}$
Stage3	$DWConv, 3\times3, 256\times\mathbf{a}$, stride 2 $\begin{bmatrix}\begin{bmatrix}[LightConv,5\times5,64\times\mathbf{b}]\times6\\ Conv,1\times1,256\times\mathbf{a}\\ Conv,1\times1,512\times\mathbf{a}\end{bmatrix}\times\mathbf{c}\end{bmatrix}\times\mathbf{d}$
Stage4	$DWConv, 3\times3, 512\times\mathbf{a}$, stride 2 $\begin{bmatrix}[LightConv,5\times5,128\times\mathbf{b}]\times6\\ Conv,1\times1,512\times\mathbf{a}\\ Conv,1\times1,1024\times\mathbf{a}\end{bmatrix}\times\mathbf{c}$

III. EXPERIMENTS

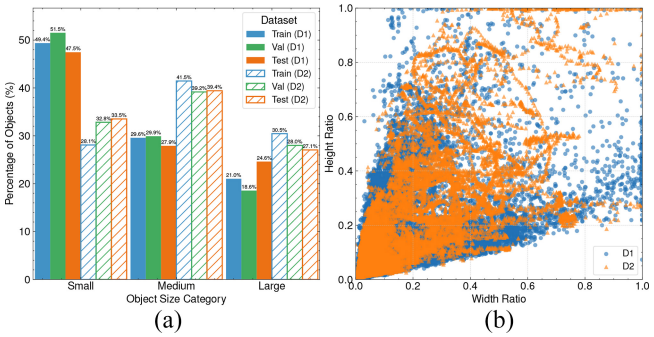
A. Setups

1) Dataset Introduction: In this study, two infrared datasets are utilized for training and evaluation: the infrared maritime dataset [22] (D1) and the multisource infrared maritime dataset (D2). Each dataset is partitioned into training, validation, and test sets with a ratio of 7:1.5:1.5, respectively.

The infrared maritime dataset is an open-source dataset available on the InfiRay Infrared Open Platform. It contains images captured at various coastal ports and docks

TABLE III
COMPARISON OF DETECTION PERFORMANCE ACROSS DIFFERENT DETECTORS ON TWO INFRARED MARITIME DATASETS (D1 AND D2)

Model	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$		$AP^{val}_{S}(\%)$		$AP^{val}_{M}(\%)$		$AP^{val}_{L}(\%)$	
				D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
Faster-RCNN-FPN	41.4	178.0	36	45.0	42.3	67.8	67.4	30.2	26.8	60.0	47.1	70.9	64.0
DETR	41.6	81.9	50	42.9	40.5	72.4	64.5	26.8	21.3	51.1	45.4	73.1	75.3
Deformable DETR	40.1	165.0	25	43.3	45.4	70.5	70.6	31.6	26.6	51.6	46.3	70.5	71.7
DAB-DETR	43.7	87.2	41	42.0	43.4	68.1	70.5	28.3	23.5	50.7	45.1	70.3	72.5
YOLOv8-L	43.7	164.9	81	54.9	47.2	76.8	73.5	40.6	29.0	64.8	54.7	74.5	76.5
YOLOv8-X	68.2	257.4	54	55.1	48.2	77.3	74.6	41.9	30.5	64.5	55.1	76.5	77.5
DINO	47.6	235.0	16	50.9	47.1	77.4	73.2	39.6	27.9	57.5	51.7	67.3	72.3
YOLOv9-C	51.0	236.8	50	54.1	50.7	77.1	74.3	39.8	31.5	65.8	54.5	77.9	77.4
YOLOv9-E	68.6	240.8	39	55.9	51.3	77.5	75.4	42.3	32.0	66.9	55.6	78.6	78.5
YOLOv10-L	25.8	126.4	87	52.8	49.3	75.6	73.1	38.3	28.7	64.4	52.8	76.2	77.3
YOLOv10-X	31.6	169.9	60	54.1	50.1	76.7	73.8	39.4	30.4	64.8	53.4	76.4	77.5
RT-DETR-L	35.1	110.5	72	56.4	49.7	83.4	73.7	44.2	31.5	65.1	52.2	76.2	77.1
RT-DETR-X	65.5	222.5	49	57.3	52.3	84.2	77.0	45.1	32.9	66.3	55.4	**78.4**	78.3
GLAF-DETR	**25.1**	**68.3**	**92**	**58.5**	**53.4**	**86.1**	**78.2**	**46.8**	**34.1**	**67.2**	**57.1**	78.1	**78.6**



(a) (b)

Fig. 5. Analysis of the Infrared Maritime Dataset (D1) and MIRMD (D2). (a) Percentage of objects in different size categories (Small, Medium, and Large) across the training, validation, and test sets. (b) Distribution of object aspect ratios. The abscissa and ordinate represent the proportion of the object width and height in the image, respectively.

using multiple infrared cameras across different times and conditions. To ensure D1 dataset reliability, we corrected labeling errors, removed problematic images, and adjusted some annotations to enhance the dataset's overall accuracy and diversity. The D1 dataset comprises 9400 infrared images with a total of 46,033 target instances. It includes 12 categories of maritime targets: liner, sailboat, speedboat, warship, bulk carrier, container ship, fishing boat, yacht, jet ski, raft, paddleboard, and buoy.

The multisource infrared maritime dataset (MIRMD) is a curated benchmark designed to address the shortage of high-quality datasets for infrared maritime object detection. D2 is constructed by selecting and integrating maritime images from three public infrared datasets: 1) LSOTB-TIR [23]; 2) MassMIND [24]; and 3) Kayak [25]. To adapt these datasets for detection, we extracted representative frames, restructured the data, and manually annotated bounding boxes for maritime targets. The dataset comprises 9500 infrared images with 38,731 annotated instances across 13 categories: speedboat, person, kayak, sailboat, passenger boat, cutter, yacht, bulk carrier, buoy, swimmer, life buoy, fisheries patrol vessel, and jet ski.

A detailed analysis of the size and aspect ratio distribution for both D1 and D2 is provided in Fig. 5. Fig. 5(a) shows the classification of Small, Medium, and Large targets based on MS COCO [26] definitions, highlighting the datasets' diversity in target sizes. Fig. 5(b) reveals a scattered distribution of target scales, indicating wide variations in aspect ratios and the presence of objects with different shapes. These datasets offer a challenging and diverse set of scenarios for infrared maritime target detection, making them well-suited to evaluate the proposed model's robustness and adaptability.

2) Experimental Environment: All experiments in this study were carried out on an operating system running Ubuntu 20.04, equipped with a 13th Gen Intel Core i9-13900K CPU and an NVIDIA GeForce RTX 3090 GPU with 24GB of memory. The algorithms were implemented using the PyTorch 1.12 framework, and all computations were executed with CUDA 11.6 and CUDNN 8.9.4 to leverage GPU acceleration. Python 3.9 was utilized for scripting and model development.

3) Implementation Details: During the training phase, the GLAF-DETR detector was optimized using the AdamW optimizer with a base learning rate of 0.0001, weight decay of 0.0001, and a global gradient clip norm of 0.1. To ensure stable training, a linear warm-up strategy was applied over 2,000 steps. Data augmentation techniques included random color adjustments, image expansion, cropping, flipping, and resizing. For the uncertainty-minimal query selection method, the initial object queries for the decoder were initialized using the first 300 encoder features.

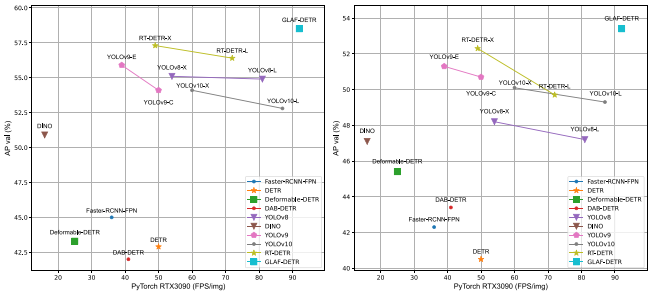
B. Comparison With the State-of-the-Art Object Detectors

We conducted comparisons with several state-of-the-art detection models to evaluate the effectiveness of GLAF-DETR. The comparison includes Faster R-CNN [27], real-time detectors, such as YOLOv8 [28], YOLOv9 [29], and YOLOv10 [30], as well as end-to-end architectures like DETR [11], Deformable DETR [31], DAB-DETR [32], and DINO [33]. All models were trained and tested under consistent experimental conditions using the same infrared maritime dataset. For consistent evaluation, the standard COCO average precision (AP) metric was used as the benchmark.

1) Comparison of Experimental Results Data: The test results for GLAF-DETR and other algorithms on the two infrared maritime dataset are presented in Table III. The

TABLE IV
COMPARISON OF DETECTION PERFORMANCE FOR DIFFERENT MODULE COMBINATIONS

Model	GLAF	DAMFF	HGNetv2-IRLight	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$		$AP^{val}_{S}(\%)$		$AP^{val}_{M}(\%)$		$AP^{val}_{L}(\%)$	
							D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
RT-DETR-L				35.1	110.5	72	56.4	49.7	83.4	73.7	44.2	31.5	65.1	52.2	76.2	77.1
Model a1			√	19.9	62.7	110	55.3	48.4	82.3	72.3	43.4	30.5	63.8	51.3	73.9	74.6
Model a2	√			38.3	111.7	60	58.0	52.6	85.2	76.7	46.7	32.6	66.3	54.7	77.7	77.6
Model a3	√		√	20.7	62.9	102	57.7	52.3	85.1	76.3	46.2	32.8	64.6	54.3	78.3	76.6
Model a4		√		38.2	115.8	58	57.6	53.0	85.1	76.8	46.0	33.1	66.1	55.3	77.5	78.8
Model a5		√	√	24.1	68.0	99	56.9	52.7	84.8	76.5	45.4	32.9	65.8	55.0	77.2	77.5
Model a6	√	√		42.5	117.1	53	59.1	54.2	86.3	78.8	48.1	34.3	67.3	58.3	78.6	79.3
GLAF-DETR	√	√	√	25.1	68.3	92	58.5	53.4	86.1	78.2	46.8	34.1	67.2	57.1	78.1	78.6



(a) (b)

Fig. 6. Comparison of AP^{val} and FPS for object detection models on two infrared maritime datasets. (a) Results on infrared maritime dataset (D1). (b) Results on MIRMD dataset (D2).

frames per second (FPS) were calculated with a batch size of 1. On both datasets, GLAF-DETR demonstrates superior performance in terms of both accuracy and efficiency. Specifically, it achieves 86.1% AP^{val}_{50} and 92 FPS on D1, outperforming other state-of-the-art detectors. GLAF-DETR performs well across objects of all sizes. It also leads on the multi-source dataset D2. This highlights its strong generalization and robustness in various infrared maritime environments.

Compared with real-time detectors YOLOv9-E and YOLOv10-L, GLAF-DETR improves AP^{val}_{50} on D1 by 8.6% and 10.5%, and on D2 by 2.8% and 5.1%, respectively. In terms of inference speed, it increases FPS by 135.9% and 5.7% compared to the two models. This highlights the enhanced performance and real-time capabilities of GLAF-DETR for infrared maritime object detection over the YOLO series.

When compared with RT-DETR-L, GLAF-DETR improves AP^{val}_{50} by 2.7% on D1 and 4.5% on D2, while achieving a 27.8% increase in FPS. In addition, for object detection at different scales, GLAF-DETR outperforms RT-DETR-L on D1 with gains of 2.6%, 2.1%, and 1.9% in AP^{val}_{S}, AP^{val}_{M}, and AP^{val}_{L}, respectively. On D2, corresponding improvements of 2.6%, 4.9%, and 1.5% are observed. These results confirm the effectiveness of GLAF-DETR in detecting infrared maritime targets of various sizes under diverse conditions.

As shown in Fig. 6, higher FPS and AP^{val} values indicate better performance. GLAF-DETR achieves consistently superior results, confirming its effectiveness in balancing speed and accuracy. This demonstrates the advantages of the proposed GLAF attention mechanism, DAMFF multiscale feature fusion module, and the lightweight HGNetv2-IRLight backbone.

2) Visualization of Experimental Results: Random images from different scenes in the infrared maritime test set were selected for comparative detection, as shown in Fig. 7 and Fig. 8. In the open sea scenes [Figs. 7(a) and 8(b)], multiple targets of varying scales are present. For large and medium targets, detectors, such as YOLOv10-L often produced incorrect predictions or exhibited low classification confidence. Most algorithms also failed to detect small and medium targets effectively. In contrast, GLAF-DETR achieved higher accuracy for large and medium objects and successfully detected more small maritime targets. In the near-coastal scenes [Figs. 7(b) and 8(a)], overlapping ships of varying sizes and the close proximity of humans and vessels pose considerable challenges for object detectors. GLAF-DETR achieves higher classification confidence and more precise localization than other methods, demonstrating its robustness in fine-grained infrared maritime detection. In the complex dock scene [Figs. 7(c) and 8(c)], GLAF-DETR performed better at detecting overlapping ships within the dock and successfully identified many small targets in the sea outside the dock.

The heatmaps generated using Grad-CAM provide a visual comparison of the performance of different detectors in infrared maritime object detection, as illustrated in Fig. 9. GLAF-DETR demonstrates superior feature capture capabilities compared to other detectors. For large and medium-sized target areas, GLAF-DETR's heatmap displays deeper, more concentrated colors, indicating its ability to effectively suppress irrelevant background information and accurately focus on target regions. Additionally, in regions containing distant small and medium targets, GLAF-DETR's heatmap shows stronger emphasis, further demonstrating its excellent performance in detecting small and medium-sized targets.

C. Ablation Study of Component Integration

A comprehensive ablation study was conducted to assess the contribution of each design component in GLAF-DETR. The detailed results are summarized in Table IV. Beginning with the baseline RT-DETR-L, GLAF, DAMFF, and HGNetv2-IRLight were progressively integrated. This stepwise analysis enabled an in-depth evaluation of each component's impact on detection performance, computational complexity, and inference speed.

In Model a1, the backbone of RT-DETR-L (HGNetv2-L) was replaced with the lightweight HGNetv2-IRLight, yielding the RT-DETR-IRLight model. This reduced parameters and GFLOPs by 43.3%, and improved FPS by 52.8%. Although AP^{val}_{50} dropped slightly by 1.1% on D1 and 1.4% on D2, HGNetv2-IRLight retained sufficient feature extraction capacity while enhancing efficiency and real-time performance. In Models a2 and a3, incorporating the GLAF attention

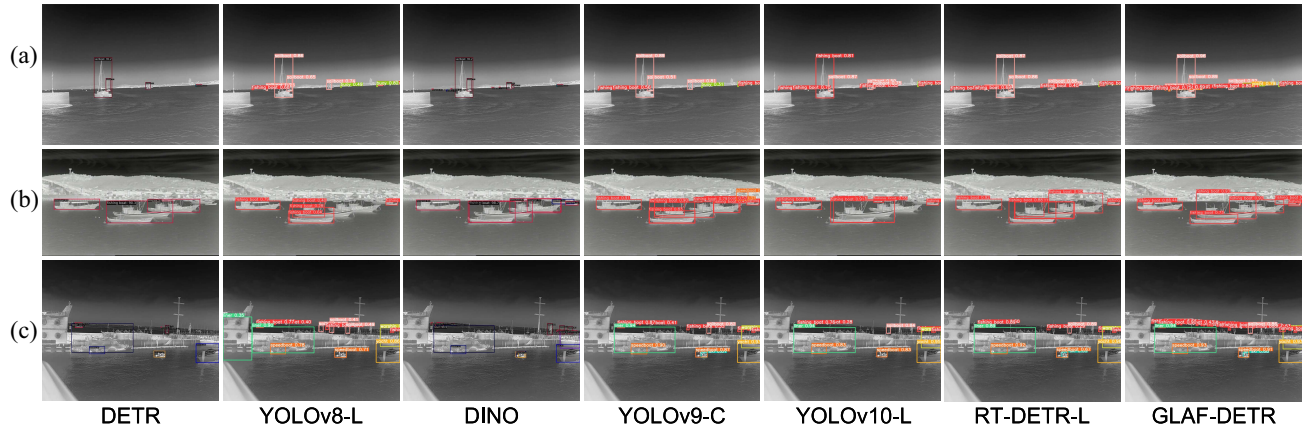


Fig. 7. Detection results of different detectors on infrared maritime dataset (D1). (a) Open-sea scene. (b) Near-coastal scene. (c) Dock scene.

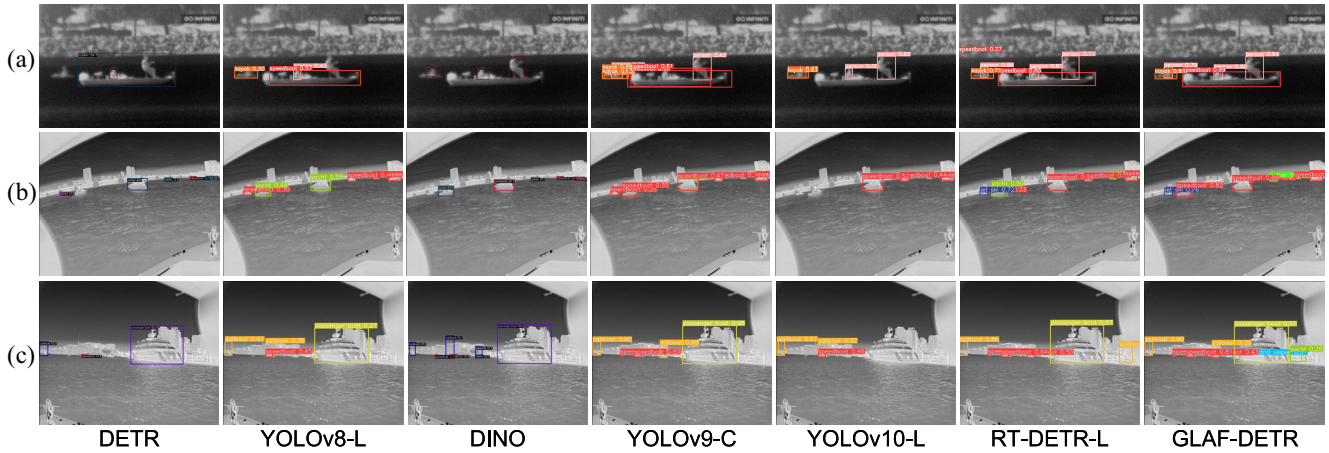


Fig. 8. Detection results of different detectors on MIRMD dataset (D2). (a) Near-coastal scene. (b) Open-sea scene. (c) Dock scene.

mechanism into RT-DETR-L and RT-DETR-IRLight led to AP_{50}^{val} gains of 1.8% and 2.8% on D1, and 3.0% and 4.9% on D2, respectively. Both models achieved higher accuracy across small, medium, and large targets, highlighting GLAF's effectiveness in multiscale detection. While GFLOPs increased marginally by 1.1% and 0.3%, FPS only decreased by 16.7% and 7.3%, demonstrating that GLAF adds minimal overhead while maintaining real-time performance. Models a4 and a5 integrated the DAMFF module, achieving AP_{50}^{val} gains of 1.7% and 2.5% on D1, and 3.1% and 4.2% on D2. Detection performance across scales also improved, validating the DAMFF's contribution to multiscale feature fusion. Model a6 combined both GLAF and DAMFF with the original backbone, achieving the highest AP_{50}^{val} of 86.3% on D1 and 78.2% on D2. However, the absence of a lightweight backbone significantly increased computational cost, reducing FPS to 53.

Finally, GLAF-DETR, integrating all three modules (GLAF, DAMFF, and HGNetv2-IRLight), achieved an optimal trade-off between accuracy and efficiency. Compared to the baseline, it demonstrated superior multiscale detection performance while maintaining real-time capability.

D. Ablation Study on Attention Mechanism

To evaluate the GLAF attention mechanism's effectiveness, we conduct four ablation experiments. First, we test GLAF's placement in different backbone stages to find the optimal strategy. Second, we assess adaptive weighting and directional enhancement components' contributions. Third, we compare GLAF with representative attention mechanisms to show its superiority. Finally, we integrate GLAF into SOTA detection frameworks to confirm its versatility. Results show GLAF enhances detection accuracy with low computational cost, proving its robustness in complex maritime scenarios.

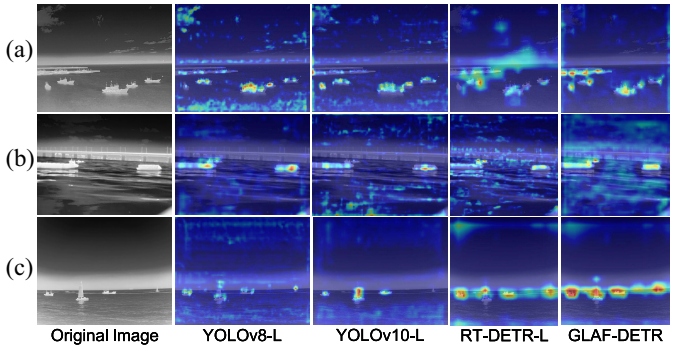


Fig. 9. Comparison of feature information heatmaps of different detectors. (a) Large/medium targets in open sea. (b) Distant small/medium targets. (c) Mixed-scale vessel scene.

1) Comparison of GLAF Attention Placement: To assess the impact of GLAF attention mechanisms at different stages

TABLE V
COMPARISON OF DETECTION PERFORMANCE FOR DIFFERENT GLAF PLACEMENT COMBINATIONS

Model	GLAF Placement	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$	
					D1	D2	D1	D2
RT-DETR-IRLight	None	19.9	62.7	110	55.3	48.4	82.3	72.3
Model s1	Stage 1	19.9	62.7	108	56.5	50.6	83.3	73.8
Model s2	Stage 2	19.9	62.7	106	57.1	51.3	85.1	75.8
Model s3	Stage 3	20.2	62.8	104	56.9	51.3	84.6	75.3
Model s4	Stage 4	20.5	62.7	105	57.3	51.7	84.4	75.9
Model s5	Stages 1 & 2	19.9	62.7	104	56.6	50.5	84.1	74.9
Model s6	Stages 1 & 3	20.2	62.8	102	56.1	49.6	83.6	74.2
Model s7	Stages 1 & 4	20.5	62.8	104	55.9	49.7	84.0	74.3
Model s8	Stages 2 & 3	20.2	62.8	103	57.0	51.2	84.8	75.9
Model s9	Stages 2 & 4	20.5	62.8	104	57.4	51.9	85.0	76.2
Model s10	Stages 3 & 4	20.8	62.9	102	57.3	51.7	84.6	75.7
Model s11	Stages 1, 2 & 3	20.2	62.9	102	55.7	51.0	84.5	75.0
Model s12	Stages 1, 2 & 4	20.5	62.8	103	57.1	51.6	84.2	75.4
Model s13	Stages 1, 3 & 4	20.8	62.9	102	56.7	51.2	84.2	75.3
Model s14	Stages 2, 3 & 4	20.7	62.9	102	**57.7**	**52.3**	**85.1**	**76.3**
Model s15	All Stages	20.8	62.9	100	56.5	51.4	84.6	75.6

of the backbone network, we conducted a series of experiments, with the results presented in Table V. The baseline RT-DETR-IRLight does not include GLAF.

Model s1, which applies GLAF at stage 1, demonstrated the lowest performance. This is because stage 1 mainly extracts shallow features, containing noise and irrelevant background. Attention applied at this level amplifies such noise, negatively affecting detection. Consequently, models including GLAF at stage 1 (e.g., s5, s6, s7, and s11) also underperformed. Model s4, applying GLAF solely at stage 4, showed suboptimal results. Although stage 4 captures abstract high-level features, neglecting intermediate stages limits the model's ability to harness multiscale context. Combinations involving stage 4 without mid-level stages (e.g., s7, s9, s12) further illustrate this limitation. Model s15, which includes GLAF at all stages, did not achieve the best results, likely due to redundant attention and overfitting from excessive feature enhancement.

In contrast, Model s14, applying GLAF at stages 2, 3, and 4, achieved the best overall performance. This configuration strikes an effective balance, enhancing both mid-level and high-level representations while avoiding shallow feature noise and attention redundancy. These results show that appropriately placed attention significantly improves detection across varying object scales in complex infrared maritime scenes.

2) Component Ablation Within GLAF: The GLAF mechanism comprises four core components: global and local information extraction, adaptive fusion, DE, and attention generation. This section focuses on evaluating the impact of adaptive fusion and DE, as shown in Table VI.

Adaptive fusion is implemented through the DWM, which adjusts fusion weights based on input features. We compared three fusion strategies: direct addition (DA), weighted addition (WA), and DWM. GLAF Variant 1 uses DA for feature fusion. It yields the weakest performance because all features are treated equally, failing to emphasize those most relevant at different scales. GLAF Variant 2 uses WA, with fixed weights learned during training. While it improves over DA, its static nature makes it less responsive to variations in object scale, background complexity, or thermal contrast that frequently occur in maritime infrared scenes. To evaluate the DE module, GLAF Variant 4 replaces DE with a 1×1 pointwise convolution. The 1×1 convolution cannot effectively capture directional dependencies. As a result, its performance degrades in scenarios with varied target orientations, highlighting the need for capturing directional information to improve detection.

GLAF Variant 3, integrating both DWM and DE, achieves the best results across all metrics. DWM contributes by adaptively selecting and weighting features carrying the most relevant spatial context, while DE strengthens the representation of direction-sensitive features. Together, they enable the model to better handle the complex, multiscale, and multi-orientation nature of maritime scenes. Overall, GLAF adaptively adjusts fusion weights and focuses on relevant regions. It enhances multi-directional feature extraction. This improves detection for targets of different sizes in complex maritime scenarios.

3) Comparison of Different Attention Mechanisms: Comparative experiments were conducted to evaluate the effectiveness of GLAF against several classic attention mechanisms. The results are presented in Table VII, using RT-DETR-IRLight as the baseline model.

The experiments showed that the SE and ECA modules slightly improved AP. However, their focus on global information extraction, without adequately modeling spatial structures, limited their ability to capture fine details of infrared maritime objects. In contrast, the CBAM and CA modules introduced spatial information, which enhanced detection performance. This highlights the importance of spatial feature fusion for improving the representation of infrared targets. Nevertheless, the inclusion of CBAM significantly increased the model's parameters, reducing overall efficiency. The SPA module further improved detection by integrating local and global cues, achieving a 1.6% and 2.7% AP^{val} gain on D1 and D2, respectively. Yet, its limited spatial modeling restricted further enhancement.

In comparison, GLAF consistently outperformed all compared attention mechanisms, increasing AP^{val}_{50} by 2.8% on D1 and 4.0% on D2. These results highlight GLAF's ability to adaptively fuse global and local features and dynamically focus on informative regions. By balancing long-range semantics with local structure, GLAF significantly enhances both accuracy and robustness in infrared maritime detection.

Fig. 10 shows detection results for different attention mechanisms, providing a direct visual comparison of their performance. In Fig. 10(a) and (c), which display wide maritime environments, GLAF detects more small and medium-sized targets than other mechanisms. In the complex dock scene depicted in Fig. 10(b), GLAF demonstrates more accurate target localization, with bounding boxes more closely matching the objects. In general, GLAF reduces false positives and missed detections in infrared maritime object detection.

Fig. 11 compares feature heatmaps across attention mechanisms, further highlighting their focus areas. In Fig. 11(a), GLAF excels at distinguishing closely positioned ships, enhancing the feature representation for each target. In Fig. 11(b), GLAF shows a stronger focus on small and medium-sized targets, confirming its superior performance in detecting such objects compared to other mechanisms.

4) Integration of GLAF Into Other Architectures: The results further validate the adaptability and generalization of the GLAF attention mechanism across various detection

TABLE VI
COMPARISON OF DIFFERENT GLAF COMPONENT VARIANTS. IN THE TABLE, DA REPRESENTS DA, WA REPRESENTS WA, DWM REPRESENTS DWM, AND DE REPRESENTS DE

Model	DA	WA	DWM	DE	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$		$AP^{val}_{S}(\%)$		$AP^{val}_{M}(\%)$		$AP^{val}_{L}(\%)$	
								D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
RT-DETR-IRLight					19.9	62.7	110	55.3	48.4	82.3	72.3	43.4	30.5	63.8	51.3	73.9	74.6
+GLAF Variant 1	√			√	20.7	62.7	105	56.2	50.2	83.8	74.4	44.1	31.3	64.3	52.8	76.5	75.8
+GLAF Variant 2		√		√	20.7	62.7	104	56.9	51.1	84.1	75.2	45.2	32.1	63.6	53.0	76.9	76.4
+GLAF Variant 3			√	√	20.7	62.9	102	**57.7**	**52.3**	**85.1**	**76.3**	**46.2**	**32.8**	**64.6**	**54.3**	**78.3**	**76.6**
+GLAF Variant 4			√		20.7	62.9	103	57.1	51.0	84.1	75.4	45.3	32.0	63.0	53.2	77.4	75.5

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT ATTENTION MECHANISMS FOR INFRARED MARITIME OBJECT DETECTION

Model	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$	
				D1	D2	D1	D2
RT-DETR-IRLight	19.9	62.7	110	55.3	48.4	82.3	72.3
+SE	20.1	62.7	107	55.9(+0.6)	50.2(+1.8)	82.8(+0.5)	73.9(+1.6)
+ECA	19.9	62.7	106	56.3(+1.0)	49.9(+1.5)	83.1(+0.8)	74.2(+1.9)
+CBAM	27.6	62.9	99	56.6(+1.3)	51.4(+3.0)	83.5(+1.2)	75.1(+2.8)
+CA	20.2	62.7	104	56.3(+1.0)	50.8(+2.4)	83.6(+1.3)	74.8(+2.5)
+SPA	22.1	62.7	105	56.9(+1.6)	51.1(+2.7)	84.1(+1.8)	74.9(+2.6)
+GLAF	20.7	62.9	102	**57.7(+2.4)**	**52.3(+3.9)**	**85.1(+2.8)**	**76.3(+4.0)**

architectures. GLAF was integrated into the backbone network blocks of each detector, consistently enhancing performance, as shown in Table VIII.

GLAF improved overall AP and detection precision for small, medium, and large objects, demonstrating its effectiveness for multiscale target detection. For instance, in the YOLOv10-L model, it increased AP^{val}_{S}, AP^{val}_{M}, and AP^{val}_{L} by 3.3%, 1.1%, and 2.3% on D1, respectively. Another key advantage of GLAF is its lightweight design. Despite minimal increases in parameters and GFLOPS, it achieved significant accuracy improvements. In the YOLOv8-L model, GLAF added only 0.3M parameters and 0.4 GFLOPs, yet improved AP^{val} by 1.8% and 3.1% on D1 and D2, respectively, demonstrating a balance between efficiency and accuracy. This balance between performance gains and computational overhead makes GLAF highly suitable for real-time infrared maritime object detection in complex environments.

E. Ablation Study on Multiscale Feature Fusion Module

This section evaluates the effectiveness of the proposed DAMFF module. First, we compare different attention mechanisms within the fusion module, including a baseline without attention, to verify the benefits of attention integration. Second, we compare DAMFF with other fusion methods. Results show DAMFF improves multiscale feature fusion and detection accuracy in complex infrared maritime scenes.

1) Comparative Study of Attention Mechanisms Within DAMFF: In DAMFF, the GLAF attention mechanism guides dynamic adaptive fusion by generating attention weights for each scale feature. To validate its effectiveness, we conducted experiments where the attention guidance was either removed or replaced with alternative mechanisms, as shown in Table IX.

The baseline model, RT-DETR-IRLight, combined with DAMFF (No Attention), achieved a slight AP increase. DAMFF (No Attention) uses simple multiscale concatenation without attention guidance. It yields minor improvements, confirming the benefit of multiscale fusion but revealing limited performance without adaptive weighting.

SE and ECA provided moderate improvements, confirming the validity of attention guidance. However, their inability to model local and spatial structure information constrained further enhancements. CBAM improved spatial structure information, and SPA introduced local feature representation, both achieving notable performance gains. GLAF outperformed all other methods by effectively combining global and local feature fusion while enhancing spatial structure understanding. It dynamically adjusts attention weights, resulting in more accurate feature representations. These findings confirm the superiority of GLAF in processing multiscale features, particularly when integrated into DAMFF. DAMFF ensures the balanced allocation of attention weights across different scales, leading to superior detection accuracy, especially in complex maritime scenarios.

2) Comparative Analysis of DAMFF With Other Multiscale Feature Fusion Methods: To evaluate its performance compared to traditional multiscale feature fusion methods, we conducted a comparative analysis with BiFPN and ASFF, as shown in Table X.

BiFPN shows moderate performance improvements, but its gains are limited. This is primarily due to its fixed weights for feature fusion, which are determined during training and remain unchanged. Consequently, BiFPN lacks adaptability to varying feature importance across scales, restricting its effectiveness. ASFF outperforms BiFPN by dynamically adjusting the fusion weights across scales, thereby enhancing detection accuracy. However, the increased network complexity of ASFF leads to significantly higher computational costs. Moreover, ASFF directly employs multiscale features to compute weights, which can include irrelevant background information, potentially causing inefficient weight distribution.

In contrast, DAMFF achieves an optimal balance between performance and computational cost, resulting in the highest AP improvement. DAMFF uses the GLAF attention mechanism to guide adaptive weighting, focusing on essential regions while suppressing background noise. Furthermore, DAMFF dynamically redistributes attention weights across scales, facilitating effective global information exchange between channels. This mechanism allows DAMFF to emphasize the most relevant features of multiscale targets, surpassing other methods and proving its effectiveness for multiscale object detection in complex environments.

F. Ablation Study on Lightweight Architecture Parameters

To optimize HGNetv2 for infrared maritime target detection, we conduct an ablation study on various $[a, b, c, d]$ configurations, as defined in Table I. The goal was to analyze the

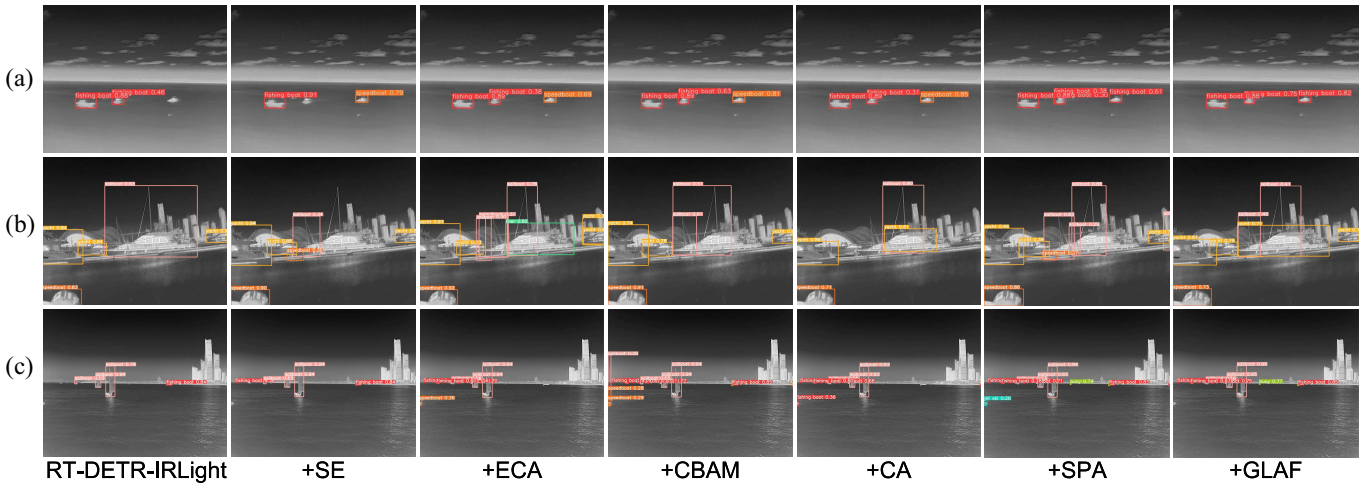


Fig. 10. Detection results of different attention mechanisms on infrared maritime targets. (a) Wide maritime scene with scattered vessels. (b) Dock scene with close ships. (c) Wide scene with small/medium vessels.

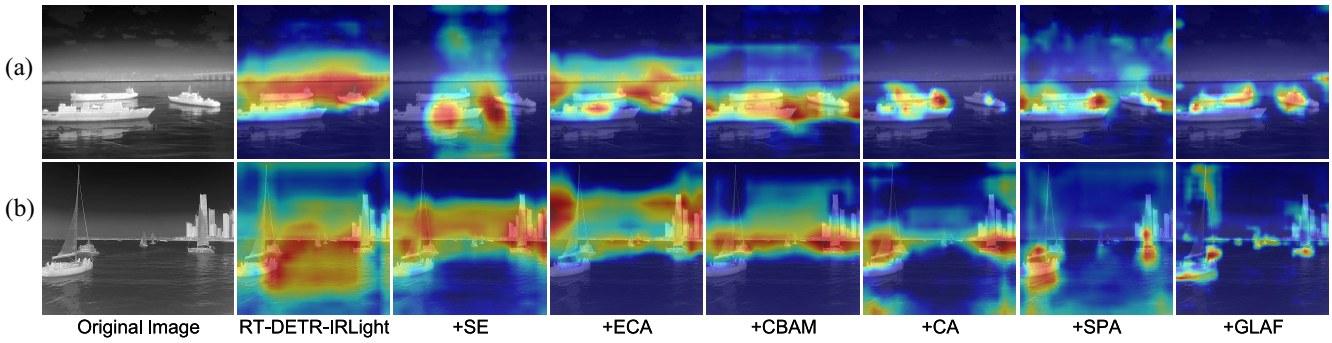


Fig. 11. Comparison of feature information heatmaps of different attention mechanisms. (a) Large/medium targets in open sea. (b) Mixed-scale vessel scene.

TABLE VIII
COMPARATIVE ANALYSIS OF GLAF INTEGRATION IN DIFFERENT DETECTORS

Model	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$		$AP^{val}_{S}(\%)$		$AP^{val}_{M}(\%)$		$AP^{val}_{L}(\%)$	
				D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
RT-DETR-IRLight	19.9	62.7	110	55.3	48.4	82.3	72.3	43.4	30.5	63.8	51.3	73.9	74.6
+GLAF	20.7	62.9	102	57.7(+2.4)	52.3(+3.9)	85.1(+2.8)	76.3(+4.0)	46.2(+2.8)	32.8(+2.3)	64.6(+0.8)	54.3(+3.0)	78.3(+4.4)	76.6(+2.0)
YOLOv8-L	43.7	164.9	81	54.9	47.2	76.8	73.5	40.6	29.0	64.8	54.7	74.5	76.5
+GLAF	44.0	165.3	69	56.7(+1.8)	50.3(+3.1)	78.4(+1.6)	76.6(+3.1)	43.2(+2.6)	31.0(+2.0)	65.7(+0.9)	57.2(+2.5)	78.0(+4.5)	78.1(+1.6)
YOLOv8-X	68.2	257.4	54	55.1	48.2	77.3	74.6	41.9	30.5	64.5	55.1	76.5	77.5
+GLAF	68.8	258.0	46	56.8(+1.7)	50.8(+2.6)	78.7(+1.4)	77.3(+2.7)	43.9(+2.0)	31.8(+1.3)	65.2(+0.7)	57.2(+2.1)	78.9(+2.4)	78.5(+1.0)
YOLOv10-L	25.8	126.4	87	52.8	49.3	75.6	73.1	38.3	28.7	64.4	52.8	76.2	77.3
+GLAF	26.2	126.6	74	55.0(+2.2)	53.0(+3.7)	78.1(+2.5)	77.2(+4.1)	41.6(+3.3)	31.2(+2.5)	65.5(+1.1)	55.4(+2.6)	78.5(+2.3)	78.7(+1.4)
YOLOv10-X	31.6	169.9	60	54.1	50.1	76.7	73.8	39.4	30.4	64.8	53.4	76.4	77.5
+GLAF	32.2	170.2	53	55.4(+1.3)	53.1(+3.0)	78.5(+1.8)	77.0(+3.2)	41.3(+1.9)	32.2(+1.8)	65.4(+0.6)	55.7(+2.3)	79.6(+3.2)	78.7(+1.2)
RT-DETR-L	35.1	110.5	72	56.4	49.7	83.4	73.7	44.2	31.5	65.1	52.2	76.2	77.1
+GLAF	38.3	111.7	60	58.0(+1.6)	52.6(+2.9)	85.2(+1.8)	76.7(+3.0)	46.3(+2.1)	32.9(+1.4)	65.7(+0.6)	54.7(+2.5)	79.3(+3.1)	77.6(+0.5)
RT-DETR-X	65.5	222.5	49	57.3	52.3	84.2	77.0	45.1	32.9	66.3	55.4	78.4	78.3
+GLAF	73.5	224.7	42	58.6(+1.3)	54.4(+2.1)	85.6(+1.4)	79.4(+2.4)	46.7(+1.6)	34.0(+1.1)	67.6(+1.3)	57.5(+2.1)	80.6(+2.2)	79.4(+1.1)

TABLE IX
PERFORMANCE COMPARISON OF ATTENTION MECHANISMS WITHIN DAMFF

Model	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$		$AP^{val}_{S}(\%)$		$AP^{val}_{M}(\%)$		$AP^{val}_{L}(\%)$	
				D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
RT-DETR-IRLight	19.9	62.7	110	55.3	48.4	82.3	72.3	43.4	30.5	63.8	51.3	73.9	74.6
+DAMFF (**No Attention**)	24.0	68.0	106	55.5(+0.2)	49.6 (+1.2)	83.3(+1.0)	73.9(+1.6)	44.1(+0.7)	31.4(+0.9)	63.3(-0.5)	53.0(+1.7)	75.7(+1.8)	76.7(+2.1)
+DAMFF (**SE**)	24.0	68.0	104	56.0(+0.7)	50.3(+1.9)	83.8(+1.5)	74.3(+2.0)	45.1(+1.7)	31.8(+1.3)	62.3(-1.5)	53.5(+2.2)	74.3(+0.4)	77.1(+2.5)
+DAMFF (**ECA**)	24.0	68.0	102	55.9(+0.6)	50.6(+2.2)	83.9(+1.6)	74.4(+2.1)	44.6(+1.2)	32.0(+1.5)	64.1(+0.3)	53.4(+2.1)	75.4(+1.5)	75.9(+1.3)
+DAMFF (**CBAM**)	36.1	108.8	74	56.2(+0.9)	51.9(+3.5)	84.1(+1.8)	75.5(+3.2)	45.3(+1.9)	32.4(+1.9)	64.4(+0.6)	54.4(+3.1)	76.2(+2.3)	77.9(+3.3)
+DAMFF (**SPA**)	24.1	68.0	101	56.2(+0.9)	51.5(+3.1)	84.2(+1.9)	75.2(+2.9)	45.6(+2.2)	32.6(+2.1)	64.3(+0.5)	54.1(+2.8)	76.8(+2.9)	76.6(+2.0)
+DAMFF (**GLAF**)	24.1	68.0	99	56.9(+1.6)	52.7(+4.3)	84.8(+2.5)	76.5(+4.2)	45.4(+2.0)	32.9(+2.4)	65.8(+2.0)	55.0(+3.7)	77.2(+3.3)	77.5(+2.9)

relationship between network redundancy and the trade-off between feature representation and efficiency.

We first evaluated the effect of reducing channel widths [a,b] while maintaining fixed block depth. As shown in Table XI, RT-DETR-L1 to L3 progressively narrow the channels from [2, 2] to [1, 1], resulting in notable reductions in parameters and GFLOPs. Despite these efficiency gains, AP drops only slightly, suggesting that wider channels introduce computational redundancy with limited benefit for infrared feature extraction. These results demonstrate that reducing channel

TABLE X
PERFORMANCE COMPARISON OF MULTISCALE FEATURE FUSION METHODS

Model	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$	
				D1	D2	D1	D2
RT-DETR-IRLight	19.9	62.7	110	55.3	48.4	82.3	72.3
+BiFPN	20.6	62.7	106	56.2(+0.9)	51.3(+2.9)	83.1(+0.8)	75.4(+3.1)
+ASFF	23.7	75.9	97	56.7(+1.4)	51.8(+3.4)	83.8(+1.5)	75.9(+3.6)
+DAMFF	24.1	68.0	99	**56.9(+1.6)**	**52.7(+4.3)**	**84.8(+2.5)**	**76.5(+4.2)**

TABLE XI
ABLATION STUDY OF DIFFERENT HGNETV2 LIGHTWEIGHT CONFIGURATIONS ON TWO INFRARED MARITIME DATASETS (D1 AND D2)

Model	[a,b,c,d]	Params(M)	GFLOPs	$FPS_{bs=1}$	$AP^{val}(\%)$		$AP^{val}_{50}(\%)$	
					D1	D2	D1	D2
RT-DETR-X	[2,4,6,5]	65.5	222.5	49	57.3	52.3	84.2	77.0
RT-DETR-L	[2,3,6,3]	35.1	110.5	72	56.4	49.7	83.4	73.7
RT-DETR-L1	[2,2,6,3]	29.2	90.9	78	56.1	48.9	82.8	73.2
RT-DETR-L2	[1,2,6,3]	22.3	72.7	94	55.4	48.1	82.4	72.0
RT-DETR-L3	[1,1,6,3]	20.7	65.6	98	55.6	48.4	82.5	72.5
RT-DETR-L4	[2,3,6,1]	27.9	90.4	80	55.8	48.7	83.0	72.9
RT-DETR-L5	[2,3,3,3]	28.7	90.0	79	56.0	48.8	82.6	73.1
RT-DETR-L6	[2,3,3,2]	27.0	85.1	85	55.9	48.6	82.6	73.0
RT-DETR-L7	[2,3,3,1]	25.4	79.9	90	55.7	48.6	82.4	72.7
RT-DETR-L8	[2,3,1,1]	23.8	73.0	93	54.3	47.7	81.8	71.5
RT-DETR-L9	[1,1,3,2]	19.9	62.7	110	55.3	48.4	82.3	72.3
RT-DETR-L10	[1,1,3,1]	19.5	61.5	113	54.2	47.4	81.5	71.2

widths can significantly improve efficiency with minimal performance loss. Next, we fix the channel widths and reduce the number of repeated blocks [c,d] to control network depth. RT-DETR-L4 to L8 show that decreasing block depth further reduces computational cost while maintaining reasonable performance. However, overly shallow structures lead to noticeable accuracy degradation, highlighting the need for a balanced depth to avoid underfitting.

Based on these findings, we propose two joint configurations, RT-DETR-L9 and RT-DETR-L10, combining reduced channels with moderate depth. L9 outperforms L10 by 1.1% AP on D1 and 1.0% on D2, demonstrating that a slightly deeper decoder ($d = 2$) enhances feature representation. Compared to the baseline RT-DETR-L, L9 reduces parameters and GFLOPs by 86% and 81.6%, respectively, with only a 0.7% AP drop on D1. Moreover, it achieves the fastest inference speed of 110 FPS. These results confirm that the proposed HGNetv2-IRLight (L9) achieves an optimal balance between accuracy and efficiency, making it a highly effective backbone for real-time infrared maritime detection within the GLAF-DETR framework.

IV. CONCLUSION

This article presents a novel infrared maritime object detection method, GLAF-DETR, which integrates the GLAF attention mechanism. GLAF effectively fuses global and local features, dynamically adjusting the fusion of global and local information. The experimental results demonstrate that GLAF significantly enhances detection accuracy, particularly for small and medium-sized objects, outperforming conventional attention mechanisms. The proposed DAMFF module further strengthens detection performance by dynamically weighting multiscale features guided by GLAF, improving detection across a wide range of object sizes. Additionally, the introduction of the lightweight HGNetv2-IRLight backbone ensures efficient infrared feature extraction with minimal computational overhead. Extensive comparisons with state-of-the-art methods verify that GLAF-DETR offers superior detection accuracy and real-time performance, especially in handling multiscale objects in complex maritime environments. This research provides a robust solution for infrared object detection in challenging maritime scenarios. Future work will explore deploying this method in maritime IoT systems, further advancing the intelligence of marine systems.

REFERENCES

[1] Y. Duan, Z. Li, X. Tao, Q. Li, S. Hu, and J. Lu, "EEG-based maritime object detection for IoT-driven surveillance systems in smart ocean," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9678–9687, Oct. 2020.

[2] S. Jung, S. Jeong, J. Kang, and J. Kang, "Marine IoT systems with space–air–sea integrated networks: Hybrid LEO and UAV edge computing," *IEEE Internet Things J.*, vol. 10, no. 23, pp. 20498–20510, Dec. 2023.

[3] B. Wang, E. Benli, Y. Motai, L. Dong, and W. Xu, "Robust detection of infrared maritime targets for autonomous navigation," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 635–648, Dec. 2020.

[4] Z. Gao, Y. Zhang, and S. Wang, "Lightweight small ship detection algorithm combined with infrared characteristic analysis for autonomous navigation," *J. Mar. Sci. Eng.*, vol. 11, no. 6, p. 1114, 2023.

[5] D. Ma, L. Dong, R. Gao, and W. Xu, "Recent advancements in long-distance marine infrared target detection: Latest method and future perspectives," *Infr. Phys. Technol.*, vol. 133, Sep. 2023, Art. no. 104729.

[6] R. Miao, H. Jiang, and F. Tian, "Robust ship detection in infrared images through multiscale feature extraction and lightweight CNN," *Sensors*, vol. 22, no. 3, p. 1226, Feb. 2022.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.

[8] J. Ye, Z. Yuan, C. Qian, and X. Li, "CAA-YOLO: Combined-attention-augmented YOLO for infrared ocean ships detection," *Sensors*, vol. 22, no. 10, p. 3782, May. 2022.

[9] Y. Wang, B. Wang, L. Huo, and Y. Fan, "GT-YOLO: Nearshore infrared ship detection based on infrared images," *J. Mar. Sci. Eng.*, vol. 12, no. 2, p. 213, Jan. 2024.

[10] H. Deng and Y. Zhang, "FMR-YOLO: Infrared ship rotating target detection based on synthetic fog and multiscale weighted feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–17, 2024.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2020, pp. 213–229.

[12] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2024, pp. 16965–16974.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[14] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 11531–11539.

[15] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 3–19.

[16] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 13708–13717.

[17] J. Guo et al., "SpaNet: Spatial pyramid attention network for enhanced image recognition," in *Proc. IEEE Int. Conf. Multimed. Expo. (ICME)*, London, U.K., Jul. 2020, pp. 1–6.

[18] T. -Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 936–944.

[19] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 8759–8768.

[20] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 10778–10787.

[21] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.
[22] (Yantai IRay Technol. Co., Ltd., Shandong, China). *InfiRay Infrared Open-Source Platform, Infrared Ship Target Detection Database*. Oct. 2024. [Online]. Available: http://openai.iraytek.com/apply/Sea_shipping.html/
[23] Q. Liu, X. Li, D. Yuan, C. Yang, X. Chang, and Z. He, "LSOTB-TIR: A large-scale high-diversity thermal infrared single object tracking benchmark," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9844–9857, Jul. 2024.
[24] S. Nirgudkar, M. DeFilippo, M. Sacarny, M. Benjamin, and P. Robinette, "Massmind: Massachusetts maritime infrared dataset," *Int. J. Robot. Res.*, vol. 42, no. 1, pp. 21–32, Mar. 2023.
[25] A. Toet, "Detection of dim point targets in cluttered maritime backgrounds through multisensor image fusion," in *Proc. SPIE*, 2002, pp. 118–129.
[26] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 740–755.
[27] 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.
[28] G. Jocher, A. Chaurasia, and J. Qiu. "Ultralytics YOLOv8." Oct. 2024. [Online]. Available: https://github.com/ultralytics/ultralytics
[29] 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*.
[30] A. Wang et al., "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.
[31] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–16.
[32] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–20.
[33] H. Zhang et al., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–19.

Wenbo Zhang received the M.S. degree in computer technology from Henan University of Science and Technology, Luoyang, China, in 2022. He is currently pursuing the Ph.D. degree in electronic information with the School of Information and Communication Engineering, Hainan University, Haikou, China.

His current research interests include computer vision, deep learning, and object detection and tracking.

Dongsheng Guo (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in communication and information systems from Sun Yat-sen University, Guangzhou, China, in 2010 and 2015, respectively.

Then, he joined Huaqiao University, Quanzhou, China, as Associate Professor. From 2018 to 2019, he was Visiting Scholar with the National University of Singapore, Singapore. He is currently Professor with the School of Information and Communication Engineering, Hainan University, Haikou, China. His current research interests include neural networks, robotics, and optimization.

Yilin Shang received the M.S. degree in control science and engineering from Henan University of Science and Technology, Luoyang, China, in 2022. She is currently pursuing the Ph.D. degree in information and communication engineering with Hainan University, Haikou, China.

Her research interests include nonlinear system control, reinforcement learning, and intelligent systems.

Weidong Zhang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 1990, 1993, and 1996, respectively.

He worked as a Postdoctoral Fellow with Shanghai Jiaotong University, Shanghai, China. He joined Shanghai Jiaotong University in 1998 as an Associate Professor and has been a Full Professor since 1999. From 2003 to 2004 he worked with the University of Stuttgart, Stuttgart, Germany, as an Alexander von Humboldt Fellow. From 2013 to 2017 he serviced as Deputy Dean of the Department of Automation, Shanghai Jiaotong University. He serving as a part-time Professor with Hainan University, Haikou, China, from 2021. He is currently Director of the Engineering Research Center of Marine Automation, Shanghai Municipal Education Commission, China. He is the author of more than 300 papers and one book, and has been recognized as an Elsevier most cited researcher. His research interests include control theory, machine learning theory, and their applications in industry and robots.

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

He has been a Professor and a Doctorial Tutor with the School of Information and Communication Engineering, Hainan University since 2020. He is currently a high-level talent in Hainan Province. His current research interests include artificial intelligence and signal and information processing.

Prof. Hu acted as a reviewer for IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, IEEE/ACM TRANSACTIONS ON NETWORKING, *Engineering Applications of Artificial Intelligence*, *Knowledge-Based Systems*, *Frontiers in Marine Science*, *Remote Sensing*, and *Applied Artificial Intelligence*. He is a Senior Member of CCF.