electronics

MDPI

Article

AQSFormer: Adaptive Query Selection Transformer for Real-Time Ship Detection from Visual Images

Wei Yang [1], Yueqiu Jiang [2,*], Hongwei Gao [3], Xue Bai [1], Bo Liu [1] and Caifeng Xia [1]

[1] Graduate School, Shenyang Ligong University, Shenyang 110159, China; weiyang@sylu.edu.cn (W.Y.); snowwhite224@163.com (X.B.); liubosylg@163.com (B.L.); xiacaifeng0528@163.com (C.X.)
[2] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China
[3] School of International Education, Shenyang Ligong University, Shenyang 110159, China; ghw1978@sohu.com
* Correspondence: yueqiujiang@sylu.edu.cn

Abstract: The Internet of Things (IoT) has emerged as a popular topic in both industrial and academic research. IoT devices are often equipped with rapid response capabilities to ensure seamless communication and interoperability, showing significant potential for IoT-based maritime traffic monitoring and navigation safety tasks. However, this also presents major challenges for maritime surveillance systems. The diversity of IoT devices and variability in collected data are substantial. Visual image ship detection is crucial for maritime tasks, yet it must contend with environmental challenges such as haze and waves that can obscure ship details. To address these challenges, we propose an adaptive query selection transformer (AQSFormer) that utilizes two-dimensional rotational position encoding for absolute positioning and integrates relative positions into the self-attention mechanism to overcome insensitivity to the position. Additionally, the introduced deformable attention module focuses on ship edges, enhancing the feature space resolution. The adaptive query selection module ensures a high recall rate and a high end-to-end processing efficiency. Our method improves the mean average precision to 0.779 and achieves a processing speed of 31.3 frames per second, significantly enhancing both the real-time capabilities and accuracy, proving its effectiveness in ship detection.

Keywords: image processing; ship detection; Internet of Things; visual image; rotary position embedding; deformable attention

check for updates

Citation: Yang, W.; Jiang, Y.; Gao, H.; Bai, X.; Liu, B.; Xia, C. AQSFormer: Adaptive Query Selection Transformer for Real-Time Ship Detection from Visual Images. *Electronics* **2024**, *13*, 4591. https://doi.org/10.3390/electronics13234591

Academic Editor: Silvia Liberata Ullo

Received: 18 October 2024
Revised: 11 November 2024
Accepted: 13 November 2024
Published: 21 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The growth in the shipping industry and the depletion of marine resources emphasize the need for effective ship monitoring [1–3]. Technologies such as synthetic aperture radar (SAR) and optical remote sensing (ORS) are essential for maritime safety and oceanography, given their comprehensive coverage and reliable performance in various weather conditions [4–6].

The workflow of the maritime ship monitoring system is depicted in Figure 1. A significant advantage of this paradigm is its ability to effectively cover vast areas without the need for expensive infrastructure development. The terminal devices can be considered image collectors, and the use of a central server facilitates user access to computing services, making it an attractive option for deploying applications. The images acquired from IoT systems are becoming more critical for maritime surveillance, offering advantages over SARs and ORSs in terms of cost, speed, and real-time processing [7,8]. This shift toward more accessible maritime monitoring tools is a significant trend despite challenges such as complex backgrounds and environmental disturbances affecting the accuracy of visual image-based ship detection [9,10]. Convolutional neural networks (CNNs), while successful in image recognition, struggle with complex visual tasks due to their local receptive fields and inefficiency in capturing global image context, especially in dynamic marine environments [11,12]. The transformer model, known for its success in natural language processing (NLP), has advantages for visual tasks [13,14]. Its self-attention mechanism can

capture long-range dependencies and perform global feature analysis [15]. It is also crucial for detecting subtle features and relationships in ship detection. However, transformer-based methods face challenges such as large parameter sizes, complex convergence, and high computational demands [16].

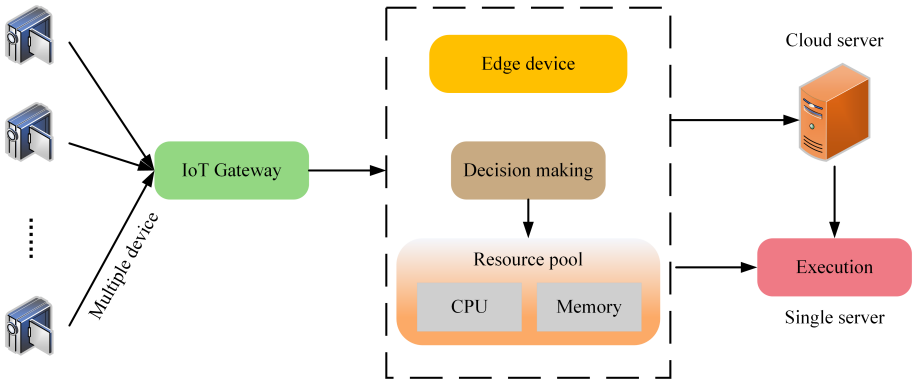


Figure 1. The maritime ship monitoring system workflow.

We propose the use of the adaptive query selection transformer (AQSFormer) for real-time ship detection from visual images to overcome the above issues. AQSFormer seeks to optimize computational efficiency and achieve faster convergence without compromising detection quality, offering a balanced solution for maritime surveillance. The contributions of this study can be summarized as follows:

(1) This study introduces 2D rotational position encoding (2D-RoPE) [17] to enhance transformer models for ship detection. 2D-RoPE enables the model to precisely interpret spatial relationships, including distances, angles, and directions between ships, by combining relative and absolute positions. This enriched positional awareness significantly improves the precision of the model in identifying ship targets, advancing spatial analysis capabilities in maritime surveillance, and providing a sophisticated solution for ship detection tasks.

(2) This paper enhances ship detection models with a deformable attention (DA) module, focusing on crucial sampling points near reference points in feature maps. This selective attention prioritizes important contour points, serving as a preliminary filter that highlights critical image features. It directs attention to essential ship details through edge contours, effectively gathering crucial edge information. This method overcomes the challenges posed by vague ship contours against complex backgrounds such as waves or fog by combining sparse sampling with transformer global relationship modeling. This improvement increases the capacity of the model to distinguish ships, enhancing its effectiveness in ship detection tasks.

(3) We propose an adaptive query selection (AQS) module to autonomously choose positive and negative training samples based on the statistical characteristics of ship targets. This module evaluates bbox quality and consistency, distinguishing highly similar queries. It offers flexible object query selection, adapts to diverse ship sizes and detection scenarios, reduces incorrect filtering, improves recall rates, and guarantees the pipeline end-to-end. This strategic improvement significantly enhances the ship detection capabilities of the model, improving efficiency and effectiveness.

The remainder of this paper is organized as follows. Section 2 reviews the ship detection literature and outlines the workflow steps. Section 3 presents the AQSFormer algorithm, including its framework and pseudocode, for a clear understanding of its operations. Section 4 details the experimental evaluations of AQSFormer, demonstrating its effectiveness. Section 5 summarizes the findings and suggests future research directions, building on the foundation of this work.

2. Related Works

2.1. Ship Detection Methods Based on CNNs

In the early stages of ship detection research, SAR and ORS images were the primary sources for detecting maritime targets [18,19]. The all-weather capability of SAR technology makes it essential for maritime object detection [20,21]. Innovations such as CenterNet++ proposed by Guo et al., which introduced feature refinement and enhancement modules, significantly advanced the field [22]. CenterNet++ was validated across three public datasets, proving its ship detection effectiveness. Leng et al. explored ship detection within range-compressed SAR data, leveraging the statistical characteristics and distance trajectories of ship targets [23]. Their method reduced the need for azimuth focusing, resulting in a more streamlined detection process. Zhu et al. addressed the detection of dim and small ships in extensive SAR images with an anchor-free method built on fully convolutional one-stage ship detection (FCOS), enhancing the small ship detection accuracy across large-scale SAR images [24]. Zhang et al. presented a novel perspective by integrating saliency detection with CNN networks to create an end-to-end ship detection framework [25]. This approach notably enhanced ship detection accuracy in SAR images by leveraging high-quality slices generated via saliency detection. Breaking conventional barriers, Leng et al. introduced a method to detect ships directly from raw SAR echo data, eliminating the traditional need to focus on SAR images [26]. This innovation opened avenues for ship detection without the imaging prerequisite. Furthermore, Zhang et al. responded to ship target diversity and similarity challenges in SAR images by developing a new scatter-point-guided region proposal network (RPN) [27]. They achieved improved performance by incorporating scatter characteristics into the network, facilitating efficient ship detection. This body of research collectively represents significant strides in evolving ship detection methodologies, highlighting the continuous pursuit of accuracy and efficiency in this domain.

ORS images are essential for offshore ship detection because of their high-resolution capabilities. Ren et al. utilized saliency information with feature enhancement structures (FESs) and saliency prediction branches (SPBs) to improve detection in complex environments [28]. Li et al. enhanced small ship detection in complex backgrounds by integrating multilevel features with a rotational RPN [29]. Wang et al. developed a convolutional network using Gaussian heatmap regression to map ship areas, addressing overfitting directly [30]. Dong et al. applied a brain-inspired visual attention mechanism with vector field filters for accurate detection in complex maritime conditions [31]. Hu et al. introduced a framework for detecting various-sized ships through multiscale supervision and attention guidance [32]. Xiao et al. proposed a method for detecting ships directly via compressive sensing (CS) measurements, bypassing scene reconstruction [33]. Cui et al. proposed an anchor-free framework that uses key points, orthogonal pooling, and soft-NMS modules for enhanced detection [34]. Nevertheless, SAR and ORS technologies are constrained by high costs, low real-time performance, and diminished effectiveness under adverse weather conditions.

Advances in computer vision have positioned visual image-based ship detection as a leading approach due to its cost-effectiveness and rapid response capabilities. The use of CNNs has significantly improved detection precision and speed. Zheng et al. proposed a small-scale ship detection method based on the attention mechanism, which emphasizes key features through a local attention module and combines them with high-level feature prediction layers to enhance small target detection ability [35]. Ngo et al. improved background differentiation and introduced uncertainties during training with techniques such as information bottleneck and reparameterization, achieving superior results on the SeaShips dataset and demonstrating the effectiveness of visual image-based detection in maritime security [36]. Liu et al. enhanced the YOLO v5 model with an adaptive image dehazing module, optimizing dehazing and detection simultaneously for more explicit ship recognition in hazy conditions [37]. Zhou et al. introduced multiscale feature-weighted fusion with cooperative attention and a bidirectional weighted feature pyramid network

(FPN) into YOLO v5, achieving precise detection in complex maritime environments [38]. Yi et al. proposed an efficient 1D convolutional neural network architecture [39], featuring the Multi-Kernel Temporal Block (MKTB) for multi-scale temporal response modeling and the Global Refinement Block (GRB) for refining global temporal features based on cross-channel similarity. By combining these modules, the proposed architecture effectively explores spatiotemporal features within tolerable computational cost and achieves state-of-the-art performance in hand gesture recognition tasks. Jiang et al. proposed a spatial–temporal interleaved network for efficient action recognition [40], which enhances 3D convolutional neural networks by introducing an interleaved feature interaction module, a boosted parallel pseudo-3D module, and a spatial–temporal differential attention mechanism. Zheng et al. proposed a binocular stereo vision-based method [41] for inland river ship recognition and depth estimation, utilizing MobileNetV1 as the feature extraction module of the YOLOv4 model to reduce computation while maintaining accuracy. Additionally, a feature point detection and matching algorithm based on the ORB algorithm enhances depth estimation precision, meeting the navigation safety requirements of inland river ships. CNNs effectively enhance ship detection capability against complex maritime backgrounds by learning deep features in images. Nonetheless, limitations in processing long-distance dependencies and global context information in images remain.

2.2. Ship Detection Methods Based on Transformers

With its self-attention mechanism, the transformer surpasses CNNs in object detection by effectively capturing long-distance dependencies and excelling in processing global information. This makes the transformer ideal for complex scenes such as ship detection. Shi et al. proposed a transformer-based model that integrates local sparse information aggregation, using explicit contour guidance to improve ship detection performance in SAR images [42]. Zhang et al. proposed the efficient ship detection transformer (ESDT) model, which accelerates model convergence through a feature distillation mechanism, enhancing ship detection efficiency and accuracy [43]. Chen et al. proposed an anchor-free method based on a transformer that detects ship targets with arbitrary orientations in SAR images [44]. Zhou et al. established the pyramid vision transformer (PVT) paradigm for multiscale feature representations in SAR images (PVT-SAR) [45]. Introducing the PVT paradigm allows the detection framework to overcome CNN limitations, enhancing detection capabilities in coastal and offshore scenes and improving ship detection performance in SAR images. Chen et al. proposed TSDet, an end-to-end SAR ship detection method based on a transformer, which employs a perceptron enhancement transformer (PET) [46]. Suppressing background noise and enhancing salient features improves ship detection performance in SAR images. Chen et al. proposed a multi-attention network for compressed video referring object segmentation [47], which consists of dual-path dual-attention module and a query-based cross-modal Transformer module. The query-based cross-modal Transformer firstly models the correlation between linguistic and visual modalities, and then the fused multi-modality features are used to guide object queries to generate a content-aware dynamic kernel and to predict final segmentation masks. Ge et al. proposed a NEural Attention Learning approach (NEAL) [48], which generates attention response maps by calculating partial derivatives of the classification output and combines them with the original detection loss for end-to-end training without additional network structures. Phan et al. introduced the UNet Structured Transformer (UNest) architecture [49], which incorporates structural inductive biases for unpaired medical image synthesis by leveraging the Segment-Anything model to precisely extract foreground structures and perform structural attention.

Although transformer-based models are effective for ship detection in SAR and ORS images, their application in visual image-based detection encounters challenges due to complex maritime environments, such as sea clutter and haze, that obscure ship outlines and limit performance. Furthermore, their extensive computational requirements and numerous parameters restrict their real-time application. To address these issues, AQSFormer

is proposed for high-precision, real-time ship detection from visual images. It focuses on overcoming the challenges faced by existing transformer-based models, enhancing ship detection in complex maritime environments, and improving computational efficiency for rapid deployment.

3. Proposed Method

This section explores the architecture of AQSFormer and key components for improving ship detection efficiency and accuracy from visual images. Unlike traditional detection transformer (DETR) networks, AQSFormer integrates novel modules such as 2D-RoPE, DA, and the AQS module, as illustrated in Figure 2. The following subsections provide detailed discussions on these key modules, highlighting their contributions to refining the model for accurate ship detection in challenging maritime environments.

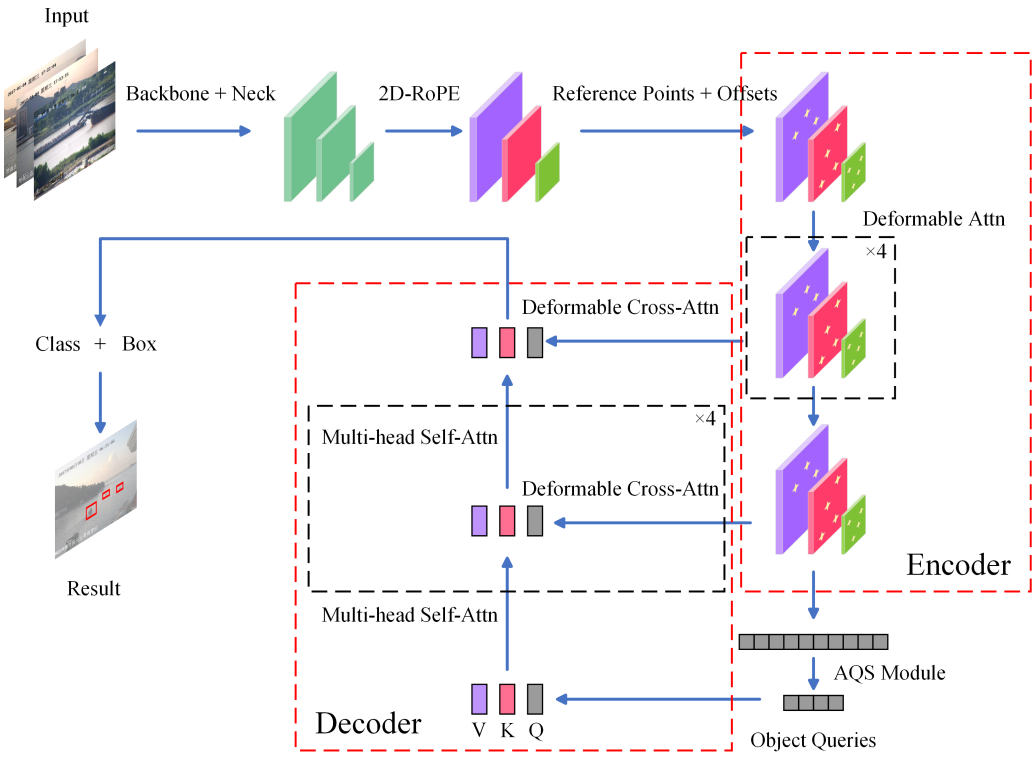


Figure 2. The architecture overview of the proposed approach AQSFormer. It mainly contains an encoder with deformable attention, a decoder with deformable cross-attention, and an adaptive query selection module.

3.1. The 2D-RoPE Module

Vision transformers (ViTs) process images by splitting them into small patches and applying positional encoding before feeding them into the encoder to maintain spatial relationships among these patches. This approach leverages the advantage of a transformer in analyzing sequential data and comprehending complex relationships, which is crucial for understanding the global context. ViTs excel in capturing long-distance dependencies because of their self-attention mechanism, unlike CNNs, which rely on predefined convolutional kernels. This capability allows ViTs to provide insights beyond CNN local receptive fields, making them exceptionally useful for tasks requiring comprehensive scene understanding. This distinct advantage of ViTs is highlighted by their ability to surpass traditional methods in capturing complexities in visual images. The process of the traditional self-attention mechanism is illustrated in Equation (1), demonstrating the technical foundation that enables ViTs to analyze images effectively.

$$\begin{cases} q_m = (x_m + p_m)W_Q \\ k_m = (x_n + p_n)W_K \\ v_j = (x_n + p_n)W_V \\ a_{i,j} = \dfrac{\exp\left(\frac{q_m^\top k_n}{\sqrt{d}}\right)}{\sum_{j=1}^{N}\exp\left(\frac{q_m^\top k_j}{\sqrt{d}}\right)} \\ o_m = \sum\limits_{n=1}^{N} a_{m,n} v_n \end{cases} \tag{1}$$

where x_m and x_n are the d-dimensional inputs at positions m and n, respectively; p_m and p_n are the d-dimensional positional encoding vectors for positions m and n; and W_Q, W_K, and W_V are the weight matrices for the query, key, and value, respectively. Positional encoding $p_{(*)}$ typically uses trigonometric functions, as shown in Equation (2).

$$\begin{cases} p_{2m} = \sin(k/10{,}000^{2m/d}) \\ p_{2m+1} = \cos(k/10{,}000^{2m/d}) \end{cases} \tag{2}$$

Trigonometric positional encoding employs sine and cosine functions at varying frequencies to embed positional information, enabling models to comprehend absolute positions within sequences. This approach is advantageous for identifying long-distance dependencies and accommodating sequences of any length. However, its inability to encode relative positions can hinder the understanding of complex relationships between sequence elements, such as the position, angle, and direction of ships, in ship detection tasks. To address this, our approach posits that the inner product between query vector q_m and key vector k_n can be encapsulated by a specific function g, as detailed in Equation (3). We aim to incorporate relative positional information, enhancing the capacity of the model to interpret intricate spatial relationships.

$$\langle f_q(x_m, m), f_k(x_n, n)\rangle = g(x_m, x_n, m-n) \tag{3}$$

where $(m-n)$ represents the relative position between x_m and x_n. The functions $f(*)$ and $g(*)$ that satisfy the above constraints are shown in Equation (4).

$$\begin{cases} f_q(x_m, m) = (W_q x_m)e^{im\theta} \\ f_k(x_n, n) = (W_k x_n)e^{in\theta} \\ g(x_m, x_n, m-n) = \mathrm{Re}[(W_q x_m)(W_k x_n)^* e^{i(m-n)\theta}] \end{cases} \tag{4}$$

where $\mathrm{Re}(*)$ represents the real part of a complex number, $(W_k x_n)^*$ denotes the conjugate of the complex number $(W_k x_n)$, and $\theta \in \mathbb{R}$ is a nonzero constant. According to the Euler formula, $f_{\{q,k\}}$ can be expanded as shown in Equation (5).

$$f_{\{q,k\}}(x_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} W_{\{q,k\}} \begin{pmatrix} x_m^{(1)} \\ x_m^{(2)} \end{pmatrix} \tag{5}$$

where $\left(x_m^{(1)}, x_m^{(2)}\right)$ represents x_m in two-dimensional coordinates. If Equation (5) is generalized to any $x_i \in \mathbb{R}$ and d is even, $f_{\{q,k\}}$ is shown in Equation (6).

$$f_{\{q,k\}}(x_m, m) = R^d_{\Theta,m} W_{\{q,k\}} x_m \tag{6}$$

where $R^d_{\Theta,m}$ is shown in Equation (7).

$$R^{d}_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix} \tag{7}$$

where $\Theta = \left\{\theta_i = 10{,}000^{-2(i-1)/d}, i \in [1, 2, \cdots, d/2]\right\}$. However, for solving 2D image data, the relative positional relationship of (x, y) should be considered; thus, 2D-RoPE is shown in Equation (8).

$$f_{\{q,k\}}[(x_m, y_n), (m, n)] = R^{c,d}_{\Theta,(m,n)} W_{\{q,k\}}(x_m, y_n) \tag{8}$$

where c and d represent the width and height of the image, respectively, and $R^{c,d}_{\Theta,(m,n)}$ is shown in Equation (9).

$$R^{c,d}_{\Theta,(m,n)} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos m\theta_{c/2} & -\sin m\theta_{c/2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \sin m\theta_{c/2} & \cos m\theta_{c/2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cos n\theta_1 & -\sin n\theta_1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \sin n\theta_1 & \cos n\theta_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & \cos n\theta_{d/2} & -\sin n\theta_{d/2} \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & \sin n\theta_{d/2} & \cos n\theta_{d/2} \end{pmatrix} \tag{9}$$

The implementation of 2D-RoPE is shown in Figure 3.

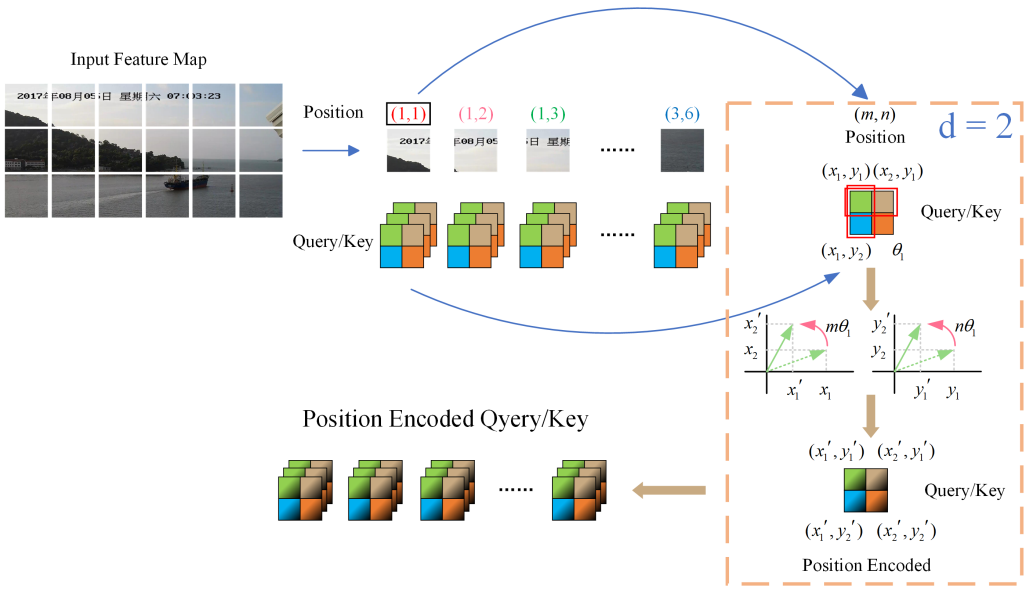


Figure 3. The workflow of the two-dimensional rotational position encoding method. The input feature map is patchified as tokens, and then the absolute and relative position information is introduced in the two-dimensional rotational position encoding method.

2D-RoPE enhances positional encoding by combining the absolute position and relative distance of pixels within image sequences, thereby improving the comprehension of positional context by the model. This approach both enriches the understanding of positional nuances and improves the ship positioning task precision and efficiency. By incorporating details such as the distance, angle, and direction between pixels, 2D-RoPE significantly advances the ability of the model to accurately determine ship locations.

3.2. The Deformable Attention Module

During the early stages of processing image features, the transformer algorithm tends to distribute attention weights almost evenly across all pixel positions, which necessitates extensive training to refine its focus on relevant areas. Additionally, the computational complexity of the self-attention module in the transformer encoder increases quadratically with the size of the spatial feature map, significantly increasing both computational and memory demands for high-resolution images. DA is introduced, which samples spatially sparse locations fused with the ability of the transformer to model contextual relationships to address the slow convergence and high complexity challenges of DETR. This approach is embodied in the design of the DA module, detailed in Figure 4, which offers a strategy to balance efficiency and effectiveness in feature processing.

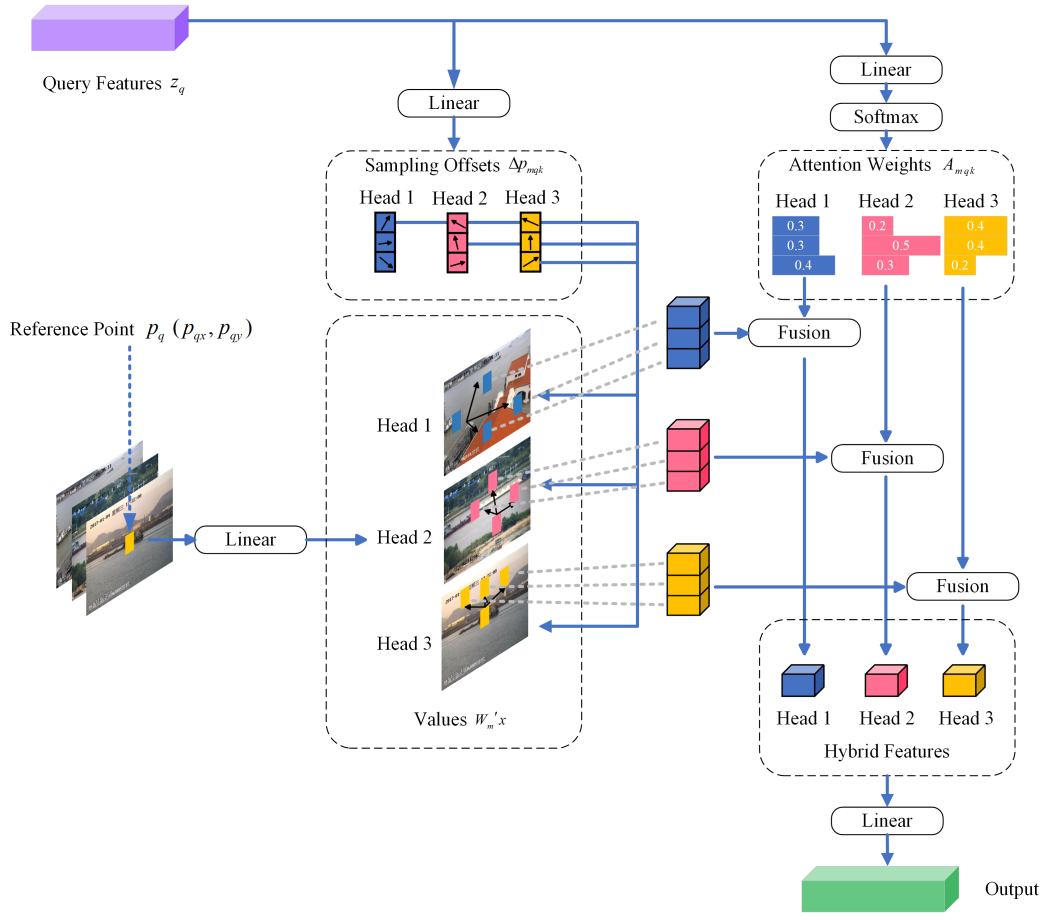


Figure 4. An illustration of our deformable attention mechanism. We present the workflow of the deformable attention mechanism. A group of reference points is sampled uniformly on the feature map, whose offsets are learned from the queries by the linear layer. Then, the deformed features are projected from the sampled reference points. The relative position bias is also computed by the deformed points, enhancing the multihead attention, which outputs the transformed features. We show only one reference point and three deformed points for clarity.

For a given input feature map $x \in \mathbb{R}^{C \times H \times W}$, an element q in the query, characterized by content features z_q and a 2D reference point p_q, has its DeformAttn features calculated as specified in Equation (10). This method allows the model to dynamically adjust its focus within the feature map based on the specific attributes and spatial reference of q, enhancing the ability of the model to extract relevant information from complex inputs.

$$DA(z_q, p_q, x) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk} \cdot W'_m x(p_q + \Delta p_{mqk}) \right] \tag{10}$$

where $m \in M$ represents the number of attention heads, and $k \in K \ll HW$ represents the number of sampled keypoints. The symbols Δp_{mqk} and A_{mqk} denote the sampling offset and the attention weight for the k_{th} sampling point of the m_{th} attention head, respectively, with the scalar attention weight A_{mqk} normalized to $A_{mqk} \in [0, 1]$ by $\sum_{k=1}^{K} A_{mqk} = 1$. $W_m \in \mathbb{R}^{C \times C_v}$ and $W'_m \in \mathbb{R}^{C_v \times C}$ are learnable weights, and C indicates the channel dimension of the feature map, with $C_v = C/M$. $\Delta p_{mqk} \in \mathbb{R}^{H \times W}$ is defined as a 2D vector within the feature map's range. Given that the offset position $(p_q + \Delta p_{mqk})$ might not be an integer, bilinear interpolation, as proposed by Dai et al. [50], is used to compute the value of x at position $(p_q + \Delta p_{mqk})$. Drawing from deformable convolution concepts, sampling positions are adaptively adjusted through offset predictions, encircling the reference point and modifying based on the ship objects' actual shapes. This method both diminishes the sensitivity of the model to background noise and improves its ability to discern ship features.

Let N_q represent the total number of query elements. In the standard DETR model, for an input feature map $x \in \mathbb{R}^{C \times H \times W}$, the number of query elements is $N_q = HW$, where H and W are the height and width of the feature map, respectively. Consequently, the computational complexity of the attention module, as detailed in Equation (1), is illustrated in Equation (11). This relationship highlights the direct impact of the spatial dimensions of the feature map on the computational demands of the attention mechanism of the DETR model.

$$O_{DETR} = (N_q C^2)^2 = (HWC^2)^2 \tag{11}$$

Equation (11) revealed that the computational complexity of the traditional DETR algorithm increases quadratically with the spatial size of the input feature map. In contrast, within the DA module referenced by Equation (10), the complexity involved in computing the sampling coordinate offsets Δp_{mqk} and their corresponding attention weights $A_{\Delta p_{mqk}}$ is also detailed in Equation (12). This highlights the efficiency improvements offered by DA mechanisms, which aim to optimize computational resources by adjusting to the most informative parts of the input feature map, potentially reducing the overall computational load compared to the traditional DETR approach.

$$\begin{cases} O_{\Delta p_{mqk}} = (2MKN_qC) \\ O_{A_{\Delta p_{mqk}}} = (MKN_qC) \end{cases} \tag{12}$$

The first $2MK$ channels are dedicated to encoding the (x_{offset}, y_{offset}) of the sampling offset Δp_{mqk}; in contrast, the remaining MK channels focus on extracting features, which are then fed into the softmax operator to compute the attention weight $A_{\Delta p_{mqk}}$. Given that $A_{\Delta p_{mqk}}$ is a scalar, the computational complexity, as referenced in Equation (10), is detailed in Equation (13). This structure efficiently separates the position encoding and attention weight calculation tasks, streamlining the process and potentially reducing the computational burden associated with these operations.

$$O_{DA} = (N_q C^2 + N_q K C^2 + 5N_q KC + 3MKN_q C) \tag{13}$$

The coefficient 5 in $5N_qKC$ originates from the calculations involved in bilinear interpolation. The default parameters are set as $M = 6$, $K = 4$, and $C = 256$ for the experiments conducted in this study. Given that $5K + 3MK < C$, the computational complexity of the

DA module is more precisely depicted by Equation (14). This formulation demonstrates the efficiency of the DA mechanism, particularly in terms of its computational requirements, by leveraging the specificities of bilinear interpolation and the chosen parameter settings to optimize processing.

$$\begin{aligned} O_{DA} &= [N_q C^2 + K N_q C^2 + N_q C(5K + 3MK)] \\ &= (6 N_q C^2) \end{aligned} \tag{14}$$

where $N_q = N$ denotes the number of object queries. According to Equation (14), the DA module benefits from a sparse sampling strategy that renders its computational complexity independent of the spatial size of the feature map. This approach leads to a significant reduction in computational complexity, demonstrating the efficiency of the DA module in processing images. The module optimizes resource usage without compromising the attention and feature extraction quality by focusing on a limited set of strategically sampled points.

The DA module selects several key sampling points around the reference point instead of all the pixels in the feature map, optimizing computational and memory usage by focusing on significant positions. This strategy enhances ship recognition efficiency and accuracy, as it emphasizes areas around ship targets, particularly edge contours. By dynamically adjusting its focus, the module prioritizes informative areas for ship detection. It enables the model to accurately capture crucial details from ship edges, improving the detection of small or partially obscured ships against complex backgrounds.

3.3. The Adaptive Query Selection Module

The DA mechanism significantly enhances object detection performance, where an effective sample selection strategy is vital for accuracy. Traditional anchor-based methods, which rely on the preset dimensions and ratios of anchors distributed densely across the feature map and a fixed intersection over union (IoU) threshold for filtering, can sometimes overlook edge targets. DETR revolutionizes this approach by utilizing "object queries" to directly predict target categories and bboxes, bypassing the need for anchor-based preselection. These queries, which represent detection targets, enable the model to predict all targets concurrently through parallel processing, moving away from traditional stepwise inference. Unlike predefined anchors, object queries are learned, making DETR adaptable to various detection tasks and scenarios.

This paper introduces the AQS module, which enhances DETR's object query concept by dynamically adjusting query selection criteria to suit complex scenarios, thus refining bbox selection. This adaptive mechanism addresses the shortcomings of conventional methods, offering more flexible and accurate object detection, particularly for ship recognition in challenging maritime environments. Ma et al. proposed a two-stage adaptive query method for style-aware video captioning [51], which adjusts caption styles flexibly through style learning in the first stage and adaptive style guidance in the second. However, the core of the AQS module lies in its dynamic adjustment of the IoU threshold. This adjustment enables the model to select the most relevant queries based on the complexity and features of the training samples. By using IoU and $L2$ distance for query selection, the AQS module avoids the fixed threshold approach common in traditional detection methods, thereby improving the model's performance in complex scenarios. The pseudocode of the proposed AQS module is shown in Algorithm 1.

Algorithm 1 Adaptive query selection module.

INPUTS:
GT is a set of ground-truth boxes on the image
Q is a set of object queries
OUTPUTS:
P is a set of positive samples
N is a set of negative samples

for each ground truth $gt \in GT$ **do**
build an empty set: $F_{gt} \leftarrow \emptyset$;
$X \leftarrow$ select 9 object queries from Q;
$F_{gt} = F_{gt} \cup X$;
compute IoU: $T_{gt} = IoU(F_{gt}, gt)$;
compute mean of T_{gt}: $m_{gt} = Mean(T_{gt})$;
compute standard deviation of T_{gt}: $s_{gt} = Std(T_{gt})$;
compute IoU threshold: $v_{gt} = m_{gt} + s_{gt}$;
for each candidate anchor $f \in F_{gt}$ **do**
if ($IoU(f, gt) \geq v_{gt}$ and center of f in gt) **then**
$P = P \cup f$;
end if
end for
end for
$N = Q - P$;
return P, N

We propose that not all training samples are equally important for model learning. A faster training process and greater accuracy with fewer computational resources can be achieved by focusing on those samples that are most critical for improving model performance. In Algorithm 1, we select the 9 bboxes closest to the ground truth gt based on the sum of $L2$ distances between the object query and the corner point of gt. We set the number of bboxes to 9 to match anchors of three different sizes and three aspect ratios. Then, we calculated the IoU between the selected object queries and the ground truth gt, from which we derived their average m_{gt} and standard deviation s_{gt}. Accordingly, we set the IoU threshold v_{gt} for the ground truth gt as m_{gt} plus s_{gt}. Finally, object queries with an IoU greater than or equal to the threshold v_{gt} and whose centers were inside the ground truth gt were selected as positive samples. For a single object query assigned to multiple gt bboxes, we chose the one with the highest IoU as the positive sample; the rest were deemed negative samples.

In object detection tasks, the IoU serves as a key metric for evaluating the matching degree between the predicted and actual bboxes. It is also applicable for assessing the matching effectiveness of predefined anchors. A high value indicates greater consistency between bboxes, which may make it difficult for the model to effectively distinguish between valid predictions and background noise, affecting the detection accuracy and potentially confusing label assignments during training, thereby reducing the generalization performance of the model. Therefore, adaptively adjusting the IoU threshold based on the values of m and s is crucial to ensure the selection of high-quality object queries and optimize detection outcomes. For difficult-to-detect or marginalized targets, the detection sensitivity can be increased by lowering the IoU threshold; in contrast, the IoU threshold can be increased to reduce false positives for easily recognizable targets.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

Our study utilized the SeaShips dataset, which was sourced from surveillance footage around Hengqin Island, Zhuhai, China [52]. The SeaShips dataset, which features 7000 high-resolution (1920 × 1080) images, represents various real-world maritime scenarios, enhancing the practical application capabilities of the model. It encompasses images of six ship types bulk carriers, general cargo ships, container ships, fishing vessels, and passenger ships under diverse environmental settings with differing perspectives, scales, and backgrounds. Each image is meticulously annotated with ship types and bounding boxes. We allocated 6000 images for training and 1000 for testing. The dataset is further exemplified in Figure 5 with ship-type variations, Figure 6 with scale variations, Figure 7 with lighting variations, and Figure 8 with complex background variations.



Figure 5. Examples of typical images of six ship types. (**a**–**f**): ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship.



Figure 6. Three different scales of ship samples.



Figure 7. Ship targets under three distinct illumination conditions.

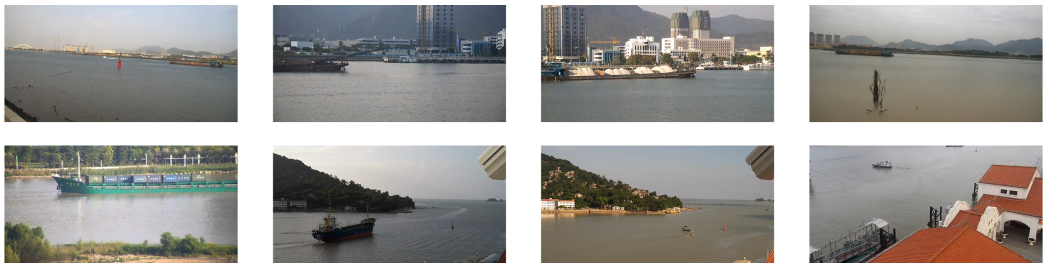


Figure 8. Ship targets across diverse backgrounds.

4.1.2. Settings

Experiments were conducted in this study on a server with an NVIDIA RTX 3090 GPU. We implemented the AQSFormer algorithm using PyTorch v1.12.0 and Python v3.8.0. During training, we set the batch size to 2, and each epoch consisted of 3000 iterations. We chose ResNet 50 as the backbone network. We employed the AdamW optimizer, setting the learning rate and weight decay coefficient to 1×10^{-4} . Our data augmentation included random flipping and cropping, with probabilities and scale ranges set to enhance variability. Unlike typical approaches, we trained all models from scratch to ensure unbiased adaptation to ship detection and to evaluate baseline convergence speeds. The training concluded after 14 epochs to optimize model performance while ensuring experimental accuracy and reproducibility.

4.1.3. Evaluation Metrics

To clearly and objectively evaluate the improvements of our proposed algorithm, we selected seven evaluation metrics: AP@0.5, mAP@0.5, mAP@0.5:0.95, memory usage, number of parameters, Giga Floating Point Operations (GFLOPs), and frames per second (FPS). These metrics provide a comprehensive assessment of the model's performance. Average Precision (AP) and mean Average Precision (mAP) are standard accuracy metrics in object detection, measuring detection precision for individual classes and all classes, respectively. Specifically, AP@0.5 and mAP@0.5 represent the average precision for single class and all classes at an Intersection over Union (IoU) threshold of 0.5. In contrast, mAP@0.5:0.95 denotes the mAP across IoU thresholds from 0.5 to 0.95 in increments of 0.05, offering a more thorough reflection of detection accuracy. The memory usage metric assesses the maximum memory required by the model during inference, which is essential for evaluating resource consumption in practical applications. The number of parameters indicates the quantity of learnable parameters in the model, typically correlating positively with the model's complexity. GFLOPs measure the number of floating-point operations the model performs per second during inference, reflecting its computational complexity. FPS refers to the number of frames the model can process per second, which is crucial for assessing real-time prediction capability. By integrating these metrics, we can comprehensively evaluate the model's performance in terms of accuracy, efficiency, and computational complexity, providing a robust quantitative assessment of the proposed AQSFormer algorithm's effectiveness.

4.2. Comparison with State-of-the-Art Methods

To provide a comprehensive assessment, we evaluated the performance of the AQSFormer algorithm against current state-of-the-art algorithms using the SeaShips dataset. To clearly and intuitively demonstrate the superior performance of our method, we compared six types of ships with multi-scale scenarios, as shown in Figure 9. It shows that, our proposed AQSFormer accurately detects ships of various types and scales. In contrast, deformable DETR and DDQ DETR suffer from missed detections and false alarms due to their use of absolute positional encoding in transformer-based models, which neglects the importance of relative positions during patch sequence encoding, leading to limitations. Similarly, YOLOX, YOLOF, TOOD, RTMDet, and GLIP encounter false detections, missed detections, and inaccurate localization. These issues likely stem from confusion with complex backgrounds like coastlines and challenges in detecting small-scale objects. These results highlight the effectiveness of our introduced modules: 2D-RoPE, deformable attention, and adaptive query selection. The quantitative comparison results are detailed in Table 1.



Figure 9. A performance comparison of our proposed algorithm and the state-of-the-art algorithms across six ship types in the SeaShips [52] dataset.

As shown in Table 1, our proposed AQSFormer outperforms other SOTA algorithms with an mAP@0.5 of 0.956 and an mAP@0.5:0.95 of 0.779. While AQSFormer demonstrates excellent average accuracy across the four ship categories, it does not achieve the highest performance in all categories. This limitation may stem from the fact that our approach is primarily designed to address challenges related to complex backgrounds and multi-scale detection, which may slightly compromise performance in detecting small-scaled or low-sample ships. Nevertheless, AQSFormer consistently delivers stable performance across all ship categories, showcasing its robustness under various conditions. This stability highlights its strong adaptability for IoT-based ship detection tasks. In contrast, the deformable DETR, ATSS, and YOLOF algorithms perform relatively poorly in ship detection. Although deformable DETR excels at handling deformed objects, its performance suffers significantly

in complex backgrounds, particularly when there are large scale differences among ship categories. ATSS, despite its strong adaptability, struggles with localization errors and false negatives, particularly in multi-scale and low-contrast scenarios. YOLOF, on the other hand, faces challenges in detecting small objects and handling low-sample cases, resulting in reduced accuracy for certain ship categories. Therefore, while these algorithms perform well in specific tasks, their effectiveness in ship detection is limited, especially in complex environments and with diverse ship types.

Table 1. A comparison with the SOTA algorithms on the SeaShips dataset. The best and second-best results are in **bold** and underlined, respectively.

Methods	AP@0.5						mAP @0.5	mAP @0.5:0.95
	Ore Ship	General Cargo Ship	Bulk Carrier	Fishing Boat	Passenger Ship	Container Ship		
(CVPR 2020) ATSS [53]	0.935	0.942	0.906	0.871	0.844	0.925	0.904	0.661
(ArXiv 2021) YOLOX [54]	0.914	0.925	0.927	0.896	0.852	0.923	0.906	0.642
(CVPR 2021) YOLOF [55]	0.921	0.933	0.917	0.912	0.847	0.915	0.908	0.693
(ICLR 2021) Deformable DETR [56]	0.888	0.932	0.896	0.830	0.828	0.920	0.882	0.617
(ICCV 2021) TOOD [57]	0.927	0.891	0.930	0.885	0.838	0.917	0.898	0.625
(ArXiv 2022) RTMDet [58]	0.902	0.915	0.933	0.862	0.833	0.924	0.895	0.624
(CVPR 2022) GLIP [59]	0.942	0.967	0.956	0.917	0.866	0.938	0.931	0.728
(CVPR 2023) DDQ DETR [60]	0.937	0.944	0.951	0.923	0.849	0.933	0.923	0.705
(ICLR 2023) DINO [61]	0.949	0.962	**0.963**	0.915	**0.923**	0.956	0.945	0.752
(ICCV 2023) CO-DETR [62]	0.936	0.958	0.962	0.908	0.901	0.942	0.935	0.749
(ours) AQSFormer	**0.963**	**0.975**	0.959	**0.936**	0.918	**0.985**	**0.956**	**0.779**

In Table 2, the proposed AQSFormer algorithm, while not the fastest, boasts a FPS of 31.3, balancing speed and its superior detection capabilities. It operates at a frame rate conducive to real-time ship detection, demonstrating its practical utility. The AQSFormer shows efficient memory usage at 6.67 G, with 37.4 M parameters and 67.855 GFLOPs. In contrast, YOLOX, despite its lower parameter count (9.0 M) and reduced computational complexity (33.777 GFLOPs), does not reach the anticipated processing speed. This discrepancy can be attributed to the fundamental architectural differences between YOLOX, an anchor-based single-stage detector, and transformer-based detection algorithms such as DETR, which forgo anchors for a direct object query approach to achieve end-to-end detection. This architectural divergence emphasizes the importance of considering more than just parameter count and computational demands when selecting object detection algorithms. Understanding the inherent differences between architectures and their impact on performance is crucial. The effectiveness of AQSFormer in distinguishing between various ship types is further illustrated in the confusion matrix presented in Figure 10.

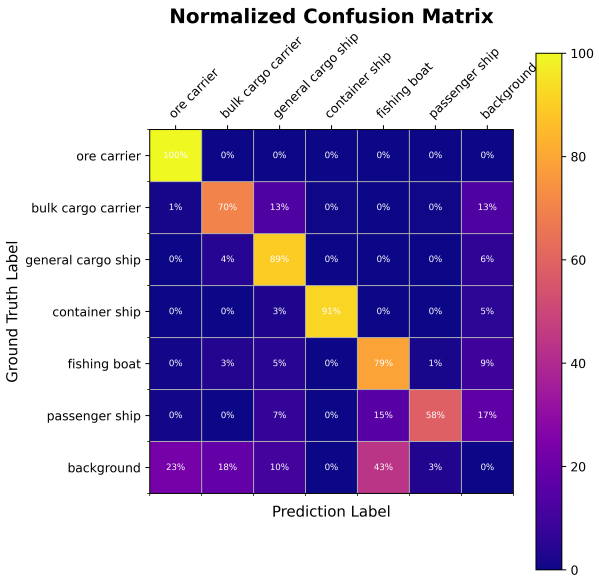


Figure 10. Normalized confusion matrix of AQSFormer.

Table 2. Experimental results of memory usage, number of parameters, GFLOPs and FPS. The best and second-best results are in **bold** and underlined, respectively. In the indicator, the arrows ↑/↓ mean that the higher/lower the value, the better.

Methods	Memory (G) ↓	Params (M) ↓	GFLOPs ↓	FPS ↑
(CVPR 2020) ATSS [53]	**6.60**	32.3	211.968	29.5
(ArXiv 2021) YOLOX [54]	6.71	**9.0**	**33.777**	29.2
(CVPR 2021) YOLOF [55]	7.27	44.2	103.424	29.1
(ICLR 2021) Deformable DETR [56]	6.72	40.1	197.632	29.4
(ICCV 2021) TOOD [57]	6.74	32.2	207.872	29.4
(ArXiv 2022) RTMDet [58]	6.75	52.3	80.121	**46.0**
(CVPR 2022) GLIP [59]	7.02	153.6	122.88	29.2
(CVPR 2023) DDQ DETR [60]	7.04	48.3	280.576	29.4
(ICLR 2023) DINO [61]	7.06	47.6	263.483	29.4
(ICCV 2023) CO-DETR [62]	7.28	65.8	374.735	29.5
(ours) AQSFormer	6.67	37.4	67.855	31.3

Figure 10 shows that AQSFormer excels in classifying ore carriers, general cargo ships, container ships, and fishing boats, with correct prediction ratios of 100%, 89%, 91%, and 79%, respectively. However, it faces challenges with bulk carriers and passenger ships, achieving 70% and 58% accuracy, respectively. The lower accuracy for bulk carriers is attributed to potential confusion with similar categories, while passenger ships are often mistaken due to background similarities. Despite these challenges, AQSFormer outperforms other SOTA algorithms in terms of critical metrics such as mAP@0.5, mAP@0.5:0.95, demonstrating its ability to predict various object scales. Although it may not lead in some hardware efficiency metrics, such as FPS, its superior detection capabilities and moderate computational resource usage make meaningful contributions to object detection technology.

4.3. Ablation Experiments and Sensitivity Analysis

To validate the effectiveness and parameter setup of AQSFormer, we conducted ablation experiments on the SeaShips dataset, and maintaining parameters consistent with those of the baseline. We explored the components of AQSFormer and their influence on performance. Specifically, we examined the impact of using different backbone networks, including ResNet50 [63], ResNet101, and Swin-L [64], on maritime ship image processing. We also assessed the enhancement in performance through test time augmentation (TTA), which enhances model generalizability on new samples by applying diverse augmentation methods during testing. The results are detailed in Table 3.

Table 3. The ablation results of AQSFormer with different backbones and test time augmentation (TTA). In the indicator, the marks ×/✓ mean that without/with TTA.

Backbone	TTA	mAP@0.5:0.95 (%)	mAP@0.5 (%)	AP_s@0.5 (%)	AP_m@0.5 (%)	AP_l@0.5 (%)
ResNet50 [63]	×	77.9	95.6	65.7	85.4	98.6
ResNet101 [63]	×	78.1 (+0.2)	96.0 (+0.4)	67.6 (+1.9)	85.8 (+0.4)	98.9 (+0.3)
Swin-L [64]	×	78.4 (+0.5)	96.8 (+1.2)	68.7 (+3.0)	85.6 (+0.2)	99.0 (+0.4)
Swin-L [64]	✓	78.7 (+0.8)	98.2 (+2.6)	70.5 (+4.8)	86.5 (+1.1)	99.6 (+1.0)

As shown in Table 3, using Swin-L as the backbone network in combination with test time augmentation (TTA) achieves the best performance on the mAP@0.5:0.95 metric, particularly excelling in small-scale object detection tasks. This result demonstrates the effectiveness and adaptability of Swin-L in complex scenarios. As a vision transformer-based network, Swin-L offers superior feature extraction capabilities and a deeper architecture, allowing it to capture fine object details more effectively, which is especially beneficial for small-scale object detection. Meanwhile, TTA improves model performance by applying

diverse transformations (e.g., rotation, scaling, flipping) to the input images during inference, performing independent inferences on the transformed images, and then combining the results. This approach reduces the bias introduced by a single input, enhancing the model's robustness and accuracy. However, it is important to note that incorporating Swin-L also increases parameters, which may limit its applicability in real-time tasks. Additionally, we evaluated the impact of the 2D-RoPE and DA modules, both designed to improve the model's ability to perceive spatial structures and adapt to various object shapes. The experimental results are detailed in Table 4.

Table 4. The ablation results of 2D-RoPE and DA modules.

Pos Embed	Attention	mAP@0.5:0.95 (%)	mAP@0.5 (%)	AP_s@0.5 (%)	AP_m@0.5 (%)	AP_l@0.5 (%)
Sine	Self-Attn	74.5	89.6	49.8	80.4	97.3
2D-RoPE	Self-Attn	76.1 (+1.6)	90.7 (+1.1)	55.9 (+6.1)	81.3 (+0.9)	98.1 (+0.8)
Sine	DA	76.6 (+2.1)	94.3 (+4.7)	59.1 (+9.3)	82.7 (+2.3)	98.1 (+0.8)
2D-RoPE	DA	77.9 (+3.4)	95.6 (+6.0)	65.7 (+15.9)	85.4 (+5.0)	98.6 (+1.3)

Table 4 demonstrates that incorporating 2D-RoPE and DA into AQSFormer significantly improves the performance. Adding 2D-RoPE notably enhances the detection accuracy for small ships by improving the spatial awareness of the model. DA contributes to better performance in detecting ships of various shapes by effectively capturing edge contours, thus improving scene understanding. We also examined various data augmentation strategies. Specifically, we selected the Flip, Resize, Crop, and Local Similarity Jitter techniques for the following reasons:

- Flip: Horizontal flipping increases data diversity and helps the model adapt to different directions, improving robustness in real-world scenarios.
- Resize: Resizing helps the model learn scale-invariant features, ensuring good performance regardless of object size in object detection tasks.
- Crop: Random cropping enables the model to focus on different areas of the image, improving recognition in dense scenes or with partial occlusions.
- Local Similarity Jitter (LSJ): This technique introduces small perturbations to simulate noise, enhancing the model's robustness to subtle variations in real-world conditions.

These data augmentation strategies were selected to improve our AQSFormer's generalization ability and help it better adapt to various scenes and data distributions. The outcomes are presented in Table 5.

Table 5. Results of different data augmentation strategies on AQSFormer. LSJ refers to Local Similarity Jitter.

Data Aug.	mAP@0.5:0.95 (%)	mAP@0.5 (%)	AP_s@0.5 (%)	AP_m@0.5 (%)	AP_l@0.5 (%)
×	74.5	87.5	52.5	80.8	97.1
Flip & Resize & Crop	76.5 (+2.0)	92.6 (+5.1)	59.8 (+7.3)	84.8 (+4.0)	98.3 (+1.2)
LSJ	76.0 (+1.5)	91.1 (+3.6)	59.4 (+6.9)	82.0 (+1.2)	97.8 (+0.7)
Flip & Resize & Crop & LSJ	77.9 (+3.4)	95.6 (+8.1)	65.7 (+13.2)	85.4 (+4.6)	98.6 (+1.5)

Table 5 shows that the combination of Flip, Resize, Crop, and LSJ yielded the greatest improvement in performance, especially for small-scale object detection. This result indicates that diverse data augmentation techniques during training significantly boost the ability of the model to detect ships of various sizes and shapes, enhancing its effectiveness.

5. Conclusions

This study introduces AQSFormer, a method designed for real-time visual image ship detection. It was tested on the SeaShips dataset to confirm its superiority and parameter efficiency. AQSFormer enhances spatial understanding by integrating 2D-RoPE, which merges absolute and relative positional data, improving ship position and orientation representation. The DeformAttn module refines the focus of the model on crucial ship features

by accurately aggregating edge key points, enhancing feature space resolution. The method dynamically updates the sampling strategy, optimizing training and inference, making it a flexible and efficient approach. It outperformed state-of-the-art models, increasing mAP by 2.7%, and achieved 31.3 FPS, showing its real-time capabilities and precision improvement. This research paves the way for advancements in visual image-based ship detection, supporting maritime surveillance, autonomous navigation, and rescue operations. Future work will explore multimodal data fusion for greater accuracy and speed, addressing broader application demands.

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

Funding: This research was funded by the Basic Research Funds for Undergraduate Universities in Liaoning Province (SYLUGXTD07), and Defense Industrial Technology Development Program (JCKY2022410C002).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy reasons.

Acknowledgments: We sincerely thank all contributors of the open-source datasets used in this study. We also appreciate the support and funding from the Educational Department of Liaoning Province and the State Administration of Science, Technology and Industry for National Defence for our work.

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

References

1. Si, J.; Song, B.; Wu, J.; Lin, W.; Huang, W.; Chen, S. Maritime ship detection method for satellite images based on multiscale feature fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6642–6655. [CrossRef]
2. Zhang, D.; Wang, C.; Fu, Q. Ofcos: An oriented anchor-free detector for ship detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6004005. [CrossRef]
3. Wang, P.; Liu, B.; Li, Y.; Chen, P.; Liu, P. IceRegionShip: Optical Remote Sensing Dataset for Ship Detection in Ice-Infested Waters. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1007–1020. [CrossRef]
4. Zhang, Y.; Lu, D.; Qiu, X.; Li, F. Scattering point topology for few-shot ship classification in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 10326–10343. [CrossRef]
5. Yin, Y.; Cheng, X.; Shi, F.; Liu, X.; Huo, H.; Chen, S. High-order spatial interactions enhanced lightweight model for optical remote sensing image-based small ship detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4201416. [CrossRef]
6. Yuan, Y.; Rao, Z.; Lin, C.; Huang, Y.; Ding, X. Adaptive ship detection from optical to SAR images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3508205. [CrossRef]
7. Zhao, J.; Chen, Y.; Zhou, Z.; Zhao, J.; Wang, S.; Chen, X. Multiship speed measurement method based on machine vision and drone images. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2513112. [CrossRef]
8. Zhao, J.; Shi, B.; Huang, T. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE Trans. Comput. Imaging* **2021**, *8*, 12–27. [CrossRef]
9. Huang, Q.; Sun, H.; Wang, Y.; Yuan, Y.; Guo, X.; Gao, Q. Ship detection based on YOLO algorithm for visible images. *IET Image Process.* **2023**, *18*, 481–492. [CrossRef]
10. Yang, D.; Solihin, M.I.; Zhao, Y.; Yao, B.; Chen, C.; Cai, B.; Machmudah, A. A review of intelligent ship marine object detection based on RGB camera. *IET Image Process.* **2023**, *18*, 281–297. [CrossRef]
11. Assani, N.; Matić, P.; Kaštelan, N.; Čavka, I.R. A review of artificial neural networks applications in maritime industry. *IEEE Access* **2023**, *11*, 139823–139848. [CrossRef]
12. Er, M.J.; Zhang, Y.; Chen, J.; Gao, W. Ship detection with deep learning: A survey. *Artif. Intell. Rev.* **2023**, *56*, 11825–11865. [CrossRef]
13. Vaswani, A. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5999–6099.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Maurício, J.; Domingues, I.; Bernardino, J. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]

16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
17. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
18. Ren, Z.; Tang, Y.; Yang, Y.; Zhang, W. SASOD: Saliency-Aware Ship Object Detection in High-Resolution Optical Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611115. [CrossRef]
19. Zhang, J.; Xing, M.; Sun, G.C.; Li, N. Oriented Gaussian function-based box boundary-aware vectors for oriented ship detection in multiresolution SAR imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5211015. [CrossRef]
20. Hu, Q.; Hu, S.; Liu, S.; Xu, S.; Zhang, Y.D. FINet: A feature interaction network for SAR ship object-level and pixel-level detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5239215. [CrossRef]
21. Yu, H.; Yang, S.; Zhou, S.; Sun, Y. Vs-lsdet: A multiscale ship detector for spaceborne sar images based on visual saliency and lightweight cnn. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1137–1154. [CrossRef]
22. Guo, H.; Yang, X.; Wang, N.; Gao, X. A CenterNet++ model for ship detection in SAR images. *Pattern Recognit.* **2021**, *112*, 107787. [CrossRef]
23. Leng, X.; Wang, J.; Ji, K.; Kuang, G. Ship detection in range-compressed SAR data. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2135–2138.
24. Zhu, M.; Hu, G.; Zhou, H.; Wang, S.; Feng, Z.; Yue, S. A ship detection method via redesigned FCOS in large-scale SAR images. *Remote Sens.* **2022**, *14*, 1153. [CrossRef]
25. Zhang, C.; Liu, P.; Wang, H.; Jin, Y. Saliency-based centernet for ship detection in sar images. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 1552–1555.
26. Leng, X.; Ji, K.; Kuang, G. Ship detection from raw SAR echo data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5207811. [CrossRef]
27. Zhang, Y.; Lu, D.; Qiu, X.; Li, F. Scattering-point-guided RPN for oriented ship detection in SAR images. *Remote Sens.* **2023**, *15*, 1411. [CrossRef]
28. Ren, Z.; Tang, Y.; He, Z.; Tian, L.; Yang, Y.; Zhang, W. Ship detection in high-resolution optical remote sensing images aided by saliency information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623616. [CrossRef]
29. Li, X.; Li, Z.; Lv, S.; Cao, J.; Pan, M.; Ma, Q.; Yu, H. Ship detection of optical remote sensing image in multiple scenes. *Int. J. Remote Sens.* **2021**, *43*, 5709–5737. [CrossRef]
30. Wang, Z.; Zhou, Y.; Wang, F.; Wang, S.; Xu, Z. SDGH-Net: Ship detection in optical remote sensing images based on Gaussian heatmap regression. *Remote Sens.* **2021**, *13*, 499. [CrossRef]
31. Dong, Y.; Chen, F.; Han, S.; Liu, H. Ship object detection of remote sensing image based on visual attention. *Remote Sens.* **2021**, *13*, 3192. [CrossRef]
32. Hu, J.; Zhi, X.; Jiang, S.; Tang, H.; Zhang, W.; Bruzzone, L. Supervised multi-scale attention-guided ship detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5630514. [CrossRef]
33. Xiao, S.; Zhang, Y.; Chang, X. Ship detection based on compressive sensing measurements of optical remote sensing scenes. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8632–8649. [CrossRef]
34. Cui, Z.; Leng, J.; Liu, Y.; Zhang, T.; Quan, P.; Zhao, W. SKNet: Detecting rotated ships as keypoints in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8826–8840. [CrossRef]
35. Zheng, J.; Liu, Y. A study on small-scale ship detection based on attention mechanism. *IEEE Access* **2022**, *10*, 77940–77949. [CrossRef]
36. Ngo, D.D.; Vo, V.L.; Nguyen, T.; Nguyen, M.H.; Le, M.H. Image-based ship detection using deep variational information bottleneck. *Sensors* **2023**, *23*, 8093. [CrossRef] [PubMed]
37. Liu, T.; Zhang, Z.; Lei, Z.; Huo, Y.; Wang, S.; Zhao, J.; Zhang, J.; Jin, X.; Zhang, X. An approach to ship target detection based on combined optimization model of dehazing and detection. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107332. [CrossRef]
38. Zhou, W.; Peng, Y. Ship detection based on multi-scale weighted fusion. *Displays* **2023**, *78*, 102448. . [CrossRef]
39. Yi, Y.; Ni, F.; Ma, Y.; Zhu, X.; Qi, Y.; Qiu, R.; Zhao, S.; Li, F.; Wang, Y. High Performance Gesture Recognition via Effective and Efficient Temporal Modeling. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, China, 10–16 August 2019; pp. 1003–1009.
40. Jiang, S.; Zhang, H.; Qi, Y.; Liu, Q. Spatial-temporal interleaved network for efficient action recognition. *IEEE Trans. Industr. Inform.* 2024, 1–10, *Early Access*. [CrossRef]
41. Zheng, Y.; Liu, P.; Qian, L.; Qin, S.; Liu, X.; Ma, Y.; Cheng, G. Recognition and depth estimation of ships based on binocular stereo vision. *J. Mar. Sci. Eng.* **2022**, *10*, 1153. [CrossRef]
42. Shi, H.; Chai, B.; Wang, Y.; Chen, L. A local-sparse-information-aggregation transformer with explicit contour guidance for SAR ship detection. *Remote Sens.* **2022**, *14*, 5247. [CrossRef]
43. Zhang, Y.; Er, M.J.; Gao, W.; Wu, J. High performance ship detection via transformer and feature distillation. In Proceedings of the 2022 5th International Conference on Intelligent Autonomous Systems (ICoIAS), Dalian, China, 23–25 September 2022; pp. 31–36.
44. Chen, B.; Yu, C.; Zhao, S.; Song, H. An anchor-free method based on transformers and adaptive features for arbitrarily oriented ship detection in SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 2012–2028. [CrossRef]

45. Zhou, Y.; Jiang, X.; Xu, G.; Yang, X.; Liu, X.; Li, Z. PVT-SAR: An arbitrarily oriented SAR ship detector with pyramid vision transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 291–305. [CrossRef]
46. Chen, Y.; Xia, Z.; Liu, J.; Wu, C. TSDet: End-to-end method with transformer for SAR ship detection. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8.
47. Chen, W.; Hong, D.; Qi, Y.; Han, Z.; Wang, S.; Qing, L.; Huang, Q.; Li, G. Multi-attention network for compressed video referring object segmentation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 4416–4425.
48. Ge, C.; Song, Y.; Ma, C.; Qi, Y.; Luo, P. Rethinking attentive object detection via neural attention learning. *IEEE Trans. Image Process.* **2023**, *33*, 1726–1739. [CrossRef]
49. Phan, V.M.H.; Xie, Y.; Zhang, B.; Qi, Y.; Liao, Z.; Perperidis, A.; Phung, S.L.; Verjans, J.W.; To, M.S. Structural attention: Rethinking transformer for unpaired medical image synthesis. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Marrakesh, Morocco, 6–10 October 2024; pp. 690–700.
50. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
51. Ma, Y.; Zhu, Z.; Qi, Y.; Beheshti, A.; Li, Y.; Qing, L.; Li, G. Style-aware two-stage learning framework for video captioning. *Knowl. Based Syst.* **2024**, *301*, 112258. [CrossRef]
52. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. Seaships: A large-scale precisely annotated dataset for ship detection. *IEEE Trans. Multimedia* **2018**, *20*, 2593–2604. [CrossRef]
53. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
54. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
55. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13039–13048.
56. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the Ninth International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
57. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3490–3499.
58. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An empirical study of designing real-time object detectors. *arXiv* **2022**, arXiv:2212.07784.
59. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10965–10975.
60. Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; Chen, K. Dense distinct query for end-to-end object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7329–7338.
61. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
62. Zong, Z.; Song, G.; Liu, Y. DETRs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 6748–6758.
63. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
64. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.