remote sensing

MDPI

Article

An Enhanced Shuffle Attention with Context Decoupling Head with Wise IoU Loss for SAR Ship Detection

Yunshan Tang [1,2], Yue Zhang [1,2], Jiarong Xiao [1], Yue Cao [1] and Zhongjun Yu [1,2,*]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: yuzj@ucas.ac.cn

Abstract: Synthetic Aperture Radar (SAR) imagery is widely utilized in military and civilian applications. Recent deep learning advancements have led to improved ship detection algorithms, enhancing accuracy and speed over traditional Constant False-Alarm Rate (CFAR) methods. However, challenges remain with complex backgrounds and multi-scale ship targets amidst significant interference. This paper introduces a novel method that features a context-based decoupled head, leveraging positioning and semantic information, and incorporates shuffle attention to enhance feature map interpretation. Additionally, we propose a new loss function with a dynamic non-monotonic focus mechanism to tackle these issues. Experimental results on the HRSID and SAR-Ship-Dataset demonstrate that our approach significantly improves detection performance over the original YOLOv5 algorithm and other existing methods.

Keywords: ship detection; synthetic aperture radar (SAR); decoupled head; attention mechanism; YOLOv5

Citation: Tang, Y.; Zhang, Y.; Xiao, J.; Cao, Y.; Yu, Z. An Enhanced Shuffle Attention with Context Decoupling Head with Wise IoU Loss for SAR Ship Detection. *Remote Sens.* **2024**, *16*, 4128. https://doi.org/10.3390/rs16224128

Academic Editor: Domenico Velotto

Received: 20 September 2024
Revised: 2 November 2024
Accepted: 4 November 2024
Published: 5 November 2024

1. Introduction

Synthetic Aperture Radar (SAR) is a microwave sensor that is unaffected by external environmental factors such as clouds, fog, snow, and night situations. It is capable of continuously monitoring local terrain scenes, possessing strong penetration capabilities and high-resolution imaging characteristics, enabling accurate detection of obscured or camouflaged targets [1]. It finds widespread applications in civilian and military sectors including topographic mapping, disaster assessment, environmental monitoring, target reconnaissance, and target localization. Among these applications, marine target detection is a significant subdivision of SAR object detection, with ship target detection being a primary focus within marine target detection.

In traditional ship detection algorithms, CFAR [2,3] and other adaptive algorithms are widely utilized due to their capability of adaptively scanning images. The CFAR method analyzes input noise to establish thresholds, thereby identifying the presence of a target when the energy of the input signal surpasses these thresholds. To cater to the diverse requirements of various SAR image applications, multiple statistical models have been proposed, encompassing Gaussian, gamma, Weibull, log-normal, G0, and K distributions [4,5]. Moreover, enhancements and variations of CFAR algorithms continually emerge [6–8]. Nevertheless, these approaches often require the manual configuration of features, which is laborious, and exhibit limited transfer ability. While these methods excel in scenarios involving single-class ships and locally uniform background noise, their efficacy wanes in scenarios such as nearshore ship detection with intense interference, as well as multi-scale ship detection [9,10]. Additionally, they lack the capability to process targets end-to-end. Hence, there exists an imperative need for more sophisticated and robust algorithms to tackle these challenges.

After AlexNet [11] achieved significant acclaim in the 2012 ImageNet competition, convolutional neural networks (CNNs) have seen a resurgence in importance within the domain of image processing. Represented by R-CNN [12], CNN-based algorithms have been employed in object detection, pioneering the development of two-stage object detection. Subsequent advancements such as SPPNet [13], Fast R-CNN [14], and Faster R-CNN [15] have further refined two-stage detection algorithms, achieving real-time processing improvements in both accuracy and speed. The evolution of two-stage detection algorithms has led to the emergence of models such as Feature Pyramid Networks (FPNs) [16], Cascade R-CNN [17], Mask R-CNN [18], and Libra R-CNN [19], among others [20]. The two-stage algorithm first proposes a region proposal, then proceeds to classify it and refine the bounding box through the subsequent stage network. While more accurate than one-stage algorithms, it suffers from much slower processing speeds.

The two-stage algorithms still face bottlenecks in speed, and there is still a certain gap in real-time image object detection. Addressing such issues, the You Only Look Once (YOLO) [21] algorithm was proposed. As the pioneering work of single-stage detection algorithms, it no longer needs to generate region proposals and process them in two steps, but directly produces the output results for bounding boxes and class, achieving a nearly 10-fold speedup compared to the previous two-stage algorithms. Wei Liu proposed Single Shot MultiBox Detector (SSD) [22], which introduces the concept of multi-scale and multi-resolution detection. Subsequently, the YOLOv2 [23] and YOLOv3 [24] algorithms address the poor accuracy issue of single-stage algorithms by incorporating ideas such as multi-box detection, feature fusion, and multi-scale outputs into the network. While maintaining fast processing speeds, these enhancements lead to a significant increase in accuracy. Following RetinaNet [25], single-stage networks have surpassed the accuracy of the best two-stage object detection networks at the time. CornerNet [26] and CenterNet [27] further introduce the concepts of corner points and center points in deep learning. YOLOv4 [28] integrates numerous contemporary ideas such as Complement IoU (CIoU) [29], PANet [30], and Mix up data augmentation [31] to achieve both fast processing speeds and higher accuracy in object detection algorithms. YOLOX [32] introduces decoupled head into object detection, achieving better results on top of existing algorithms. This paper selects the YOLOv5 [33] framework as the baseline for experimentation.

While existing networks have achieved good results in optical images, there are still notable cases of false alarms and missed detections in SAR ship detection, particularly in scenarios with strong interference near shorelines and in situations involving multi-scale and small targets, as depicted in Figure 1. Therefore, there is an urgent need for algorithmic improvements tailored to SAR images.

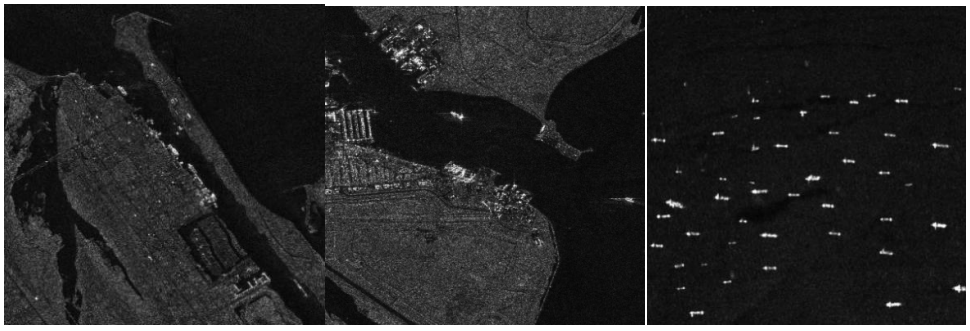


Figure 1. Several typical examples of situations with small vessel targets and an inshore background.

With the introduction of the SAR Ship Detection Dataset (SSDD) [34] and the emergence of more SAR target detection datasets [35,36], a plethora of papers on SAR domain object detection have been proposed [37]. The earliest works typically employed classical networks such as Faster R-CNN [34], SSD [38], and YOLOv2 [39], without improvements specifically tailored to SAR ship target problems, resulting in a relatively mediocre performance.

Attention mechanisms, by weighting key feature maps and spatial regions of importance, are commonly employed for the deep mining of multi-scale and small object information, serving as a means to address targets in complex nearshore scenes effectively [40–50].

In earlier endeavors, the integration of the Squeeze and Excitation (SE) attention mechanism with Faster R-CNN has demonstrated excellent detection results on the early version of the SSDD dataset [40]. Zhao et al. [41] proposed utilizing the Convolutional Block Attention Module (CBAM) and Receptive Fields Block (RFB) to address detection and recognition challenges on top of YOLOv5. Wang et al. [42] introduced the sim attention mechanism and C3 channel shuffling to tackle multi-scale ship detection issues in complex scenarios. Li et al. [43] presented coordinate attention to enhance the performance of detecting small objects. Tang et al. [44] devised a Multiscale Receptive Field Convolution Block with Attention Mechanism (AMMRF) to leverage positional information in feature maps, accurately capturing regions crucial for detection in feature maps, as well as capturing relationships between feature map channels to better understand the ship–background dynamics. A study [45] proposed the United Attention Module (UAM) and Global Context-guided Feature Balanced Pyramid (GC-FBP) to enhance ship detection performance. Wu et al. [46] introduced a method based on the coordinate attention (CA) mechanism and Asymptotic Feature Fusion (AFF) to alleviate the problem of small object position loss and enhance the model's ability to detect multi-scale targets. Hu et al. [47] put forward a Balance Attention Network (BANet), employing both Local Attention Module (LAM) and Non-Local Attention Module (NLAM) to respectively capture the local information of ships, strengthen network robustness, and equilibrate local and non-local features. Ren [48] proposed incorporating the Channel and Position Enhancement Attention (CPEA) module to enhance the precision of target localization by utilizing positional data. DSF-Net [49] incorporated the Pixel-wise Shuffle Attention module (PWSA) to boost feature extraction capabilities and employed Non-Local Shuffle Attention (NLSA) to enhance the long-term dependency of features, thereby promoting information exchange. Cui et al. [50] proposed the addition of a Spatial Shuffle-Group Enhance (SSE) attention module to the CenterNet network to enhance its performance. Cai et al. [51] introduced FS-YOLO, which incorporates a Feature Enhancement Module (FEM) and a Spatial Channel Pooling Module (ESPPCSPC) on top of the original YOLO backbone, thereby improving network performance. Wang et al. [52] integrated the Global Context-Aware Subimage Selection (GCSS) module with the Local Context-Aware False Alarms Suppression (LCFS) module to enhance the network's adaptability to duplicated scenes. Cheng et al. [53] improved the YOLOX backbone by proposing the S2D network, which better integrates information from the neck component and enhances the network's performance in detecting small objects. Additionally, Zhang et al. [54] discovered the modulation effects of target motion on polarization and Doppler. Meanwhile, Gao et al. [55] employed the dualistic cascade convolutional method to enhance the performance of ship target detection.

Many papers have also focused on improving the loss function to enhance object detection performance. Zhang et al. [56] introduced the center loss to ensure an equitable allocation of loss contributions among different factors and reduce the sensitivity of object detection to changes in ground truth box shapes. YOLO-Lite [48] utilized a confidence loss function to improve the accuracy of ship object detection. DSF-Net [49] employed an R-tradeoff loss to improve small detects, accelerate training efficiency, and reduce false positive rates. Zhou [57] developed a loss function that employs a dual Euclidean distance approach, leveraging the corner coordinates of predicted and ground truth boxes, which accurately describes various overlapping scenarios. Zhang [58] used global average precision loss (GAP loss) to enable the model to quickly differentiate between positive and negative samples to enhance accuracy. The paper [59] utilized a KLD loss function to improve accuracy. Chen [60] used the SIoU loss to aid the training process of the network.

These loss functions enhance the detection capability for small objects to some degree, accelerate training convergence, and elevate accuracy. However, they do not consider

the impediment caused by inferior instances to the learning ability of the object detection model, resulting in limited performance improvement.

Many articles have also explored the use of decoupled heads [43,47,61] to decouple the semantic information head and bounding box information head, preventing interference between different features and achieving better results. However, these simple decoupled heads only provide limited performance improvements as they do not consider the differences in semantic and bounding box information.

Therefore, in this paper, based on the YOLOv5 backbone, we propose the SAR Ship Context Decoupled Head (SSCDH), which is based on the characteristics of localization and semantic information. We use shuffle attention to enhance the focus on understanding complex backgrounds. Additionally, we introduce a new Wise IoU loss grounded in a dynamic non-monotonic focus framework and designed to utilize the degree of anomaly. The goal is to improve the accuracy of ship detection. Hence, the primary advancements of this paper include the following:

1. In order to enhance the effectiveness of the original decoupling head model, we design dedicated decoupling heads that align with the specific characteristics of positioning and semantic information.
2. To improve the model's capability in detecting objects of varying scales, we incorporate a shuffle attention module into the larger feature layers of the original model's neck.
3. To boost the accuracy of object detection, we utilize the Wise IoU loss function, which leverages attention-based bounding box regression loss and a dynamic non-monotonic focus mechanism.
4. To demonstrate the effectiveness of the proposed technique, we conduct extensive experiments using the HRSID dataset and the SAR-Ship-Dataset.

The first part of this paper served as an introduction, which presents the background, related works pertinent to this study, and the identified issues. The second part focuses on the methods, describing the network structure and the design approach for each module. The third part presents the experimental details and results. The fourth part discusses the effectiveness of our chosen head and attention mechanism. Finally, the fifth part concludes the entire paper.

2. Methods

This section introduces the method of the proposed SSCDH. The first part provides an overview of the architecture of the proposed model. The second part discusses the shuffle attention module utilized in our model, along with its principles of spatial and channel attention mechanisms. The third part introduces the decoupled heads based on contextual information from ships. Lastly, the fourth part describes the Wise IoU loss function employed.

2.1. Network Architecture

The network is based on YOLOv5 architecture [33]. The overall structure of the proposed method is shown in Figure 2. The input RGB image size is $H \times W \times 3$, where H represents the height of the image and W represents the width of the image. The input image passes through 1 large convolutional module and 2 convolutional and residual convolutional modules, resulting in a feature map of size $\frac{H}{8} \times \frac{W}{8} \times 256$ after 3 downsampling operations. Subsequently, another convolutional and residual module produces a feature map of size $\frac{H}{16} \times \frac{W}{16} \times 512$, followed by another similar module yielding a feature map of size $\frac{H}{32} \times \frac{W}{32} \times 1024$. These feature maps are then forwarded to the SPP bottleneck module and subsequently to the neck module, still retaining the dimensions $\frac{H}{32} \times \frac{W}{32} \times 1024$.

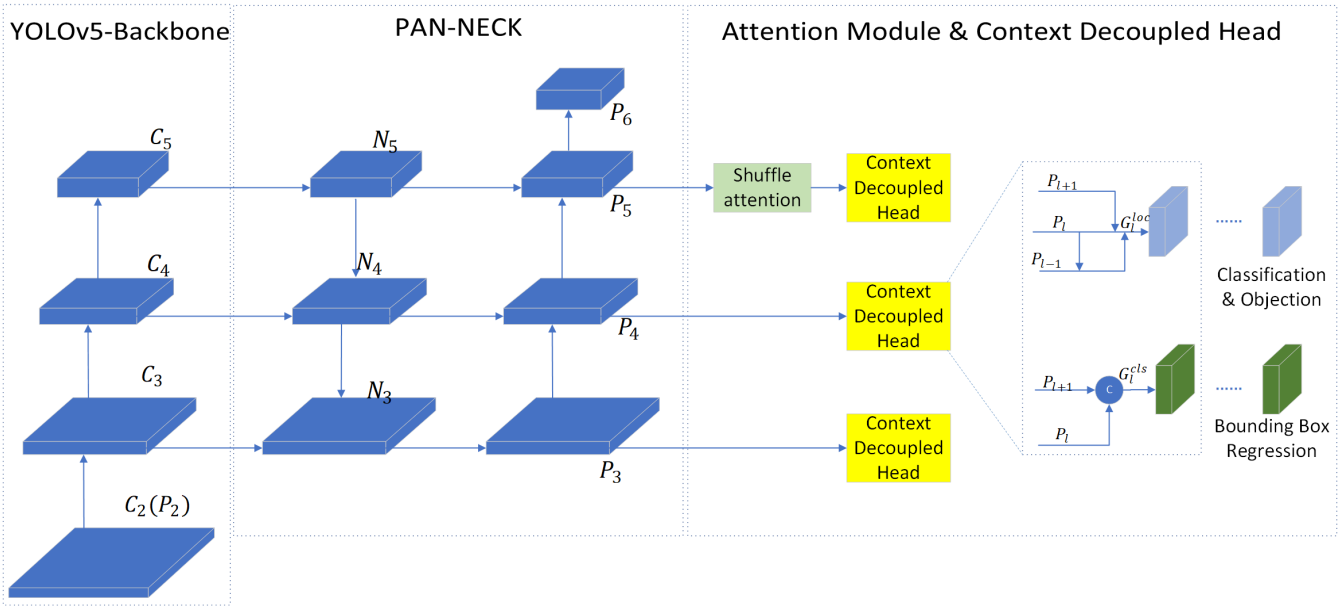


Figure 2. Overview of the proposed method's structure. We used the backbone of YOLOv5 and neck of PAN for the network, while the shuffle attention module and Context Decoupled Head added in this paper are in the Attention Module and Context Decoupled Head part of this figure.

The feature map of size $\frac{H}{32} \times \frac{W}{32} \times 1024$ is processed through a 512-channel 1×1 convolutional layer, resulting in a feature map of size $\frac{H}{32} \times \frac{W}{32} \times 512$. This is then upsampled twice and concatenated with another feature map. The feature map obtained after the first upsampling, $\frac{H}{16} \times \frac{W}{16} \times 512$, is concatenated with the feature map from the backbone, resulting in a feature map of size $\frac{H}{16} \times \frac{W}{16} \times 1024$. This is followed by another convolutional layer to obtain a feature map measuring $\frac{H}{16} \times \frac{W}{16} \times 512$, which is then upsampled to obtain a feature map of size $\frac{H}{8} \times \frac{W}{8} \times 512$. A 256-channel 1×1 convolution is applied to obtain the P_3 feature map of size $\frac{H}{8} \times \frac{W}{8} \times 256$.

Additionally, the feature map of size $\frac{H}{8} \times \frac{W}{8} \times 256$ undergoes downsampling using a 256-channel convolution with a kernel size of 3, padding of 1, and a stride of 2, resulting in a feature map of size $\frac{H}{16} \times \frac{W}{16} \times 256$. This is concatenated with the output of the second convolution, resulting in a feature map of size $\frac{H}{16} \times \frac{W}{16} \times 512$, which is then passed through a convolutional residual block to obtain the P_4 feature map measuring $\frac{H}{16} \times \frac{W}{16} \times 512$.

Similarly, the feature map of size $\frac{H}{16} \times \frac{W}{16} \times 512$ undergoes downsampling using a 512-channel convolution with a kernel size of 3, padding of 1, and a stride of 2, resulting in a feature map of size $\frac{H}{32} \times \frac{W}{32} \times 512$. This is concatenated with the output of the second convolution, resulting in a feature map measuring $\frac{H}{32} \times \frac{W}{32} \times 1024$. Another convolutional residual block is applied to obtain the feature map P_5 measuring $\frac{H}{32} \times \frac{W}{32} \times 1024$. A shuffle attention module is then applied to this feature map to enhance feature extraction.

Subsequently, the model undergoes another convolution operation with 1024 channels, a stride of 2, a kernel dimension of 3, along with a padding of 1. The generated feature map is then directed to the next C3 module, yielding the feature map P_6 of size $\frac{H}{64} \times \frac{W}{64} \times 1024$.

Finally, the SAR Ship Context Decoupled Head is utilized to fuse features from multiple hierarchical levels. The feature map measuring $\frac{H}{4} \times \frac{W}{4} \times 128$ obtained after the second downsampling is used as P_2, the feature map. Consequently, $P_3^{'}$ is derived by incorporating information from P_2, P_3, and P_4 feature maps. Similarly, $P_4^{'}$ incorporates information from P_3, P_4, and P_5 feature maps, and $P_5^{'}$ incorporates information from P_4, P_5, and P_6 feature maps. This process ultimately yields the final bounding box positions and confidence scores for target classification.

2.2. Shuffle Attention Module

The application of the SE [62] mechanism considers the crucial role of channel attention in target recognition and detection, which has found widespread application in object detection. CBAM [63] combines both channel attention and spatial attention mechanisms, resulting in a notable enhancement in the accuracy of computation. The shuffle attention (SA) module [64] also integrates channel attention and spatial attention mechanisms while incorporating the concept of group convolutional kernel channel rearrangement. This achieves superior results compared to other attention mechanisms. In this proposed method, we chose to integrate the shuffle attention component after 32× downsampling layers, aiming to elevate the understanding of the semantic and channel information for the final layer, thereby achieving more accurate detection capabilities for complex scenes, small targets, and multi-scale objects. Figure 3 illustrates the shuffle attention process framework.

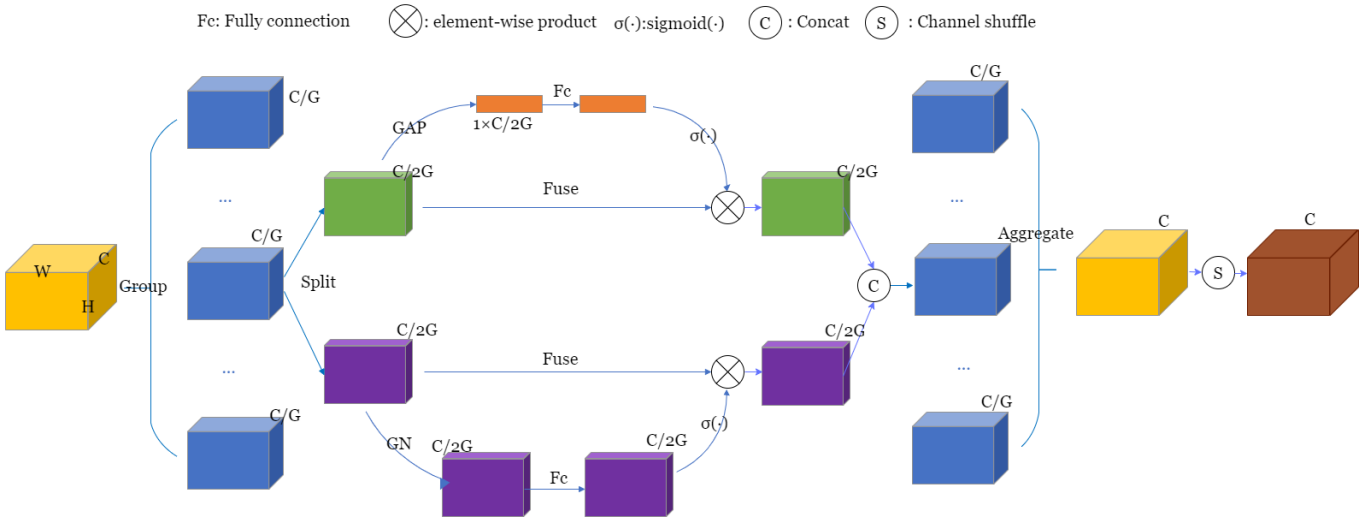


Figure 3. The structure of the shuffle attention process.

First, shuffle attention employs "channel partitioning" to concurrently process sub-features for each group. Next, in the channel attention pathway, global average pooling is utilized to compute statistics at the channel level. This is followed by the application of a pair of parameters to adjust the scaling and shifting of the channel vectors. For the spatial attention pathway, group normalization (GN) is utilized to derive statistics at the spatial level, resulting in a condensed feature representation similar to that of the channel pathway. Then, these two pathways are combined. Following this, all the derived sub-features are consolidated and, ultimately, the channel shuffle technique is applied to enhance the data exchange between the various sub-features.

Shuffle attention achieves the grouping of features, initially, by partitioning the feature maps of a given size $C \times H \times W$ into G groups. Here, C indicates the total number of channels, while H signifies the vertical dimension of the feature map, and W corresponds to its horizontal dimension. Specifically, shuffle attention divides the feature maps of $\boldsymbol{X}$ as G clusters, denoted as $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G]$, where each $\boldsymbol{X}_k$ is of the size $\frac{C}{G} \times H \times W$. Consequently, during training, every individual component map $\boldsymbol{X}_k$ progressively captures different interpretive insights.

Subsequently, an attention module is used to generate the corresponding significance weights for each component map. In detail, each attention unit processes the input feature map $\boldsymbol{X}_k$ by splitting it into two separate pathways, denoted as $\boldsymbol{X}_{k1}$ and $\boldsymbol{X}_{k2}$, each of size $\frac{C}{2G} \times H \times W$. One branch, $\boldsymbol{X}_{k1}$, is used to create channel attention maps using connections between channels to improve channel effectiveness. Meanwhile, the other branch, $\boldsymbol{X}_{k2}$, produces spatial attention maps using connections between spatial features to identify more useful spatial characteristics.

First, we extract channel-level statistical information from the input $\boldsymbol{X}_{k1}$ by utilizing global average pooling (GAP), embedding global information into $\boldsymbol{s}$ of size $\frac{C}{2G} \times 1 \times 1$. $\boldsymbol{s}$ can be obtained by performing spatial average pooling with dimensions $H \times W$, defined as

$$\boldsymbol{s} = \mathrm{GAP}(\boldsymbol{X}_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{X}_{k1}(i, j). \tag{1}$$

Next, employing a basic gating function combined with a Sigmoid activation, we construct a compact feature to precisely and adaptively select. The ultimate result of channel attention $\boldsymbol{X}_{k1}^{'}$ can be derived as follows:

$$\boldsymbol{X}_{k1}^{'} = \sigma(F_c(\boldsymbol{s}))\boldsymbol{X}_{k1} = \sigma(\boldsymbol{W}_1\boldsymbol{s} + \boldsymbol{b}_1)\boldsymbol{X}_{k1}, \tag{2}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ are parameters of size $\frac{C}{2G} \times 1 \times 1$ and used for the fully connected and bias term $\boldsymbol{s}$, $F_c(\cdot)$ represents the full collection operation, and $\sigma(\cdot)$ represents the Sigmoid activation function.

Simultaneously, we process data to obtain spatial-level statistical information, enhancing the representation through a Group Norm (GN) operation. The ultimate result of spatial attention can be derived as follows:

$$\boldsymbol{X}_{k2}^{'} = \sigma(\boldsymbol{W}_2\mathrm{GN}(\boldsymbol{X}_{k2}) + \boldsymbol{b}_2)\boldsymbol{X}_{k2}, \tag{3}$$

where $\boldsymbol{W}_2$ and $\boldsymbol{b}_2$ are parameters of size $\frac{C}{2G} \times 1 \times 1$.

Finally, we merge the passways of the channel and spatial attention to obtain the output of the same size as the input, $\boldsymbol{X}_k^{'} = [\boldsymbol{X}_{k1}^{'}, \boldsymbol{X}_{k2}^{'}]$, with dimensions $\frac{C}{G} \times H \times W$. Subsequently, all components are aggregated. Lastly, we employ a "channel shuffle" that enhances the flow of information between groups across channel dimensions. The final output of the SA module matches the size of the input $\boldsymbol{X}$.

2.3. SAR Ship Context Decoupled Head

The preference inconsistency towards feature context between classification and localization is strong. Specifically, localization tends to emphasize boundary features for accurate bounding box regression, whereas object classification leans towards semantic context. Existing methods like YOLOX utilize decoupled heads to handle different feature contexts for various tasks. However, since these heads work with the same input features, there is an imbalance between classification and localization.

Based on the structure and principles of Task-Specific Context Decoupling (TSCODE) [65], we separately manage the encoding of features for categorization and positioning, known as context decoupling, to selectively employ more suitable semantic contexts for specific tasks. For the classification branch, rich semantic contextual features present in the image are typically required to infer object categories. Therefore, we use feature encoding that is broad but captures strong semantic details. For the localization branch, which requires precise boundary information, we offer high-resolution feature maps to better define object edges.

While classification in object detection is less detailed and focuses on identifying objects within a bounding box, using downsampled feature maps for classification does not significantly impact performance but does lower computational costs. On the other hand, object categories can be inferred from their surrounding environments; for instance, ship targets are likely to appear on the sea surface or docked at port edges. Employing broad insights derived from detailed semantic information improves classification performance.

Building on these findings, we developed Semantic Context Encoding (SCE) to enhance classification efficiency and accuracy. As illustrated in Figure 4, SCE uses two levels of feature maps, P_l and P_{l+1}, at each pyramid level l to produce a feature map with rich semantic information for classification.

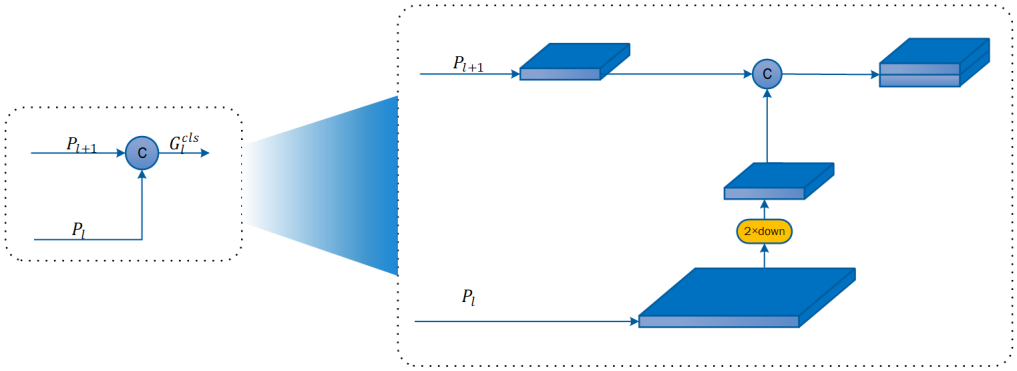


Figure 4. Semantic Context Encoding (SCE).

Initially, we downsample P_l by a factor of two and then concatenate it with P_{l+1}, to yield the final classification feature map, G_l^{cls}:

$$G_l^{cls} = \text{Concat}(\text{DConv}(P_l), P_{l+1}), \tag{4}$$

where Concat((•)) signifies a concatenation operation, and DConv(•) refers to a shared convolutional layer used for downsampling. It is noteworthy that the resolution of G_l^{cls} is half of P_l.

Subsequently, G_l^{cls} is passed through to $F_c(\cdot) = \{f_{cls}(\cdot), C(\cdot)\}$ to predict classification scores, where $f_{cls}(\cdot)$ represents the classification loss function and $C(\cdot)$ represents further classification and the Objection Operation. We employ $f_{cls}(\cdot)$, consisting of two convolutional layers with 512 channels. Given that G_l^{cls} is downsampled by a factor of 2 compared to P_l, at each position (x, y) in G_l^{cls}, the predicted classification scores of its four nearest neighbors in P_l are computed, denoted as $\tilde{C} \in R^{H_{l+1} \times W_{l+1} \times 4N}$, where N is the number of classes, and H_{l+1} and W_{l+1} represent the height and width of the feature map. Subsequently, $\tilde{C}$ is reshaped to $\tilde{C} \in R^{H_l \times W_l \times N}$ to recover the resolution

$$\tilde{C}[2x + i, 2y + j, c] = \tilde{C}[x, y, (2i + 2j)c], \forall i, j \in \{0, 1\}. \tag{5}$$

This approach not only leverages the sparse key features from P_l but also incorporates the rich semantic information from higher levels on the pyramid as P_{l+1}.

Localization is more complex than classification, needing additional details for keypoint prediction. Methods usually use a one-scale feature map P_l, though lower pyramid levels often have stronger responses to object contours, edges, and fine textures. Nevertheless, higher-level feature maps are crucial for localization as they facilitate the comprehensive observation of the entire object, thus giving more details to understand the complete shape of the object.

Based on these findings, we recommend Detail Preserving Encoding (DPE) for accurate localization. At each layer l of the pyramid, our DPE integrates feature maps from three layers: P_{l-1}, P_l, and P_{l+1}. P_{l-1} supplies detailed edge features, whereas P_{l+1} gives a broader object view.

Figure 5 shows the DPE structure. The feature map on P_l is first upsampled by a factor of 2 and then aggregated with P_{l-1}. Subsequently, it is downsampled to the resolution of P_l through a 3 × 3 convolutional layer with a stride of 2. Finally, P_{l+1} is upsampled and combined to produce the final classification feature map, G_l^{loc}. The computation process is as follows:

$$G_l^{loc} = P_l + \mu(P_{l+1}) + \text{DConv}(P_{l-1} + \mu(P_l)). \tag{6}$$

Here, $\mu(\bullet)$ signifies upsampling, while DConv(•) indicates a shared convolutional layer for downsampling. Specifically, we compute G_3^{loc} using C_2, P_3, and P_4. Subsequently, further bounding box predictions at the l-th pyramid level are performed through $F_r(\cdot) =$

$\{f_{los}(\cdot), R(\cdot)\}$, where $f_{los}(\cdot)$ represents the locational loss function and $R(\cdot)$ represents the further bounding box regression operation.

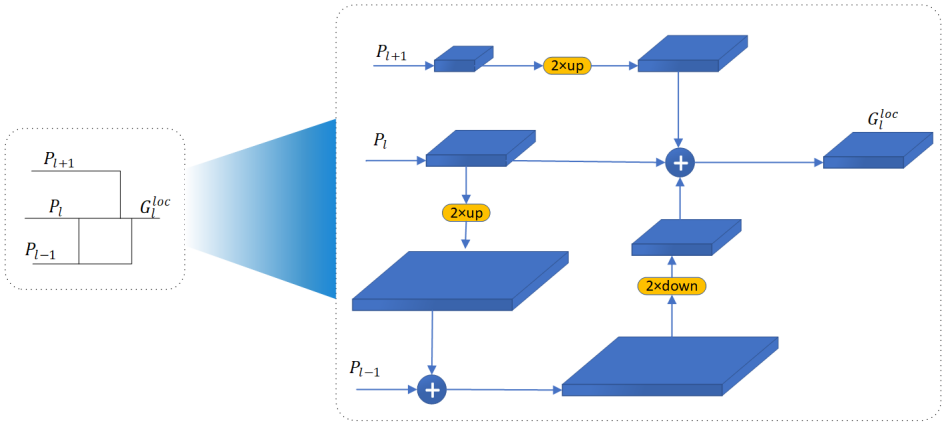


Figure 5. Detail Preserving Encoding (DPE).

2.4. Wise IoU Loss

In the field of object detection, Intersection over Union (IoU) evaluates the overlap between anchor boxes and target boxes. Compared to employing the norm as the bounding box loss function, IoU loss effectively mitigates interference from the proportional representation of bounding box sizes, which allows the model to efficiently balance learning for both large and small objects when IoU loss is utilized for bounding box regression. IoU loss is defined as

$$L_{IoU} = 1 - \text{IoU}. \tag{7}$$

However, when IoU is zero (i.e., $W_i = 0$ or $H_i = 0$), the gradient of the IoU loss $\frac{\partial L_{IoU}}{\partial W_i} = 0$, resulting in the disappearance of gradients during back-propagation and the failure to update the overlapping distance W_i.

To address this issue, existing research accounts for various geometric aspects of bounding boxes and incorporates a penalty term R_i. The existing bounding box regression (BBR) loss follows the paradigm

$$L_i = L_{IoU} + R_i. \tag{8}$$

The Generalized Intersection over Union (GIoU) loss function extends the standard IoU loss by incorporating a penalty term. Unlike traditional IoU, which only assesses the overlap between boxes, GIoU also evaluates the surrounding non-overlapping regions. However, when one box is fully enclosed within another, GIoU cannot differentiate its relative positional relationships.

To address the limitations of GIoU, Distance-IoU (DIoU) [29] adjusts the penalty term by maximizing the overlap area. This is achieved through minimizing the normalized distance between the center points of two bounding boxes. This modification aims to prevent divergence issues that can occur during the training process when using IoU loss and GIoU loss.

DIoU is defined as the relative spacing between the centers of two bounding boxes:

$$R_{DIoU} = \frac{\rho^2(b, b^{gt})}{c^2} \tag{9}$$

where b and b^{gt} are the centers of the predicted and ground truth bounding boxes, respectively. The term ρ represents the Euclidean distance between these centers, while c refers to the diagonal length of the minimal bounding rectangle that can enclose both the predicted and actual boxes.

This method effectively addresses the gradient vanishing issue encountered with L_{IoU} and incorporates a geometric aspect. By utilizing R_{IoU}, DIoU can make more intuitive selections when faced with anchor boxes that have identical L_{IoU} values.

Furthermore, considering the aspect ratio in addition to DIoU leads to the proposed CIoU:

$$R_{CIoU} = R_{DIoU} + \alpha v, \tag{10}$$

where

$$\alpha = \frac{v}{L_{IoU} + v} \tag{11}$$

and v describes the consistency of aspect ratios:

$$v = \frac{4}{\pi^2}(tan^{-1}\frac{w}{h} - tan^{-1}\frac{w_{gt}}{h_{gt}})^2. \tag{12}$$

Here, w and w_{gt} denote the widths of the prediction box and the ground truth box, while h and h_{gt} represent the heights of the prediction box and the ground truth box, respectively. Because the unavoidable presence of poor-quality instances in the dataset leads to increased penalties, especially when influenced by factors like geometry, distance, and aspect ratio, thus diminishing the model's generalization performance. In order to reduce the effects of geometry when anchor boxes align closely to target boxes, while intervening less during training to elevate the model's ability to generalize, we construct WIoU v1 [66] as

$$L_{WIoUv1} = R_{WIoU} L_{IoU}. \tag{13}$$

The IoU score $L_{IoU} \in [0, 1]$ significantly diminishes the penalization for high-quality anchor boxes in R_{WIoU}, emphasizing the gap between center points when anchor boxes closely match with target boxes, where $R_{WIoU} \in [1, e)$ is the term amplifying L_{IoU} for regular quality anchor boxes.

$$R_{WIoU} = exp(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{W_g^2 + H_g^2}). \tag{14}$$

Here, W_g and H_g denote the size of the minimum bounding box, while the numerator represents the l_2 distance between the prediction box and ground truth. For the purpose of stopping R_{WIoU} from causing gradients hindering optimization, W_g and H_g are excluded from the computation framework and the computation is denoted by the superscript *. This effectively eliminates factors hindering convergence, thus avoiding the introduction of new metrics like the aspect ratio.

Inspired by focal loss, which concentrates model attention on challenging samples, improving classification performance, we introduce a monotonic focusing coefficient $L_{IoU}^{\gamma*}$ for L_{WIoUv1}:

$$L_{WIoUv2} = L_{IoU}^{\gamma*} L_{WIoUv1}, \gamma > 0. \tag{15}$$

The introduction of the focusing coefficient alters the gradient propagation of WIoU v2:

$$\frac{\partial L_{WIoUv2}}{\partial L_{IoU}} = L_{IoU}^{\gamma*} \frac{\partial L_{WIoUv1}}{\partial L_{IoU}}, \gamma > 0. \tag{16}$$

It is noteworthy that the gradient gain $r = L_{IoU}^{\gamma*} \in [0, 1]$. During model training, as L_{IoU} decreases, the gradient gain also diminishes, resulting in diminished efficiency in the final training phases. Thus, we introduce the average of L_{IoU} as a normalization factor:

$$L_{WIoUv2} = (\frac{L_{IoU}^{\gamma*}}{\overline{L_{IoU}}})^{\gamma} L_{WIoUv1}. \tag{17}$$

Here, $\overline{L_{IoU}}$ denotes the exponentially weighted momentum-weighted moving average with parameter m. Dynamic adjusting of the normalization parameter maintains the gradient improvement $r = (\frac{L_{IoU}^{\gamma*}}{\overline{L_{IoU}}})^{\gamma}$ on a more elevated perspective overall, thus dealing with the challenge of reduced convergence speed in later training phases.

The abnormality of anchor boxes is distinguished by the proportion of L_{IoU} to $\overline{L_{IoU}}$:

$$\beta = \frac{L_{IoU}^{\gamma*}}{\overline{L_{IoU}}} \in [0, +\infty). \tag{18}$$

Lower abnormality implies a higher quality of anchor boxes. We assign smaller gradient improvement to them, focusing the regression on anchor boxes of normal quality. Additionally, assigning reduced gradient improvement to anchor boxes with higher abnormality effectively prevents large gradients from low-quality samples. We construct a non-monotonic focusing coefficient apply it to WIoU v1:

$$L_{WIoUv3} = rL_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}}. \tag{19}$$

Here, when $\beta = \delta$, r = 1. When the abnormality of anchor boxes satisfies $\beta = C$, where C represents a constant, the reference box will obtain the maximum gradient benefit. Since $\overline{L_{IoU}}$ is variable, the standards for categorizing anchor box quality are, likewise, flexible, enabling WIoU v3 to adopt the most suitable gradient gain distribution method at each moment.

3. Experiment and Results

3.1. Experiment Setup

The experiment was carried out on PyTorch 1.13.1, CUDA 12.0, on a system equipped with an NVIDIA Quadro P5000 GPU and Windows 10. The model started with weights that were previously trained provided by ImageNet, and trained with the stochastic gradient descent algorithm for 400 epochs, with a starting learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. Additionally, a warm-up of weights was performed for the first 3 epochs, with a momentum of 0.8 during the warm-up phase. Furthermore, batch sizes of 64 and 16 were used for HRSID and SAR-Ship-Dataset, respectively. All remaining parameters were aligned with the initial YOLOv5 setup. The same settings were utilized in every experiment that involved alternative techniques to ensure a fair comparison. Table 1 presents the setup for the experiment.

Table 1. Table of experiment setup.

Experiment Details	
PyTorch Version	1.13.1
CUDA Version	12.0
GPU	NVIDIA Quadro P5000
Operating System	Windows 10
Batch Size (HRSID)	64
Batch Size (SAR-Ship-Dataset)	16

3.2. Dataset

3.2.1. HRSID

The HRSID dataset, annotated and publicly released by Wei et al. [35], comprises 5604 SAR image samples from Germany's TerraSAR-X, and TanDEM, the Sentinel-1 satellite of the European Space Agency that includes 16,951 annotated ship targets. Images are divided into patches of 800 × 800 pixels and have resolutions of 0.5 m, 1 m, and 3 m. They cover international maritime routes such as those in São Paulo, Barcelona, Chittagong, and Bangladesh. The dataset encompasses diverse ship environments, ranging from good

to poor sea conditions, coastal scenes, and simple offshore scenes. Given the variety of complex scenes in the HRSID dataset, it is appropriate for evaluating SAR detection performance in challenging environments. The dataset creators partitioned HRSID, allocating 65% for training and 35% for validation. All experiments conducted in this paper on HRSID were trained and tested using this partitioning.

3.2.2. SAR-Ship-Dataset

To address the issue of network training relying on large amounts of data, Wang et al. [36] built a dataset named SAR-Ship-Dataset. The SAR-Ship-Dataset comprises 43,819 images and 59,535 ship targets, sourced from 108 Sentinel-1 images and 102 Gaofen-3 SAR images. The images are cropped into 256 × 256 patches, with resolutions of 3 m, 5 m, 8 m, and 10 m. The original authors did not provide an official partitioning of training and validation sets. We randomly partitioned and selected the experimental data based on a proportion of 4:1 for the training and testing sets.

3.2.3. Analysis of the Two Datasets

The SAR-Ship-Dataset has a large scale, containing 43,819 images, including a significant number of high-noise images, which enhances the robustness of models trained on this dataset for real-world applications. However, the slices of the SAR-Ship-Dataset are 256 × 256 pixels, which is relatively small. This limitation may pose some challenges to the generalization capability of the dataset during training. Because of the small slice size, various models generally achieve high AP50 results on this dataset. However, the smaller slice dimensions result in lower AP50-95 scores, which require higher accuracy.

In contrast, the slices of the HRSID dataset are 800 × 800 pixels, which allows for a more substantial inclusion of land information and accommodates a variety of ship target sizes at different scales, as well as a greater range of complex dense scenes and nearshore environments. This larger slice size is advantageous for distinguishing multi-scale ship targets in images and for effectively addressing nearshore conditions. Although the larger slice size results in slightly lower AP50 scores across different models, the AP50-95 scores of the models are relatively higher. However, it is worth noting that the HRSID dataset has a relatively limited number of images, with only 5604 available, which could somewhat influence the model's overall capabilities. Additionally, the increased clarity of the HRSID images might lead to some challenges in maintaining robustness in scenarios that involve significant noise.

3.3. Evaluation Metrics

To evaluate ship detection systems, we utilized metrics including precision (P), recall rate (R), F1 Score, and average precision (AP). The formulas for precision and recall are outlined below:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{20}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{21}$$

In these formulas, true positive (TP) refers to instances correctly identified as positive, while false positive (FP) indicates cases incorrectly classified as positive. False negative (FN) refers to ship targets missed due to misclassification as background. Precision indicates the likelihood of correct predictions, while recall measures the probability of successfully identifying true positive samples.

The F1 Score assesses the balance between precision and recall and is calculated using

$$\text{F1 Score} = 2 \times \frac{P \times R}{P + R}. \tag{22}$$

Because precision and recall are mutually influenced, a high precision often implies a low recall and vice versa. Their relationship is represented by the P-R curve. The formula for average precision (AP) is as follows:

$$\mathrm{AP} = \int_0^1 P(R)dR. \tag{23}$$

When the IoU threshold is defined as 0.5, we obtain the result for AP50. AP50-95 is the average of AP values computed across different instances as the IoU threshold varies between 0.5 and 0.95 in increments of 0.05.

3.4. Ablation Study

This part investigates the impact of various enhancements on object detection performance through an ablation study conducted on two datasets: HRSID and SAR-Ship-Dataset. Modifications to the baseline YOLOv5 model, including Wise IoU loss, shuffle attention, and Context Decoupled Head, individually and in combination, are evaluated.

Table 2 summarizes the performance improvements achieved by different enhancements on the HRSID dataset. The baseline model attains a precision of 91.4% and a recall of 86.5%, with AP50 and AP50-95 scores of 93.4% and 68.1%. Integrating Wise IoU loss slightly improves recall and AP50 by 1.0% and 0.4%, respectively, with AP50-95 increasing by 1.4%. Adding shuffle attention results in improved precision, recall, and AP50 by 0.8%, 0.8%, and 0.4%, while increasing AP50-95 to 69.3%. Combining both enhancements results in a further increase in precision to 92.8% and recall to 88.0%, with notable improvements in AP50 to 94.2% and AP50-95 to 70.5%. Incorporating the Context Decoupled Head yields notable improvements across all metrics, with precision, recall, AP50, and AP50-95 increasing by 0.9%, 1.8%, 0.7%, and 2.6%, respectively. Combining Wise IoU loss and shuffle attention with the Context Decoupled Head further enhances performance. The highest overall improvements are observed in the model incorporating all three enhancements, with AP50 increasing by 1.1% and AP50-95 by 4.0% compared to the baseline model.

Table 2. Detection results on HRSID.

Baseline	+Wise IoU Loss	+Shuffle Attention	+Context Decoupled Head	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
✓				91.4	86.5	88.9	93.4	68.1
✓	✓			91.4	87.5	89.4	93.8 (+0.4)	69.5
✓		✓		92.2	87.3	89.7	93.8 (+0.4)	69.3
✓			✓	92.3	88.3	90.3	94.1 (+0.7)	70.7
✓	✓	✓		92.8	88.0	90.4	94.2 (+0.8)	70.5
✓	✓		✓	92.3	88.9	90.6	94.3 (+0.9)	71.3
✓		✓	✓	92.5	88.7	90.6	94.3 (+0.9)	71.1
✓	✓	✓	✓	92.4	89.4	91.0	94.5 (+1.1)	72.1

Similar performance improvements can also be seen in results from the SAR-Ship-Dataset in Table 3. The baseline YOLOv5 attains a precision of 90.6%, recall of 89.8%, AP50 of 94.7%, and AP50-95 of 56.1%. Adding Wise IoU loss slightly improves precision, recall, and AP50 by 0.1%, 0.4%, and 0.3%, respectively. Incorporating shuffle attention results in improvements across all metrics, with AP50 and AP50-95 increasing by 0.2% and 0.5%. Context Decoupled Head integration yields significant improvements, with precision, recall, AP50, and AP50-95 all increasing by 1.3%, 0.6%, 0.4%, and 1.0%, respectively. Combining Wise IoU loss and shuffle attention with the Context Decoupled Head further enhances performance. The highest overall improvements are observed in the model incorporating all three enhancements in our proposed method, with AP50 increasing by 0.8% and AP50-95 by 2.2% in comparison to the baseline network.

Table 3. Detection results on SAR-Ship-Dataset.

Baseline	+Wise IoU Loss	+Shuffle Attention	+Context Decoupled Head	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
✓				90.6	89.8	90.3	94.7	56.1
✓	✓			90.7	90.2	90.5	95.0 (+0.3)	56.5
✓		✓		91.2	89.7	90.4	94.9 (+0.2)	56.6
✓			✓	91.9	90.4	91.1	95.1 (+0.4)	57.1
✓	✓	✓		91.5	89.7	90.6	95.2 (+0.5)	56.9
✓	✓		✓	92.0	90.5	91.2	95.3 (+0.6)	57.7
✓		✓	✓	92.2	90.3	91.2	95.2 (+0.5)	57.4
✓	✓	✓	✓	92.5	90.5	91.5	95.5 (+0.8)	58.3

To summarize, the ablation study illustrates the cumulative effect of integrating Wise IoU loss, shuffle attention, and the Context Decoupled Head on enhancing object detection performance across both datasets, resulting in notable improvements in precision, recall, and AP scores.

3.5. Comparative Experiments

The comparative experiment result on the HRSID dataset is shown in Table 4. Based on the comparative experiments on the HRSID dataset, we focused on the performance of various object detection models across key metrics including F1 Score, AP50, and AP50-95. YOLOv5, serving as the baseline model, demonstrates a strong performance, with an F1 Score of 88.9%, AP50 of 93.4%, and AP50-95 of 68.1%. In contrast, classic methods like Faster R-CNN and SSD show comparatively less impressive results on these metrics. YOLOv3 exhibits high performance but it is lower than the baseline model YOLOv5. The results of CenterNet are lower than those of YOLOv3. YOLOv4 performs better than YOLOv3, but it is still below our baseline, YOLOv5. YOLOX, another emerging method, exhibits a notable performance for AP50, AP50-95, and F1 Score, at 89.5%, 93.1%, and 67.7%, respectively, albeit slightly below the baseline model YOLOv5. However, our proposed method showcases superior overall performance across all key metrics, achieving an F1 Score of 90.9%, AP50 of 94.5%, and AP50-95 of 72.1%, significantly outperforming all other models. This underscores the significant advantages of our approach in object detection tasks, particularly in enhancing detection accuracy, recall, and stability.

Table 4. Detection results on HRSID.

Method	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
Faster R-CNN	81.7	81.6	81.6	84.1	53.4
SSD	86.3	80.8	83.5	87.1	57.8
YOLOv3	91.5	85.7	88.5	92.7	66.5
CenterNet	90.1	84.3	87.1	91.4	63.1
CenterNet+SSE	91.1	86.2	88.6	93.0	65.0
YOLOv4	91.1	85.9	88.4	92.9	67.2
YOLOv5	91.4	86.4	88.9	93.4	68.1
FS-YOLO	92.0	87.1	89.5	93.7	68.6
GLC-DET	91.6	87.9	89.7	93.9	69.0
YOLOX	92.7	86.6	89.5	93.1	67.7
S2D	92.7	87.6	90.1	94.0	69.7
Proposed Method	92.4	89.4	90.9	94.5	72.1

Furthermore, our method achieves remarkable results compared to several other SAR image processing approaches. The core metrics of CenterNet + SSE [50], such as AP50 and AP50-95, while superior to the results of CenterNet, still fall short of those achieved by our proposed method. Although FS-YOLO [51], GLC-DET [52], and S2D [53] have shown improvements based on their chosen YOLO backbone, their performance still does not

match that of our approach. Therefore, in comparison with the latest SAR object detection methods, our method continues to deliver outstanding results.

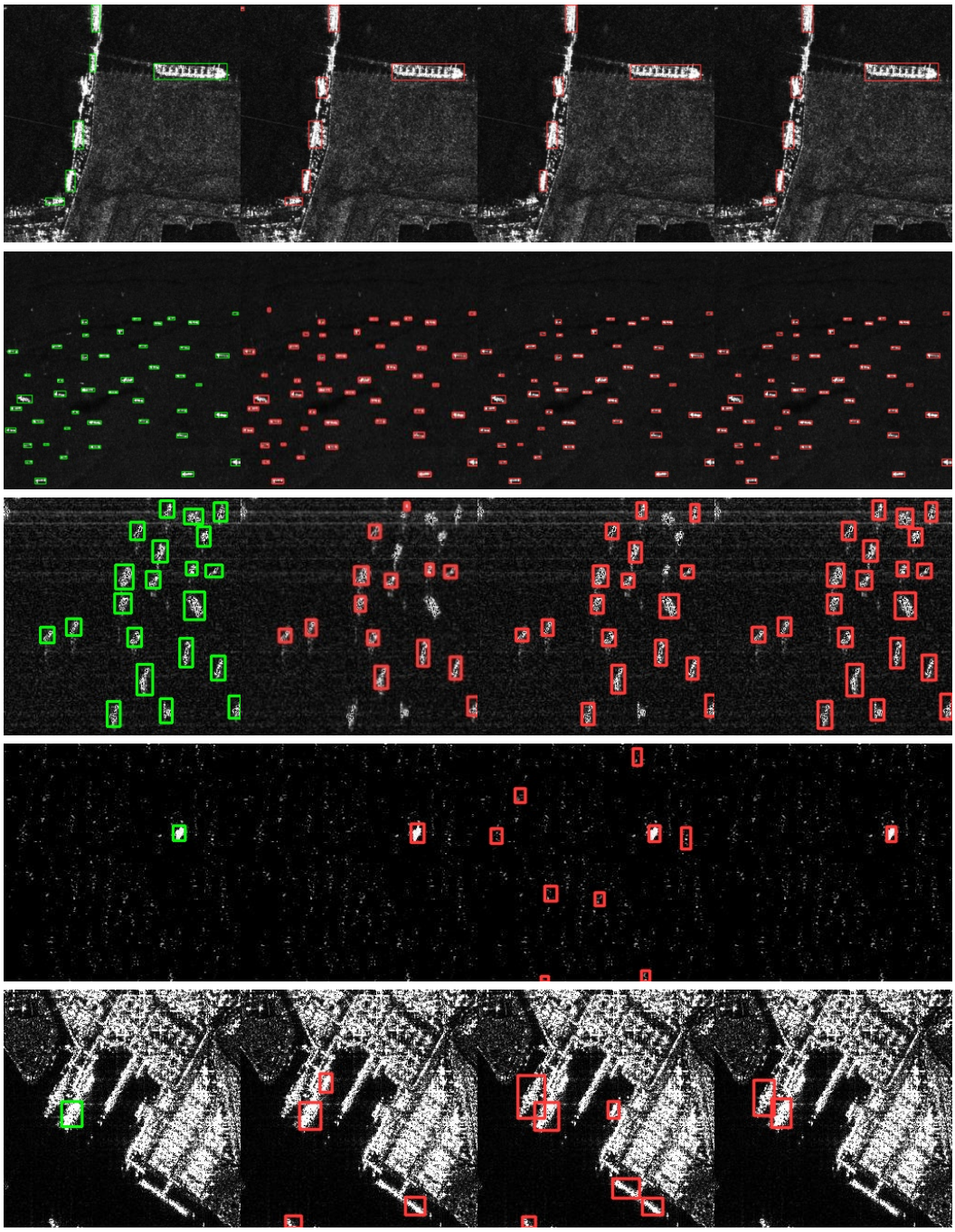
Similarly, based on the comparative experiments on the SAR-Ship-Dataset shown in Table 5, we focused on different object detection models' performance metrics such as precision, recall, F1 Score, AP50, and AP50-95. Traditional methods like Faster R-CNN and SSD demonstrate stable performance but fall short compared to the YOLO series, achieving an AP50 of 90.6% and 92.3%, respectively. Modern methods including CenterNet, YOLOv3, YOLOv4, YOLOv5, and YOLOX exhibit higher performance levels, achieving an AP50 of 92.6%, 93.9%, 94.2%, 94.7%, and 94.4%, respectively. However, our proposed method outperforms all others across all key metrics, achieving 92.5% precision, 90.3% recall, 91.5% F1 Score, as well as an AP50 of 95.4% and AP50-95 of 58.3%, significantly surpassing all other models. Meanwhile, the proposed method surpasses the latest SAR object detection methods [50–53] across a range of metrics, including F1 Score, AP50, and AP50-95. This highlights the exceptional performance of our method on the SAR-Ship-Dataset.

Table 5. Detection results on SAR-Ship-Dataset.

Method	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
Faster R-CNN	85.2	88.1	86.6	90.6	47.2
SSD	87.3	87.7	87.5	92.3	49.8
YOLOv3	89.8	88.7	89.2	93.9	54.4
CenterNet	88.1	87.9	88.0	92.6	54.2
CenterNet+SSE	89.3	88.4	88.8	93.5	55.1
YOLOv4	90.2	89.3	89.7	94.2	55.4
YOLOv5	90.6	89.8	90.2	94.7	56.1
FS-YOLO	91.2	90.0	90.6	94.9	56.9
GLC-DET	92.0	89.7	90.8	95.0	57.1
YOLOX	90.7	90.2	90.4	94.4	56.6
S2D	91.4	90.3	90.8	95.0	57.4
Proposed Method	92.5	90.3	91.5	95.4	58.3

3.6. Comparison Experiment Visualization

Figure 6 below compares the performance of YOLOX, baseline YOLOv5, and the proposed algorithm in dense and complex scenes, highlighting distinct advantages of the proposed algorithm. The first row of the images depicts results from complex dock scenes in the HRSID dataset, where many port facilities resemble ships in shape and exhibit strong electromagnetic scattering, leading to false alarms and missed detections. All three algorithms incorrectly identify a ship facility as a ship, but besides that mistake, YOLOX also detects a noise signal false alarm as a ship target, while YOLOv5 misses a small ship in the bottom left corner. The second row shows detection in dense target scenes within the HRSID dataset, where the proposed algorithm exhibits fewer false alarms compared to YOLOv5 and YOLOX. In the SAR-Ship-Dataset, the advantages of the proposed algorithm are more pronounced. In the third row, YOLOv5 and YOLOX show severely missed detections in dense ship scenes, whereas the proposed algorithm achieves a better detection of dense vessels. In the fourth row, amid noisy conditions, YOLOv5 as the baseline incorrectly identifies many noise signals as ships. Lastly, in the complex port data, the proposed algorithm demonstrates the least false alarms and mistaken detections. This comparative demonstration has proved the effectiveness of our proposed approach to be superior to both the baseline YOLOv5 and methods such as YOLOX.



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

Figure 6. Comparison figures of algorithm detection performance for SAR ship targets with various algorithms: (**a**) column represents the ground truth (GT), (**b**) column shows the performance of YOLOX algorithm, (**c**) column shows the performance of YOLOv5 as the baseline algorithm and (**d**) column displays the effectiveness of the proposed approach. Here the green box represents the targets of GT, while the red box represents the detected targets.

3.7. Visualization of Test Results in Complex Situations

Further tests are conducted to assess the robustness of the proposed method in complex scenarios, including an analysis of the model's robustness under challenging conditions. We selected high noise situations, dense ship scenarios, and complex background cases. The visualization of the experimental test results is shown below. From Figure 7, it can be seen that our method achieves excellent results in complex scenarios. The first row depicts high noise conditions; by comparing it with the ground truth, we find that our method

can overcome high noise interference and correctly detect the targets. The second row illustrates dense and small target situations. From the comparison of (e) and (f), we can see that, out of 120 ship targets, we only miss one, and this missed detection was due to two ship targets being too close to distinguish. In (g) and (h), our main errors are also due to the excessive density of ship targets, making it difficult to discern the exact number of targets. Additionally, some targets are too small to differentiate from floating objects in the river, contributing to some of our errors. Nevertheless, our method successfully detects the vast majority of targets (79 targets, with 75 correctly detected and 1 false alarm). In such overly complex situations, corresponding optical remote sensing images are needed for assistance, which will be a focus of our future research. The third row depicts a situation where targets of varying sizes coexist in a complex nearshore environment, and our method successfully and accurately detects all ship targets here. The superior performance of our method in complex scenarios also demonstrates its strong robustness in handling such conditions.

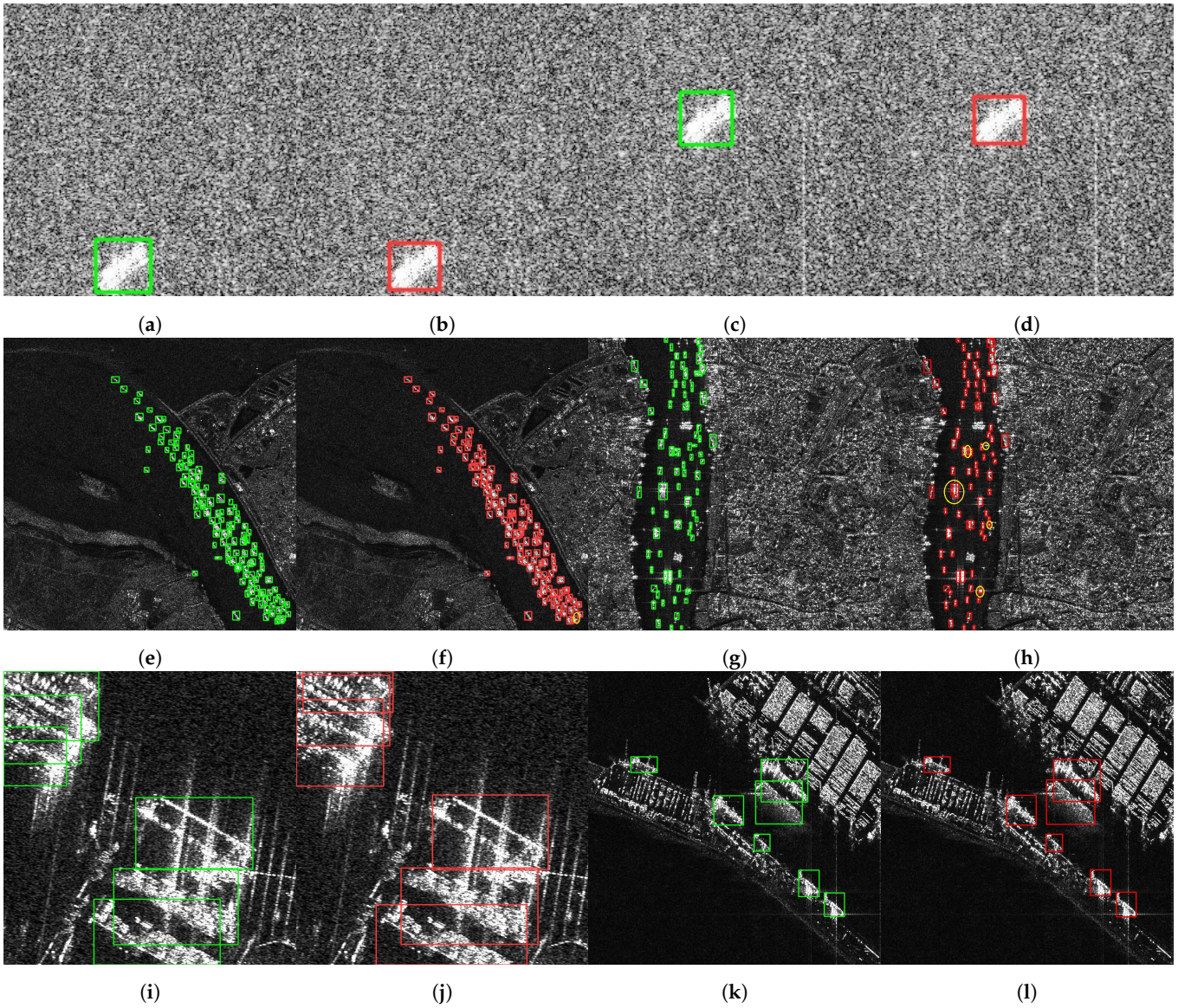


Figure 7. Test results displayed in complex scenarios. The first row shows high noise conditions, where (**a**,**c**) are the ground truth, and (**b**,**d**) are the corresponding test results; the second row presents dense and small target situations, with (**e**,**g**) as the ground truth, and (**f**,**h**) as the corresponding test results; the third row illustrates complex scenarios with multiple scales, where (**i**,**k**) are the ground truth, and (**j**,**l**) are the corresponding test results. Here the green and the red box represents the target of GT and the detected target, while the yellow circle represents the missed or incorrect detection.

4. Discussion

4.1. Attention Mechanism

The integration of shuffle attention serves as a critical enhancement in feature representation. Unlike traditional attention mechanisms that often prioritize spatial or channel-wise features in isolation, shuffle attention dynamically adjusts the attention weights across both dimensions simultaneously. This dual approach enables the model to effectively capture contextual relationships among objects and their surroundings, which is particularly beneficial in cluttered environments. By concentrating on relevant spatial features while maintaining a holistic view of the input data, the model's ability to infer object categories and their contextual significance is markedly improved. Furthermore, the adaptability of shuffle attention to multi-scale objects allows for a more nuanced understanding of features, thereby enhancing the model's overall performance across varying object sizes.

In this part, we conducted extensive experiments applying various attention mechanisms on the HRSID dataset and the SAR-Ship-Dataset, analyzing their effectiveness in object detection tasks.

Concerning the HRSID dataset, the comparative experiment results are shown in Table 6. Among the various attention mechanisms examined, shuffle attention demonstrated outstanding performance in enhancing recall, getting precision and recall rates of 92.4% and 89.4%, along with an F1 Score of 90.9%. Furthermore, it attained high levels of 94.5% and 72.1% on the AP50 and AP50-95 evaluation metrics, respectively. These results indicate that, compared with many other attention mechanisms, shuffle attention effectively elevates the network's capabilities to identify ship targets in object detection tasks.

Table 6. Detection results on HRSID.

Method	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
+SE	93.7	87.4	90.4	94.1	71.5
+CBAM	92.7	87.7	90.1	94.2	71.2
+ECA	93.1	87.6	90.3	94.1	71.1
+Coordinate attention	93.2	86.9	89.9	94.3	71.3
+sim attention	92.5	87.2	89.8	94.1	70.9
+shuffle attention	92.4	89.4	90.9	94.5	72.1

Apart from shuffle attention, other attention mechanisms exhibited relatively weaker performances in recall. For instance, SE, CBAM, and Efficient Channel Attention (ECA) achieved recall rates of 87.4%, 87.7%, and 87.6%, respectively, much lower than shuffle attention's 89.4%. Additionally, coordinate attention and sim attention achieved recall rates of 86.9% and 87.2%, respectively, also lower than shuffle attention. Besides recall, other performance metrics (F1 Score, AP50, and AP50-95) also failed to surpass shuffle attention. Specifically, shuffle attention achieved relatively high levels of 90.9%, 94.5%, and 72.1% on the F1 Score, AP50, and AP50-95, respectively. In comparison, the performance of other attention mechanisms on these metrics was slightly inferior. For instance, the performance of SE, CBAM, and ECA on these metrics were 90.4%, 90.1%, and 90.3% (F1 Score), 94.1%, 94.2%, and 94.1% (AP50), and 71.5%, 71.2%, and 71.1% (AP50-95), respectively. Although their performance remains respectable, they cannot match the overall performance of shuffle attention. Thus, shuffle attention not only excels in recall rate but also achieves high levels on other crucial performance metrics, further demonstrating its superiority in object detection tasks.

Furthermore, the results in Table 7 indicate that shuffle attention also performs optimally on the SAR-Ship-Dataset. It surpasses other attention mechanisms in key performance indicators such as precision (92.5%), recall (90.5%), F1 Score (91.5%), AP50 (95.5%), and AP50-95 (58.3%). This underscores the significant advantage of shuffle attention in object detection tasks, particularly in improving recall and overall performance.

Table 7. Detection results on SAR-Ship-Dataset.

Method	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
+SE	91.7	89.6	90.6	94.8	56.7
+CBAM	91.8	89.6	90.7	94.9	57.3
+ECA	91.9	89.8	90.8	95.1	56.7
+Coordinate attention	91.2	90.1	90.7	94.8	56.4
+Sim attention	92.3	90.2	91.2	95.2	58.0
+Shuffle attention	92.5	90.5	91.5	95.5	58.3

Consequently, we conclude that shuffle attention is the optimal choice among many attention mechanisms for achieving object detection on the SAR-Ship-Dataset.

4.2. Decoupled Head

In object detection, classification and localization are two main sub-tasks, but there is an inconsistency in their requirements for feature context. The localization task focuses more on boundary features to accurately regress bounding boxes, while the classification task tends to rely on a rich semantic context. Existing methods typically employ decoupled heads to address this issue, attempting to learn different feature contexts for each task. However, these decoupled heads still operate based on the same input features, resulting in an unsatisfactory balance between classification and localization. Specifically, bounding box regression requires more texture details and edge information to precisely locate the object's boundaries, whereas the classification task necessitates a stronger semantic context to identify the object's category.

This situation means that traditional decoupled head detectors cannot effectively meet the demands of these two tasks because they still share the same input feature maps, limiting their ability to select task-specific contexts. Although traditional decoupling designs achieve parameter decoupling by learning independent parameters, they still fail to fully resolve the issue, as the semantic context is largely determined by the shared input features. This leads to the phenomenon of feature redundancy in the classification task, while the localization task relies on more detailed texture and boundary information, making it difficult to achieve accurate corner predictions.

In order to demonstrate that designing decoupled heads based on different contextual semantics for classification and regression branches achieves better target detection results in SAR ship target detection than simple decoupled heads, we conducted comparative experiments using the simple decoupled head and Context Decoupled head.

The Tables 8 and 9 below present the performance metrics of the two different heads, simple decoupled head and Context Decoupled head, on the HRSID and SAR-Ship-Datasets. These methods were evaluated based on precision (Pre), recall (Rec), AP50, AP50-95, and Giga Floating-point Operations (GFLOPs).

Table 8. Comparative detection result on HRSID.

Method	Pre (%)	Rec (%)	AP50 (%)	AP50-95 (%)	GFLOPs
+simple decoupled head	91.6	88.4	94.2	70.1	7.1
+Context Decoupled head	92.4	89.4	94.5	72.1	9.8

Table 9. Comparative detection result on SAR-Ship-Dataset.

Method	Pre (%)	Rec (%)	AP50 (%)	AP50-95 (%)	GFLOPs
+simple decoupled head	91.3	90.2	94.8	57.1	7.1
+Context Decoupled head	92.5	90.5	95.5	58.3	9.8

For the simple decoupled head method, on the HRSID dataset, its precision is 91.6%, recall is 88.4%, AP50 is 94.2%, AP50-95 is 70.1%, and computational complexity is 7.1 GFLOPs.

On the SAR-Ship-Dataset, its precision is 91.3%, recall is 90.2%, AP50 is 94.8%, and AP50-95 is 57.1%, with computational complexity remaining at 7.1 GFLOPs. In contrast, the Context Decoupled head method demonstrates superior performance on both datasets. On the HRSID dataset, its precision is 92.4%, rate of recall is 89.4%, AP50 is 94.5%, AP50-95 is 72.1%, and computational complexity is 9.8 GFLOPs. On the SAR-Ship-Dataset, its precision is 92.5%, recall is 90.5%, AP50 is 95.5%, and AP50-95 is 58.3%, with computational complexity still at 9.8 GFLOPs.

These results show that the Context Decoupled head approach outperforms the simple decoupled head method regarding precision, recall, and AP on both datasets, albeit with slightly higher computational complexity.

4.3. Wise IoU Loss

The Wise IoU loss introduces a sophisticated mechanism to mitigate the negative impact of low-quality samples during training. Traditional loss functions often penalize the model heavily for geometric discrepancies, which can disproportionately affect generalization, especially in datasets with noisy annotations. By employing a distance attention mechanism alongside a dynamic focus mechanism, our loss function alleviates the penalty on well-aligned anchor boxes while downplaying the influence of poorly aligned ones. This novel approach not only fosters better training dynamics but also enhances the model's robustness against false positives and negatives. The result is a model that excels in precise localization, particularly in challenging scenarios where object overlap and occlusion are prevalent.

The comparison experiments of the loss functions on HRSID and SAR-Ship-Dataset are shown in Table 10 and Table 11, respectively.

The loss function used in the original baseline method is the CIoU loss function, while the loss function used in this paper is the Wise IoU loss. We conducted comparative experiments on the HRSID and SAR-Ship-Dataset, demonstrating the superiority of the Wise IoU algorithm.

Table 10. Detection results on HRSID.

Method	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
Baseline (CIoU Loss)	91.4	86.5	88.9	93.4	68.1
+Wise IoU Loss	91.4	87.5	89.4	93.8 (+0.4)	69.5

Table 11. Detection results on SAR-Ship-Dataset.

Method	Precision (%)	Recall (%)	F1 Score (%)	AP50 (%)	AP50-95 (%)
Baseline (CIoU Loss)	90.6	89.8	90.3	94.7	56.1
+Wise IoU Loss	90.7	90.2	90.5	95.0 (+0.3)	56.5

The results from the experiments clearly demonstrate that the use of Wise IoU leads to improvements in various aspects of object detection on the HRSID and SAR-Ship-Dataset.

5. Conclusions

To sum up, this work introduces an innovative approach for ship detection in SAR imagery, addressing key challenges faced by existing methods. The proposed SAR Ship Context Decoupled Head leverages both positioning and semantic information, enhancing the network's ability to recognize multi-scale objects with greater accuracy. Also by incorporating a shuffle attention module and a Wise IoU loss function, the proposed method attains superior performance in object detection tasks, as demonstrated through extensive experiments on benchmark datasets. These contributions represent significant advancements in SAR-based ship detection algorithms, with promising implications for applications in maritime surveillance and security. While our method demonstrates promising results,

it is worth noting that our proposed method comes with a higher computational cost. In later studies, we will delve into more lightweight network designs to mitigate this issue. Additionally, considerations for deploying the network on hardware devices should also be incorporated into future research efforts.

Author Contributions: Methodology, Y.T.; Software, Y.T. and J.X.; Validation, Y.T.; Investigation, Y.T. and Y.Z.; Writing—original draft, Y.T.; Writing—review & editing, Y.Z.; Supervision, Y.C.; Project administration, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Project named Three Dimensional Cross Band Multi Frequency Composite Antenna Microsystem Technology with grant number E3Z221030F.

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

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

References

1. Moreira, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Hajnsek, I.; Papathanassiou, K.P. A tutorial on synthetic aperture radar. *IEEE Geosci. Remote. Sens. Mag.* **2013**, *1*, 6–43. [CrossRef]
2. Eldhuset, K. An automatic ship and ship wake detection system for spaceborne SAR images in coastal regions. *IEEE Trans. Geosci. Remote Sens.* **1996**, *34*, 1010–1019. [CrossRef]
3. Robey, F.C.; Fuhrmann, D.R.; Kelly, E.J.; Nitzberg, R. A CFAR adaptive matched filter detector. *IEEE Trans. Aerosp. Electron. Syst.* **1992**, *28*, 208–216. [CrossRef]
4. Henschel, M.D.; Rey, M.T.; Campbell, J.W.M.; Petrovic, D. Comparison of probability statistics for automated ship detection in SAR imagery. In Proceedings of the International Conference on Applications of Photonic Technology III: Closing the Gap between Theory, Development, and Applications, Ottawa, ON, Canada, 4 December 1998; pp. 986–991.
5. Frery, C.; Müller, H.-J.; Yanasse, C.C.F.; Sant'Anna, S.J.S. A model for extremely heterogeneous clutter. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*, 648–659. [CrossRef]
6. Schwegmann, P.; Kleynhans, W.; Salmon, B.P. Manifold adaptation for constant false alarm rate ship detection in South African oceans. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2015**, *8*, 3329–3337. [CrossRef]
7. Qin, X.; Zhou, S.; Zou, H.; Gao, G. A CFAR detection algorithm for generalized gamma distributed background in high-resolution SAR images. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 806–810.
8. He, J.; Wang, Y.; Liu, H.; Wang, N.; Wang, J. A novel automatic PolSAR ship detection method based on superpixel-level local information measurement. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 384–388. [CrossRef]
9. Colone, F.; Filippini, F.; Pastina, D. Passive Radar: Past, Present, and Future Challenges. *IEEE Aerosp. Electron. Syst. Mag.* **2023**, *38*, 54–69. [CrossRef]
10. Li, J.; Xu, C.; Su, H.; Gao, L.; Wang, T. Deep Learning for SAR Ship Detection: Past, Present and Future. *Remote Sens.* **2022**, *14*, 2712. [CrossRef]
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
14. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
16. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017; pp. 936–944. [CrossRef]
17. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162. [CrossRef]
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
19. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.

20. Li, Z.; Peng, C.; Yu, G.; Zhang, X.Y.; Deng, Y.D.; Sun, J. Light-head r-cnn: In defense of two-stage object detector. *arXiv* **2017**, arXiv:1711.07264.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
23. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
24. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
26. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2018**, *128*, 642–656. [CrossRef]
27. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577. [CrossRef]
28. Bochkovskiy, A.; Wang, C.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
29. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000. [CrossRef]
30. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]
31. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. MixUp: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
32. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
33. Jocher, G.; Nishimura, K.; Mineeva, T.; Vilariño, R. YOLOv5 by Ultralytics. Code Repository. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 4 October 2022).
34. Li, J.; Qu, C.; Shao, J. Ship detection in SAR images based on an improved faster R-CNN. In Proceedings of the Sar in Big Data Era: Models, Methods & Applications, Beijing, China, 13–14 November 2017.
35. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S.L. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
36. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A High-Resolution SAR Images Dataset for Ship Detection and Instance Segmentation. *IEEE Access* **2020**, *8*, 2031. [CrossRef]
37. Zhang, C.; Zhang, X.; Gao, G.; Lang, H.; Liu, G.; Cao, C.; Song, Y.; Guan, Y.; Dai, Y. Development and Application of Ship Detection and Classification Datasets: A Review. *IEEE Geosci. Remote Sens. Mag.* **2024**, 2–36. [CrossRef]
38. Wang, Y.; Wang, C.; Zhang, H.; Zhang, C.; Fu, Q. Combing Single Shot Multibox Detector with transfer learning for ship detection using Chinese Gaofen-3 images. In Proceedings of the 2017 Progress in Electromagnetics Research Symposium-Fall (PIERS-FALL), Singapore, 19–22 November 2017.
39. Khan, H.M.; Cai, Y. Ship detection in SAR Image using YOLOv2. In Proceedings of the 2018 37th Chinese Control Conference (CCC), Wuhan, China, 25–27 July 2018.
40. Lin, Z.; Ji, K.; Leng, X.; Kuang, G. Squeeze and Excitation Rank Faster R-CNN for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 751–755. [CrossRef]
41. Zhao, W.; Syafrudin, M.; Fitriyani, N.L. CRAS-YOLO: A Novel Multi-Category Vessel Detection and Classification Model Based on YOLOv5s Algorithm. *IEEE Access* **2023**, *11*, 11463–11478. [CrossRef]
42. Wang, Z.; Hou, G.; Xin, Z.; Liao, G.; Huang, P.; Tai, Y. Detection of SAR Image Multiscale Ship Targets in Complex Inshore Scenes Based on Improved YOLOv5. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5804–5823. [CrossRef]
43. Li, Q.; Xiao, D.; Shi, F. A Decoupled Head and Coordinate Attention Detection Method for Ship Targets in SAR Images. *IEEE Access* **2022**, *10*, 128562–128578. [CrossRef]
44. Tang, H.; Gao, S.; Li, S.; Wang, P.; Liu, J.; Wang, S.; Qian, J. A Lightweight SAR Image Ship Detection Method Based on Improved Convolution and YOLOv7. *Remote Sens.* **2024**, *16*, 486. [CrossRef]
45. Bai, L.; Yao, C.; Ye, Z.; Xue, D.; Lin, X.; Hui, M. A Novel Anchor-Free Detector Using Global Context-Guide Feature Balance Pyramid and United Attention for SAR Ship Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4003005. [CrossRef]
46. Wu, K.; Zhang, Z.; Chen, Z.; Liu, G. Object-Enhanced YOLO Networks for Synthetic Aperture Radar Ship Detection. *Remote Sens.* **2024**, *16*, 1001. [CrossRef]
47. Hu, Q.; Hu, S.; Liu, S. BANet: A Balance Attention Network for Anchor-Free Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5222212. [CrossRef]
48. Ren, X.; Bai, Y.; Liu, G.; Zhang, P. YOLO-Lite: An Efficient Lightweight Network for SAR Ship Detection. *Remote Sens.* **2023**, *15*, 3771. [CrossRef]

49. Xu, Z.; Zhai, J.; Huang, K.; Liu, K. DSF-Net: A Dual Feature Shuffle Guided Multi-Field Fusion Network for SAR Small Ship Target Detection. *Remote Sens.* **2023**, *15*, 4546. [CrossRef]
50. Cui, Z.; Wang, X.; Liu, N.; Cao, Z.; Yang, J. Ship Detection in Large-Scale SAR Images Via Spatial Shuffle-Group Enhance Attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 379–391. [CrossRef]
51. Cai, S.; Meng, H.; Yuan, M.; Wu, J. FS-YOLO: A multi-scale SAR ship detection network in complex scenes. *Signal Image Video Process.* **2024**, *18*, 5017–5027. [CrossRef]
52. Wang, Z.; Wang, R.; Ai, J.; Zou, H.; Li, J. Global and Local Context-Aware Ship Detector for High-Resolution SAR Images. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4159–4167. [CrossRef]
53. Cheng, P. Improve the Performance of SAR Ship Detectors by Small Object Detection Strategies. *Remote Sens.* **2024**, *16*, 3338. [CrossRef]
54. Zhang, X.; Gao, G.; Chen, S.-W. Polarimetric Autocorrelation Matrix: A New Tool for Joint Characterizing of Target Polarization and Doppler Scattering Mechanism. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5213522. [CrossRef]
55. Gao, G.; Bai, Q.; Zhang, C.; Zhang, L.; Yao, L. Dualistic Cascade Convolutional Neural Network Dedicated to Fully PolSAR Image Ship Detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 663–681. [CrossRef]
56. Zhang, C.; Gao, G.; Liu, J.; Duan, D. Oriented Ship Detection Based on Soft Thresholding and Context Information in SAR Images of Complex Scenes. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5200615. [CrossRef]
57. Zhou, Y.; Liu, H.; Ma, F.; Pan, Z.; Zhang, F. A Sidelobe-Aware Small Ship Detection Network for Synthetic Aperture Radar Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5205516. [CrossRef]
58. Zhang, L.; Liu, Y.; Qu, L.; Cai, J.; Fang, J. A Spatial Cross-Scale Attention Network and Global Average Accuracy Loss for SAR Ship Detection. *Remote Sens.* **2023**, *15*, 350. [CrossRef]
59. Liu, Y.; Ma, Y.; Chen, F.; Shang, E.; Yao, W.; Zhang, S.; Yang, J. YOLOv7oSAR: A Lightweight High-Precision Ship Detection Model for SAR Images Based on the YOLOv7 Algorithm. *Remote Sens.* **2024**, *16*, 913. [CrossRef]
60. Chen, Z.; Liu, C.; Filaretov, V.F.; Yukhimets, D.A. Multi-Scale Ship Detection Algorithm Based on YOLOv7 for Complex Scene SAR Images. *Remote Sens.* **2023**, *15*, 2071. [CrossRef]
61. Yu, W.; Wang, Z.; Li, J.; Luo, Y.; Yu, Z. A Lightweight Network Based on One-Level Feature for Ship Detection in SAR Images. *Remote Sens.* **2022**, *14*, 3321. [CrossRef]
62. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
63. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
64. Zhang, Q.-L.; Yang, Y.-B. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239. [CrossRef]
65. Zhuang, J.; Qin, Z.; Yu, H.; Chen, X. Task-Specific Context Decoupling for Object Detection. *arXiv* **2023**, arXiv:2303.01047.
66. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.