remote sensing

MDPI

Article

WDFA-YOLOX: A Wavelet-Driven and Feature-Enhanced Attention YOLOX Network for Ship Detection in SAR Images

Falin Wu [1], Tianyang Hu [1,*], Yu Xia [1], Boyi Ma [1], Saddam Sarwar [1] and Chunxiao Zhang [2]

[1] SNARS Laboratory, School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China; falin.wu@buaa.edu.cn (F.W.); xia_yu@buaa.edu.cn (Y.X.); maboyi@buaa.edu.cn (B.M.); saddamsarwar@buaa.edu.cn (S.S.)

[2] Beijing Institute of Space Mechanics and Electricity, Beijing 100094, China; chun_xiao_zhang@163.com

* Correspondence: huty_11@buaa.edu.cn; Tel.: +86-10-82313929

Abstract: Ships are important targets for modern naval warfare detection and reconnaissance. The accurate detection of ships contributes to the maintenance of maritime rights and interests and the realisation of naval strategy. Synthetic Aperture Radar (SAR) image detection tasks play a vital role in ship detection, which has consistently been a research hotspot in the field of SAR processing. Although significant progress has been achieved in SAR ship detection techniques using deep learning methods, some challenges still persist. Natural images and SAR images significantly diverge in imaging mechanisms and scattering characteristics. In complex background environments, ships exhibit multiscale variations and dense arrangements, and numerous small-sized ships may be present, culminating in false or missed detections. To address these issues, we propose a novel SAR ship detection network, namely, a Wavelet-Driven Feature-Enhanced Attention–You Only Look Once X (WDFA-YOLOX) network. Firstly, we propose a Wavelet Cascade Residual (WCR) module based on the traditional image processing technique wavelet transform, which is embedded within an improved Spatial Pyramid Pooling (SPP) module, culminating in the formation of the effective wavelet transform-based SPP module (WSPP). The WSPP compensates for the loss of fine-grained feature information during pooling, enhancing the capability of the network to detect ships amidst complex background interference. Secondly, a Global and Local Feature Attention Enhancement (GLFAE) module is proposed, leveraging a parallel structure that combines convolutional modules with transformer modules to reduce the effect of irrelevant information and effectively strengthens valid features associated with small-sized ships, resulting in a reduction in false negatives in small-sized ship detection. Finally, a novel loss function, the Chebyshev distance-generalised IoU loss function, is proposed to significantly enhance both the precision of the detection box and the network convergence speed. To support our approach, we performed thorough experiments on the SSDD and HRSID, achieving an average precision (AP) of 99.11% and 96.20%, respectively, in ship detection. The experimental results demonstrate that WDFA-YOLOX has significant advantages in terms of detection accuracy, generalisation capability, and detection speed and can effectively realise more accurate detection in SAR images, consistently exhibiting superior performance and application value in SAR ship detection.

Keywords: deep learning; synthetic aperture radar (SAR); ship detection; you only look once (YOLO)

Citation: Wu, F.; Hu, T.; Xia, Y.; Ma, B.; Sarwar, S.; Zhang, C. WDFA-YOLOX: A Wavelet-Driven and Feature-Enhanced Attention YOLOX Network for Ship Detection in SAR Images. *Remote Sens.* **2024**, *16*, 1760. https://doi.org/10.3390/rs16101760

Academic Editor: Timo Balz

Received: 27 February 2024
Revised: 11 May 2024
Accepted: 13 May 2024
Published: 15 May 2024

1. Introduction

Synthetic Aperture Radar (SAR), a microwave remote sensing sensor, operates on the principle of the synthetic aperture. Imaging is achieved by transmitting phase-encoded pulses from a radar beam in a direction almost perpendicular to the sensor's motion vector [1]. The echo signal returned from the surface is then received and recorded. As SAR has continued to advance, its utilisation for ship detection has proven its indispensable value and found extensive practical applications in both military and civilian domains. It

provides robust data support and technical assistance for maritime ship detection. However, SAR ship detection still encounters numerous challenges. Unlike optical images, SAR images are vulnerable to system noise and background clutter. In some coastal areas and islands, the reflection characteristics in SAR images can resemble ships, resulting in false positives. Ship targets in SAR images vary in scale and are often densely clustered in coastal regions, making their accurate detection more challenging [2–6].

Traditional SAR image ship target detection algorithms primarily rely on contrast disparities between targets and background clutter for detection. In relatively straightforward scenarios, these traditional methods can yield reasonably accurate results. Commonly used target detection approaches for SAR images include the Constant False Alarm Rate (CFAR) [7,8], template matching [9], entropy [10], and wavelet transform for ship target detection [11]. However, in complex scenarios with less pronounced features and minor contrast disparities, such as coastal regions featuring small islands or rocky outcrops, traditional SAR ship detection algorithms necessitate the manual selection of the most suitable target feature set and demonstrate limited generalisation capabilities.

The advancement of artificial intelligence has led to significant progress in target detection algorithms for visible-light images using the deep learning method, specifically the convolutional neural network (CNN) framework [12]. This approach offers improved detection accuracy and stronger generalisation. Deep learning approaches encompass two types. One type is two-stage detection networks, such as region-CNN (R-CNN) [4]. Based on R-CNNs, more typical area-based object detection networks have been proposed, including Fast R-CNN [13], Faster R-CNN [14], and Mask R-CNN [15]. The primary principle is to employ selective search techniques to produce the recommended region, which is subsequently subjected to regression classification. One-stage object detection algorithms are the other kind. They reduce detection difficulties to regression problems and just need convolutional neural networks to extract the class probability and target position coordinates. Examples of representative algorithms are You Only Look Once (YOLO) and YOLO series [16–21], single-shot multibox detector (SSD) [22], Retina-Net [23], and so on. YOLO-series algorithms are generally faster than other algorithms and have a good effect on small object detection.

Some researchers are dedicated to creating SAR image datasets specifically for deep learning research on ship detection. Notable examples include the SSDD [24], the SAR-Ship dataset [25], the LS-SSDDv1.0 dataset [26], and the HRSID [27]. Researchers have extensively explored target detection algorithms for SAR images using transfer learning, which enables the algorithms to effectively address specific detection tasks. Researchers have proposed numerous advanced techniques to address challenges in SAR image identification using these datasets. For example, Zhang et al. [28] introduced a quad-feature pyramid network (FPN) to address the challenges of complicated backdrops and multiscale features in ship detection. This approach significantly enhanced the accuracy of ship detection by combining four distinct FPN modules in a cascading manner. For multiscale SAR ship detection in complex scenes, Fu et al. [29] proposed an anchor-free method based on FBR-Net. They created an ABP to balance the multiscale features across levels semantically, and they suggested an FR module to address feature misalignment, which helps to increase localisation accuracy. Hu et al. [30] proposed an anchor-free approach utilising a balanced attention network to improve the ability to detect ships at different scales. A local attention module was incorporated into this network to further augment its robustness. An additional nonlocal attention module was implemented to efficiently extract nonlocal features from SAR imagery. Zhang et al. [31] proposed an FSF module, drawing inspiration from the filtering mechanisms of the human brain. This module efficiently sifts through information, swiftly excluding irrelevant data while retaining pertinent information related to the target. On AIR-SARship-1.0, the FSF fully utilises its features to segregate the target and interference regions using neural networks with robustness. Zhu et al. [32] introduced a new anchor-free network for SAR image target recognition. The network, based on FCOS, incorporates deformable convolution (Dconv) and an improved residual network (IRN)

to enhance the network's capacity to extract features to obtain a low computational cost in SAR ship detection. In order to remove broad stretches of ocean and coastline background from SAR images taken at various levels, Zhang et al. [33] proposed a new SAR ship identification network called MLBR-YOLOX. This network uses the SSPD and DSFD modules. Huang et al. [34] proposed a brand-new two-stage detector called CViTF-Net, which combines visual transformers and CNNs in a novel way with three cutting-edge parts: a level-sync attention mechanism (LSAM), a Gaussian prior discrepancy (GPD) assigner, and a CViT backbone. As a result, the feature map visualisation demonstrates that the detector can reduce background noise while more precisely focusing on the positions of small ship targets. Zhang et al. [35] introduced a one-stage anchor-free SCSA-Net, to which an SCSA module and a GAP loss were added to enhance the network's capacity for feature extraction and lessen the nearshore background's interference with ship targets. The two-stage detection algorithm efficiently increases detection efficiency but at the cost of memory and computational overhead. Particularly for small targets, the one-stage YOLO-series algorithm further increases detection accuracy while significantly reducing processing complexity. The current study primarily employs attention mechanisms, feature pyramid network architectures, and similar techniques to enhance target features. However, due to the dense clustering of ship targets and significant scale variations, effective ship detection in SAR images remains challenging. Additionally, this may lead to a slow detection time and a high computational cost.

In order to address these issues, this paper presents an effective SAR ship detection network named the Wavelet-Driven Feature-Enhanced Attention–You Only Look Once X (WDFA-YOLOX) network, which was inspired by the wavelet algorithm. It can leverage the inherent correlation between spatial- and frequency-domain feature information to improve the frequency domain, as well as the local and global information to more effectively capture ship features while lowering the computational complexity of the model. The main contributions of our work are as follows:

1. Addressing the complexities inherent in SAR images, including complex background interference, limited available feature information, and the dense arrangement of ships in coastal areas, we propose a novel Wavelet Cascade Residual (WCR) module. This module is integrated into the Spatial Pyramid Pooling (SPP) module to propose a new wavelet transform-based SPP module (WSPP). By incorporating the wavelet transform into a CNN, spatial- and frequency-domain features are captured. It not only compensates for the loss of fine-grained feature information during pooling but also extends the receptive field of feature maps, ultimately reducing false positives.
2. In response to the prevalence of numerous small-sized ships, which have weak representation capabilities in SAR datasets, we propose a Global and Local Feature Attention Enhancement (GLFAE) module. Through a parallel structure, we fuse the outputs of channel and spatial attention mechanisms with those of the transformer module, which assigns greater importance to regions of interest while suppressing unnecessary features and enables the capturing of both local and global information related to ships.
3. To address the issue of slow convergence speed in the model, which adversely affects model performance, we replace the loss function with GIOU and introduce the Chebyshev distance with dual penalty terms on top of that, which is called the Chebyshev distance-based generalised IoU loss function. It improves the ability to accurately align and match bounding boxes, which helps with the convergence and stability of the training process and strikes a balance between model accuracy and speed.

The rest of this paper is organised as follows. Section 2 presents the structure and details of the entire network. Section 3 gives the experimental results and performance analysis, as well as ablation experiments on different modules. Section 4 discusses the limitations of the proposed method and suggests future research directions. Finally, Section 5 gives a short conclusion.

2. Methodology

2.1. Overall Network Structure

We chose YOLOX with a straightforward structure and stable detection accuracy as the baseline network. It is the first network in the YOLO series to utilise an anchor-free structure, which proves to be more suitable for ship detection compared to anchor-based approaches, given the multiscale and sparse characteristics of ships in SAR images. Furthermore, YOLOX introduces the SimOTA algorithm to enhance the allocation of ambiguous samples, which is particularly beneficial in allocating prediction samples for ship targets amidst complex backgrounds. This approach offers a trade-off between accuracy and speed, ultimately enhancing the network model's detection performance. The architecture of our proposed WDFA-YOLOX is shown in Figure 1, consisting of three sections: the Darknet53 backbone, the neck network, and three decoupled prediction heads. These components collectively enable ship detection in SAR images [36–38].

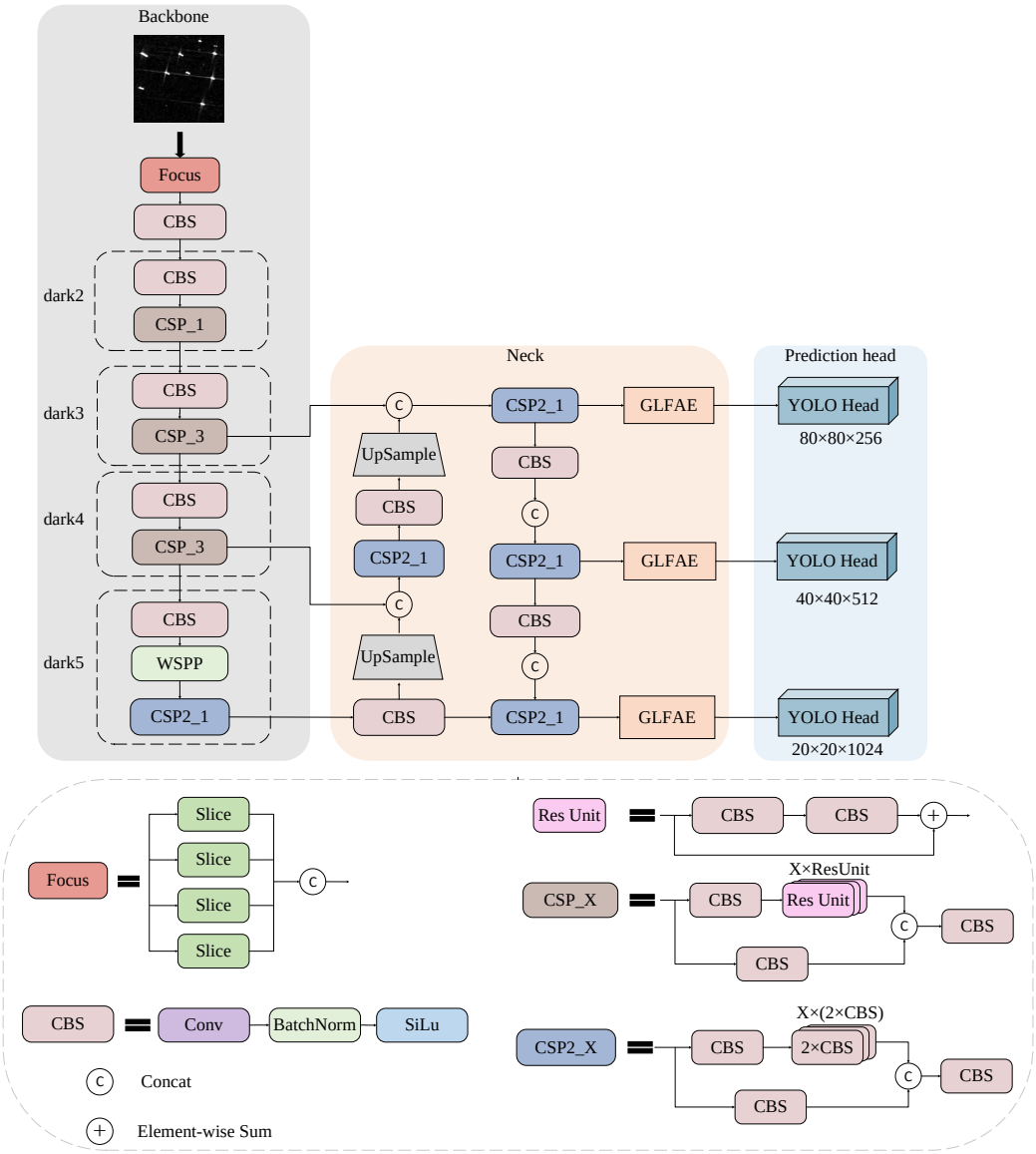

Figure 1. The structure of our proposed WDFA-YOLOX.

First, we introduce the wavelet transform algorithm to propose the Wavelet Cascade Residual (WCR) module. This module is able to capture image details from all directions and effectively extract texture features. To extend the receptive field of the network,

the WCR is integrated into the Spatial Pyramid Pooling (SPP) module [39] within the fourth layer of the backbone, dark5, to propose the novel wavelet transform-based SPP module (WSPP). Second, the output of Darknet53 is fed to the pyramid feature extraction network, known as PA-FPN, to extract contextual information. The Global and Local Feature Attention Enhancement (GLFAE) module, which we designed, is incorporated at the end of each output layer of PA-FPN, specifically preceding the three decoupled heads. GLFAE suppresses irrelevant features by assigning increased weights to regions of interest through global spatial and local channel attention mechanisms. Due to the organic integration of wavelets and attention mechanisms, the network can concentrate on relevant texture information while mitigating discrepancies among images in various domains. Subsequently, the decoupled head employs an anchor-free structure for category and position prediction. This approach enhances adaptability to fluctuations in ship sizes and mitigates problems associated with inaccurate predictions stemming from significant variations in ship size. Lastly, the original loss function is enhanced to form the Chebyshev distance-based generalised IoU loss function ($Loss_{CGIOU}$). This modification leads to improved detection frame accuracy and faster network convergence. WDFA-YOLOX is applied for ship detection using an open-source dataset of SAR ships, resulting in notable performance enhancements.

Sections 2.2 and 2.3 introduce the WCR and WSPP, which are employed to classify the regions of targets and backgrounds in the SAR images, respectively. The GLAFE is presented in Section 2.4. In Section 2.5, the Chebyshev distance-based generalised IoU loss function is described in detail.

2.2. Wavelet Cascade Residual Module

Contemporary ship detection networks frequently overlook various intermediate structural and texture-related cues, and they inadequately investigate frequency-domain information. As a consequence, existing methods suffer from performance limitations. To mitigate these limitations, we introduce the wavelet transform algorithm and present a novel WCR. The WCR leverages both frequency- and spatial-domain information to enhance the extracted features, resulting in improved structural and edge representations. The wavelet transform systematically refines the signal by applying translation and scaling operations, decomposing the image into a series of sub-band signals with distinct frequency characteristics [40,41]. We use the 2D discrete wavelet transform (DWT) to process the ship image, which produces a single low-frequency component and three high-frequency components for each decomposition layer while retaining the subject and detail information. The component combination can achieve a more effective balance between preserving fine details and ensuring adequate noise immunity performance [42]. After wavelet multiscale decomposition, the decomposed image set requires reconstruction. The existing literature predominantly employs reconstruction methods that focus on low-frequency components. While effective in suppressing noise, this approach often leads to the loss of crucial detail information. Additionally, decomposed images at various scales demonstrate intercorrelations and redundancy. High-frequency information from one layer can be embedded within the low-frequency content of the preceding layer.

The WCR comprises fundamental 2D wavelet decomposition and reconstruction processes. Additionally, it incorporates compact modules for enhancing features, aimed at improving contour information and retaining the detail information within both the low- and high-frequency components. The DWT employs the Haar wavelet basis with multi-resolution [43]. We use the input $X_i \in \mathbb{R}^{C\times H\times W}$ to convolve with four horizontal and vertical filters, f_{LL}, f_{LH}, f_{HL}, and f_{HH}, characterised by fixed parameters and a step size of two. The filters are as follows:

$$f_{LL} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad f_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \quad f_{HL} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \quad f_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \quad (1)$$

Subsequently, downsampling is performed, completing the decomposition of the original image to obtain four sub-bands: I_{LL}, I_{LH}, I_{HL}, and I_{HH}. The formula is expressed as follows:

$$I_n = X_i * f_n \downarrow 2, \quad n = LL, LH, HL, HH, \tag{2}$$

where * represents the convolution operation, and $\downarrow 2$ represents downsampling. I_{LL} is a component of low-frequency information that represents the overall information of the image. It enhances the global information of the feature map and primarily consists of the original image's information. I_{LH} represents the wavelet coefficient obtained from low-pass filtering in the horizontal direction and high-pass filtering in the vertical direction. It mostly captures the characteristics present in the horizontal direction. I_{HL} represents the wavelet coefficient obtained from high-pass filtering in the horizontal direction and low-pass filtering in the vertical direction, which primarily captures the features present in the vertical direction. I_{HH} represents the wavelet coefficient obtained by high-pass filtering in both the horizontal and vertical directions and carries the minimum amount of information. I_{LH}, I_{HL}, and I_{HH} pertain to high-frequency data, capturing intricate details of the image that are crucial in enhancing texture, edges, and other fine features.

The wavelet transform utilises its reversible nature and downsampling properties; the receptive field can be expanded to mitigate any loss of information. In Figure 2, the DWT procedure and outcomes are shown, displaying the basic tower structure after performing a one-layer DWT decomposition on the SAR images. The input X_i can be recovered using the IDWT due to the DWT's bi-orthogonal nature. In this way, the inverse discrete wavelet transform (IDWT) is the transposed convolution operation of the DWT, which is a convolution operation with a kernel size of 2×2, a step size of two, and fixed weights.

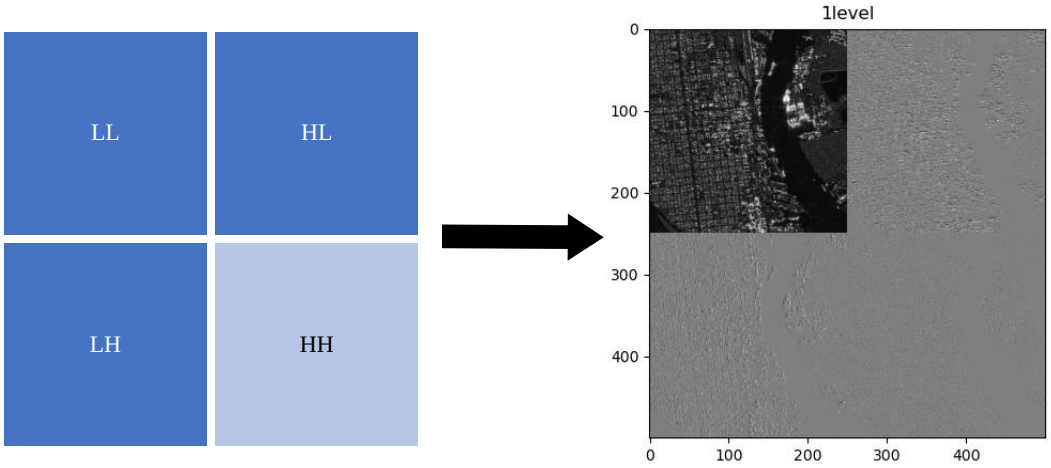

Figure 2. Tower structure in wavelet transform on SAR images.

We obtain the sub-bands I_{LH}, I_{HL}, and I_{HH}, which represent the high-frequency details of the image in different directions after the DWT. In the sub-highband, we propose the convolution residual block (CRB), taking the high-frequency details, I_{LH}, I_{HL}, and I_{HH}, as input. It allows the model to focus more on spatial information features of the high-frequency information and conduct deeper feature extraction. Figure 3 displays the structure of the CRB. There are four basic blocks, with each block consisting of 3×3 convolution and batch normalisation [44] in the CRB. The output of the basic block is fed to the ReLU activation function. The output F_{Ho1} is derived by applying two basic blocks. The input $F_{Hi} \in \mathbb{R}^{C \times H \times W}$ is then added directly to F_{Ho1} to yield the final residual output F_{Ho2}. Considering the measurement and parameters of the model, two consecutive feature extractions are carried out to yield F_{Ho3}. Subsequently, comprehensive feature ex-

traction and fusion are achieved through the average pooling layer and the fully connected layer, FC, resulting in the output F_{H_m} of the CRB, which can be formulated as follows:

$$F_{H_m} = fc(\text{Avg}(\text{RELU}(F_{Ho2}))), \quad m = 1,2,3, \tag{3}$$
$$F_{Ho2} = \text{RELU}(F_{Ho1}) + \phi(\text{RELU}(F_{Ho1})), \tag{4}$$
$$F_{Ho1} = F_{Hi} + \phi(F_{Hi}), \tag{5}$$
$$\phi = \text{ReLU}(\delta(\text{conv}_{3\times3}(\text{ReLU}(\delta(\text{conv}_{3\times3}(F_{Hi})))))), \tag{6}$$

where $conv_{3\times3}$ denotes a convolutional operation with a kernel of 3×3, and δ denotes batch normalisation. $F_{H_m}(m = 1,2,3)$ denotes the output of the mth CRB. fc$(\cdot)$ is the fully connected layer. ϕ is the operation of two basic block operations. There are three consecutive CRBs connected in series, and deep features F_{H_η} are ultimately extracted by the residual structure in the sub-highband.

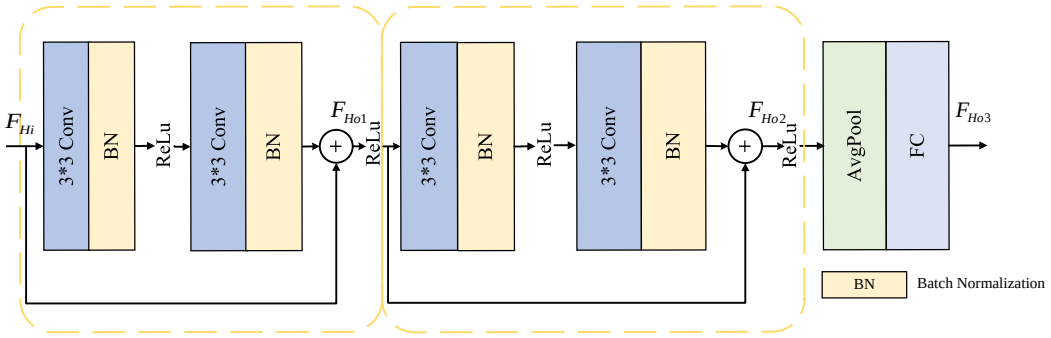

Figure 3. The structure of the CRB.

The low-frequency detail I_{LL} is acquired using the DWT, encompassing the primary ship construction information, as well as the background details of the image. We propose the gated residual convolution block (GRCB) for the sub-lowband, as described in Figure 4, which utilises the low-frequency detail as input and incorporates a gating mechanism in the sub-lowband. The gating mechanism employs a dual-branch structure, utilising two 3×3 convolutions to expand the channels of the layer-normalised features simultaneously by a factor of two. One of these convolutions is subsequently followed by GELUs (Gaussian Error Linear Units) [45], which can be formulated as follows:

$$GELU(x) = x * P((X \leq x) = x * \Phi(x), \tag{7}$$

where $\Phi(x)$ refers to the cumulative function of the Gaussian normal distribution of x. It has both linear and nonlinear transformation characteristics. In comparison to the ReLU activation function, the GELU exhibits a higher likelihood in its output, resembling a Gaussian distribution near zero. Its derivatives exist throughout the entire real number domain, ensuring smooth continuity. This aids the model in adapting to continuous variables and helps mitigate issues like vanishing and exploding gradients. This feature makes the GRCB more adaptable when handling different types of tasks. Finally, the resulting output is multiplied element-wise to leverage more intricate local features. The channels are subsequently diminished to the initial input dimensions 1×1 by convolution to achieve the result. The structure of the GRCB is illustrated in Figure 4, which can be formulated as follows for the input $F_{Li} \in \mathbb{R}^{C\times H\times W}$:

$$F_{L_n} = \text{conv}_{1\times1}(\text{Gating}(F_{Li})) + F_{Li}, \quad n = 1,2, \tag{8}$$
$$\text{Gating}(F_{Li}) = \text{conv}_{3\times3}(\psi(F_{Li})) \odot GELU(\text{conv}_{3\times3}(\psi(F_{Li}))), \tag{9}$$
$$\psi\left(F_{Li}^{l}\right) = \frac{F_{Li}^{l} - \mu^{l}}{\sqrt{\left(\sigma^{l}\right)^{2} + \varepsilon}} g^{l} + b^{l}, \tag{10}$$

where $\text{conv}_{1\times1}$ and $\text{conv}_{3\times3}$ denote convolutional operations with kernels of 1×1 and 3×3, respectively. $F_{L_n}(n = 1, 2)$ denotes the output of the nth GRCB. ψ is layer normalisation, and $\odot$ is element-wise multiplication. F_{Li}^{l} denotes the 1st channel of the input tensor, μ^l and $\left(\sigma^l\right)^2$ are the mean and variance of X_{in}^{l}, ε is a small constant that prevents the denominator from being zero, and g^l and b^l are two learnable parameters. In general, due to the utilisation of global residual learning, the GRCB enables us to selectively propagate specific properties to the subsequent layer of the network. For ship detection, this enables the transmission of ship-related data while filtering out information from areas with a lot of noise, which enhances accuracy in identifying ship regions. The shallow features F_{L_n} are ultimately extracted by the residual structure in the sub-lowband.

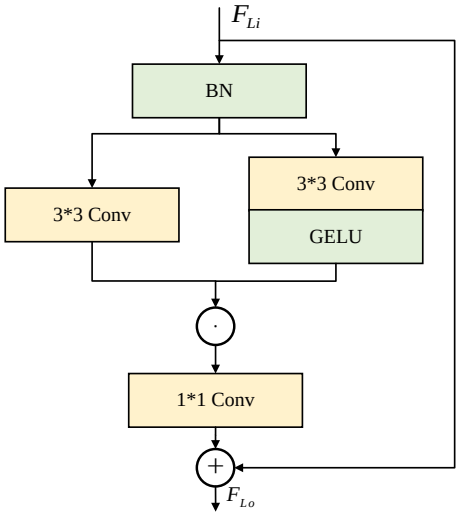

Figure 4. The structure of the GRCB.

The high-frequency features derived from the sub-highband have a low density. Adding more layers to the DWT to capture additional high-frequency details will lead to increased computational complexity for the network. To reduce the length of the feature information transfer path, the feature $F_{L_n}(n = 1, 2)$ extracted from the GRCB in the sub-lowband is added to the feature $F_{H_m}(m = 1, 2, 3)$ extracted from the CRB in the sub-highband using a shortcut connection. The red solid line in Figure 5 illustrates the transmission of low-frequency feature information to high-frequency features. Hence, the WCR that we propose not only possesses the inherent benefits of the wavelet transform but also facilitates a more comprehensive fusion of deep features and shallow features through the inclusion of a multi-layer cascade of high-frequency and low-frequency bands. This approach ensures more comprehensive feature fusion by utilising the feature information from each layer, enhances the representation of small targets in the feature map, and enhances the performance of detection. Afterwards, the deep features F_{H_3} and shallow features F_{L_2} are combined, resulting in the same number of channels. Ultimately, we restore the combination to their initial dimensions by executing the IDWT and handle the resulting information flexibly using 1×1 convolutional layers to obtain the output feature of the WCR, denoted by X_{out} . The complete structure of the WCR is depicted in Figure 5.

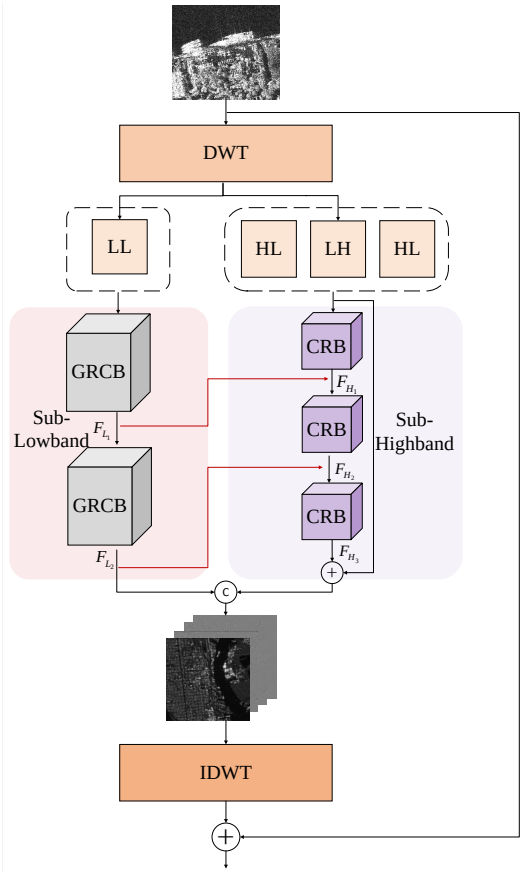

Figure 5. The structure of the WCR.

2.3. Wavelet Transform-Based Spatial Pyramid Pooling

Small targets in SAR images might experience a loss of features in the backbone network owing to their low resolution. Consequently, YOLOX employs SPP, which is positioned in the fourth layer of Darknet53 in order to expand the receptive field of the convolutional neural network (CNN). Nevertheless, the conventional average pooling and maximum pooling layers discard some amount of intricate information present in the initial feature maps, and the downsampling of the CNN can cause substantial harm to the features. The CNN is unable to retrieve the information that is lost during pooling layer operations and downsampling. This leads to the removal and fading of details in the image, resulting in a decrease in the ability of the network to distinguish and identify patterns. However, the conventional CNN used for SAR ship detection only considers the spatial-domain information, disregarding the significance of the frequency-domain information and failing to investigate the relationship between the frequency and spatial domains.

In Section 2.2, the DWT is used to generate a series of sub-band images, which are then used as the input for frequency-domain recovery. The WCR is integrated into the SPP module, forming a new module called the WSPP, which combines information from both the frequency- and spatial-domain information and utilises the spatial data from the 2D DWT to facilitate the ship detection task, hence enabling the extraction of features at several scales. During the feature fusion phase, low-level feature maps that undergo fewer downsampling iterations possess a narrower receptive field, higher resolution, and greater spatial information and retain finer details. This is beneficial for detecting smaller targets. High-level feature maps obtained by more instances of downsampling have a larger receptive field and maintain stronger semantic information, which are well suited for identifying targets of larger dimensions. Combining the characteristics of the small resolution of the images in the SAR dataset, we employed 2×2, 4×4, and 8×8 resolutions

for feature extraction in the pooling layer to improve SPP. This resulted in 4, 4, and 16 multi-dimensional feature vectors after using max pooling. Subsequently, these vectors were merged to create fused pyramid pooling information comprising 24 dimensions. The improved SPP module is depicted in Figure 6.

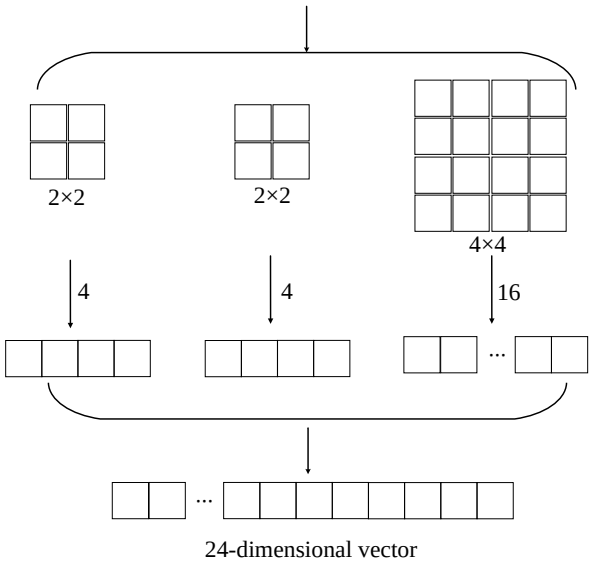

Figure 6. The structure of the improved SPP module.

The low-frequency and high-frequency components extracted from the WCR are combined with feature vectors of varying receptive field sizes of 2×2, 4×4 and 8×8 in the improved SPP module. The concat operation is then performed to generate the multiscale receptive field output, which forms the WSPP. The structure is illustrated in Figure 7. The CNN learns the spatial- and frequency-domain features separately because they have distinct characteristics. This allows for the fusion of these features at the feature map level, compensating for the loss of detailed feature information during the pooling process. The issue of information loss is partially mitigated. Within the intricate context of SAR images, the outcome of feature reconstruction preserves numerous fundamental attributes, such as shape, texture, and other characteristics of the target. Subsequently, through the integration of information from various levels, the accuracy of identifying ship details is significantly enhanced. The more concentrated region of ships is emphasised, and the outline of each ship target is preserved, enabling the differentiation of ships with identical distances. This enables the precise detection and segmentation of dense targets.

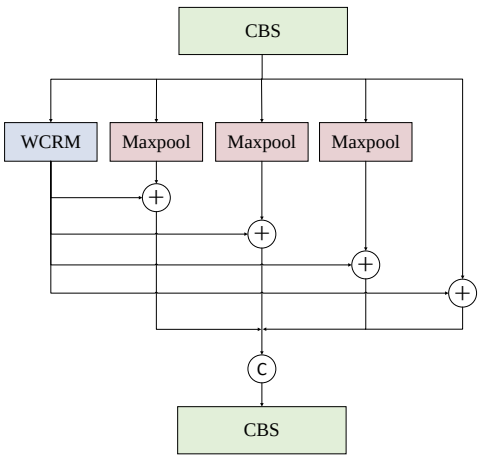

Figure 7. The structure of the WSPP.

2.4. Global and Local Feature Attention Enhancement Module

There is a high prevalence of small-sized ships in SAR images, and there is limited ability to depict small SAR ship features. Additionally, existing SAR ship detection methods that rely on CNNs for feature extraction struggle to establish long-distance dependency relationships and extract global information, leading to limited detection accuracy [46]. In order to mitigate the influence of irrelevant data on small-sized ships within the image, our developed GLFAE incorporates both a CNN and a transformer to effectively enhance the representation of features specific to small-sized ships. It maintains the fast inference speed and local feature extraction capability of a CNN while also leveraging the global sensing capability of the transformer. This helps establish stronger connections between distant pixel points and improves the algorithm's robustness and recognition accuracy. Subsequently, the global spatial attention (GSA) module in the global branch and the local channel attention (LCA) module in the local branch of GLFAE are described.

The global branch of GSA consists of a transformer [47] that utilises a self-attention mechanism to capture the global dependency between the input and output, extracting the global properties of the ship. The structure of GSA is illustrated in Figure 8. It employs spatial geometric division to independently monitor the characteristics of the ship image. GSA has the capability to simultaneously analyse every region in the entire image without dividing it into fixed-size sub-regions. This allows for easier handling of ships with varying sizes and proportions and enables the efficient modelling of the contextual relationship between smaller ships and their surrounding environments. In the local branch, our designed LCA involves filtering the channel dimension of the data. We combine global average pooling and global max pooling layers after obtaining a feature map with a height and width of one. The resulting feature maps are then merged through a splicing operation, and their dimensions are converted to normal dimensions using Depthwise separable convolution (DWConv) [48], which can reduce model parameters. Subsequently, a series of operations, including linear layers, ReLU, and the sigmoid function, are applied to reset the channel dimension and obtain the attention vector. Finally, the attention vector is multiplied by the input features. Figure 9 displays the structure of the LCA module. LCA prioritises the essential data of the target area and distinct local characteristics of the target, such as edges, texture, or shape, to facilitate precise target localisation and classification. In addition, the channels that are redundantly computed are assigned a negligible weight, hence minimising the computational workload.

GLFAE serialisation incorporates GSA and LCA into the GLFAE module, as depicted in Figure 10. The input F_i is fed into LCA to obtain the output F_{o1}. This output is then subjected to matrix multiplication with the input feature layer F_{i1}. Subsequently, GSA is utilised to derive the output F_{o2} by performing matrix multiplication with the output of the preceding module. This results in the generation of the output F_{o3}. The integration of GSA with LCA enhances the precision of target object localisation for WDFA-YOLOX. GSA enables the model to concentrate on the entire image, whereas LCA enables the model to concentrate on crucial components of the objective. This can enhance the emphasis on the positional data of compact ships, hence enhancing their precision in determining their location.

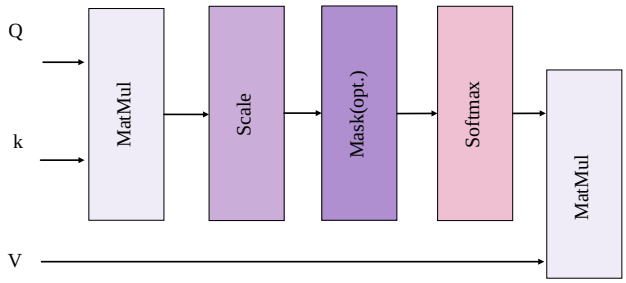

Figure 8. The structure of GSA.

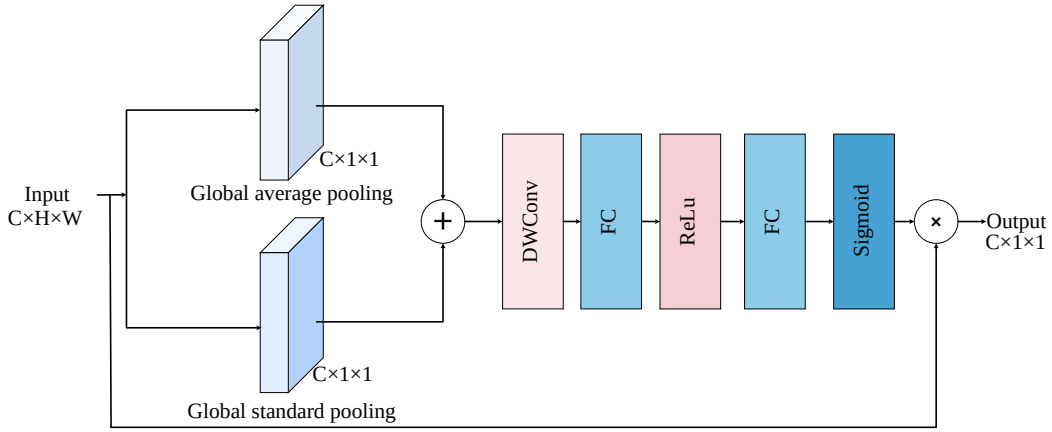

Figure 9. The structure of LCA.

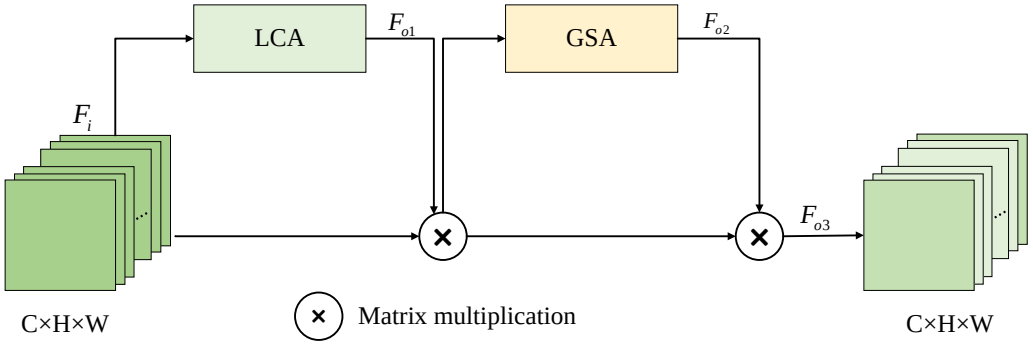

Figure 10. The structure of GLFAE.

The neck network utilises the PA-FPN to merge features from both the wavelet and spatial domains to fully exploit multi-domain features. To improve the capture and sensing of local and global information, GLFAE is incorporated at the end of each of the three output layers of PA-FPN, just before the three decoupled heads. This improves the distinguishability and robustness of the features, ultimately enhancing the discriminative ability of the network. GLFAE specifically examines the correlation between channels and spatial aspects inside the WDFA-YOLOX network to optimise the expression of effective features and inhibit the expression of ineffective features, boosting the ability to accurately represent small-sized ships.

2.5. The Chebyshev Distance-Generalised IoU Loss Function

YOLOx uses the decoupled head to separate the classification and regression tasks. It splits the loss into three components: bounding box regression loss $\left(\text{Loss}_{reg}\right)$, category loss $\left(\text{Loss}_{cls}\right)$, and object loss $\left(\text{Loss}_{obj}\right)$. The Binary Cross-Entropy Loss $\left(\text{Loss}_{BCE}\right)$ is employed during the training phase for Loss_{cls} and Loss_{obj}. Following the SimOTA label-matching process, a total of N positive samples can be selected. The target and prediction frames are then aligned, allowing for the calculation of the final loss function, which is calculated as follows:

$$\begin{aligned} \text{Loss} &= \text{Loss}_{reg} + \text{Loss}_{cls} + \text{Loss}_{obj} \\ &= \text{Loss}_{reg} + \frac{1}{N}\sum_{m=1}^{M} -\left[\hat{C}_m \log(C_m) + \left(1-\hat{C}_m\right)\log(1-C_m)\right] \\ &\quad + \frac{1}{N}\sum_{i\in Pos}\sum_{n}^{N} -\left[\hat{O}_n \log(O_n) + \left(1-\hat{O}_n\right)\log(1-O_n)\right] \end{aligned} \tag{11}$$

where N represents the count of positive case prediction boxes, while M represents the count of prediction boxes. C_m denotes the number of target species present in the mth positive case prediction box, and $\hat{C}_m$ represents the number of target species in the ground-truth box corresponding to the mth positive case prediction frame. O_n denotes the confidence score of the nth prediction box. Additionally, $\hat{O}_n$ indicates whether the nth prediction box is a positive or negative example, with a value of 1 representing a positive example and 0 representing any other case. Loss_{reg} uses Loss_{IOU}, which can be described as follows:

$$\text{Loss}_{IOU} = 1 - \frac{G \cap P}{G \cup P}, \tag{12}$$

where P and G are the prediction box and ground-truth box. $\cap$ denotes the intersection operation, and $\cup$ denotes the union operation.

Due to the requirement for multiple iterations to achieve convergence, the initial (Loss_{IOU}) in YOLOX suffers from numerical instability, resulting in the degradation of model performance. Simultaneously, the four edges of the two bounding boxes (top, bottom, left, and right) are disregarded, not only their centroids or areas. Moreover, alterations in the horizontal and vertical bounding box ratios have a specific influence on the loss, resulting in a decline in the accuracy of the predicted box. Thus, the generalised IoU loss function Loss_{GIOU} [49] is introduced and can be described as follows:

$$\text{Loss}_{GIOU} = 1 - \frac{A \cap B}{A \cup B} + \frac{|C - A \cup B|}{|C|}, \tag{13}$$

where A and B are the prediction box and ground-truth box, respectively. C is the minimum box that encompasses both A and B.

However, in contrast to Loss_{IOU}, Loss_{GIOU} considers the lack of overlap between A and B by incorporating a penalty term, and it only takes into account the spatial position of the frames and disregards the distance between them. Consequently, the penalty of GIOU is removed, and a novel penalty is introduced to reduce the distance between the centroids of the two prediction boxes. Simultaneously, another penalty is incorporated to enhance the convergence speed of the network loss by taking into account the correlation between the aspect ratios of A and B. To address the aforementioned problem, we introduce the Chebyshev distance IOU (Loss_{CIOU}) [50] to the enhanced loss function as the Chebyshev distance-generalised IoU loss function, and its formula is presented as follows:

$$\text{Loss}_{CGIOU} = 1 - \text{IOU} + \frac{C(A,B)}{L} + \left(\arctan\left(\frac{\frac{w_i}{h_i} - \frac{w_j}{h_j}}{1 + \frac{w_i w_j}{h_i h_j}} \right) \right)^2, \tag{14}$$

where the coordinates of the centre point of the prediction box(P) are represented by $P = (x_i, x_i)$, where x_i is the horizontal coordinate, and y_i is the vertical coordinate. Similarly, the coordinates of the centre point of the ground-truth box(G) are represented by $G = (x_i, x_i)$, where x_j is the horizontal coordinate, and y_j is the vertical coordinate. L represents the diagonal length of the smallest enclosing frame that contains both P and G. C represents the Chebyshev distance between the centre points of P and G. w_i, h_i, w_j, and h_j represent the length and width of P and G, respectively.

The Chebyshev distance is a continuous and differentiable metric that provides greater stability when calculating the gradient. This enhances the convergence and stability of the training process, leading to the accelerated convergence of the network. A schematic representation of Loss_{CGIOU} is depicted in Figure 11. It illustrates that as the distance between P and G increases, the value of Loss_{CGIOU} also increases. As the distance decreases, the value of Loss_{CGIOU} also decreases and approaches 0. The enhanced penalty term in the loss function increases the sensitivity to the predicted and real frames, resulting in a more precise measurement of the discrepancy between P and G. This improvement aids

the model in gaining a better understanding of the target's location and shape, ultimately leading to improved detection accuracy.

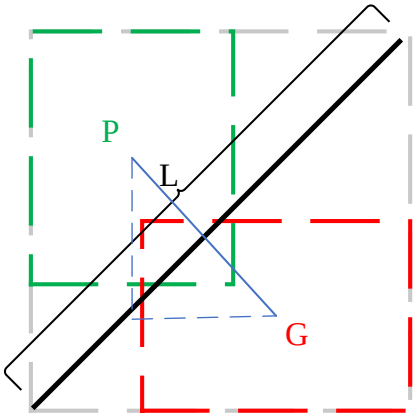

Figure 11. Schematic diagram of Loss_{CGIOU}.

3. Experiments

In this section, the performance of the proposed WDFA-YOLOX model is demonstrated through some experiments.

3.1. Implementation Details

3.1.1. Datasets

This work selected two authoritative SAR ship datasets, namely, the SAR Ship Detection Dataset (SSDD) [24] and the High-Resolution SAR Images Dataset (HRSID) [27], which are widely utilised in the field of SAR ship detection, for the experimental analysis. The first dataset in the field of SAR ship detection that was made available to the public is the SSDD. The HRSID is a collection of high-resolution SAR images created by Wei et al. [27]. The HRSID utilises 136 panoramic SAR images obtained from various satellites, which are then processed by cropping and filtering. The two datasets include basic information parameters, as listed in Table 1.

Table 1. The experimental results of different methods on the SSDD and HRSID datasets.

	SSDD	***HRSID***
Data sources	RadarSat-2, TerraSAR-X, Sentinel-1	Sentinel-1B, TerraSAR-X, TanDEM-X
Polarisation mode	HH, VV, VH, HV	HH, VV, VH, HV
Band	X and C bands	X and C bands
Resolution (m)	1–15	0.5–3
Category	ship	ship
Number (sheets)	1160	5604
Image size (pixels)	$28 \times 28 - 256 \times 256$	800×800
Ship number	2456	16,951

3.1.2. Evaluation Metrics

To assess the effectiveness of WDFA-YOLOX, we use precision (P), recall (R), average precision (AP), and frame per second (FPS), which are correlated with each other. Precision

is a statistical measure that evaluates the accuracy of prediction results. Recall is a statistical measure derived from the actual sample set. They are defined as follows:

$$\text{precision}(P) = \frac{TP}{TP + FP} \times 100\%, \quad (15)$$

$$\text{recall}(R) = \frac{TP}{TP + FN} \times 100\%, \quad (16)$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

Recall and precision can be represented as horizontal and vertical coordinates to create a precision–recall curve for each target category in the dataset. A metric called average precision (AP) gauges a model's accuracy at various recall rates. The precision–recall (PR) curve's area under the curve is used to compute it. The definitions of AP and mAP are as follows, respectively:

$$AP = \int_0^1 p(r)dr \times 100\%, \quad (17)$$

$$mAP = \frac{1}{N}\sum_{j=1}^{N} AP_j \times 100\%, \quad (18)$$

where j represents the jth category, and N denotes the total number of categories. The mean value of AP_j is mAP. Since there is only one class of ships in the dataset, $mAP = AP$. In order to calculate AP, the IoU threshold should be set to 0.5.

We use frames per second (FPS) to evaluate the detection efficiency of the model. It indicates the number of images detected per second.

3.1.3. Implementation Details

This paper's experimental setup is shown in Table 2. We divided the training, verification, and test subsets at random and established a 7:1:2 split, respectively. The batch size was set to 8, and the learning rate restart cycle was set to 24 epochs throughout the tests. For every iteration, the dataset was generated at random. We employed the Adam optimiser with a 0.0001 starting learning rate [51]. The optimiser's weight decay was set at 0.05, and cosine annealing was used to dynamically modify the learning rate with restarts. The test phase established a scoring criterion of 0.5. To guarantee the stability and dependability of the experimental data, we carried out the experiment three times and averaged the results.

Table 2. The experimental environment configuration.

Configuration	Parameter
GPU	NVIDIA GeForce GTX 3090 GPU ×3
Operating system	Ubuntu 20.04.4 LTS
Development tools	Python 3.10, Pytorch 1.13.0+cul17

3.2. Ablation Experiment

WDFA-YOLOX incorporates many enhancements in model architecture compared to the baseline YOLOX network. Consequently, it is imperative to examine the practical implications of all proposals and their interplay. We conducted ablation experiments on the SSDD and HRSID to evaluate the efficacy of the WSPP in the backbone network, GLFAE in the neck network, and the improved loss function Loss_{CGIOU}. These experiments involved various combinations of these innovative modules, resulting in a total of 16 sub-experiments, each set consisting of 8 sub-experiments. The outcomes are presented in Table 3.

Table 3. Ablation experiments on the SSDD and HRSID datasets.

Dataset	*WSPP*	*GLAFE*	Loss_{CGIOU}	*P*(%)	*R*(%)	*AP*(%)	*FPS*
SSDD				92.81	96.95	96.92	28.28
	√			94.14	95.30	97.66	27.31
		√		93.29	95.24	97.30	28.19
			√	91.63	94.83	94.08	60.53
	√	√		**95.92**	96.32	98.98	20.27
	√		√	95.01	97.43	98.03	57.11
		√	√	94.93	96.77	97.29	**58.75**
	√	√	√	95.07	**98.33**	**99.11**	58.34
HRSID				89.02	93.73	92.61	28.42
	√			92.59	95.98	94.53	27.26
		√		91.84	94.96	93.29	28.41
			√	90.85	88.32	91.29	60.33
	√	√		**93.80**	95.70	96.19	27.35
	√		√	92.90	95.62	95.89	57.28
		√	√	92.72	95.23	94.37	**59.29**
	√	√	√	93.25	**95.89**	**96.20**	59.13

The check mark "√" indicates that the technique was used in training. The bold numbers in the table represent the maximum value in this column.

We substituted SPP in the fourth layer of the backbone of YOLOX with the proposed WSPP. This proposal resulted in a 0.74% increase in AP on the SSDD and a 0.38% increase in AP on the HRSID as compared to the baseline. By incorporating the WSPP, the model successfully detected all the ships in the image, whereas the baseline model falsely detected the background as ships. The results indicate that the WSPP successfully restores ship texture that was previously lost in the feature map and enhances ship detection, making it more effective. The proposed GLFAE is incorporated into the neck network of YOLOX, resulting in a 0.38% improvement in AP on the SSDD and a 0.55% improvement in AP on the HRSID. GLFAE outperforms the baseline in detecting dense ships in images. The number of dense ships in the HRSID is greater than in the SSDD. Consequently, GLFAE has significantly more influence on AP on the HRSID than on the SSDD. This module highlights the significance of attention enhancement and feature fusion. The enhancement of the loss function results in a 0.08% improvement in AP on the SSDD and a 1.83% improvement in AP on the HRSID. Additionally, FPS sees a 32.2 improvement on the SSDD and a 31.9 improvement on the HRSID. Due to the inclusion of numerous modules in WDFA-YOLOX, the computational cost of the ship detection network may increase. However, the FPS with the addition of multiple modules to the network is not as high as the FPS with the introduction of only the improved loss function in the network.

WDFA-YOLOX, as compared to the baseline network, YOLOX, demonstrates significant improvements on the SSDD. Specifically, it enhances the model's precision by 2.26% (from 92.81% to 95.07%), its recall by 1.38% (from 96.95% to 98.33%), and its AP by 2.19% (from 96.92% to 99.11%). WDFA-YOLOX demonstrates a substantial enhancement in the effectiveness of detecting objects, as indicated by the rise in FPS from 28.3 to 58.3. On the HRSID, WDFA-YOLOX enhances precision by 4.23% (from 89.02% to 93.25%), recall by 2.16% (from 93.73% to 95.89%), and AP by 3.59% (from 92.61% to 96.20%). FPS shows a substantial improvement, ranking third highest, with an increase from 28.4 to 59.1.

The effectiveness of each enhancement is demonstrated by the significant improvement in the performance of the benchmark model when each enhancement is implemented separately. Furthermore, as the number of enhancements increases, the model's performance improves significantly. This suggests that there is minimal overlap in the performance enhancements achieved by different improvements. This is because each improvement targets a distinct problem, resulting in limited substitutability among the various enhancements.

By comparing the outcomes of experiments that include two enhancements with those that only include one enhancement, it is evident that the WSPP has the greatest impact, while combining the other two modules further improves *AP*. WDFA-YOLOX demonstrates superior performance compared to any individual module or combination of two modules, indicating its overall effectiveness.

3.3. Comparison with Other Methods

In order to explore the stability and generality of WDFA-YOLOX and to provide a more comprehensive performance evaluation, we conducted detailed experiments on the SSDD and the HRSID to validate the performance of our ship detection network in scenarios where ships are densely arranged or small-sized amidst complex background SAR images. As shown in Table 4, the first five rows represent mainstream object detection algorithms, including the two-stage algorithm Faster R-CNN [14] and the one-stage algorithms RetinaNet [23], YOLOv5 [19], YOLOv7 [20], and YOLOX [21]. Rows 6 to 8 (from MLBR-YOLOX to SCSA-Net) [33–35] show the state-of-the-art algorithms in SAR ship detection. In the last row, WDFA-YOLOX represents our proposed method. For an equitable comparison of detection performance and computational complexity across various networks, we adhered to similar parameter settings to those utilised in the compared networks while referencing the optimal experimental outcomes reported in the references of the compared methods.

Table 4. Comparisons of WDFA-YOLOX and the state-of-the-art methods on the SSDD and HRSID datasets.

Dataset	Model	*P*(%)	*R*(%)	*AP*(%)	*FPS*
SSDD	Faster-RCNN [14]	81.63	85.31	89.62	11.37
	RetinaNet [23]	93.34	87.54	92.13	23.82
	YOLOv5 [19]	95.14	90.01	96.61	98.80
	YOLOv7 [20]	91.05	84.92	93.68	51.63
	YOLOX [21]	92.81	96.95	96.92	28.28
	MLBR-YOLOX [33]	86.70	95.70	96.69	**120.71**
	CviTF-Net [34]	94.30	98.18	97.80	-
	SCSA-Net [35]	**98.19**	94.72	98.70	22.01
	WDFA-YOLOX	95.07	**98.33**	**99.11**	58.34
HRSID	Faster-RCNN [14]	83.81	72.57	77.98	11.41
	RetinaNet [23]	78.40	83.4	88.80	24.80
	YOLOv5 [19]	78.24	83.41	88.89	24.76
	YOLOv7 [20]	91.52	80.58	89.64	51.82
	YOLOX [21]	89.02	93.73	92.61	28.42
	MLBR-YOLOX [33]	92.72	88.61	92.16	**121.25**
	CviTF-Net [34]	90.95	93.69	92.98	-
	SCSA-Net [35]	**96.45**	90.02	95.40	22.29
	WDFA-YOLOX	93.25	**95.89**	**96.20**	59.13

It is challenging to fully replicate the experimental codes and findings of some SAR ship detection networks since they are not publicly available and because certain parameters and features are not clearly stated. Therefore, the "-" designation is used for these results. The bold numbers in the table represent the maximum value in this column.

Table 4 reveals that, on the SSDD, both Faster R-CNN and RetinaNet exhibit limited generalisation capability. The extracted features of the ships are insufficiently accurate, resulting in subpar detection performance characterised by lower precision, recall, and *AP*. Additionally, the overall response rate of the model is low, with *FPS* below 25. The YOLO series, including YOLOv5 and YOLOv7, demonstrate favourable outcomes in terms of precision and recall. However, *AP* still falls short of being sufficiently high, as there are instances of missed detection and false alarms for small targets. While MLBR-YOLOX and YOLOv5 have the greatest and second-highest *FPS*, respectively, they achieve this

speed at the expense of sacrificing model precision, potentially leading to false alarm issues. MLBR-YOLOX demonstrates exceptional recall at an impressive rate of 95.70%, which indicates its strong ability to accurately identify and capture targets. The recall of CviTF-Net is 98.10%, indicating its effective capture of the target. However, its precision is very low, perhaps resulting in more false alarms. The precision of SCSA-Net reaches a maximum of 98.19%, demonstrating a significant decrease in false alarms during the detection process. Nevertheless, it has a deficiency in recall, achieving only 94.72%, which suggests a constraint on its ability to effectively detect small targets. By comparison, our WDFA-YOLOX demonstrates excellent performance in precision, recall, and AP, achieving 95.07%, 98.33%, and 99.11%, respectively. While WDFA-YOLOX may not achieve the greatest FPS, its FPS is boosted from 28.28 to 58.34 compared with YOLOX. Our network has a competitive edge in overall performance when compared to the other networks. This advantage lies in its ability to effectively balance detection accuracy and speed. The findings clearly showcase the method's viability, with excellent detection performance.

In comparison to the SSDD, the background of the SAR images in the HRSID is notably more intricate, and the number of small ships is greater. Therefore, the values of precision, recall, and AP on the HRSID are all lower when compared to the SSDD. Table 4 reveals that the Faster R-CNN and RetinaNet models exhibit reduced precision, recall, and AP. Additionally, the total reaction rate of the model is diminished, with FPS below 25. The YOLO series, specifically YOLOv5 and YOLOv7, both demonstrate an impressive AP of 93.49% and 89.64%, respectively. MLBR-YOLOX exhibits a precision rate of 94.33% and a recall score of 88.61%, and it achieves a maximum FPS of 121.25. The recall of cviTF-Net is 93.69%, indicating a high level of accuracy in identifying positive instances. However, its precision is comparatively poor, suggesting a higher likelihood of false alarms. The precision of SCSA-Net reaches a maximum of 96.45%, indicating that it is able to detect ships more accurately and with fewer instances of misclassifying non-ship objects, such as fuel tanks on the coast of the port, throughout the detection process. Nevertheless, its recall stands at a mere 90.02%, suggesting a susceptibility to overlooking small-sized ships during detection. Our WDFA-YOLOX outperforms it in terms of precision, recall, and AP, achieving 93.25%, 95.89%, and 96.20%, respectively. FPS increases from 28.42 to 59.13, which ranks third in terms of FPS performance. WDFA-YOLOX demonstrates the capability to uphold a high level of precision while simultaneously achieving the comprehensive detection of targets. The results unequivocally showcase the method's viability, as it boosts the model's feature representation, resilience, and adaptability, hence enhancing the ship detection performance at the feature level.

3.4. Visualisation Comparison of Detection Results

The presence of complex backgrounds on the sea surface, such as inshore, offshore, near-island, and near-object interference, along with multiscale changes in the characteristics of ship targets on the sea surface and their dense arrangement, pose challenges for SAR ship detection. Therefore, many challenging scenes were chosen for detection in order to assess whether WDFA-YOLOX can enhance the accuracy of detection. To provide a clearer comparison of the aforementioned methodologies, Figure 12 presents the detection outcomes of various algorithms, such as Faster-RCNN, YOLOX, MLBR-YOLOX, and CviTF-Net, on the SSDD and HRSID. The ground truth of the ships is indicated by green boxes, and the detection results of Faster R-CNN, YOLOX, SCSA-Net, and WDFA-YOLOX are indicated by dark-blue, yellow, purple and light-blue boxes, respectively. The false and missed detections are marked with green triangles and red circles, respectively.

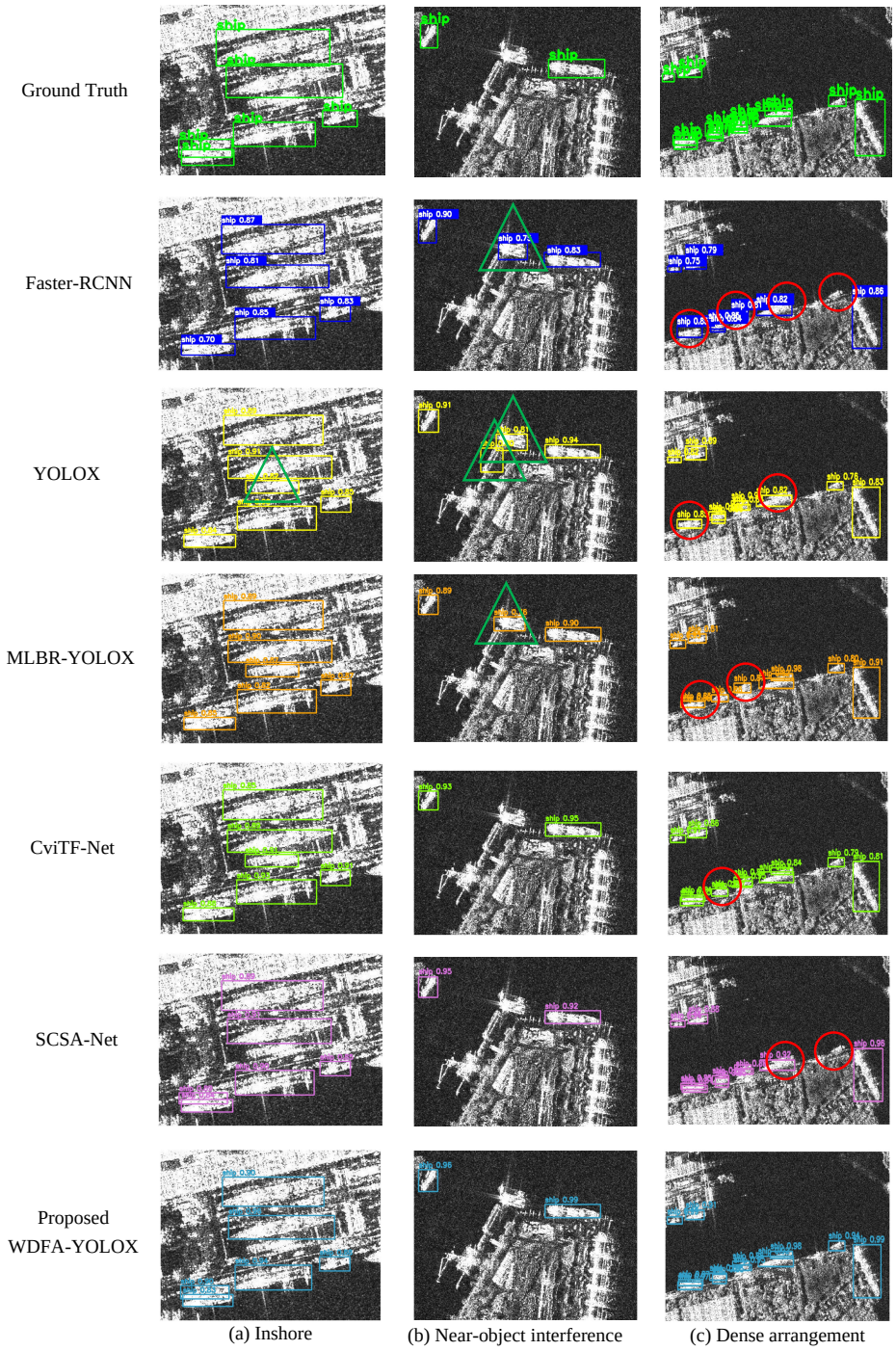

Figure 12. Experimental results of different algorithms in complex backgrounds, including inshore, near-object interference, and densely arranged scenarios.

As shown in Figure 12, the second rows of Figure 12(a–c) demonstrate that the detection results of Faster R-CNN exhibit a higher number of missed targets, particularly for small-scale targets. The network has a poor detection rate for targets and is prone to wrongly identifying items like oil tanks as ships. YOLOX enhances the detection efficiency, although it does result in some incorrect identifications of targets, which negatively impacts performance. Although MLBR-YOLOX and CviTF-Net can detect most ship targets, their confidence level is not as good as that of WDFA-YOLOX. WDFA-YOLOX demonstrates a minimal occurrence of false alarms and missed detections when applied to the sample images.

As shown in Figure 13, both datasets exhibited multiscale transformations and dense arrangements in the characteristics of ships. While several existing algorithms and WDFA-YOLOX successfully detected the corresponding targets, there were instances where the current algorithms failed to accurately identify the ships. These cases resulted in very low probability scores for the detections. However, our proposed WDFA-YOLOX algorithm achieved confidence of 91%, 93%, 96%, and 97% for detecting the four targets, respectively, as shown in Figure 13a in the seventh row. Furthermore, the confidence of the Faster R-CNN algorithm and YOLOX algorithm in detecting ships was approximately 1–5% lower than that of WDFA-YOLOX. Simultaneously, both Faster R-CNN and YOLOX exhibit 1–4 instances of missed detections in the detection results when the texture features of the ships are not prominent and occlusion occurs. However, WDFA-YOLOX effectively mitigates false alarms and reduces the number of missed detections, with only one ship remaining undetected. Furthermore, WDFA-YOLOX increases the probability fraction of detection by approximately 4%. The aforementioned result confirms the efficacy of the enhanced technique for identifying ships in SAR images. However, MLBR-YOLOX and CviTF-Net are not efficient in handling densely arranged targets in the nearshore scene. They tend to under-report and mistakenly treat multiple targets as a single target, as demonstrated in rows 2 and 3 in Figure 13b. So, WDFA-YOLOX is more proficient in dealing with this type of target.

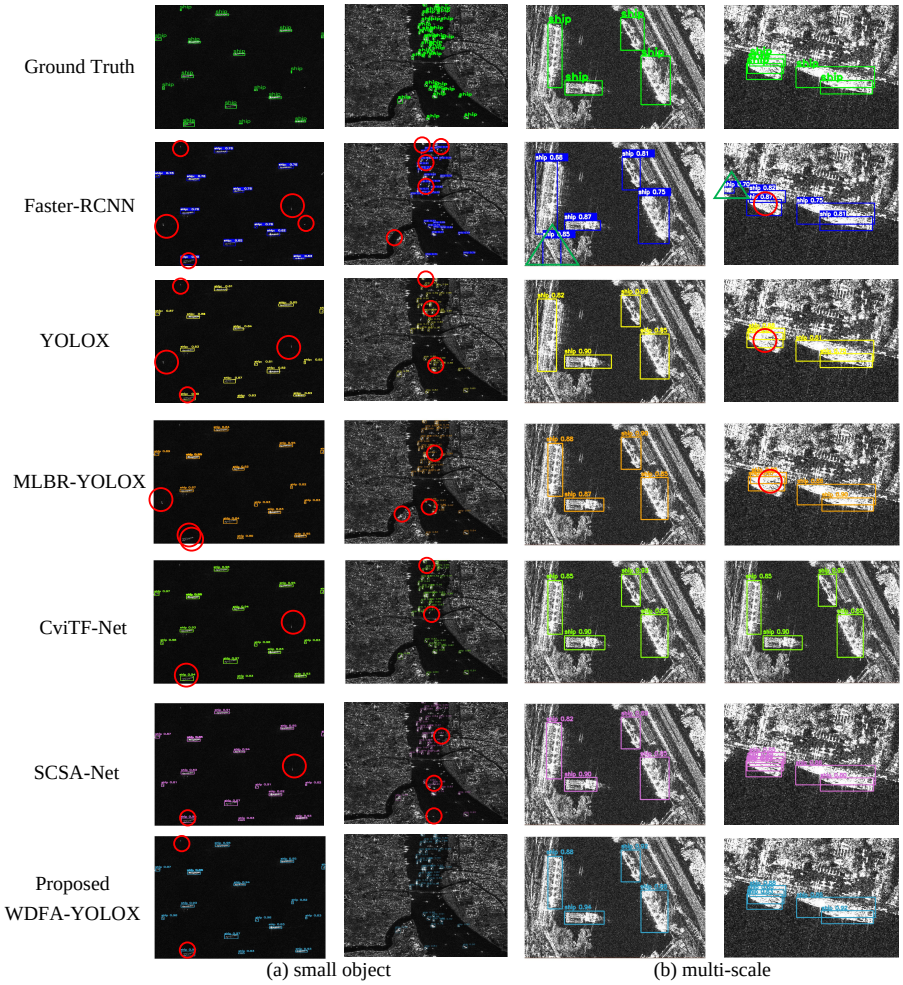

Figure 13. Experimental results of different algorithms on small, multiscale ships.

4. Discussion

The performance of algorithms was evaluated on two datasets with different scenarios, including offshore and inshore scenarios, scenarios with objects that interfere near the

ships, and scenarios with densely packed ships. In offshore scenarios, the target has a relatively simple background environment but occupies a small portion of the image, which increases the possibility of missed detections. In inshore scenarios, the target is surrounded by a complex background environment, making it challenging to distinguish between the ship target and the shore's buildings. In scenarios with object interference, such as oil tanks that resemble the ship in shape and texture, it becomes difficult to differentiate the target, leading to a higher chance of false alarms. This suggests that the proposed algorithm effectively reduces the possibility of both false and missed detections while still maintaining a high level of accuracy in detecting ships within a complex environmental context. Hence, WDFA-YOLOX has been proven to possess robust SAR ship detection capabilities across several settings, which efficiently mitigates the missed detection of ships, with favourable outcomes even in the detection of tiny SAR ships, showing that the AP of ship detection using this approach achieves 99.11% and 96.20% on the SSDD and HRSID datasets, respectively. On the basis of the current research, it is planned to design a series of experiments in a thesis extension to further test the performance of the model on randomly selected SAR images, especially on unseen scenes, to comprehensively evaluate its performance. Furthermore, it can be seen from the size of the FPS index in the experimental results that, compared with similar high-performance detection models, WDFA-YOLOX establishes a better balance between improving accuracy and efficiency, achieving the more efficient use of computing resources. But the module proposed in this article does add some computational overhead. Therefore, in future research work, we will focus on solving the core issue of balancing computing efficiency and model accuracy. In particular, effective network optimisation algorithms like quantisation compression, sparse training, and other tactics, as well as the creation of lightweight network architectures like channel pruning, knowledge distillation, and other technological methods, will be actively explored.

5. Conclusions

In complex environments such as densely packed ship formations in coastal regions like harbours, there is a high density of small-sized ships, which increases the risk of encountering challenges related to false and missed detections. Additionally, the original model network exhibits slow convergence speed. To address these challenges, a new SAR ship detection network based on YOLOX called WDFA-YOLOX, which is driven by the wavelet transform, using attention to enhance features in SAR images, is proposed. The proposed network greatly enhances the detection accuracy of ships in SAR images across various intricate scenarios and scales, particularly at smaller scales. Wavelet features are employed to guide the attention of the network towards intricate characteristics in the image, taking into account the influence of the complex background on ship detection. The WCR is proposed as the backbone to form the WSPP, with the aim of mitigating the influence of background noise and intricate backgrounds on ship detection. Subsequently, we propose GLAFE, designed to capture both the local and global information of ships. This module boosts the feature representation of small targets, mitigates the impact of irrelevant information in images with small-sized ships, and reduces missed detections. The original loss function is substituted for the Chebyshev distance-generalised IoU loss function with a dual penalty term, resulting in the enhanced accuracy of the detection frame and the better convergence speed of the network.

The performance of the proposed method was validated on two SAR ship image datasets, showing that the AP of ship detection using this approach achieves 99.11% and 96.20% on the SSDD and HRSID datasets, respectively. This shows that the proposed method outperforms other techniques in SAR ship detection. Compared to other high-performance detection models, WDFA-YOLOX establishes a better balance between increasing accuracy and efficiency, achieving the more effective use of computing resources, even though the modules provided in this paper do add some computing overhead.

In future research work, we will focus on solving the core issue of balancing computing efficiency and model accuracy. In particular, effective network optimisation algorithms like quantisation compression, sparse training, and other tactics, as well as the creation of lightweight network architectures like channel pruning, knowledge distillation, and other technological methods, will be actively explored. Furthermore, validating the effectiveness of the model on a wider set of images is an important direction for future research. On the basis of the current research, it is planned to design a series of experiments in a thesis extension to further test the performance of the model on randomly selected SAR images, especially on unseen scenes, to comprehensively evaluate its performance. This will include but not be limited to SAR images of different bands, resolutions, looking detections, and environmental conditions to ensure that the practicality of the model is fully verified.

Author Contributions: Conceptualisation, F.W. and T.H.; methodology, T.H. and Y.X.; software, Y.X. and B.M.; validation, T.H. and B.M.; formal analysis, Y.X. and S.S.; investigation, T.H. and B.M.; resources, F.W. and C.Z.; data curation, Y.X. and C.Z.; writing—original draft preparation, T.H.; writing—review and editing, F.W., T.H., and S.S.; visualisation, T.H. and S.S.; supervision, F.W. and C.Z.; project administration, F.W. and C.Z.; funding acquisition, F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This paper uses the SSDD and HRSID. Data sources: https://github.com/TianwenZhang0825/Official-SSDD and https://github.com/chaozhong2010/HRSID (accessed on 1 January 2024).

Acknowledgments: The editors and anonymous reviewers are appreciated by the authors for their insightful feedback, which significantly enhanced the manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average precision
CNN	Convolutional neural network
CRB	Convolution residual block
Dconv	Deformable convolution
DWT	Discrete wavelet transform
FPS	Frame per second
GRCB	Gated residual convolution block
GLFAE	Global and Local Feature Attention Enhancement
GSA	Global spatial attention
HRSID	High-Resolution SAR Images Dataset
LCA	Local channel attention
P	Precision
R	Recall
R-CNN	Region-convolutional neural network
SAR	Synthetic Aperture Radar
SPP	Spatial pyramid pooling
SSD	Single-shot multibox detector
SSDD	SAR Ship Detection Dataset
WCR	Wavelet cascade residual
WDFA-YOLOX	Wavelet-Driven Feature-Enhanced Attention–You Only Look Once X Network
WSPP	Wavelet transform-based SPP module
YOLO	You Only Look Once

References

1. Asiyabi, R.M.; Datcu, M.; Anghel, A.; Nies, H. Complex-Valued End-to-End Deep Network With Coherency Preservation for Complex-Valued SAR Data Reconstruction and Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5206417. [CrossRef]
2. Du, L.; Wang, Z.; Wang, Y. Survey of research progress on target detection and discrimination of single-channel SAR images for complex scenes. *J. Radars* **2020**, *9*, 34–54. [CrossRef]
3. Mullissa, A.G.; Marcos, D.; Tuia, D.; Herold, M.; Reiche, J. deSpeckNet: Generalizing Deep Learning-Based SAR Image Despeckling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5200315. [CrossRef]
4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
5. Huang, Z.; Pan, Z.; Lei, B. What, Where, and How to Transfer in SAR Target Recognition Based on Deep CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2324–2336. [CrossRef]
6. 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]
7. Leng, X.; Ji, K.; Yang, K.; Zou, H. A Bilateral CFAR Algorithm for Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1536–1540. [CrossRef]
8. Liu, T.; Zhang, J.; Gao, G.; Yang, J.; Marino, A. CFAR Ship Detection in Polarimetric Synthetic Aperture Radar Images Based on Whitening Filter. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 58–81. [CrossRef]
9. Wang, S.; Wang, M.; Yang, S.; Jiao, L. New Hierarchical Saliency Filtering for Fast Ship Detection in High-Resolution SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 351–362. [CrossRef]
10. Kapur, J.; Sahoo, P.; Wong, A. A new method for gray-level picture thresholding using the entropy of the histogram. *Comput. Vision, Graph. Image Process.* **1985**, *29*, 273–285. [CrossRef]
11. Wardlow, B.D.; Egbert, S.L.; Kastens, J.H. Analysis of time-series MODIS 250 m vegetation index data for crop classification in the US Central Great Plains. *Remote Sens. Environ.* **2007**, *108*, 290–310. [CrossRef]
12. Wu, F.; He, J.; Zhou, G.; Li, H.; Liu, Y.; Sui, X. Improved Oriented Object Detection in Remote Sensing Images Based on a Three-Point Regression Method. *Remote Sens.* **2021**, *13*, 4517. [CrossRef]
13. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1440–1448. [CrossRef]
14. 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]
15. He, K.; Gkioxari, G.; Dollar, 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]
16. 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]
17. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 30TH IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
18. Shen, L.; Tao, H.; Ni, Y.; Wang, Y.; Stojanovic, V. Improved YOLOv3 model with feature map cropping for multi-scale road object detection. *Meas. Sci. Technol.* **2023**, *34*, 045406. [CrossRef]
19. Jocher. YOLOv5 by Ultralytics. Available online: https://github.com/ultralytics/yolov5 (accessed on 8 January 2024).
20. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
21. Hou, H.; Chen, M.; Tie, Y.; Li, W. A Universal Landslide Detection Method in Optical Remote Sensing Images Based on Improved YOLOX. *Remote Sens.* **2022**, *14*, 4939. [CrossRef]
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016, PT I, Amsterdam, The Netherlands, 11–14 October 2016 ; Leibe, B.; Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 21–37. [CrossRef]
23. 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]
24. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
25. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
26. Zhang, T.; Zhang, X.; Ke, X.; Zhan, X.; Shi, J.; Wei, S.; Pan, D.; Li, J.; Su, H.; Zhou, Y.; et al. LS-SSDD-v1.0: A Deep Learning Dataset Dedicated to Small Ship Detection from Large-Scale Sentinel-1 SAR Images. *Remote Sens.* **2020**, *12*, 2997. [CrossRef]
27. 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*, 120234–120254. [CrossRef]

28. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A Novel Quad Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
29. Fu, J.; Sun, X.; Wang, Z.; Fu, K. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1331–1344. [CrossRef]
30. 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]
31. Zhang, P.; Luo, H.; Ju, M.; He, M.; Chang, Z.; Hui, B. Brain-Inspired Fast Saliency-Based Filtering Algorithm for Ship Detection in High-Resolution SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5201709. [CrossRef]
32. Zhu, M.; Hu, G.; Li, S.; Zhou, H.; Wang, S.; Feng, Z. A Novel Anchor-Free Method Based on FCOS plus ATSS for Ship Detection in SAR Images. *Remote Sens.* **2022**, *14*, 2034. [CrossRef]
33. Zhang, J.; Sheng, W.; Zhu, H.; Guo, S.; Han, Y. MLBR-YOLOX: An Efficient SAR Ship Detection Network With Multilevel Background Removing Modules. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5331–5343. [CrossRef]
34. Huang, M.; Liu, T.; Chen, Y. CViTF-Net: A Convolutional and Visual Transformer Fusion Network for Small Ship Target Detection in Synthetic Aperture Radar Images. *Remote Sens.* **2023**, *15*, 4373. [CrossRef]
35. 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]
36. Qiu, Z.; Rong, S.; Ye, L. YOLF-ShipPnet: Improved RetinaNet with Pyramid Vision Transformer. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 58. [CrossRef]
37. Xu, X.; Zhang, X.; Zhang, T. Lite-YOLOv5: A Lightweight Deep Learning Detector for On-Board Ship Detection in Large-Scene Sentinel-1 SAR Images. *Remote Sens.* **2022**, *14*, 1018. [CrossRef]
38. Zhang, Y.; Chen, C.; Hu, R.; Yu, Y. ESarDet: An Efficient SAR Ship Detection Method Based on Context Information and Large Effective Receptive Field. *Remote Sens.* **2023**, *15*, 3018. [CrossRef]
39. 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] [PubMed]
40. Hsu, W.Y.; Chang, W.C. Recurrent wavelet structure-preserving residual network for single image deraining. *Pattern Recognit.* **2023**, *137*, 109294. [CrossRef]
41. Hsu, W.Y.; Jian, P.W. Detail-Enhanced Wavelet Residual Network for Single Image Super-Resolution. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5016913. [CrossRef]
42. Sun, K.; Tian, Y. DBFNet: A Dual-Branch Fusion Network for Underwater Image Enhancement. *Remote Sens.* **2023**, *15*, 1195. [CrossRef]
43. Zi, Y.; Ding, H.; Xie, F.; Jiang, Z.; Song, X. Wavelet Integrated Convolutional Neural Network for Thin Cloud Removal in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 781. [CrossRef]
44. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15), Lille, France, 7–9 July 2015; pp. 448–456.
45. Lee, M. Mathematical Analysis and Performance Evaluation of the GELU Activation Function in Deep Learning. *J. Math.* **2023**, *2023*, 2314-4629. [CrossRef]
46. Xie, F.; Lin, B.; Liu, Y. Research on the Coordinate Attention Mechanism Fuse in a YOLOv5 Deep Learning Detector for the SAR Ship Detection Task. *Sensors* **2022**, *22*, 3370. [CrossRef]
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
48. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
49. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 658–666. [CrossRef]
50. 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 34th AAAI Conference on Artificial Intelligence/32nd Innovative Applications of Artificial Intelligence Conference/10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000. [CrossRef]
51. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.