remote sensing

MDPI

Article

Ship Detection via Multi-Scale Deformation Modeling and Fine Region Highlight-Based Loss Function

Chao Li †, Jianming Hu †, Dawei Wang *, Hanfu Li and Zhile Wang

Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China; 19b921033@stu.hit.edu.cn (C.L.); hujianming@hit.edu.cn (J.H.); 21b921023@stu.hit.edu.cn (H.L.); wangzhile@hit.edu.cn (Z.W.)

* Correspondence: wangdawei@hit.edu.cn; Tel.: +86-0451-86414883

† These authors contributed equally to this work.

Abstract: Ship detection in optical remote sensing images plays a vital role in numerous civil and military applications, encompassing maritime rescue, port management and sea area surveillance. However, the multi-scale and deformation characteristics of ships in remote sensing images, as well as complex scene interferences such as varying degrees of clouds, obvious shadows, and complex port facilities, pose challenges for ship detection performance. To address these problems, we propose a novel ship detection method by combining multi-scale deformation modeling and fine region highlight-based loss function. First, a visual saliency extraction network based on multiple receptive field and deformable convolution is proposed, which employs multiple receptive fields to mine the difference between the target and the background, and accurately extracts the complete features of the target through deformable convolution, thus improving the ability to distinguish the target from the complex background. Then, a customized loss function for the fine target region highlight is employed, which comprehensively considers the brightness, contrast and structural characteristics of ship targets, thus improving the classification performance in complex scenes with interferences. The experimental results on a high-quality ship dataset indicate that our method realizes state-of-the-art performance compared to eleven considered detection models.

Keywords: multi-scale deformation modeling; fine region highlight; loss function; ship detection; optical remote sensing

Citation: Li, C.; Hu, J.; Wang, D.; Li, H.; Wang, Z. Ship Detection via Multi-Scale Deformation Modeling and Fine Region Highlight-Based Loss Function. *Remote Sens.* **2023**, *15*, 4337. https://doi.org/10.3390/rs15174337

Academic Editors: Mohamed Lamine Mekhalfi, Yakoub Bazi, Edoardo Pasolli and Mawloud Guermoui

Received: 20 July 2023
Revised: 20 August 2023
Accepted: 25 August 2023
Published: 3 September 2023

1. Introduction

The detection of maritime ships holds immense significance in the domains of automated fishery management, port rescue operations, and marine traffic maintenance [1]. The precision of detection technology is directly proportional to the safety and promptness of both military and civilian applications. Automatic identification system (AIS) is a short-range coastal tracking system currently used on ships, which can realize real-time data exchange between ships and ships, ship and shore, as well as target detection. It is designed to ensure the safety of maritime navigation. Information provided by AIS equipment, such as unique identification, position, course, and speed, can be received by other ships or maritime authorities to track and monitor the movements of different vessels. However, for a variety of reasons, ships can turn off their AIS transceivers. Therefore, there are some limitations in using AIS data to detect ships or monitor sea areas, especially for non-cooperative ships and sea areas. In comparison, visible light imaging offers rich textures and apparent target structure features, making its information intuitive and easy to comprehend, which aligns with the daily observation habits of the human eye. Moreover, it is not limited by regional conditions and non-cooperative targets. As a result, visible light remote sensing images have emerged as a crucial data source for ship detection and identification [2]. However, remote sensing images usually cover vast areas, thus encompassing various environmental factors such as clouds and diverse port infrastructures in

the image scene [3]. Hence, detecting ships amidst complex optical image scenes remains a challenging task.

For the detection of ships in intricate scenes, scholars have proposed a diverse range of strategies and models. Among them, the visual saliency model stands out as a prime example [4]. This model emulates the mechanism of attention in the human eye and promptly pinpoints unusual regions or points in complex scenes, offering a vast potential for applications in the processing of massive data. Conventional image attributes, such as edge, contrast, and texture features [5], were employed in early visual saliency models to highlight potential areas. Subsequently, various extraction strategies, such as frequency domain analysis [6], cellular automata [7], and Bayesian theory [8], were adopted to achieve more precise contour of suspected objects. These techniques have achieved superior object detection outcomes for specific scenes. However, due to the intricate scene conditions, such as various artificial facilities and interferences from environmental elements like clouds, the saliency model is prone to producing a high number of false alarms when applied to large-scale remote sensing scenes. Consequently, there is a pressing exigency to proffer a potent tactic to redress this quandary.

As is widely recognized, convolutional neural networks (CNNs) serve as powerful tools for representing high-level semantic features in deep learning [9]. Consequently, they offer an effective framework for the detection of ships in remote sensing images [10]. The application of deep learning technology has notably enhanced the performance of region extraction and ship detection through saliency methods based on CNNs [11]. In order to enable the network to effectively detect ships of various sizes and types, feature pyramid network (FPN) [12] is widely used to solve multi-scale object detection problems in images due to its advantage of generating feature representations at different levels. Interestingly, the multi-level structure of FPN actually uses information from different receptive fields. The receptive field usually refers to the spatial range in the original image that is perceived by neurons located at different positions. By mining the context relationship between the target and the background in different receptive fields [13], the detection model can learn the different characteristics of the target and the background at different scales to a certain extent, thus achieving the correct detection of the target.

However, relevant research has proved that not all receptive field pixels contribute to the response of the output unit for different practical tasks [14]. Therefore, how to dynamically select different receptive fields and carry out information fusion to achieve the effective learning of features is one of the problems to be studied in the field of ship detection. In addition, due to the observation distance and direction of the remote sensing platform, the shape of the target in remote sensing images may undergo various degrees of deformation, and the characteristics of the target may also have certain differences, which poses new challenges to the robustness of the detection network. Consequently, it is urgent to accurately characterize the multi-scale, multi receptive fields, and deformation characteristics of the target and enable the model to learn the differences between the target and the background, so as to achieve accurate detection of ship targets in complex scenes. Furthermore, due to the inherent locality of convolution operation and the pixel-level characteristics of cross-entropy loss function [15], the current deep learning detection model is difficult to achieve uniform saliency extraction of objects, resulting in incomplete prominent areas of ships in complex scenes, which limits the performance of the detection algorithm under complex port, cloud interference, and other conditions. Therefore, it is urgent to closely combine the characteristics of ships in image scenes and then design the loss function that can precisely learn the target regions.

To address these above issues, we propose a new ship-detection framework that combines visual attention-based saliency extraction and fine region highlight-based loss function, enabling the high-accuracy detection of typical ship targets in complex scenes. Specifically, we propose a saliency extraction framework that jointly models the multi-scale and deformation characteristics of targets. This framework leverages dilated convolutions and deformable convolutions to learn the discriminative features of targets against the

background and intrinsic features of the targets themselves, thereby generating saliency extraction results for the targets. On this basis, a loss function is presented that comprehensively utilizes global and local features such as brightness, contrast, and structure. This loss function facilitates the fine and uniform enhancement of target regions, thereby improving the overall detection performance of ships. Finally, experimental results on a high-quality dataset demonstrate the accurate detection performance of the proposed method under complex interference conditions such as port facilities and cloud disturbances.

The main contributions of our work are summarized as follows:

- A ship detection framework combining multi-scale saliency extraction with semantic consistency and loss function for fine region highlight is proposed, which can achieve a high-accuracy detection performance for typical ship targets in complex port and sea scenes.
- A deformable multi-scale convolutional saliency extraction network is proposed, which can jointly model the multi-scale and deformation characteristics of the target. The network combines multiple receptive fields to improve the ability to distinguish the difference between the target and the background and employs deformable convolution to extract the complete features of the target, thus enhancing the ability to characterize the target.
- A specialized loss function for fine and overall target region highlight is proposed, which comprehensively considers the brightness, contrast, and structure characteristics of the ship target, thus improving the classification performance in complex scenes such as port facilities, cloud and shadow interference.

2. Previous Related Research

In this section, we first provide a succinct overview of the salient object detection technology evolution and explain the inadequacies of the classical methods. On this basis, we explicate the solution of our method. Furthermore, we scrutinize the utilization of the conventional edge enhancement-based loss function and clarify the distinctions between our approach and other advanced classification methods.

2.1. Salient Object Extraction Model

The issue of detecting ships in intricate scenes has garnered significant attention within the field of remote sensing. The visual attention mechanism, akin to the discerning gaze of the human eye, plays a pivotal role in the processing of visual information [9]. By effectively filtering out insignificant stimuli and swiftly honing in on crucial regions within an image scene, it expedites the search and localization of the target of interest against complex backgrounds, such as sparsely distributed ships. Drawing inspiration from this remarkable mechanism, researchers have proposed various visual saliency models tailored for ship detection. These models are designed to extract and leverage the crucial and valuable information present in the salient regions, thereby facilitating timely and accurate ship localization even in complex scenes. In past research on salient object detection, a large number of detection models have been proposed based on intrinsic cues and extrinsic cues. In [16], a saliency optimization technique is used to estimates the backgroundness probability of superpixels through geodesic saliency analysis. In [17], a graph-based model based on manifold ranking is proposed, constructing a query sequence by considering boundary nodes as either background or non-salient regions. In [18], a method via conditional random field (CRF) learning is applied to model groups of saliency features, generating detection results.

With the development of hardware computing capability, the methods based on deep learning, especially the use of convolutional neural networks (CNNs), have gradually become the dominant paradigm in the research of salient target detection. Many studies have adopted various CNN architectures, including fully convolutional networks (FCNs) [19] and residual networks (ResNets) [20], which have achieved significant advancements in accuracy and speed. Moreover, researchers have explored multi-scale and multi-modal

fusion techniques to enhance the accuracy and robustness of salient object detection. These techniques combine features from multiple scales or multiple modalities, such as color, texture, and depth, to capture salient objects from different perspectives and improve detection performance. With the rise of feature enhancement ideas, attention mechanisms have been widely used in salient object detection to improve the performance of CNNs [21]. These mechanisms allow the network to focus on the most relevant regions of the image, which can help to capture salient objects more accurately and efficiently. In addition, domain adaptation [22] and generalization have emerged as important research directions in salient object detection. These approaches aim to improve the modeling ability to generalize across different domains, such as different lighting conditions, weather conditions, and camera perspectives, which can enhance the practical applicability of salient object detection in real-world scenes. Different from other works, we focus on the joint modeling of multi-scale, multiple receptive fields and deformation characteristics and construct a deformable multi-scale convolutional neural network. Moreover, we enhance the important features of the target in channel and spatial dimensions based on attention mechanisms, improving the discriminatory ability for interested targets and complex background environments.

2.2. Salient Region Highlighting Method

Extracting the intricate contours of the target with precision is a prerequisite for achieving the accurate classification of the object. Numerous approaches have been proposed by researchers to tackle this challenge [23]. In addition to fine feature extraction, feature enhancement, and the optimization of detection head strategies, proposing an appropriate loss function is also one of the important ideas to improve training efficiency and detection accuracy. In the field of salient object detection, the development of loss functions has seen notable advancements [15]. Loss functions play a crucial role in training deep learning models for object detection tasks by quantifying the discrepancy between predicted and ground truth bounding boxes. Over the years, researchers have proposed various loss functions to address specific challenges and improve the performance of object detection systems [9].

Initially, the most commonly used loss function is the mean squared error (MSE) [24], which computes the average squared difference between predicted and ground truth bounding box coordinates. However, MSE is known to be sensitive to outliers and can lead to unstable training. To overcome these limitations, researchers introduced smooth L1 loss [24], which assigns less weight to outliers and offers a more robust alternative to MSE. Another significant development in loss functions for object detection is the introduction of focal loss [25]. Focal loss addresses the problem of class imbalance by assigning higher weights to misclassified examples, thereby focusing the model's attention on challenging samples. This loss function effectively reduces the impact of well-classified instances and allows the model to prioritize the detection of rare or difficult objects. Furthermore, recent advancements have seen the emergence of Intersection over Union (IoU) loss as an alternative to traditional regression-based losses. IoU loss directly optimizes the intersection over union metric, which measures the overlap between predicted and ground truth bounding boxes. By directly optimizing this metric, IoU loss encourages more accurate localization and precise object detection. The Lovász-Softmax loss function has been proposed for optimizing the mean Intersection over Union (mIoU) in foreground and background segmentation within image scenes [26]. Additionally, attention has been given to developing loss functions that consider object-level relationships and contextual information. Contextual loss functions take into account the spatial relationships between objects in an image, improving the detection accuracy by leveraging the contextual cues. A hierarchical saliency network is introduced that initially employs diverse global saliency cues and subsequently iteratively enhances the fine-grained details of the feature map [27]. A pyramid pooling structure combined with multi-level refinement strategy is used to optimize the spatial details and fine edges of the target in the saliency map, and has provided a good saliency performance on multiple datasets [28]. A context-aware refinement loss is utilized to con-

strain the encoder–decoder network, achieving the excellent preservation of target contours at different resolutions [29]. These loss functions aim to enhance the model's understanding of object interactions and scene understanding. Consequently, these advancements contribute to the continual improvement of object detection systems, enabling the more accurate and efficient detection of significant objects in various applications.

Different from other methods, we propose a new loss function, which comprehensively considers the brightness, contrast, structure, and other characteristics of the ship target, and constrains the predicted salient regions from both global and local aspects, thus improving the accuracy of target edge contour extraction.

3. Proposed Method

3.1. Method Overview

The structural diagram of the proposed method is illustrated in Figure 1. Similar to the classical Faster R-CNN [30], the proposed method is a two-stage network structure. In order to learn the multi-scale and deformation characteristics of ship targets, we propose a multi-scale deformation modeling-based feature extraction (MDMFE) module, which consists of two components: multi-scale feature enhanced part (MSFEP) and dual attention fusion part (DAFP). Through these two modules, we can explore the multi-scale differences between the target and the background, and simultaneously learn the deformation features of the target. By applying the region proposal network (RPN) [30] to process the extracted feature maps, preliminary salient region extraction results are obtained. On this basis, to improve the efficiency and accuracy of extracting salient regions in complex scenes, we propose a loss function to finely distinguish region differences between target and background from global and local aspects, thus realizing satisfactory highlight performance of ship target in feature maps.

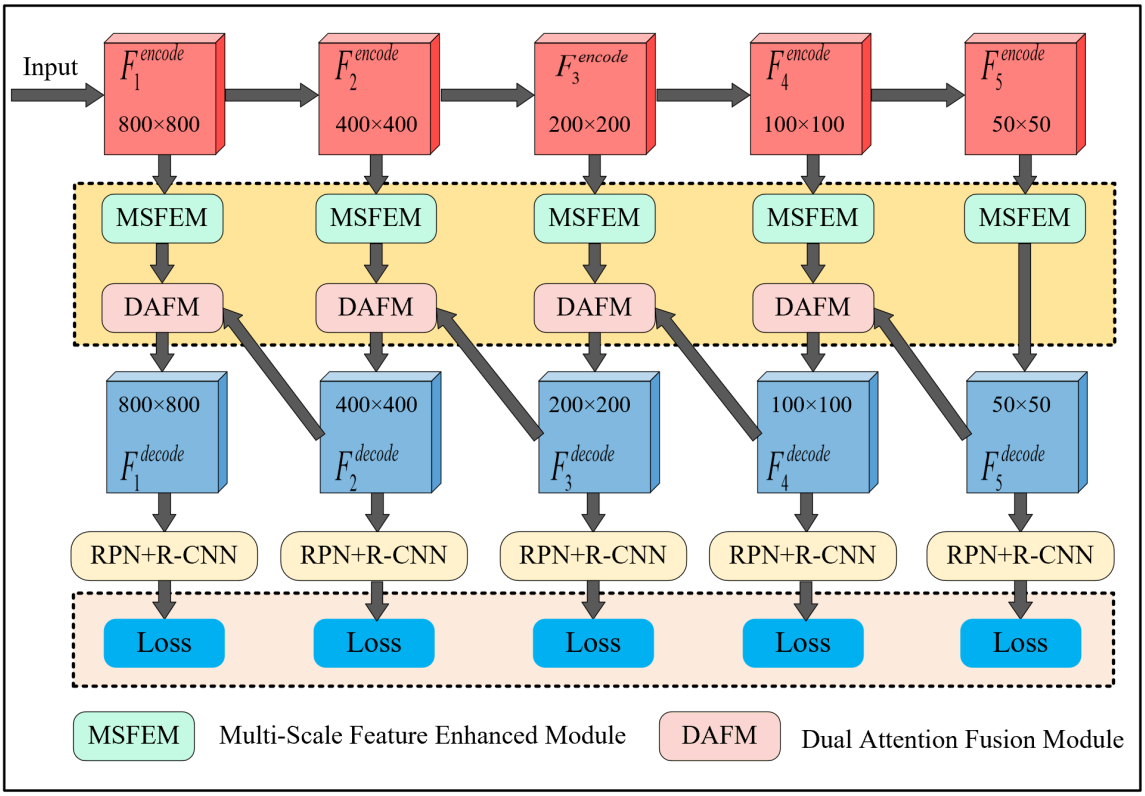

Figure 1. Flow chart of the proposed work.

3.2. Multi-Scale Deformation Modeling Based Saliency Extraction

In the realm of maritime ship detection applications, various types and scales of ships are usually distributed in wide-area remote sensing scenes. Conventional neural networks (CNNs), renowned for their remarkable feature extraction advantage, are frequently employed in the design of detection frameworks. However, CNNs inherently suffer from limitations in modeling large-scale and unknown shape transformations, stemming from the fixed sampling positions of their convolutional units on input feature maps. This results in uniform receptive fields for all activation units within the same convolutional layer, while objects at different positions may correspond to varying scales or deformations. Thus, adaptive scale and receptive field size are imperative for precise localization of targets.

In response to the above issue, we propose the multi-scale deformation modeling-based feature extraction module to enhance the modeling ability of CNN for scale and deformation. As illustrated in Figure 2, this module is mainly composed of two components: multi-scale feature enhanced part (MSFEP) and dual attention fusion part (DAFP). The MSFEP adds three feature extraction branches of atlas convolution, and applies multiple receptive field fusion to learn context relations at different scales, thus improving the learning ability of difference features between target and background. This module can easily replace the common module in typical CNN and achieve end-to-end training through backpropagation. Furthermore, by applying the additional deformable convolutional neural network, we improve the ability to model multi-scale deformation in regions or targets of interest. In addition, the DAFP is then introduced to fully fuse the extracted features along the spatial and channel dimension.

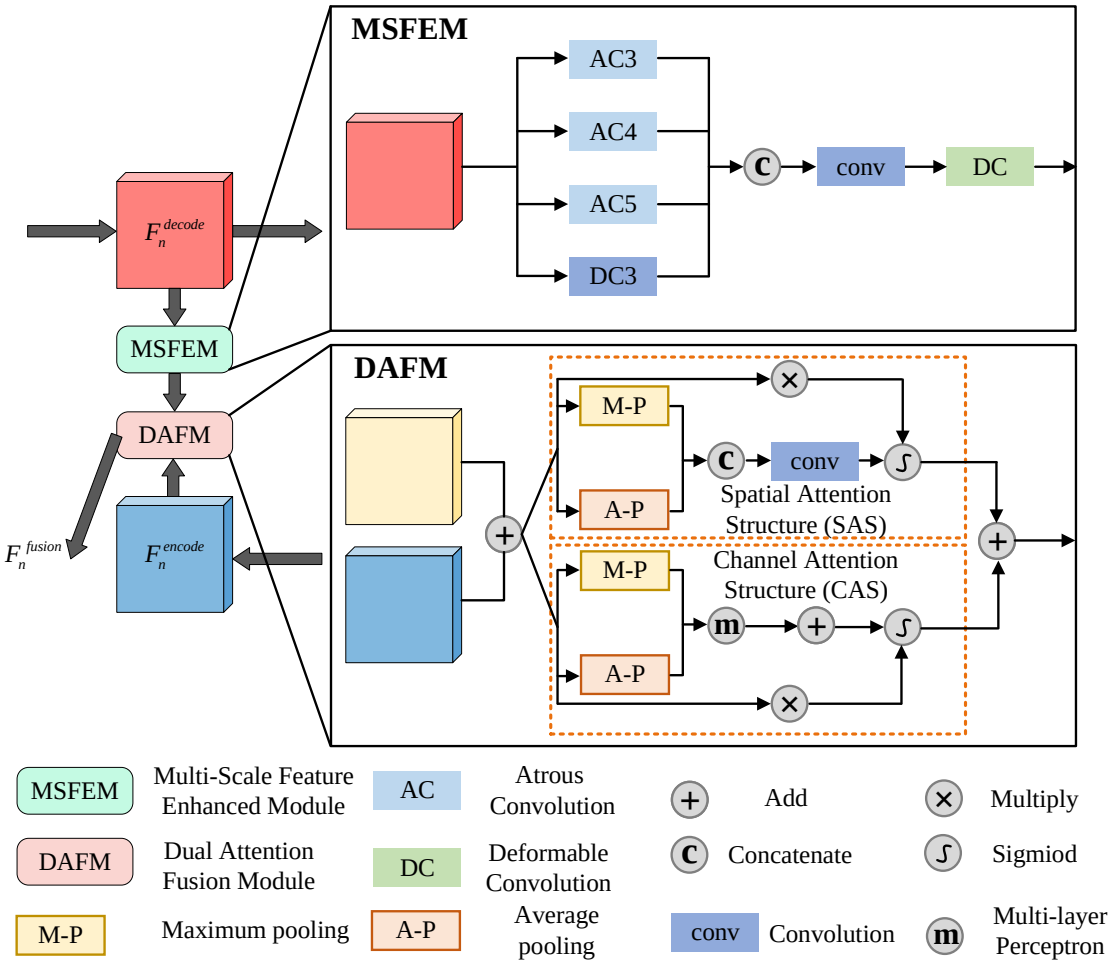

Figure 2. The structure of the MDMFE module.

Specifically, we adopt the encode–decode framework to achieve the salient object detection. Within the framework, a skip-connection structure is applied to provide additional multi-scale and multi-level features by introducing the encoding features during the decoding process. It helps to reduce the compression and disappearance of detailed information. The first multi-scale-feature-enhanced part is applied in each level during the encoding process to extract multi-scale information of the targets. Firstly, we adopt atrous convolution to obtain features at different scales, which is beneficial for capturing the scale variation of targets. By introducing the dilated rate r, atrous convolution can enlarge the receptive field of the network and enhance the ability of feature extraction without additional computation. For the n-th level of features in the encoding process, we apply the atrous convolution with dilated rates of 3, 4, and 5 in parallel to extract multi-scale features. The 1×1 convolution is also used to preserve the information at the original resolution. Finally, the multi-scale features are aggregated by channel concatenation and 1×1 convolution. Here, 1×1 convolution is used to restore the same number of channels as the input feature map. This process is formulated as:

$$F_n = \mathrm{conv}(\mathrm{concat}(\mathrm{AC}_3(F_n), \mathrm{AC}_4(F_n), \mathrm{AC}_5(F_n), \mathrm{conv}(F_n))) \tag{1}$$

where $\mathrm{AC}_n(\cdot)$ represents the atrous convolution with dilated rate of n, $\mathrm{concat}(\cdot)$ is the channel concatenation operator, and $\mathrm{conv}(\cdot)$ is the 1×1 convolution.

At the same time, to adapt to the variable shape of the targets, after extracting the multi-scale features through atrous convolution at the fixed receptive fields, we introduce the deformable convolution to automatically adjust the shape change modeling. Compared with ordinary convolution, deformable convolution additionally learns offset coordinate information, so that receptive field in the convolution process will be dynamically refined. For the input feature map F_n^1, the output at position p_0 via deformable convolution is calculated as

$$F_n^2(p_0) = \sum_{p_n \in R} w(p_n) \cdot F_n^1(p_0 + p_n + \Delta p_n) \tag{2}$$

where $R = \{(-1,-1), (-1,0), ..., (0,-1), (1,1)\}$ is the range of the convolution kernel, w represents the weights, p_n and Δp_n are the index and offset coordinates of the convolution kernel, respectively.

To improve the training efficiency of features, we propose the dual attention fusion part to fuse the encoding and decoding features by introducing the dual attention mechanism [31]. For the encoded feature map F_n^2 enhanced by MSFEP and decoded feature map F_n^{encode}, we perform preliminary fusion by addition operation. And then, we achieve more adequate fusion through channel attention structure (CAS) and spatial attention structure (SAS). The feature map obtained by the preliminary fusion is formulated as

$$F_n^3 = F_n^2 + F_n^{decode} \tag{3}$$

By modeling the importance of different feature channels, CAS assigns the weights along channel dimension to enhance effective features and suppress irrelevant interference. To achieve that, maximum pooling and average pooling are adopted to embed spatial dimension information in channel significance. Then, we use the shared multilayer perceptron (MLP) to better capture the correlation between different channels. Finally, the weights are obtained through the sigmoid activation function and assigned to the preliminary fusion feature map F_n^2. This process can be written as

$$F_n^{channel} = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(F_n^3\right)\right) + \mathrm{MLP}\left(\mathrm{MaxPool}\left(F_n^3\right)\right)\right) \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid activation function, and $\mathrm{MLP}(\cdot)$ represents the multilayer perceptron.

By generating masks along the spatial dimension, the SAS achieves the feature enhancement of specific regions while suppressing the feature representation of irrelevant

regions. The spatial correlation between features is first aggregated by maximum pooling and global pooling along the channel dimension. After performing channel concatenation, spatial saliency is generated by a 7×7 convolution. The weights are then obtained using the sigmoid activation function and multiplied with the preliminary fusion feature map F_n^2, which is formulated as

$$F_n^{spatial} = \sigma\left(\text{conv}\left(\text{concat}\left(\text{MaxPool}\left(F_n^3\right), \text{AvgPool}\left(F_n^3\right)\right)\right)\right) \tag{5}$$

where $\text{conv}(\cdot)$ represents the 7×7 convolution.

We apply the CAS and SAS in parallel to extract the enhanced features along the channel and spatial dimension in parallel. The final fusion is then achieved by addition operation.

$$F_n^{fusion} = F_n^{channel} + F_n^{spatial} \tag{6}$$

After the above steps, we apply the minimum bounding rectangle to mark all potential salient regions, thus obtaining the potential region extraction results for the first stage.

3.3. Loss Function for Fine Region Highlighting

For the salient ship detection, the binary cross entropy (BCE) loss function [15] is often applied to measure the pixel difference between the extracted suspected region and the ground truth region. The BCE loss is defined as

$$L_{BCE} = \sum_{p \in P, g \in G} -[g \log p + (1-g)\log(1-p)] \tag{7}$$

where P represents the predicted saliency map, G represents the ground truth region. $p \in (0,1)$ and $g \in (0,1)$ mean the probability of belonging to the salient area and ground truth, respectively.

In order to improve the accuracy of image training, the predicted foreground region should overlap with the ground truth as much as possible, and changes in object scale should not cause fluctuations in this phenomenon. We first measure the probability of overlapping regions between the P and G, and this dependency relationship can be characterized as

$$R(P,G) = \frac{P \cap G}{P \cup G} = \frac{2pg}{p+g} \tag{8}$$

To ensure the monotonicity of loss function constraints, we convert this overlapping probability into an offset loss function as

$$L_{Offset} = 1 - \frac{2\sum pg}{\sum p + \sum g} \tag{9}$$

Obviously, by limiting the offset loss function to a certain range, we can ensure the high consistency of the edge contour of the extracted region and the ground truth region.

The above two loss functions mainly constrain the extracted saliency regions from the structural perspective. In fact, the brightness of the predicted region and the contrast between the region and background are also key influencing factors affecting subsequent classification performance. Considering that the structural similarity index can comprehensively characterize the attributes of the image from three aspects: brightness, contrast, and structure, assuming that x and y are corresponding patches of the predicted and ground truth regions, respectively, then the structural similarity loss is applied as

$$L_{SSIM} = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)} \tag{10}$$

where μ_x, μ_y represent the mean values of the patches. σ_x and σ_y are the standard deviation. σ_{xy} means their covariance.

Actually, the BCE loss is calculated based on local pixel positions, while the offset loss is calculated for global regions. By combining the two penalty terms, the global and local constraint can be applied to image training, thus finely learning the target edge contour. Considering the constraints of brightness and contrast comprehensively, therefore the final loss can be characterized as

$$L_{IOU} = \alpha L_{BCE} + \beta L_{Offset} + \gamma L_{SSIM} \tag{11}$$

where α, β and γ are weight parameters, which are utilized to adjust the different contribution relationship. During the experiments, we optimize these parameters to focus on edge pixels, thus enhancing the ability to distinguish between the edge and the interior of the ship targets.

Since the proposed network is a two-stage network, the comprehensive loss can be defined as

$$L_{last} = L_{RPN} + L_{IOU} + L_{R-CNN} \tag{12}$$

where L_{RPN} and L_{R-CNN} represent the losses of RPN and classification module, respectively, and each term includes losses in both classification and regression aspects. Their detailed definition can be found in classical two-stage Faster R-CNN work [30].

4. Experimental Results

4.1. Experimental Settings

4.1.1. Datasets Description

To verify the correctness and robustness of the proposed algorithm, the public high-quality HRSC2016 dataset [32] is employed for performance analysis, which contains 1061 images of port scenes. Considering that in practical ship detection applications, port facilities and environmental factors such as clouds and sea clutter are interference sources, while the images collected from the HRSC dataset mainly include port facility elements, lacking examples of natural factors. To address this issue, we select 500 scene images containing natural scene elements from the Airbus ship detection challenge dataset and high-resolution commercial satellite images such as worldview-3 and Jilin-1 for experiments. The resolution of these satellite images is approximately from 0.3 m to 0.8 m. Considering the fact that the actual satellite image size is too large for training and testing, we crop the satellite image size to be between 800 × 800 and 1800 × 1800 pixels.

The dataset we used contains a total of 1561 images, and the allocation ratio of the training and testing sets is approximately 8:2. As we mainly focus on the detection of ship targets, the labels of the dataset images are all ship. Some typical samples of the new dataset are shown in Figure 3.

4.1.2. Evaluation Metrics

The precision, recall, and average-precision (*AP*) metrics, the widely accepted benchmarks, are employed to quantitatively evaluate the performance of our saliency model. The precision can describe the proportion of true positives among the predicted positives, and the recall can illustrate the proportion of true positives correctly identified. Combining these two indicators can provide a comprehensive overview of the model's performance across different operating points, enabling a nuanced analysis of its effectiveness in capturing the relevant information. Actually, the average precision is a comprehensive description for precision and recall. It is a single-value metric that summarizes the trade-off between precision and recall into a single score, which allows for a concise assessment of the overall performance of the model. By utilizing these metrics, we precisely evaluate the performance of our saliency model, enabling us to make informed conclusions about its effectiveness and suitability for the salient region extraction task. These metrics are given as

$$precision = \frac{|P \cap G|}{|P|}, recall = \frac{|P \cap G|}{|G|} \tag{13}$$

$$AP = \int_0^1 precision(recall)d(recall) \tag{14}$$

where P and G, respectively, denote the predicted saliency map and the ground truth [3].

The efficacy of the overall ship detection is evaluated through the utilization of two pivotal metrics: the average accuracy (AC) and the false alarm rate (FAR). These quantitative measures provide a comprehensive assessment of the algorithm's application potential, encapsulating the accuracy of object prediction and the probability of false classification. The definitions of these metrics are as

$$AC = \frac{\text{number of correctly detected ships}}{\text{number of real ships}} \tag{15}$$

$$FAR = \frac{\text{number of detected false alarms}}{\text{number of detected candidates}} \tag{16}$$

Figure 3. Typical samples of the employed dataset (the first row is from the HRSC2016 dataset, the second is from commercial satellite images).

4.2. Performance Analysis

4.2.1. Ablation Analysis

4.2.2. Implementation Details

The experiments are conducted on a high-performance workstation equipped with an Nvidia RTX 2080 GPU. The proposed algorithm is implemented through the renowned PyTorch framework, which is known for its flexibility and efficiency in developing deep learning models. To extract diverse and rich features from input images, we select the widely used and pre-trained ResNet-101 as the backbone for our algorithm. This choice is made based on the extensive research and empirical evidence that highlight the efficacy of ResNet-101 in various computer vision tasks. During the training process, we conduct a comprehensive investigation on various hyperparameters to ensure optimal model performance. The number of training epochs is set to 250, which can achieve a balance between model convergence and avoiding overfitting. The batch size for training is set to 4, which is set to optimize the GPU memory usage and computational efficiency. In our experiments, we employ the stochastic gradient descent (SGD) strategy to calculate the loss function during the training stage and update the model parameters. The learning rate is initialized

at 0.001, and we apply a weight decay of 0.0003 and momentum of 0.9 to further enhance the model's optimization process, ensuring robustness and stability.

In order to verify the necessity and effectiveness of the proposed modules, the ablation experiments are carried out. First, the proposed learning-based saliency model is compared with several advanced saliency extraction methods, including the Amulet [33], Pase [34], and ERPNet [35] methods. Figure 4 shows the saliency extraction results of different methods. As shown in the Figure 4, for the first scene image containing complex port facility interference (see the image in the first column of Figure 4), the Amulet method has a poor effect on distinguishing the background and foreground, and numerous backgrounds are introduced into the saliency map. Although the Pase method and ERPNet method can suppress the prominent areas of buildings such as ground houses and warehouses, they find it difficult to effectively distinguish the docks in the scene, mainly because the docks and ships both have symmetry and their sizes are relatively close. Our method can mine and fuse features at different levels, thus accurately highlighting multiple ship targets in the image. In the second scene, shadows in the image are the main interference to the extraction of ship areas. The Amulet method mistakenly treats shadows as salient regions due to the lack of prior information about ships. The Pase and ERPNet method both highlight most areas of the ships, but the Pase method does not distinguish the difference between the ship and the dock. The ERPNet method applies the suppression strategy for the low brightness area, which can better restrain the dock area but also affect the retention of some low dynamic range areas of the ships themselves. Although there are certain non-target regions in our method results, the shape and position features of the ships are most fully preserved. In view of the third scene image containing thick cloud interference, the Amulet misses the detection of ships in cloud shadows, while the Pase and ERPNet methods accurately extract the contours of the three ships, but there are still varying degrees of the highlighting of thick clouds. In contrast, our method can achieve excellent suppression of cloud regions. In the scene of fragmented cloud interference (as shown in the fourth column of Figure 4), the Amulet and Pase methods do not effectively distinguish the significance of the background and foreground, resulting in both the target and some fragmented clouds being highlighted. The ERPNet and our method can effectively extract the ship target, but the former introduces some waves around the ship. Our method can learn the target characteristics more finely and has a better inhibitory effect on the sea waves caused by navigation.

In order to more directly display the gain effect of each module on the entire algorithm, Table 1 lists the objective experimental evaluation results on the selected dataset. It can be seen that when using the Faster R-CNN original network with IOU set to 0.5, the benchmark value of the AP index is 84.2%. When configuring the MDMFE module on the benchmark network, the AP value increased by 4.5%. When the network further imposes the constraint of the proposed loss function, the AP of the overall algorithm even increases by 7.0%. The experimental results with an IOU set to 0.75 also reflect a similar growth pattern. From the above analysis, we can find that both the proposed MDMFE and loss function have a positive effect on the detection algorithm.

Table 1. Evaluation results of ablation experiments.

Methods	Loss Function	AP(%)@IOU0.5 (%)	AP(%)@IOU0.75 (%)
Faster R-CNN	The original function	84.2	81.6
Faster R-CNN + MDMFE	The original function	88.7 (+4.5)	85.4 (+3.8)
Faster R-CNN	The proposed loss function	86.3 (+2.1)	83.4 (+1.8)
Faster R-CNN + MDMFE	The proposed loss function	91.2 (+7.0)	87.9 (+6.3)

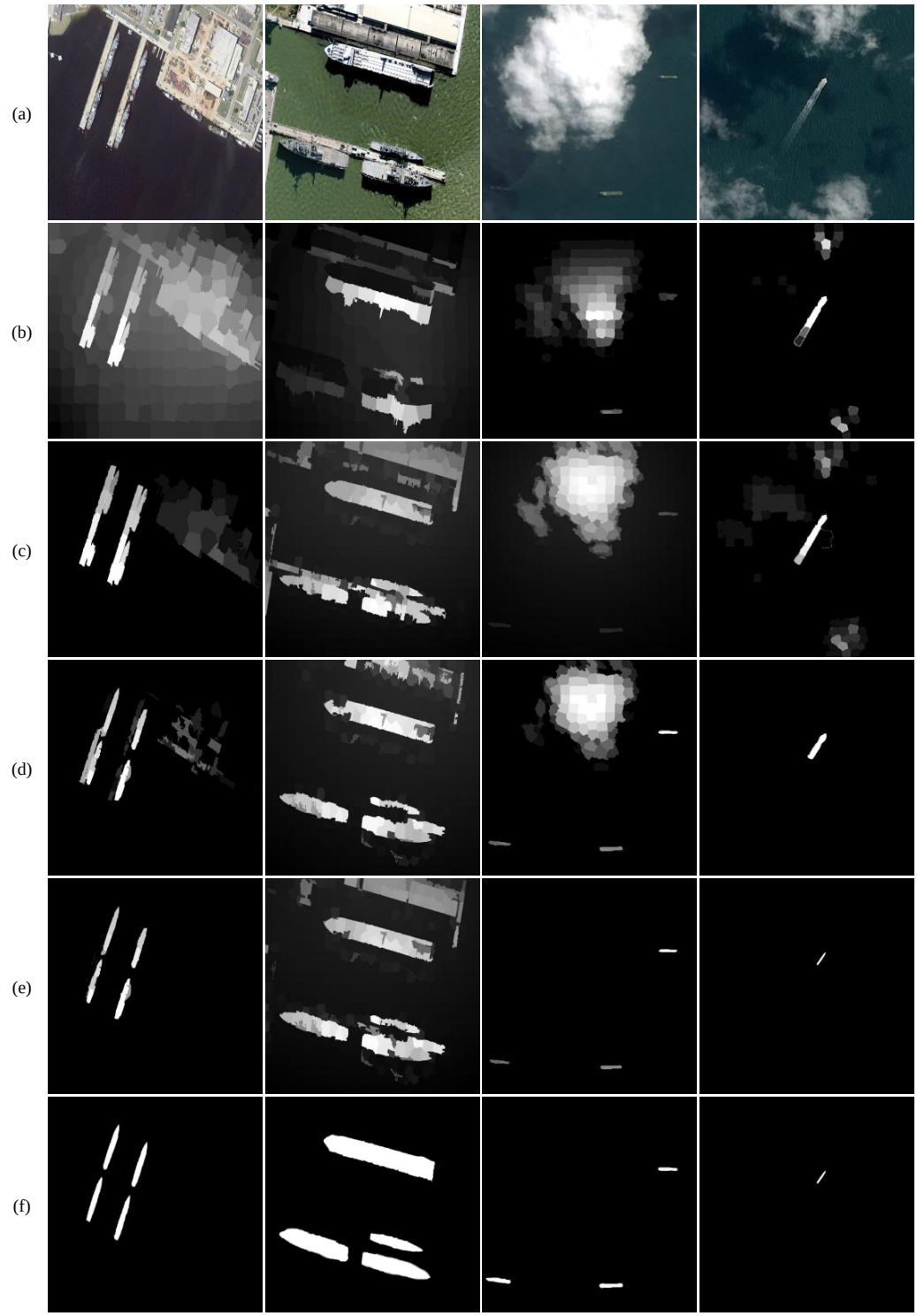

Figure 4. Comparison results of the proposed method and several advanced saliency models on the considered dataset: (**a**) input image, (**b**) Amulet, (**c**) Pase, (**d**) ERPNet, (**e**) proposed method and (**f**) ground truth.

4.2.3. Algorithm Performance Comparison

In order to assess the efficacy and robustness of the proposed model, a comparative analysis is conducted against eleven representative models for ship detection on the established dataset. The models considered in this evaluation encompass well-established approaches such as the conventional Faster R-CNN [30], YOLOv4 [36], YOLOv7 [37],

R2CNN [38], RRPN [39], Gliding Vertex [40], CenterMap-Net [41], RITSD [42], R3Det [43], FES-SPB [44], and MSSDet [45]. These models have been widely acknowledged for their excellence in object detection applications. Notably, the first three methods are network architectures designed for general object detection, while the remaining nine models are specifically tailored for target detection in remote sensing images and have gained prominence in recent years. For the purpose of fair comparison, all methods being compared have employed the same data augmentation strategy.

In order to demonstrate the applicability of the proposed method across various image scenes, we provide the detection results for several typical scene conditions in Figure 5. To enhance clarity, we have removed the confidence labels from the resulting images. We can observe that the tested image scenes encompass various background elements, such as varying degrees of clouds, obvious shadows, and complex port facilities. Furthermore, the considered scenes also include performance validation under multi-object conditions. The experimental results from diverse scenes reveal that this proposed method outperforms all comparative approaches in terms of average accuracy and false detection rate.

Specifically, for the scene of port facility interference, as shown in the first column in the Figure 5, the Faster R-CNN and Gliding Vertex methods have missed detection of a ship target and mistakenly detected multiple port facilities as targets; besides, the R3Det method can suppress the interference of environmental elements, but it detects two ships docked side-by-side as one target, resulting in missed detection. In contrast, our method can extract the effective features of the target, thus accurately detecting all ship targets from the complex environment. For the scene image with obvious shadow interference, the Faster R-CNN method has many missed and false detections, and the Gliding Vertex and R3Det methods have detected most ship targets, but the shadow interference has not been completely removed, resulting in a certain deviation between the boundary box of the detected target and the real boundary box. Our method can better suppress the influence of shadow on the location of target boundary box, thus accurately describing the contour boundary of the target. For the application of thin clouds and occluded scenes, all methods have successfully detected the ships near the center of the image, but only R3Det and our method accurately selected the targets at the edge of the image, that is, they have the ability to detect the targets under certain occluded conditions. For the scene conditions of thick clouds and shadow interference, the Faster R-CNN, Gliding Vertex, and R3Det methods all detect two ships with ideal lighting conditions. Moreover, the R3Det method detects the target in the shadow in the upper left corner of the image, but the three methods do not distinguish the ship in the other shadow in the lower right corner of the image. Our method can suppress the interference of shadows and clouds, finely distinguishing all ship targets.

The quantitative evaluation results of the utilized dataset are demonstrated in Table 2. It can be found from the evaluation index results that our algorithm can obtain the highest average accuracy and the lowest false alarm rate. In contrast, the R2CNN method obtains the lowest average accuracy, and the RRPN method shows the highest false alarm rate. The performance of the R3Det method is second only to that of the proposed method. Interestingly, this method also incorporates a feature refinement module to accurately learn the position information of the target. However, this method utilizes a typical CNN structure, which has inherent limitations in modeling the scale and shape variations of ships. In contrast, our approach leverages multi-scale deformable convolutions to capture the scale changes and shape deviations of the target, while considering the semantic consistency across different extraction branches. This enables the more efficient utilization of features and consequently achieves a better detection performance. Therefore, through the above analysis of quantitative and qualitative results, we can see that our method has better detection performance than all comparison methods, and can effectively detect targets under the interference of complex scenes such as clouds, shadows, complex port facilities, etc.

Table 2. Quantitative evaluation results of different methods on the applied dataset (all methods adopt the 2012 evaluation metric).

Methods	Backbone	AC(%)	FAR(%)
R2CNN	ResNet-101	68.35	13.73
RRPN	ResNet-101	79.63	14.28
Faster R-CNN	ResNet101	84.20	13.45
YOLOv4	CSPDarknet53	85.12	10.38
YOLOv7	ELAN-Net	87.68	7.48
Gliding Vertex	ResNet-101	88.22	8.53
CenterMap-Net	ResNet-101	92.84	5.84
RITSD	ResNet-101	92.92	6.36
MSSDet	ResNet-101	93.05	9.43
FES-SPB	ResNet-101	93.20	7.85
R3Det	ResNet-101	94.61	5.50
Proposed	ResNet101	**94.89**	**5.25**

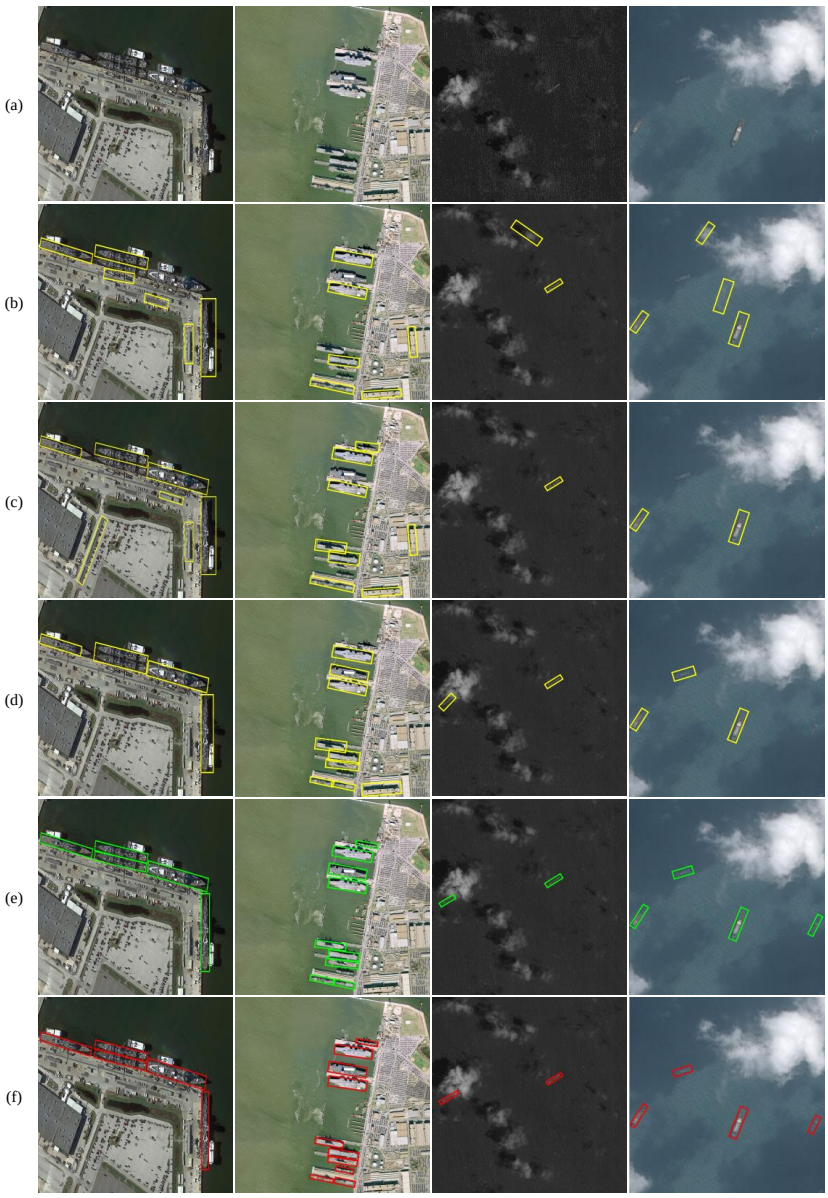

Figure 5. Detection results provided by the proposed method and the compared models: (**a**) original images, (**b**) Faster R-CNN, (**c**) Gliding Vertex, (**d**) R3Det, (**e**) the proposed method, and (**f**) ground truth.

5. Conclusions

In this paper, we propose a novel ship-detection model which can effectively detect ships of various scales and deformations in complex scenes, including intricate harbor facilities, varying degrees of cloud cover, and significant shadow interferences. To address challenges posed by complex environmental interferences, we design a saliency extraction structure that jointly models the multi-scale and deformation characteristics of targets. Leveraging dilated convolutions and deformable convolutions, we learn the distinctive features of targets and backgrounds, obtaining relatively accurate saliency maps highlighting potential ship candidates. Furthermore, we present a new loss function, incorporating brightness, contrast, and structural aspects, to refine the extraction of salient regions, thus enhancing the final classification performance of targets. The experimental results on high-quality datasets show that our method achieves the highest average accuracy and lowest false alarm rate compared to other excellent benchmark testing methods, which also proves the effectiveness and robustness of our method for ship detection in complex environments.

In the future, considering that our detection model still has slightly higher false detections, we plan to further explore more features that can more effectively represent ship targets. In addition, we will further optimize the strategy to obtain higher granularity target interpretation capabilities, such as ship recognition.

Author Contributions: Methodology, writing the original draft, C.L.; validation, revising and editing, funding acquisition, J.H.; validation, revising and editing, D.W.; investigation, H.L.; supervision, project administration, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (62305088).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zhang, T.; Quan, S.; Yang, Z.; Guo, W.; Zhang, Z.; Gan, H. A two-stage method for ship detection using PolSAR image. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
2. Kanjir, U.; Greidanus, H.; Oštir, K. Vessel detection and classification from spaceborne optical images: A literature survey. *Remote Sens. Environ.* **2018**, *207*, 1–26. [CrossRef]
3. Hu, J.; Zhi, X.; Zhang, W.; Ren, L.; Bruzzone, L. Salient Ship Detection via Background Prior and Foreground Constraint in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3370. [CrossRef]
4. Nie, T.; Han, X.; He, B.; Li, X.; Liu, H.; Bi, G. Ship detection in panchromatic optical remote sensing images based on visual saliency and multi-dimensional feature description. *Remote Sens.* **2020**, *12*, 152. [CrossRef]
5. Li, L.; Zhou, F.; Zheng, Y.; Bai, X. Saliency detection based on foreground appearance and background-prior. *Neurocomputing* **2018**, *301*, 46–61. [CrossRef]
6. Song, S.; Jia, Z.; Yang, J.; Kasabov, N. Salient detection via the fusion of background-based and multiscale frequency-domain features. *Inf. Sci.* **2022**, *618*, 53–71. [CrossRef]
7. Lv, P.; Yu, X.; Chi, J.; Wu, C. Saliency detection via absorbing Markov chain with multi-level cues. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2022**, *105*, 1010–1014. [CrossRef]
8. Ma, F.; Sun, X.; Zhang, F.; Zhou, Y.; Li, H.C. What catch your attention in SAR images: Saliency detection based on Soft-Superpixel lacunarity cue. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 1–17. [CrossRef]
9. Ji, Y.; Zhang, H.; Zhang, Z.; Liu, M. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Inf. Sci.* **2021**, *546*, 835–857. [CrossRef]
10. Hu, J.; Zhi, X.; Shi, T.; Zhang, W.; Cui, Y.; Zhao, S. PAG-YOLO: A Portable Attention-Guided YOLO Network for Small Ship Detection. *Remote Sens.* **2021**, *13*, 3059. [CrossRef]
11. Jiang, W.; Liu, M.; Peng, Y.; Wu, L.; Wang, Y. HDCB-Net: A Neural Network with the Hybrid Dilated Convolution for Pixel-Level Crack Detection on Concrete Bridges. *IEEE Trans. Ind. Inform.* **2020**, *17*, 5485–5494. [CrossRef]

12. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
13. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
14. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
15. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7479–7489.
16. Zhu, W.; Liang, S.; Wei, Y.; Sun, J. Saliency optimization from robust background detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2814–2821.
17. Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; Yang, M.H. Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3166–3173.
18. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 353–367.
19. Wang, W.; Shen, J.; Shao, L. Video salient object detection via fully convolutional networks. *IEEE Trans. Image Process.* **2017**, *27*, 38–49. [CrossRef]
20. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
21. Hu, J.; Zhi, X.; Jiang, S.; Tang, H.; Zhang, W.; Bruzzone, L. Supervised Multi-Scale Attention-Guided Ship Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
22. Song, S.; Miao, Z.; Yu, H.; Fang, J.; Zheng, K.; Ma, C.; Wang, S. Deep domain adaptation based multi-spectral salient object detection. *IEEE Trans. Multimed.* **2020**, *24*, 128–140. [CrossRef]
23. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
24. Ji, W.; Li, J.; Yu, S.; Zhang, M.; Piao, Y.; Yao, S.; Bi, Q.; Ma, K.; Zheng, Y.; Lu, H.; et al. Calibrated RGB-D salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9471–9481.
25. Wang, Y.; Zhao, X.; Hu, X.; Li, Y.; Huang, K. Focal boundary guided salient object detection. *IEEE Trans. Image Process.* **2019**, *28*, 2813–2824. [CrossRef] [PubMed]
26. Berman, M.; Triki, A.R.; Blaschko, M.B. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4413–4421.
27. Liu, N.; Han, J. Dhsnet: Deep hierarchical saliency network for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 678–686.
28. Wang, T.; Borji, A.; Zhang, L.; Zhang, P.; Lu, H. A stagewise refinement model for detecting salient objects in images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4019–4028.
29. Islam, M.A.; Kalash, M.; Rochan, M.; Bruce, N.D.; Wang, Y. Salient Object Detection using a Context-Aware Refinement Network. In Proceedings of the BMVC, London, UK, 4–7 September 2017.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
31. 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.
32. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; Volume 2, pp. 324–331.
33. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 202–211.
34. Wang, W.; Zhao, S.; Shen, J.; Hoi, S.C.; Borji, A. Salient object detection with pyramid attention and salient edges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1448–1457.
35. Zhou, X.; Shen, K.; Weng, L.; Cong, R.; Zheng, B.; Zhang, J.; Yan, C. Edge-guided recurrent positioning network for salient object detection in optical remote sensing images. *IEEE Trans. Cybern.* **2022**, *53*, 539–552. [CrossRef]
36. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
37. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.

38. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R 2 cnn: Rotational region cnn for arbitrarily-oriented scene text detection. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3610–3615.
39. Nabati, R.; Qi, H. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3093–3097.
40. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef] [PubMed]
41. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning center probability map for detecting objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]
42. Wang, T.; Li, Y. Rotation-invariant task-aware spatial disentanglement in rotated ship detection based on the three-stage method. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
43. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 3163–3171.
44. Ren, Z.; Tang, Y.; He, Z.; Tian, L.; Yang, Y.; Zhang, W. Ship detection in high-resolution optical remote sensing images aided by saliency information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
45. Chen, W.; Han, B.; Yang, Z.; Gao, X. MSSDet: Multi-Scale Ship-Detection Framework in Optical Remote-Sensing Images and New Benchmark. *Remote Sens.* **2022**, *14*, 5460. [CrossRef]