applied sciences

MDPI

Article

Saliency Guided DNL-Yolo for Optical Remote Sensing Images for Off-Shore Ship Detection

Jian Guo [1], Shuchen Wang [2] and Qizhi Xu [2,*]

[1] Beijing Institute of Technology, School of Computer Science and Technology, Beijing 100081, China; guojian501@126.com
[2] Beijing Institute of Technology, School of Mechatronical Engineering, Beijing 100081, China; scwang@bit.edu.cn
* Correspondence: qizhi@bit.edu.cn

Abstract: The complexity of changeable marine backgrounds makes ship detection from satellite remote sensing images a challenging task. The ubiquitous interference of cloud and fog led to missed detection and false-alarms when using imagery-based optical satellite remote sensing. An off-shore ship detection method with scene classification and a saliency-tuned YOLONet is proposed to solve this problem. First, the image blocks are classified into four categories by a density peak clustering algorithm (DPC) according to their grayscale histograms, i.e., cloudless areas, thin cloud areas, scattered clouds areas, and thick cloud areas. Secondly, since the ships can be regarded as salient objects in a marine background, the spectral residue saliency detection method is used to extract prominent targets from different image blocks. Finally, the saliency tuned YOLOv4 network is designed to quickly and accurately detect ships from different marine backgrounds. We validated the proposed method using more than 2000 optical remote sensing images from the GF-1 satellite. The experimental results demonstrated that the proposed method obtained a better detection performance than other state-of-the-art methods.

Keywords: optical satellite image; ship detection; convolutional neural networks; deep learning

Citation: Guo, J.; Wang, S.; Xu, Q. Saliency Guided DNL-Yolo for Optical Remote Sensing Images for Off-Shore Ship Detection. *Appl. Sci.* **2022**, *12*, 2629. https://doi.org/10.3390/app12052629

Academic Editor: Hyung-Sup Jung

Received: 27 January 2022
Accepted: 28 February 2022
Published: 3 March 2022

1. Introduction

Optical remote sensing images have attracted increasing attention, due to their wide imaging range and high resolution [1]. In existing studies, panchromatic (Pan) images are a data source that have been most widely studied. Thanks to its ability to uncover subtle information, it has been continuously developed in military and civilian applications [2,3], i.e., geological analysis, city planning, and military reconnaissance.

However, it is still difficult to detect ships from remote sensing images due to complex and changeable marine backgrounds. Furthermore, huge waves and different types of clouds and fog all can reduce the accuracy of detection, as shown in Figure 1. Clouds may cast shadows on the target, thus substantially changing the illumination of the surface. Scattered clouds and big waves may also become false alarms and confuse the detector. As a consequence, solving how to increase the detection of the target ship and reduce the interference of cloud and fog is the key to improving the ship detection performance.

In general, ship detection methods include the following steps: (i) image preprocessing, (ii) image feature extraction, and (iii) target ship detection. Early in the development of object detection, many methods focused on how to extract more features from the image and ignored the impact of image quality. However, the success of image processing relies upon the production of accurate imagery along with effective human interpretation. Consequently, the three aspects of ship detection all are important.

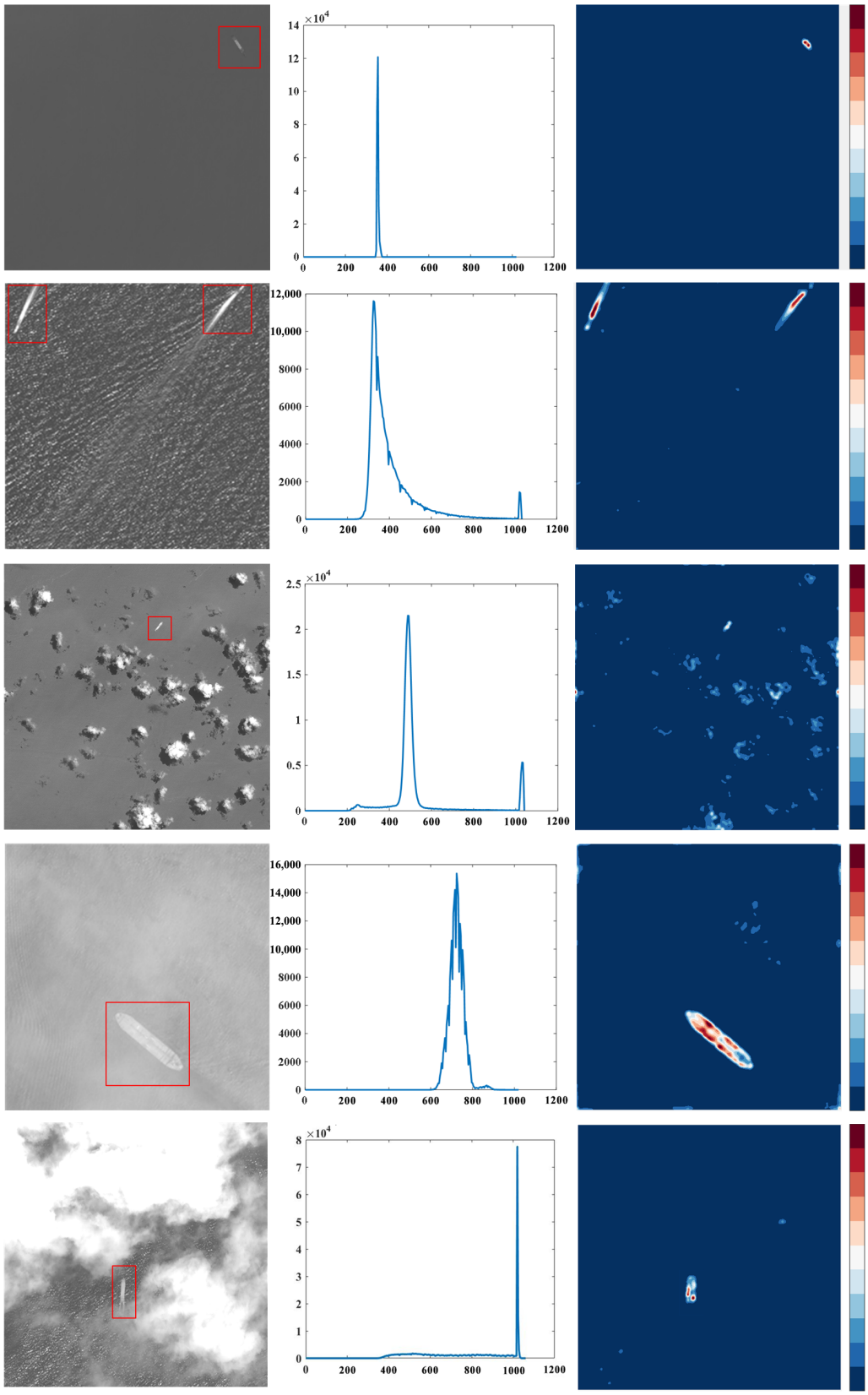

Figure 1. Marine scenes under different types of interference, namely, cloudless areas (with no waves and big waves), scattered clouds areas, thin cloud areas, and thick cloud areas. The images in the first column are the original panchromatic images from the GF-1 dataset. Ship targets are displayed in red boxes. The histograms of the grayscale statistics are illustrated in the second column. The last column contains images that show the heat maps representing the visual saliency of each scene.

1.1. Preprocessing Methods for Ship Detection

As noted earlier, remote sensing images have diverse backgrounds and complex spatial structures. Thus, different methods of image preprocessing are often required before various tasks can be performed [4,5], including image sharpening processing, contrast enhancement and cloud removal processing, etc. In [6], an N-D probability density function matching technique for the preprocessing of multitemporal images is introduced in the remote sensing domain. It can retain the data correlation structure after the probability density function matching. In addition, Ref. [7] designed a preprocessing technique to enhance the local information of the original image data. A new image structure represented by a fuzzy function is utilized to encode the information into an image. Ref. [8] proposed a preprocessing algorithm that both smooths noise and enhances edges. It consists of an improved adaptive spatially-weighted filter that can achieve both of the above functions. In [9], image preprocessing steps such as filtering, upsampling, and band registration were standardized by providing references. These steps suggest that preprocessing is an important part of many applications based on remote sensing images.

Although there are many preprocessing methods for different tasks, scene classification is still the most basic method. It can mine more targeted and valuable information, and save us from using one method to solve all problems [10]. Hence, target ship detection could benefit from this significant progress. Traditionally, scene classification methods based on object recognition are approaches that require prior information about the objects. In [11], a multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery is devised in order to classify scenes. In this method, the complementary features can be effectively combined and appropriate low-level feature descriptions can be provided for the semantic representations. It is interesting and superior to single-features methods.

Furthermore, deep learning-based methods have became the choice for solving many computer vision and remote sensing problems [12]. The authors of [13] have devised a key region or location capturing method. This method can be combined with different CNN models and prevent confusion of different categories. On the other hand, in [14], a density peak clustering algorithm (DPC) is proposed for preprocessing hyperspectral image classification. By using a density peak clustering algorithm, the HSI pixels are automatically subdivided into smaller classes, and the within-class difference caused by spectral variation is reduced. In addition, we observed that image blocks in different scenes have different grayscale histograms, as illustrated in Figure 1. This intuitive mapping relationship provides us with an idea. We take the gray histograms of all image patches as the input and then use the DPC algorithm to classify different scenes. This can quickly and effectively classify different scenes.

1.2. Ship Detection Using Feature Engineering

In the process of image processing, feature extraction and target detection are often integrated in the same algorithm. In addition, in the early stages of development, most of object detection methods are implemented by utilizing visual features (such as color, texture, local binary pattern, spatial relationships, etc.). In [15], a semisupervised model is presented to distinguish between ships and nonships. Besides common shape and texture features, this article also adds local multiple patterns to enhance feature representation. Reference [16] proposes a new detection method. It analyzes whether the sea surface is uniform through two features, and uses a new linear function to select candidate regions. Based on this, it can ignore non-candidate regions to reduce calculation time and achieve fast detection.

In [17], a robust algorithm is proposed. It combined a visual attention model with a local binary pattern (CVLBP), and was analyzed in complementary ways. This method can reduce the sensitivity of clouds and illumination, as well as having a better detection effect. Furthermore, in order to increase the separability between ships and background, reference [18] proposed to rearrange the spatially adjacent pixels in the vicinity of ships into

a vector, so that it can be endowed with some contextual information. At the same time, histograms of oriented gradients (HOGs) are used to validate real ships out of a selection of candidates. In addition, aiming to deal with complex scenes, a contour refinement and the improved generalized Hough transform (GHT)-based ship detection scheme are proposed in [19]. It can achieve a more accurate ship detection by repairing damaged bow contours and removing any erroneous candidates. However, these methods often rely upon one or more handcraft features, and they cannot find deep semantic information, so it is difficult to achieve better detection results.

In addition, since ships are often conspicuous in the background of the ocean, their salient characteristics have provided us with another research idea. We can detect the regions that represent the scene as auxiliary features for finding a ship. For example, Ref. [20] propose a spectral residual (SR) approach for visual saliency detection. It converts the input image to the spectral domain, and finds the spectral residual to obtain a saliency map. In [21], a context-aware saliency detector (CASD) was designed. The main design principle is that the context of the dominant objects is just as essential as the objects themselves. They are simple to implement, fast in processing speed, and can show good performance. However, for some complex ocean scenes the acquired saliency map may contain scattered clouds or ocean waves, so it can only be used as a reference.

1.3. Ship Detection Using Convolutional Neural Networks

With the extensive research of deep learning methods in recent years, convolutional neural networks have been increasingly used in the field of object detection [22,23]. Convolutional neural networks can automatically extract deep abstract features that are difficult to represent by hand design, so they show advantages compared to traditional machine learning algorithms. At the same time, it can also reduce manual participation and save labor costs. We describe these CNN methods from two aspects: region-based (two-stage) frameworks and unified (one-stage) frameworks [24]. The region-based framework is a two-stage cascading network. One is used to generate object proposals, and the other determines whether the desired object exists. When the successful application of deep CNNs in image classification was transferred to object detection it resulted in the first region-based CNN (RCNN) detector and showed good performance [25]. In spite of achieving high object detection quality, the RCNN still has some drawbacks. Subsequently, a series of improved algorithms were derived, such as SPPNet [26], Fast RCNN [27], Faster RCNN [28], and DeFRCN [29]. In addition, the authors of [30] developed an attention mask R-CNN to detection ships. It can accurately segment ships at the pixel level by adding a bottom-up structure to the FPN structure of a Mask R-CNN. In [31], a novel deep CNN with a hierarchical selective filtering layer is proposed to detect ships with various scales. It shows a high detection accuracy and strong robustness.

Nevertheless, region-based approaches are still slow and hard to optimize, and computationally expensive for current mobile/wearable devices. Therefore, researchers have begun to design unified detection strategies. This is different from two-stage frameworks, because one stage detection frameworks have a single method that does not separate the process of the detection proposal. All the computation will be encapsulated in a single CNN, thus making optimization easier [24]. Commonly used one-stage object detection algorithms are, DetectorNet [32], SSD [33], and YOLO series [34–37]. Compared to the methods of R-CNN series, YOLO turns the object detection problem into a regression problem. Given the input image, YOLO is able to directly return the bounding box of the target and its classification category at multiple positions of the image utilizing regression algorithm. YOLO is a convolutional neural network that can predict multiple box positions and categories at the same time. It can achieve end-to-end target detection and recognition. The fast speed of YOLO is its greatest strength. The optimized YOLO network can output multi-scale feature maps and has accurate target detection performance. Moreover, in [38], a new and improved approach CornetNet is proposed to detect objects. The paired top-left and bottom-right keypoint is utilized to substitute a set of anchor boxes. In [39], a improved

model named YOLOF is proposed. They designed two key components, dilated encoder and uniform matching, which make the method achieve considerabled effectiveness. Consequently, after comprehensive consideration, we improved the YOLOv4 network model to obtain refined ship detection results.

After the aforementioned research, we propose a saliency-tuned offshore ship detection method. We used a density peak clustering algorithm [14] as the first preprocessing step in order to classify the image blocks of different scenes. We use different saliency methods (e.g., SR [20], CASD [21], and a histogram-based contrast method (HC) [40]) to detect the saliency of each scene image. In view of image characteristics in different scenes, different saliency algorithms have different advantages and can obtain better detection results for different scene images. Then, the saliency detection results are used as the third channels and stacked with the original panchromatic image. They are treated as input data and fed into the subsequent network. Moreover, we added the disentangled non-local (DNL) module [41] into the different layers of the YOLOv4 network. The DNL module is an improved self-attention mechanism module, which can dig out the attention area and boundary information through pairwise and unary operators and enhance a network's ability to detect salient objects in an image. Finally, the extracted features are sent to the detection head to complete the detection of an offshore ship.

The main contributions of this paper are as follows: First, a DPC algorithm is introduced for scene classification of image blocks. Second, because ships can be regarded as anomaly objects, different saliency detection methods (e.g., SR, HC, and CASD) are used to extract prominent objects from different marine backgrounds. The above-mentioned first step preprocessing operation can reduce the intra-class difference of each sea surface scene category, and the second step can enhance the characteristics of the target ship. Finally, the DNL-added YOLONet is proposed to integrate the salient features of the target effectively; therefore, better detection results can be obtained. Compared with YOLO, DNL-added YOLONet enhances the ability of network context modeling by introducing an attention mechanism. We have also produced a batch of ship data sets based on the panchromatic image data from the South China Sea. For the problem of an insufficient number of ship targets in thick cloud and scattered cloud images, an experiment creating simulation samples was carried out to obtain sufficient training samples.

The remainder of this paper is organized as follows. The proposed ship detection framework is described in Section 2. The experiments and analysis are discussed in Section 3. The conclusion is drawn in Section 4.

2. Methodology

Our ship detection approach is implemented based on an DNL-added YOLONet framework with a preprocessing module. As shown in Figure 2, all image patches are divided to different classes by DPC algorithm according to their gray-histogram. Then we use scene-adapted saliency detection method to capture the salient targets of image blocks in different background. Next, integrate the saliency maps with image blocks as training samples and fed them into network. After that, the YOLOv4 network added with the DNL module (or called saliency guided DNL-YOLO) is utilized for feature extraction and target detection. These details are elaborated upon in the following subsections.

2.1. Scene Classification Using DPC Algorithm

In general, the width of remote sensing images is relatively large, so their scenes are local. Therefore, the original remote sensing image needs to be divided into sub-blocks to accurately detect targets in different scenes. Meanwhile, in order to avoid a ship being cut in half, the sub-block cutting needs to be overlapped. After obtaining the image block set, different scene images need to be classified according to certain characteristics. Furthermore, for panchromatic images, grayscale intensity distribution can accurately reflect different sea scene characteristics according to the height and position of the peaks [42]. However, the statistics-based method is not accurate enough and requires further

improvement. Density peak clustering approach, proposed on science in 2014 [14], is an algorithm that can accomplish efficient clustering of a data set. The basic idea of the algorithm is that the local density of the data cluster center is larger than that of adjacent data points, and it is far away from other data cluster centers.

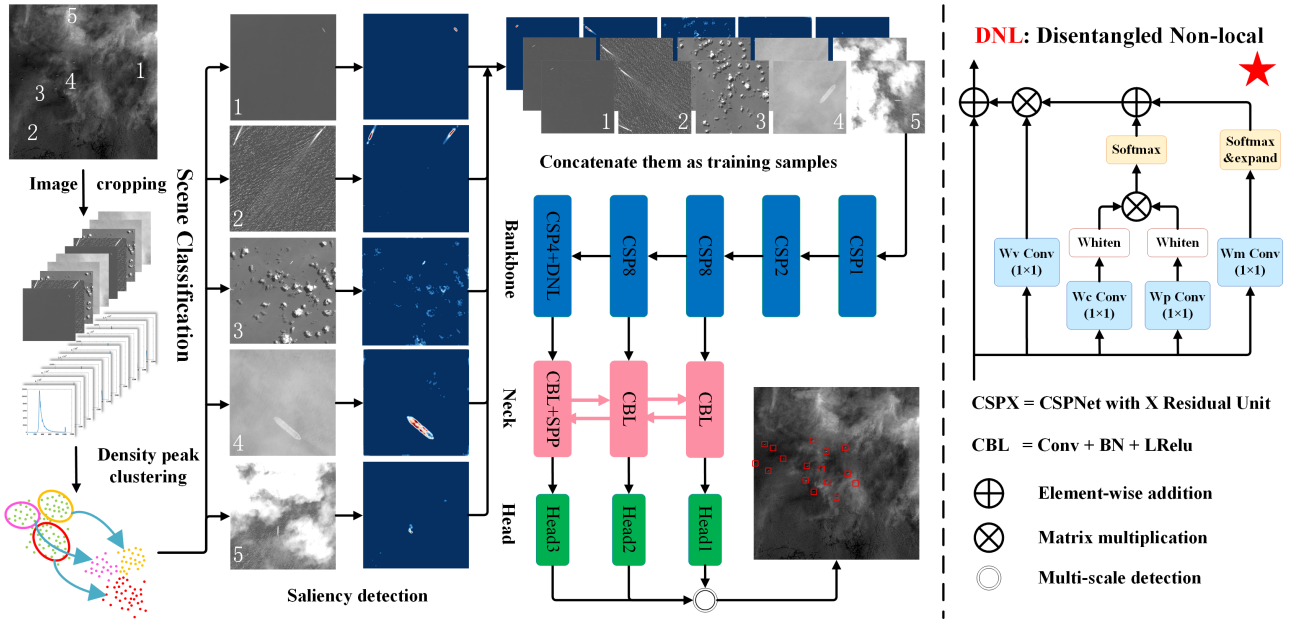

Figure 2. The overall flowchart of proposed saliency tuned YOLONet for ship detection. First, we utilized the grayscale histograms of each block to classify scenes. Second, the saliency maps were regarded as third channels. Finally, they were separately fed into the YOLO network adjusted by the DNL module.

The grayscale histograms of each sub-blocks were written as the input vector of clustering algorithm $h(i) = [h_0, h_1, \ldots, h_k, \ldots, h_{n-1}]^T$, where h_k is the quantity of pixels with each grayscale value. DPC algorithm needs to introduce 3 key parameters, namely d_{ij} (the distance between vectors), ρ_i (the local density), and δ_i (the reference distance). The distance d_{ij} between vectors are used to measure the similarity of two vectors. The more similar the two vectors, the smaller the d_{ij} is. There are different methods to calculate the distance between vectors, including Euclidean distance and Manhattan distance, etc. In contrast, we chose the Manhattan distance due to its better stability in describing similarity features and better intuitiveness in describing larger data feature differences.

$$d_{ij} = \sqrt{|h(i) - h(j)|' \sum^{-1} |h(i) - h(j)|} \tag{1}$$

The local density ρ_i was defined as follows:

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c) \tag{2}$$

where $\chi(\cdot) = 1$ when $(\cdot) < 0$, otherwise $\chi(\cdot) = 0$. d_c is a cut-off distance, defining the scope of points to be counted. In fact, ρ_i is the number of data points within the cut-off distance around the currently counted point. Thus, ρ_i can measure how much data is clustered around that point.

The reference distance δ_i was defined as follows:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{3}$$

The meaning of δ_i is the distance between data point $h(i)$ and $h(j)$ with bigger local density ρ_j and smallest distance d_{ij}. Note that for the data point with biggest global density h(m), its reference distance was defined as follows:

$$\delta_m = \max_j(d_{ij}) \tag{4}$$

We noticed that the necessary and sufficient condition for a data point to become the center of the data cluster is that the reference distance and local density are both large. In this way, isolated data points with small ρ and large δ and similar point families with a large ρ and small δ were excluded. After clustering, we divided the marine scene sub-blocks into five categories: big wave area, no wave area, thick cloud area, thin cloud area, and scattered clouds area as shown in Figure 3. For different marine scenes, we use different methods to perform the next step of saliency preprocessing in order to enhance the characteristics of the ship target.

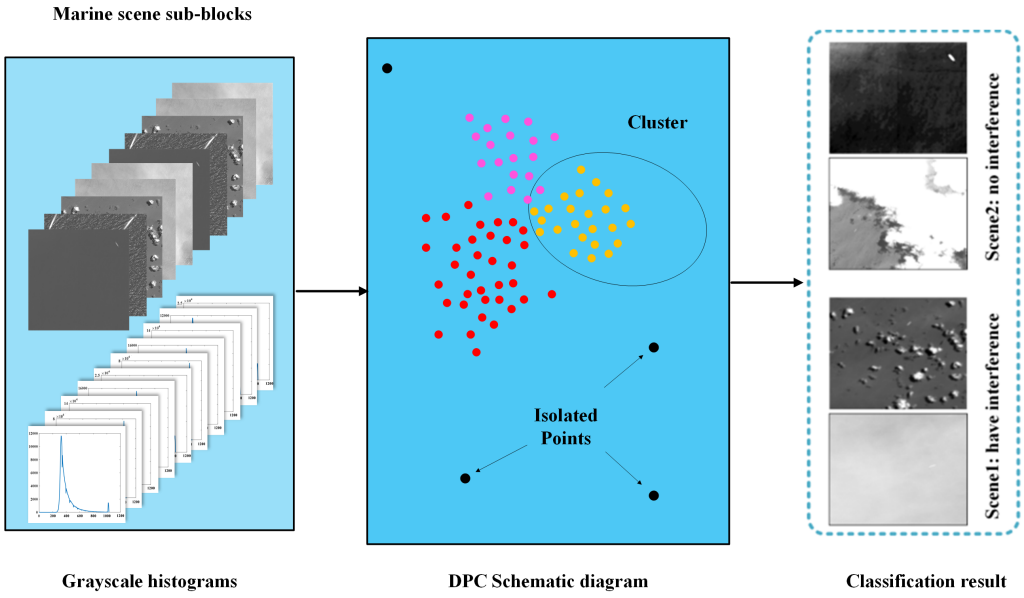

Figure 3. Marine scene classification algorithm based on density peak clustering. First, the grayscale curves of different marine scene sub-blocks are employed as the input of DPC algorithm. Then, these sub-blocks automatically gather to form 5 clusters, representing that marine scenes are classified into five categories.

2.2. Saliency Detection for Different Scenes

Due to the salient characteristics of ship targets in the ocean background, we perform saliency detection on image blocks to enhance the recognizability of ship targets. Ideally, we can perform good saliency detection on all images in different scenes using only one algorithm. However, the interference of clouds and sea waves makes this impossible. Therefore, we used different methods to process image blocks in different scenes. (i) Cloudless scene: the SR algorithm is used to extract the saliency map and the mean filter parameter is set to 200 × 200. (ii) Thin cloud scene: the above-mentioned saliency algorithm is repeated twice. The mean filter parameter is set to 400 × 400. (iii) Scattered cloud scene: the HC algorithm is used to extract the scattered cloud area and remove it. Then we used the SR algorithm to detect the salient object. The mean filter parameter is set to 400 × 400. (iv) Thick cloud scene: first, the CASD algorithm is used to detect the salient regions of the original image, and then the SR algorithm is used for a second detection. The mean filter parameter is set to 400 × 400. All images initially used median filtering to eliminate noise.

Spectral residual algorithm. This method is mainly implemented in the frequency domain. First, the image is transformed into the frequency domain through Fourier transform, and the amplitude spectrum and phase spectrum are calculated.

$$a = A(\mathfrak{F}(x)) \tag{5}$$

$$p = P(\mathfrak{F}(x)) \tag{6}$$

where $\mathfrak{F}$ represents the Fourier transform. A and P represent the acquisition of the amplitude spectrum and the phase spectrum, respectively. Then the logarithm of the amplitude spectrum is filtered through the linear space. Then, we calculated the difference between the two to obtain the spectrum residual.

$$l = \log(a) \tag{7}$$

$$r = l - h * l \tag{8}$$

where h represents the mean filtering, and r represents the calculated spectral residual. Finally, the spectral residual and phase spectrum are transformed to the space domain through inverse Fourier transform, and the saliency result map is obtained through linear filtering.

$$s = h * \mathfrak{F}^{-1}[exp(r + p)]^2 \tag{9}$$

Context-aware algorithm. This method is mainly implemented according to four major principles: local low-level considerations, global considerations, visual organization rules, and high-level factors. At the beginning, we divided the image into multiple blocks, and then we compared the block P_i corresponding to i with all other blocks P_j in the lab color space. If the P_i block has a large distance from other blocks, it is a salient block. In addition, the saliency areas are mostly clustered, thus the distance between the small blocks is relatively close. The calculation formula is as follows:

$$d(p_i, p_j) = \frac{d_c(p_i, p_j)}{1 + c \cdot d_p(p_i, p_j)} \tag{10}$$

where d_c is the Euclidean color distance of the two image blocks in Lab space, and d_p is the Euclidean position distance. For the p_i block, if the difference from any p_j block is large, it is considered a salient block. Hence, the first N (usually 65) image blocks with the smallest distance to p_i can be obtained. By calculating the difference between them and p_i block, the saliency formula is defined as follows:

$$s_i = 1 - \exp\left\{-\frac{1}{n}\sum_{n=1}^{N} d(p_i, q_n)\right\} \tag{11}$$

Meanwhile, changing N can change the saliency area scale. By taking the average of the multi-scale results, the saliency features at multiple scales can be combined to enhance the representativeness of the saliency map.

$$\bar{s}_i = \frac{1}{M}\sum_{r \in R} s_i^r \tag{12}$$

Moreover, the context correction must be added. We need to set a threshold and extract the most attended localized areas in the saliency map. The saliency value of the pixel outside the attended areas is obtained by calculating the weighted Euclidean distance between it and the nearest attended area. This is expressed as follows:

$$\hat{s}_i = \bar{s}_i\Big(1 - d_f(i)\Big) \tag{13}$$

$$d_f(i) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{14}$$

where, $d_f(i)$ is the Euclidean distance between the pixel (x_i, y_i) and the nearest focus of attention pixel (x_j, y_j). With this operation, the saliency value near the salient target can be increased, and the saliency value of the background area can be reduced, so as to achieve a better saliency detection effect.

Histogram-based contrast algorithm. This saliency detection algorithm is implemented based on color features. The saliency value of a pixel is determined by the color difference with other pixels. Assuming there are N pixels in the image, the saliency formula is expressed as:

$$s_i = \sum_{\forall j \in N} D(i, j) \tag{15}$$

when calculating by color, the formula is:

$$s_i = s(\mathfrak{c}_i) = \sum_{j=1}^{m} f_i D(\mathfrak{c}_i, \mathfrak{c}_j) \tag{16}$$

where, $\mathfrak{c}_i$ is the color value in the pixel i, m is the number of different pixel colors, and f_i is the frequency of the pixel color $\mathfrak{c}_j$ in the image. The principle of this method is simple, especially for scattered cloud scenes, it can achieve better cloud area extraction and reduce interference for saliency detection of ship targets.

2.3. Saliency Tuned YOLONet

Most of the existing methods only use convolution operations to extract features. However, non-local blocks are a popular self-attention module used to enhance the modeling ability of CNN and it shows good performance in visual applications. DNL module is an improved version of a non-local module. By using an independent softmax function and disentangling the embedding matrix, the learning complexity of the CNN is reduced, and a clearer salient area can be extracted. As a consequence, the DNL module is added on the architecture of the YOLOv4 network to combine both convolution and attention operations, which can further explore the salient features in image. Especially for ships that are prominent targets in the ocean background, it can exert a better detection performance. The main framework of the network model is illustrated in Figure 2.

DNL module. The DNL module is obtained by optimizing the non-local module, which is a popular module used to enhance the context modeling capabilities of traditional CNN. The calculation formula of the original non-local module is as follows:

$$y_i = \sum_{j \in \Omega} \omega(x_i, x_j) g(x_j) \tag{17}$$

$$g(x_j) = W_g * x_j \tag{18}$$

where Ω denotes the all pixels of input feature map. $g(\cdot)$ is the transformation function related to weight matrix W_g, which is utilized to map a point to a vector. $\omega(x_i, x_j)$ is the similarity function of pixel j (referred to as a center pixel) and other pixel i, typically instantiated by an embedded Gaussian as:

$$\omega(x_i, x_j) = \sigma\left(p_i^T c_j\right) = \frac{\exp(p_i^T c_j)}{\sum_{t \in \Omega} \exp(p_i^T c_t)} \tag{19}$$

where, $p_i = W_p x_i$ and $c_j = W_c x_j$ denote the embedding of pixel i and j, and $\sigma(\cdot)$ is the softmax function. After that, the above formula can be transformed by a whitening dot product denoted as:

$$\omega(x_i, x_i) = \sigma\left(\underbrace{(p_i - \mu_p)^T (c_j - \mu_c)}_{pairwise} + \underbrace{\mu_p^T c_j}_{unary}\right) \tag{20}$$

where the μ_p and μ_c are the averaged embedding over all of the pixels. The first *whitened* dot product term represents the pure *pairwise* relation between a pixel i and a pixel j, and the second term represents the *unary* relation where a pixel j has the same impact on all pixels i. As a consequence, the DNL module divides the original formula into two terms, paired terms and unary terms, by disentangling matrix. One of them then learns the relationship within the area, and the other learns the salient boundary. The separation of these two items can then reduce the mutual interference, thereby achieving a better feature extraction effect.

The specific frame is shown in the dashed box in Figure 2. For an input feature map of $C \times H \times W$, two 1×1 convolutions w_k, w_q is utilized to extract features firstly. Then they are calculated via a *whitening* dot product. Finally, the output features of $HW \times HW$ are obtained through the softmax function. In another branch, an independent 1×1 convolution w_m is used to extract features, and the softmax with expand operations are connected to obtain the same size feature as the above. At the end, we added the previous two results, and then a dot product with a 1×1 convolution w_m of the original input feature in order to obtain the final output.

DNL-added YOLONet. The YOLOv4 network is a recent developed optimization framework in the YOLO family for target detection. The YOLOv4 network model mainly includes five basic components: CBM, CBL, Res Unit, CSPX, and SPP, as shown in Figure 4. The CBM component is composed of convolution, batchnorm, and mish activation functions, and is the most basic structural unit. The CBL component is composed of convolution, batchnorm, and leaky_rule activation functions. The Res Unit component draws on the residual structure in the Resnet network, allowing the network to be built deeper. The CSPX component draws on the CSPNet network structure and consists of convolutional layers and X Res unit modules. The spatial pyramid pooling (SPP) component contains a multi-scale pooling structure, which can then complete multi-scale feature fusion.

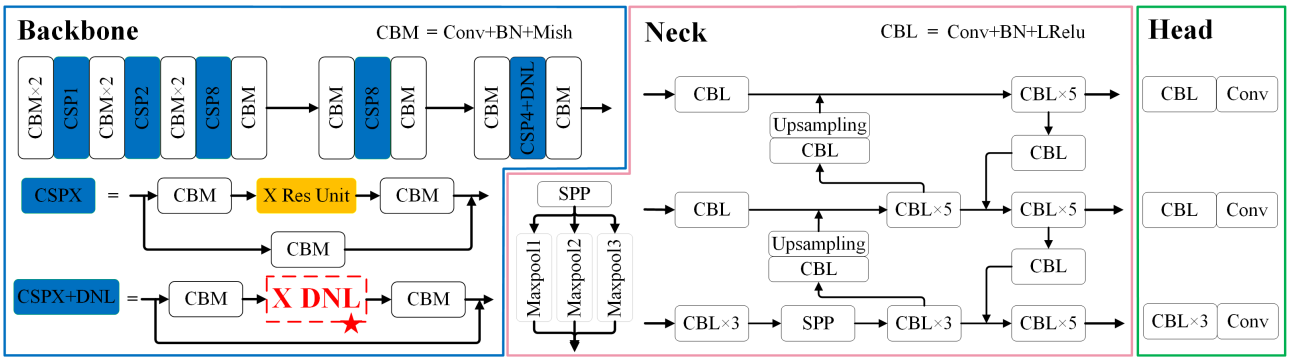

Figure 4. The specific details module of the YOLOv4 network with the DNL module added. The YOLOv4 network model consists of three parts: backbone, neck, and detection head. This article mainly modified the backbone module, adding the DNL module to the some CSPX modules. Therefore, it can extract more salient feature information.

The overall framework is the same as other versions of the YOLO network, including three parts: backbone, neck, and detection heads. Of these, the backbone network combines many new methods, such as CSPX, Mish activation function, and dropblock. In the neck part, it mainly includes SPP module, FPN, and a PAN structure. For the prediction part, it includes three multi-scale detection heads and uses an improved CIOU_loss loss function. In addition, we introduced the DNL module in the last two Res units of the CSPX structure. For example, in the CPS8 module, the first six residual units only contain two CBM blocks,

and the last two residual units are added to the DNL. The DNL module can reduce the difficulty of joint learning in the original non-local module. At the same time, it can model two different visual cues, regional relations and salient boundaries, to improve the accuracy of salient feature extraction. Thus, the improved YOLONet has a greater advantage for object detection.

3. Experiments

In this section, we demonstrate the effectiveness of our method on the Nan-Hai dataset and compare our method with state-of-the-art methods. The model is implemented using the Python language, and the network is built using the PyTorch deep learning framework. All programs were run on an NVIDIA GeForce GTX 1080 Ti graphics card with 12 GB of onboard memory.

3.1. Datasets

We verify the performance of the above method on the South China Sea dataset. This dataset is produced from images returned by the Gaofen-1 remote sensing satellite. The size of each image is 18,192 × 18,000 pixels, and the spatial resolution is 2 m. The spectrum ranges from 450 to 900 nm. We selected some images in different scenes and cropped them to a size of 512 × 512 pixels. As shown in Table 1, 2344 image blocks are converted into one-dimensional grayscale vectors as training samples for scene classification. In addition, the training samples for object detection contain more than 2000 image patches, and the ratio of positive and negative sample images is 1:2. Moreover, there are many types of ship targets in training samples, and the size of the target ships in 2-m resolution images varies from 8 pixels to 157 pixels. They are illustrated in Figure 5.

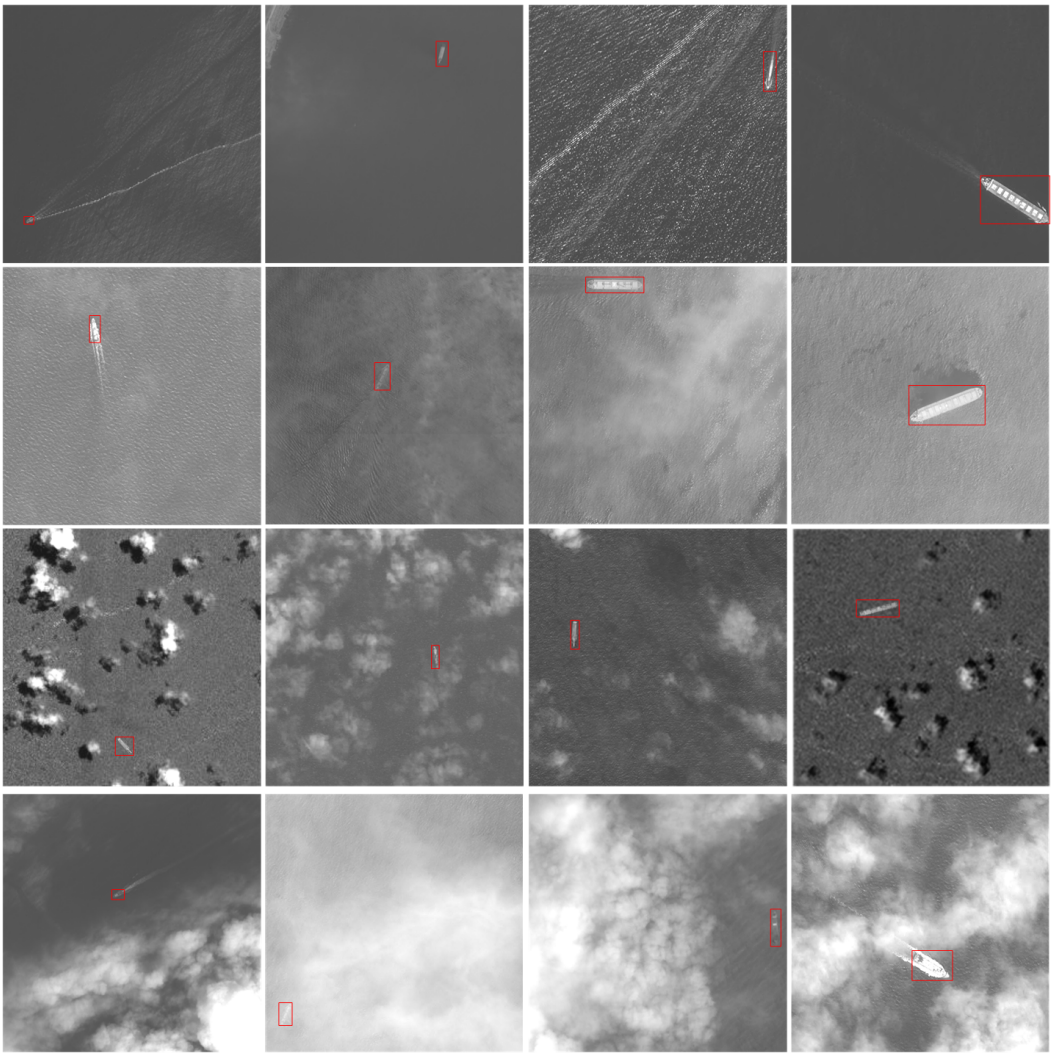

Figure 5. Schematic diagram of training sample blocks. The data set includes various types and sizes of ship targets in different scenes, as well as static targets and dynamic targets. At the same time, simulation samples are added to make up for the lack of samples in some scenes.

Table 1. Dataset overview.

–	**Scene Classification**	**Object Detection**
Data size	1 × 256	512 × 512
Data type	Gray vector	Panchromatic
Total samples	2344	2721
Train samples	2110	2448
Test samples	234	273
Original samples	–	2245
Augment samples	–	476

Moreover, for image samples in scenes such as thick clouds, there are fewer target samples due to a large amount of cloud and fog occlusion. Therefore, we used the simulation method usually used for small object augmentation [43] to create simulation samples, which is shown in Figure 6. Ship samples of different types and sizes are cut out and randomly added to the background image. Then, the boundary is blurred by multi-size Gaussian filters so that the target and the background can be integrated better. After that, we re-screened the obtained simulation samples and removed the images with poor simulation results to obtain the final data set.

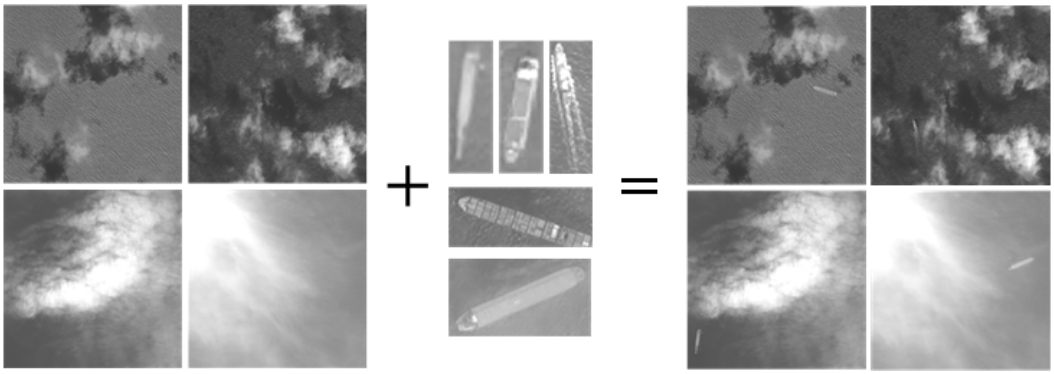

Figure 6. Schematic diagram of simulation sample production with randomly added different types of ships in the background in order to obtain augment samples.

3.2. Experimental Analysis

In general, we evaluated the performance of object detection methods through multiple evaluation indicators, such as precision, recall, and AP value. Precision, also written as P, indicates how many samples of the predicted result are correct. When P reaches 100%, it means that there is no false detection. Similarly, recall usually remembers R, indicating how many positive samples in the predicted results have been correctly detected. When R reaches 100%, it means that there are no missed targets. To obtain these indicators, the calculation of intersection over union (IoU) is necessary.

$$\text{IoU} = A_o / A_u \tag{21}$$

where A_o and A_u represent the overlapping and union area of prediction boxes and the ground truth box. Therefore, we need an IoU threshold to determine whether the detection box is correct. Then, true positives (*TP*s), false positives (*FP*), false negatives (*FN*), and true negatives (*TN*) in detection results can be found. Consequently, the precision and recall can be calculated; the formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{22}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{23}$$

Moreover, it is not enough to only evaluate the performance of the model with precision and recall. Because it is possible that when the precision of model A is higher than that of model B, its recall rate is lower than that of model B. The solution to this problem is to combine precision and recall to calculate another indicator, mean average precision (AP) score—where m means the average of multiple categories. Quantitatively, AP means the average value of precision for each object category when recall varies from 0 to 1. It can characterize the area under the precision–recall curve. Compared to F1, AP can more accurately and intuitively reflect the performance of the detection model. For the ship detection task discussed in this paper—because only a ship needs to be detected—mAP is equal to AP.

Analysis of different schemes. This section focus on verifying the superiority of the proposed schemes, including scene classification, saliency detection, and disentangled non-local modules. As presented in Table 2, methods 1–5 are constructed models with different schemes. The first row is the baseline method Yolov4, and the second method is the baseline algorithm with saliency detection module. Other methods add different modules in turn to verify the validity of each scheme. It can be seen that each method performs better than the previous one. When all three schemes are included in a model, the AP can be improved to 87.05%. Consequently, the proposed schemes in this paper are all effective.

Table 2. Quantitative evaluation of the proposed schemes. (SC: scene classification, SD: saliency detection).

Methods	Baseline	SC	SD	DNL	P(%)	R(%)	AP(%)
1	✓	-	-	-	90.77	80.30	85.21
2	✓	-	✓	-	91.96	80.65	85.65
3	✓	-	-	✓	92.37	81.33	85.82
4	✓	-	✓	✓	92.74	81.88	86.59
5	✓	✓	✓	✓	**93.23**	**82.44**	**87.05**

Number of DNL analysis. The DNL module can extract more salient features, but it also means that the network is prone to overfitting when a large number of modules are added. Thus, in order to obtain a better detection result, the number of DNL added into the YOLO network needs to be adjusted. Figure 7 displays the relationship between the number of DNL in the network and the average precision. We can see that the average precision is highest when the number of DNL modules is 12. After that, as the number of DNL increases, average precision will decrease instead. Simultaneously, we have observed experimentally that as DNL increases, the loss of verification dataset will also increase. It indicates that the network learning has overfitted for this training dataset.

Confidence score analysis. As a key parameter, the confidence score has a great influence on the final result. A higher confidence score usually means less FPs; it may increase precision, but also decrease the recall. Thus, to obtain a better trade-off between detection and location accuracies, the confidence score threshold in this paper is set 0.5. The IoU threshold is also set 0.5, according to experiments. As Figure 8 shows, it displays precision trends of three different object detection methods in a series of confidence scores. It can be seen that the proposed method can obtain better detection results compared with other networks. This indicates our approach is capable of learning more significant features for ship detection tasks.

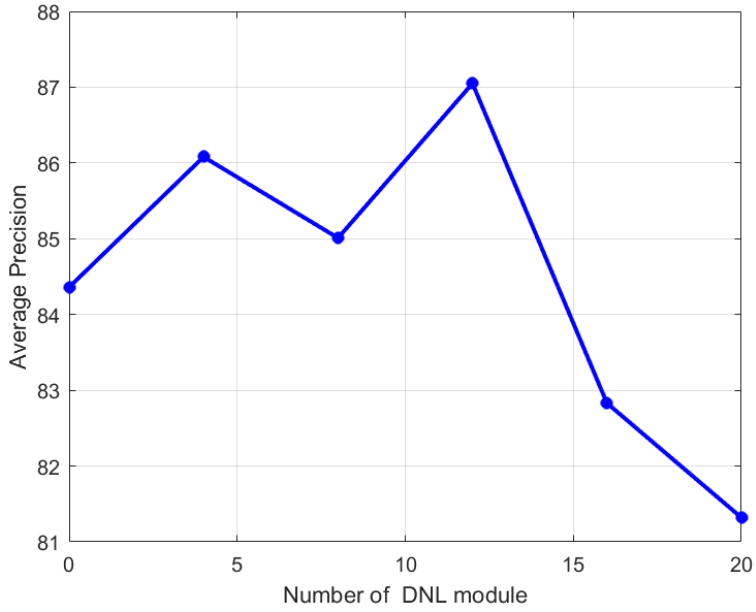

Figure 7. Overall average precision (%) versus different numbers of DNL module in each CSPX structure based on the Nan-Hai dataset.

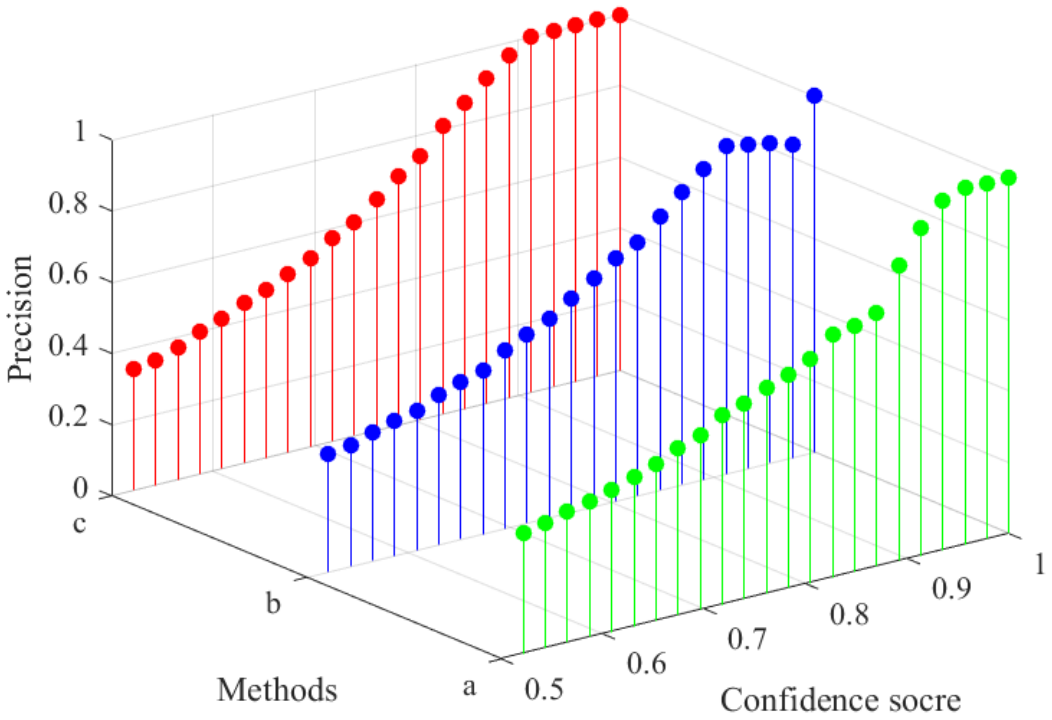

Figure 8. Ship detection precision (%) versus confidence score for different detection methods. (**a**) YOLOv4, (**b**) Faster RCNN, and (**c**) our method.

3.3. Comparisons on Detection Task with Other Methods

To illustrate the performance of the proposed saliency-tuned YOLONet ship detection method, we conducted comparative experiments and compared them with some other state-of-the-art object detection approaches. These included two one-stage methods, YOLOv4 citeyolov4 and SSD [33], and two-stage methods, Faster-RCNN [28] and SPPNet [26]. Because the variation of the object size is large and some ships are small, the IoU and confidence threshold were set to 0.2 and 0.5, respectively, to obtain better results. We trained the data samples over 100 epochs, and all other models were tuned to optimal parameters in order to achieve the best results.

Precision–recall curve analysis. The precision–recall curves and AP scores of different methods on the Nan-Hai dataset are shown in Figure 9 and Table 3, respectively. The recall threshold is set from 0 to 1. In addition, to ensure the fairness of the experiment, we try our best to ensure that all models are trained optimally. Table 3 shows that the detection precision rate and recall rate of the proposed method is higher than the compared methods. For the comprehensive index AP, the evaluation index demonstrates that it can reach 87.05%. This AP is significantly higher than that of the other detectors investigated. Moreover, the red line in Figure 9 can more intuitively show that our method has a larger area under the line. This can be attributed to the fact that the method proposed in this paper indeed have better detection performance.

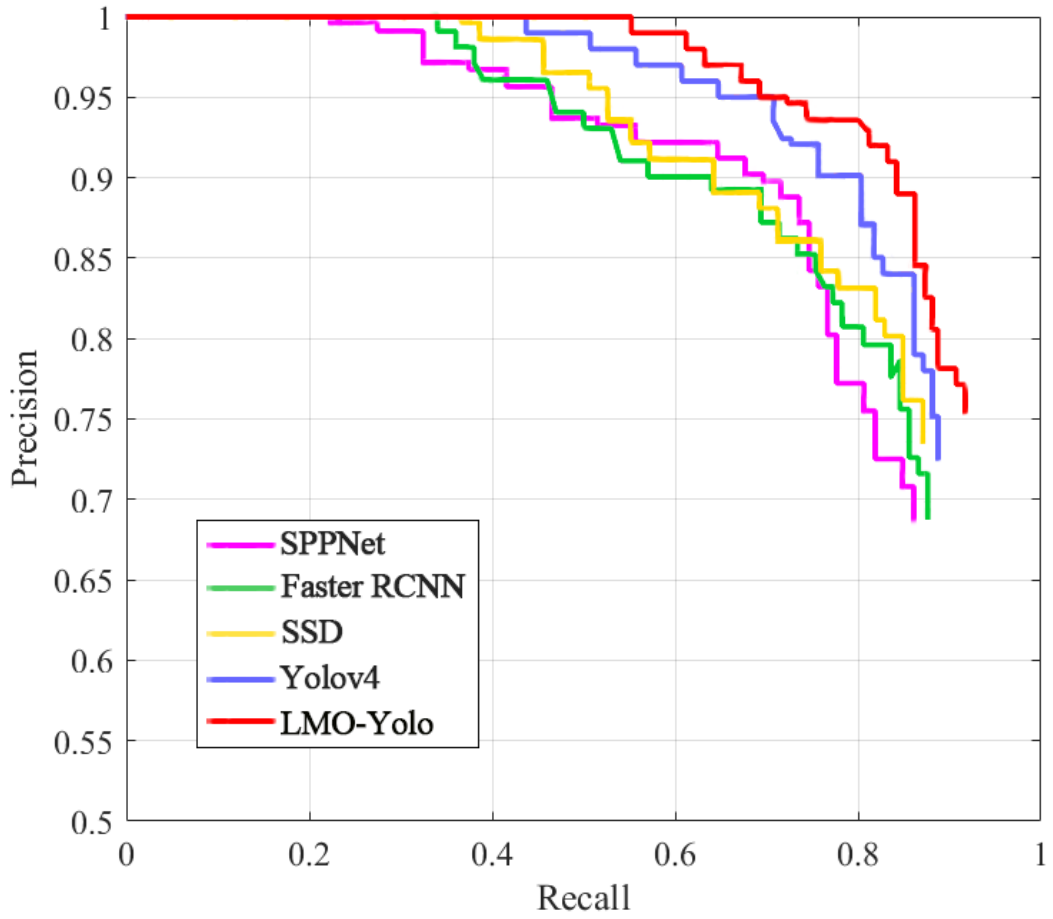

Figure 9. Ship detection precision (%) versus recall (%) for different detection methods. The larger the area under the curve, the better the model performance.

Comprehensive analysis. We tested the model with images of size 2048 × 2048. Large-format images can better reflect the performance of network detection in various sea scenes. At the same time, in order to verify the generalization ability of the detection method, simulation samples of different types and sizes are added in some images. The comparative experimental results of the different methods are illustrated in Figure 10. Among them, the green box represents the ship targets in the original images, and the red box represents the detection results. It can seen from this figure that all methods can obtain good detection results for cloudless scenes. However, when the scene is more complex, such as those containing cloud interference and occlusion, SPPNet has a higher false alarm rate. This is partly because the network does not extract enough features, which leads to false detections. However, compared to the other two methods, YOLONet and Faster RCNN, it can detect relatively more ship targets. For our method, although it adds some false alarms, more accurate detection results can also be obtained. Over all, the network model proposed in this paper has a better detection performance.

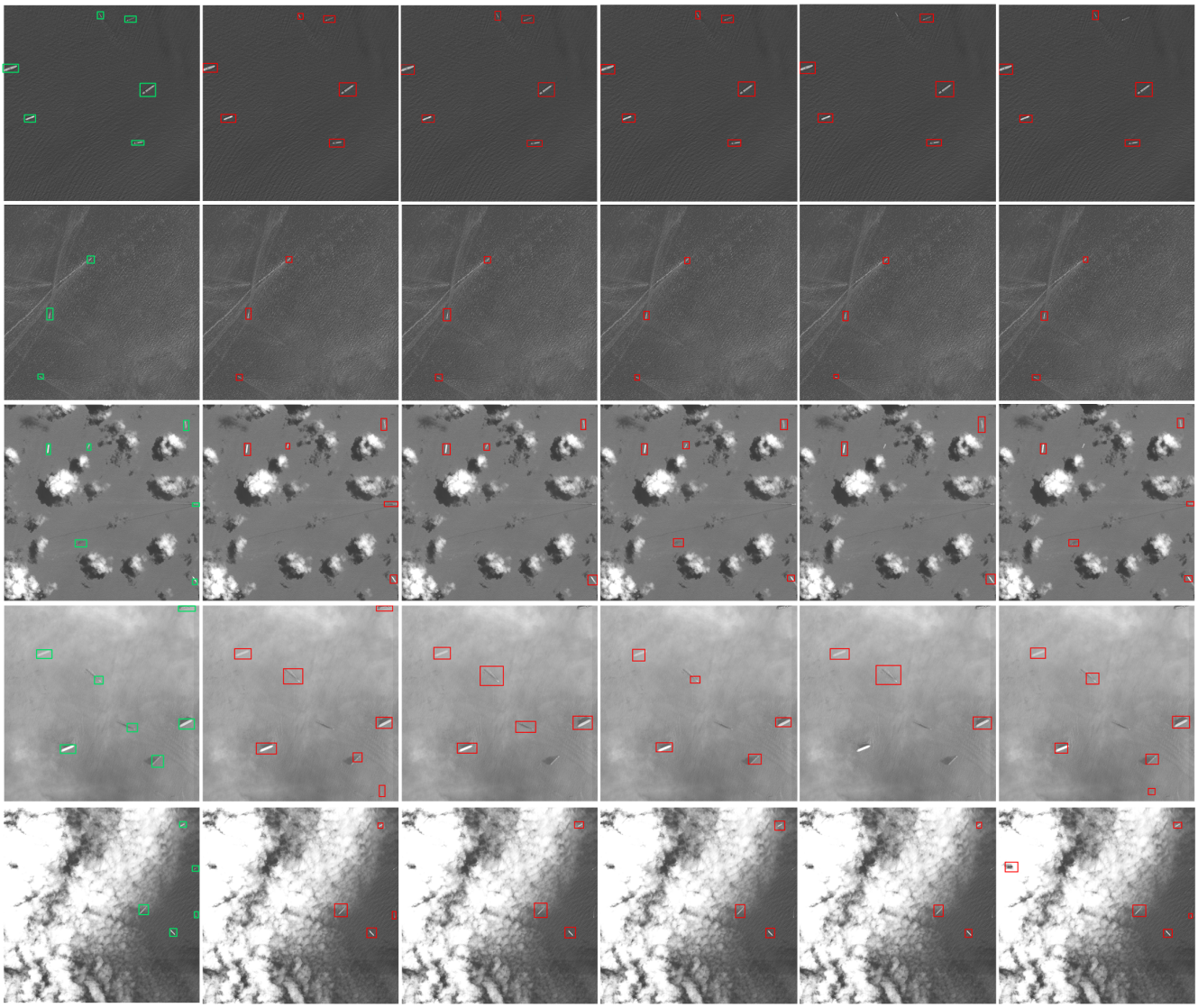

Figure 10. Detection maps for the Nan-Hai dataset based on different methods. Among them, each row represents different scenes: cloudless, big waves, scattered clouds, thin clouds, and thick clouds. The first column represents the original image, columns 2–6, represent the results of our method, YOLOv4, Faster RCNN, SSD, and SPPNet, respectively. Furthermore, the green boxes represent the true ship targets; red boxes represent detected result targets.

Table 3. Performance results of different methods (P: precision, R: recall).

Algorithm	P(%)	R(%)	AP(%)
Our approach	**93.23**	**82.44**	**87.05**
YOLOv4	90.77	80.30	85.21
SSD	84.98	77.76	81.33
Faster-RCNN	83.59	76.49	80.59
SPPNet	88.32	75.22	79.46

3.4. Discussion

With the help of scene classification and saliency detection, the proposed method can promote the performance many off-shore ship detection methods in optical satellite images. Unfortunately, the proposed method does not consider the amount of model parameters,

resulting in a relatively long running time of the system. From the perspective of practical application, the lightweight network model will be our next research goal. In addition, the introduced saliency detection method is not good enough for images with a complex background. It still has some noise in the obtained saliency map, especially for the scenes with scattered clouds. It is very important to design a good saliency detection method for ocean backgrounds. Furthermore, it can also be considered from the perspective of multi-frame continuous images. For marine scenes dominated by moving ships, ship targets will appear in a series in multiple frames of continuous images that are different from the background.

4. Conclusions

In this article, a saliency-tuned YOLONet is proposed for off-shore ship detection from the optical satellite image. First, the density peak clustering is introduced to classify the different scene blocks by their histograms. The purpose of the scene classification is to reduce the intra-class variance of the dataset for targeted image enhancement processing. Second, since the ship can be regarded as a salient target, different saliency region extraction methods (e.g., SR, HC, and CASD) are utilized to extract prominent targets in different scene categories. Scene classification and targeted saliency detection can significantly reduce the interference of clouds and fog on the sea surface. Finally, a DNL-added YOLONet is proposed. By introducing an attention mechanism, the DNL module enables the feature maps in order to add more global information to the original local features. Thus, the neural network enhances its attention to salient targets and can capture more salient features. The advantages of our approach come from two aspects: On the one hand, the scene classification can effectively reduce differences within a class and pertinently process images in different scenes. On the other hand, the additions of the saliency map and the DNL module improve the saliency detection capabilities of the network model.

The experimental results based on the South China Sea dataset returned by the Gaofen-1 satellite show that the detection accuracy of the proposed off-shore ship detection method outperforms other state-of-the-art methods. However, due to the noise of the saliency map, the recall rate is not greatly improved. This problem will be addressed in our future research.

Author Contributions: Conceptualization, Q.X. and J.G.; data curation, S.W. and Q.X.; methodology, Q.X. and S.W.; supervision, J.G. and Q.X.; validation, S.W.; writing—original draft, S.W.; writing—review and editing, S.W. and Q.X. 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 under Grant 61972021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to confidentiality.

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

References

1. You, Y.; Ran, B.; Meng, G.; Li, Z.; Liu, F.; Li, Z. OPD-Net: Prow Detection Based on Feature Enhancement and Improved Regression Model in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6121–6137.
2. Hou, B.; Ren, Z.; Zhao, W.; Wu, Q.; Jiao, L. Object Detection in High-Resolution Panchromatic Images Using Deep Models and Spatial Template Matching. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 956–970. [CrossRef]
3. Segl, K.; Kaufmann, H. Detection of small objects from high-resolution panchromatic satellite imagery based on supervised image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 2080–2083. [CrossRef]

4. Audebert, N.; Le Saux, B.; Lefèvre, S. How useful is region-based classification of remote sensing images in a deep learning framework? In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5091–5094. [CrossRef]
5. Wenxiu, W.; Yutian, F.; Feng, D.; Feng, L. Remote sensing ship detection technology based on DoG preprocessing and shape features. In Proceedings of the IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 1702–1706. [CrossRef]
6. Inamdar, S.; Bovolo, F.; Bruzzone, L.; Chaudhuri, S. Multidimensional Probability Density Function Matching for Preprocessing of Multitemporal Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1243–1252. [CrossRef]
7. Hurtik, P.; Molek, V.; Hula, J. Data Preprocessing Technique for Neural Networks Based on Image Represented by a Fuzzy Function. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 1195–1204. [CrossRef]
8. Zheng, L.; Xu, W. An Improved Adaptive Spatial Preprocessing Method for Remote Sensing Images. *Sensors* **2021**, *21*, 5684. [CrossRef] [PubMed]
9. Vivone, G.; Dalla Mura, M.; Garzelli, A.; Pacifici, F. A Benchmarking Protocol for Pansharpening: Dataset, Preprocessing, and Quality Assessment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6102–6118. [CrossRef]
10. Yao, J.; Fidler, S.; Urtasun, R. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 702–709. [CrossRef]
11. Zhong, Y.; Zhu, Q.; Zhang, L. Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [CrossRef]
12. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
13. Li, F.; Feng, R.; Han, W.; Wang, L. High-Resolution Remote Sensing Image Scene Classification via Key Filter Bank Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8077–8092. [CrossRef]
14. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef] [PubMed]
15. Zhu, C.; Zhou, H.; Wang, R.; Guo, J. A Novel Hierarchical Method of Ship Detection from Spaceborne Optical Image Based on Shape and Texture Features. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3446–3456. [CrossRef]
16. Yang, G.; Li, B.; Ji, S.; Gao, F.; Xu, Q. Ship Detection From Optical Satellite Images Based on Sea Surface Analysis. *IEEE Trans. Geosci. Remote Sens. Lett.* **2014**, *11*, 641–645. [CrossRef]
17. Song, Z.; Sui, H.; Wang, Y. Automatic ship detection for optical satellite images based on visual attention model and LBP. In Proceedings of the IEEE Workshop on Electronics, Computer and Applications, IWECA, Ottawa, ON, Canada, 8–9 May 2014; pp. 722–725. [CrossRef]
18. Shi, Z.; Yu, X.; Jiang, Z.; Li, B. Ship Detection in High-Resolution Optical Imagery Based on Anomaly Detector and Local Shape Feature. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 4511–4523. [CrossRef]
19. Chen, H.; Gao, T.; Chen, W.; Zhang, Y.; Zhao, J. Contour Refinement and EG-GHT-Based Inshore Ship Detection in Optical Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8458–8478. [CrossRef]
20. Hou, X.; Zhang, L. Saliency Detection: A Spectral Residual Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [CrossRef]
21. Goferman, S.; Zelnik-Manor, L.; Tal, A. Context-aware saliency detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2376–2383. [CrossRef]
22. Tang, J.; Deng, C.; Huang, G.; Zhao, B. Compressed-Domain Ship Detection on Spaceborne Optical Image Using Deep Neural Network and Extreme Learning Machine. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1174–1185. [CrossRef]
23. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
24. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikainen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2018**, *128*, 261–318. [CrossRef]
25. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef]
27. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster RCNN: Towards real time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
29. Quiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
30. Nie, X.; Duan, M.; Ding, H.; Hu, B.; Wong, E.K. Attention Mask R-CNN for Ship Detection and Segmentation From Remote Sensing Images. *IEEE Access* **2020**, *8*, 9325–9334. [CrossRef]

31. Li, Q.; Mou, L.; Liu, Q.; Wang, Y.; Zhu, X.X. HSF-Net: Multiscale Deep Feature Embedding for Ship Detection in Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7147–7161. [CrossRef]
32. Szegedy, C.; Toshev, A.; Erhan, D. Deep neural networks for object detection. In Proceedings of the Neural Information Processing Systems Conference, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2553–2561.
33. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A. SSD: Single shot multibox detector. In *Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; pp. 21–37.
34. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
35. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
36. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
37. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
38. Law, H.; Deng, J. CornerNet: Detecting objects as paired keypoints. In *Computer Vision—ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; pp. 756–781.
39. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.I.; Cheng, J.; Sun, J. You Only Look One-level Feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
40. Cheng, M.-M.; Mitra, N.J.; Huang, X.; Torr, P.H.S.; Hu, S.-M. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, 37, 569–582. [CrossRef]
41. Yin, M.; Yao, Z.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; Hu, H. Disentangled non-local neural networks. In *Computer Vision—ECCV 2020*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020.
42. Zheng, J.; Xu, Q.; Chen, J.; Zhang, C. The on-orbit noncloud-covered water region extraction for ship detection based on relative spectral reflectance. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 818–822. [CrossRef]
43. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.