Ships’ Small Target Detection Based on the CBAM-YOLOX Algorithm

Yuchao Wang, Jingdong Li, Zeming Chen and Chenglong Wang *

College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
* Correspondence: wangchenglong@hrbeu.edu.cn

Abstract: In order to solve the problem of low accuracy of small target detection in traditional target detection algorithms, the YOLOX algorithm combined with Convolutional Block Attention Module (CBAM) is proposed. The algorithm first uses CBAM on the shallow feature map to better focus on small target information, and the Focal loss function is used to regress the confidence of the target to overcome the positive and negative sample imbalance problem of the one-stage target detection algorithm. Finally, the Soft Non-Maximum Suppression (SNMS) algorithm is used for post-processing to solve the problem of missed detection in close range ship target detection. The experimental results show that the average accuracy of the proposed CBAM-YOLOX network target detection is improved by 4.01% and the recall rate is improved by 8.81% compared with the traditional YOLOX network, which verifies the effectiveness of the proposed algorithm.

Keywords: deep learning; small target detection; convolutional neural network; attention mechanism

1. Introduction

As the main carrier of transportation of marine activities, ships are widely used in marine transportation, marine resources exploration, and other fields. Automatic detection of ship targets can monitor the maritime traffic in specific ports or sea areas, assist in rescuing ships in distress, and cooperate with safety management departments to monitor and combat illegal fishing and other illegal activities. Therefore, it is very necessary to carry out the research of ship target detection on the sea.

With the rapid development of the deep convolutional neural network, target detection technology has gradually changed from traditional manual feature design method to automatic feature extraction. The deep learning method based on a convolutional neural network can learn various semantic information in images well by designing a network model and choosing appropriate training methods.

Target detection algorithms based on convolutional neural networks are mainly divided into two-stage and one-stage algorithms. The two-stage target detection algorithm firstly extracts the feature information of the candidate region, and then sends it into the detection network to complete the prediction and recognition of the candidate target category and location, such as Fast R-CNN [1] and Faster R-CNN [2]. The two-stage target detection algorithm has high accuracy but slow detection speed, which is not suitable for tasks that require high real-time performance. One-stage target detection algorithms directly use detection network to detect the category and location of targets, such as SSD algorithm [3] and YOLO series algorithm [4–6]. Compared with the two-stage target detection algorithm, the one-stage target detection algorithm directly extracts the features of the input image, and then performs the bounding box regression on the extracted feature image, which greatly improves the detection speed. However, because small targets have lower resolution and less visual information, after repeated convolution and downsampling, further reducing the resolution of the small target, the characteristic information contained is gradually reduced, which increases the difficulty of small target detection. Therefore,
the current target detection network still has the problems of low detection accuracy and missing detection.

He et al. proposed a residual network that can effectively deepen the number of network layers and obtain richer features, while being able to solve gradient disappearance and gradient explosion problems [7]. Huang et al. proposed dense convolution to ensure the maximum exchange of information between layers in the network. Each layer accepts the output from all previous layers as input, and with the same number of layers, the dense convolution achieves better performance with fewer parameters [8]. Lin et al. proposed feature pyramid networks that obtained high resolution information from the shallow network and stronger semantic information from the deep network, and constructed a feature pyramid with strong semantic information on all scales on a single image, which better solves the problem of small targets missing detection [9]. Woo et al. used deconvolution instead of bilinear interpolation to upsample the deep feature maps and highly accurate detection results were obtained [10]. Raghunandan et al. improved the SSD network and proposed the Deconvolutional Single Shot Detector (DSSD) network, which deconvolutes the deep feature map while multiplying element-by-element with the shallow feature map, which can better suppress irrelevant information and highlight important information in the target region [11]. Shrivastava et al. designed a top-down feature fusion module. The model fuses low-level features with high resolution and high-level features with rich semantic information, improving the detection ability of small targets [12]. Li et al. proposed an expanded convolutional module, which increases the perceptual field of the network while retaining higher resolution and better recognition ability for small targets [13]. To make the network learn different scale targets in the input image, Li et al. designed the receptive field of three parallel branches, which further improved the effect of multi-scale feature fusion [14]. Zhang et al. redesigned the Feature Pyramid Networks (FPN) structure by fusing feature maps with different expansion rates to enhance the detection accuracy of small targets with the same computational effort [15]. Cai et al. introduced contextual information into the multi-scale candidate region extraction network, which effectively improved the small target detection ability [16]. Guan et al. adopted a semantic context-aware network with cross-application of maximum pooling and average pooling, effectively balancing precision and recall [17]. YOLOv2 used the K-means algorithm to cluster the true bounding box of the data in the training set to reduce the difficulty of small target localization [5]. In addition, in YOLOX, the Anchor-Free method was chosen to change the original method of setting anchor box into a key point estimation method, which improves the generalization ability of target detection [18]. Kisantal et al. proposed an oversampling method for small targets, which effectively increases the number of small targets in the data set and facilitates the model to better focus on small targets [19]. Chen et al. proposed a new data enhancement method by stitching four images into a regular size and scaling the large and small targets at the same scale to ensure a reasonable distribution of the dataset [20]. The Mosaic data enhancement method used in YOLOv4, on the other hand, allows the four images to be of different sizes, further improving the data enhancement strategy [21]. Zhang et al. proposed the Mixup data enhancement method, which blends different classes of images, effectively expanding the dataset and providing some improvement for small target detection [22]. Inspired by human vision, attention mechanisms have been widely used in convolutional neural networks in recent years to enhance target detection. Hu et al. proposed the channel attention mechanism Squeeze-and-Excitation Networks (SENet), which better investigates the relationship between channel information in convolutional operations and preserves valuable features [23]. Jaderberg et al. proposed a spatial attention mechanism that transforms the spatial information in the original image to another space, preserving the key part of the information well [24]. Woo et al. proposed a hybrid domain attention mechanism, which focuses on the detected object from both spatial and channel dimensions, and the features cover more parts of the object to be recognized and the detection ability of small targets is improved compared to the single channel attention mechanism and spatial attention mechanism [25]. In 2016, Redmon et al. transformed the
object detection problem into a regression problem and proposed the YOLO algorithm [4]. In subsequent studies, YOLO algorithm has been widely applied in various fields. Liu et al. applied YOLOv4 in ship target detection, and Zhou et al. improved the YOLOv5 algorithm for ship target detection at sea [26,27].

The traditional YOLO algorithm implements end-to-end detection, with high detection speed but low detection accuracy. Because the YOLO algorithm does not use regional sampling, it has a good performance for global information, but it has some shortcomings in small range information. Especially for small target detection, the traditional YOLO algorithm faces problems such as low resolution of small target, weak network focusing ability, and unbalanced positive and negative samples. Therefore, based on the YOLO algorithm, the CBAM-YOLOX algorithm for a ship’s small target detection is proposed in this paper. The CBAM attention mechanism is introduced to enhance network focusing ability from spatial and channel dimensions. CIoU Loss is selected as the loss function of location prediction to improve the accuracy of small target location prediction. Selecting the Focal Loss function to regress the target confidence to solve the problem of unbalanced positive and negative samples. Finally, the SNMS algorithm is used for post-processing to further improve the detection ability of small targets. In practical applications, images are captured by visual cameras and input to the CBAM-YOLOX network to complete target detection.

2. Small Target Detection of Ships Based on the CBAM-YOLOX Network

2.1. CBAM-YOLOX Network Structure

Based on the YOLOX network, a new CBAM-YOLOX network is formed by adding the CBAM attention mechanism to the YOLOX backbone network. The backbone network of CBAM-YOLOX is the CSPDarkNet53 structure, and feature extraction is performed on the input images to obtain three effective feature layers for detecting large, medium, and small targets, respectively. The FPN structure and Pixel Aggregation Network (PAN) is used in the enhancement feature extraction stage to enhance the feature extraction of the three feature layers obtained from the backbone network, whereas the up-sampling and down-sampling methods are used to achieve bi-directional feature fusion and complete the fusion of shallow semantic information and deep semantic information. The YOLO head structure serves as the classifier and regressor of YOLOX, and completes the target detection by judging the feature layers that have been extracted with enhanced features. The CBAM-YOLOX network is shown in Figure 1.

![CBAM-YOLOX network structure](image)

Figure 1. CBAM-YOLOX network structure.

As can be seen in Figure 1, in the backbone network of CBAM-YOLOX, the input image size is 640 × 640 and the input image is transformed to a 320 × 320 size by the Focus structure. The Focus module obtains four images to complement each other by slicing the image and taking a value every other pixel in an image. The information of W and H is concentrated in the channel space, the input channel is expanded by four times, and the stitched picture becomes 12 channels compared to the original RGB three channels.
The obtained new image is then convolved, and finally a double-sampled feature map is obtained without information loss. Although a smaller number of parameters is added compared to traditional down-sampling operations, information loss is avoided.

In the feature extraction stage, the convolutional layer, the Batch Normalization (BN) layer, and the SiLU activation function are used to form the main structure. Among them, the convolution layer is a residual block composed of $1 \times 1$ convolution and $3 \times 3$ convolution, whereas the Cross Stage Partial Network (CSPNet) structure is used to continue stacking the original residual blocks, and the residual edges are directly connected to the end of the residual blocks. The Spatial Pyramid Pooling (SPP) structure is added at the end of the backbone network, and the maximum pooling operation is performed by pooling kernels of different structures to effectively improve the receptive field of the network. The Dark3, Dark4, and Dark5 modules of the backbone network output three scales of $80 \times 80$, $40 \times 40$, and $20 \times 20$ feature maps for detecting small, medium, and large targets, respectively.

The available features of small targets are fewer. Compared with large and medium targets, the resolution of small targets is lower and the information is less. It is difficult to extract distinguishing features. Therefore, the CBAM attention mechanism is added to the backbone network, and features are abstracted as attention weights by using global average pooling, and then the weights are added to the original space or channel features. Compared with the original network, the network pays more attention to the relevant features of small targets, which can avoid missing detection. Since the detection of small targets is characterized by low resolution and less visual information, it is more difficult to identify compared to large targets. Therefore, the CBAM module is added to the Dark3 module of the shallow network, and the attention weights are inferred sequentially from both spatial and channel dimensions, and finally multiplied with the $80 \times 80$ scale feature map, which makes the feature response of small targets further enhanced. The structure of the CBAM module is shown in Figure 2.

![Figure 2. CBAM module structure.](image)

The CBAM module includes the Channel Attention Module (CAM) and Spatial Attention Module (SAM). First, the input feature map $F$ of size $H \times W \times C$ is processed by the CAM module, and then the obtained feature map $F'$ is input to the SAM module for processing to obtain the final attention feature map $F''$.

The CAM module focuses on the feature information of the small target from the channel domain, and then obtains the channel attention feature map $F'$. The structure of the CAM module is shown in Figure 3.

![Figure 3. CAM module structure.](image)
As can be seen from Figure 3, the CAM module first performs maximum pooling and average pooling in the spatial domain for the input feature map \( F \) of size \( H \times W \times C \) to obtain two \( 1 \times 1 \times C \) channel information, and then inputs them into the multilayer perceptron (MLP) and sums them separately. Finally, obtain a weight coefficient \( M_c \) after the sigmoid activation function and multiply the weight coefficient \( M_c \) with the original features to obtain the channel attention feature map \( F' \), as shown in Equations (1) and (2).

\[
M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{1}
\]

\[
F' = M_c(F) \otimes F \tag{2}
\]

In Equations (1) and (2), \( F \) is the input feature map, \( \text{AvgPool} \) is the average pooling operation, \( \text{MaxPool} \) is the maximum pooling operation, \( \text{MLP} \) is the multilayer perceptron, \( \sigma \) is the Sigmoid activation function, \( M_c(F) \) is the channel attention weight factor, and \( F' \) is the channel attention feature map.

In the CBAM module, the feature map \( F \) is processed by the CAM module to obtain the channel attention feature map \( F' \), and then \( F' \) is input to the SAM module, which in turn obtains the mixed-domain attention feature map \( F'' \) based on the channel and spatial. The structure of the SAM module is shown in Figure 4.

As can be seen from Figure 4, the maximum pooling operation and the average pooling operation in the channel domain are first performed in the SAM module to obtain two \( H \times W \times 1 \) spatial information, after a \( 7 \times 7 \) convolutional layer and the sigmoid activation function, the weight coefficients \( M_s \) are obtained. Multiply \( M_s \) with the input feature map to obtain the final mixed-domain attention feature map, as shown in Equations (3) and (4).

\[
M_s(F') = \sigma(f^{7 \times 7}([\text{AvgPool}(F'); \text{MaxPool}(F')]))) \tag{3}
\]

\[
F'' = M_s(F') \otimes F' \tag{4}
\]

In Equations (3) and (4), \( F' \) is the channel attention feature map, \( \text{AvgPool} \) is the average pooling operation, \( \text{MaxPool} \) is the maximum pooling operation, \( f^{7 \times 7} \) is the \( 7 \times 7 \) convolution operation, \( \sigma \) is the Sigmoid activation function, \( M_s(F') \) is the spatial attention weight coefficient, and \( F'' \) is the final hybrid domain attention feature map.

2.2. Loss Function

The loss value is the comparison between the network prediction results and the real value. In the CBAM-YOLOX network, the loss of the network consists of Regression (Reg), Object (Obj), and Classification (Cls). The Reg is used to regress the location of the detected target, the Obj is used to regress the confidence of the detected target, and the Cls is the categories of feature point containing objects.

For the regression problem of target location, CIoU Loss is selected as the loss function for location regression because of the high requirement of location prediction accuracy for small targets. The CIoU Loss is a loss function that considers three factors: coverage area, centroid distance, and aspect ratio, and can avoid the gradient disappearance problem caused by the non-intersection of the prediction box and the ground truth box. CIoU Loss
is based on the Intersection over Union (IoU). IoU as shown in Equation (5) and CIoU Loss as shown in Equation (6), where $\alpha$ is the scaling factor and $\upsilon$ is used to measure the consistency of the aspect ratio, $\alpha$ and $\upsilon$ are calculated as shown in Equations (7) and (8).

$$IoU = \frac{B \cap B^t}{B \cup B^t}$$  \hspace{1cm} (5)$$

$$Loss_{CIoU} = 1 - IoU + \frac{\upsilon^2 (d - d^t)}{c^2} + \alpha \upsilon$$  \hspace{1cm} (6)$$

$$\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon}$$  \hspace{1cm} (7)$$

$$\upsilon = \frac{4}{\pi} (\arctan \frac{\omega}{h^t} - \arctan \frac{\omega}{h})^2$$  \hspace{1cm} (8)$$

In Equations (5)–(8), $B$ denotes the prediction box, $B^t$ denotes the ground truth box, $d$ and $d^t$ denote the centroids of $B$ and $B^t$, respectively. $\omega$ and $h$ represent the length and width of the predicted box, respectively, and $\omega^t$ and $h^t$ represent the length and width of the ground truth box, respectively. $c$ is the diagonal distance of the minimum outer rectangle of $B$ and $B^t$, and $\rho(\cdot)$ denotes the Euclidean distance.

For the confidence regression problem, since CBAM-YOLOX is a one-stage target detector, the same positive and negative sample imbalance problem exists. Many easily distinguishable negative samples dominate in the loss function, which brings difficulties to small object detection. To better alleviate the problem of unbalanced positive and negative samples, the Focal Loss function is used to regress the confidence of the detected targets. By reducing the weight of easy-to-classify samples, the model is trained to focus more on difficult-to-classify samples. The Focal Loss function is shown in Equation (9).

$$Loss_{conf} = \begin{cases} -\alpha_1 (1 - p_t)^\gamma \cdot \log(p_t), & y^t_{gt} = 1 \\ -\alpha_1 (1 - p_t)^\gamma \cdot \log(1 - p_t), & y^t_{gt} = 0 \end{cases}$$  \hspace{1cm} (9)$$

where $y^t_{gt}$ is the true value of the target confidence, $p_t$ is the confidence of the predicted target, $\alpha_1$ and $\gamma$ are hyperparameters, which are experimentally adjusted and set to 0.35 and 2 [28], respectively.

The BCE Loss function is used to judge the category of objects in the feature points, as shown in Equation (10).

$$Loss_{cls} = -w(C^t_{gt} \log(C_p) + (1 - C^t_{gt}) \log(1 - C_p))$$  \hspace{1cm} (10)$$

where $w$ is the weight of different sample loss values in each training batch, $C_p$ is the predicted value of object type, and $C_{gt}$ is the true value of object category.

From Equations (6), (9) and (10), the total loss function of the CBAM-YOLOX network is calculated as shown in Equation (11).

$$Loss_{all} = Loss_{CIoU} + Loss_{conf} + Loss_{cls}$$  \hspace{1cm} (11)$$

where $Loss_{all}$ is the total loss function of CBAM-YOLOX network, $Loss_{CIoU}$ represents the loss function of target location regression, $Loss_{conf}$ represents the loss function of target confidence regression, and $Loss_{cls}$ represents the loss function of target classification. Referring to the method of setting the loss function parameters by the authors of YOLOX and combining with the actual situation in the training process of the network, the $Loss_{CIoU}$ weight is set to 5, and the weights of both $Loss_{conf}$ and $Loss_{cls}$ are set to 1.

2.3. Soft Non-Maximum Suppression Algorithm

The Non-Maximum Suppression (NMS) algorithm is an important part of the post-processing part of the target detection algorithm. The NMS algorithm filters the suggestion
box according to the predicted probability. Firstly, the confidence of the prediction box of the same category is ranked, the box with the highest confidence is taken out, and then the IoU with another box is calculated. If the IoU value is greater than the set suppression threshold, the prediction box will be suppressed. The NMS is calculated as shown in Equation (12).

\[
S_i = \begin{cases} 
S_i, & \text{if } \text{IoU}(M, b_i) < N_i \\
0, & \text{if } \text{IoU}(M, b_i) \geq N_i 
\end{cases} \quad (12)
\]

where \( S_i \) represents the specific score of each edge, \( M \) is the current highest scoring box, \( b_i \) is any one of the remaining boxes, and \( N_i \) is the set threshold value.

However, in the process of actual detection, the distance between different ships may be close. According to the principle of NMS algorithm, if the small target is within the preset repetition threshold, it will be ignored, and the small target detection cannot be completed. For better post-processing operations, SNMS is selected to replace the non-maximum suppression. SNMS does not require additional training and can be well embedded into the existing network. SNMS is calculated as shown in Equation (13).

\[
S_i = \begin{cases} 
S_i, & \text{if } \text{IoU}(M, b_i) < N_i \\
S_i(1 - \text{IoU}(M, b_i)), & \text{if } \text{IoU}(M, b_i) \geq N_i 
\end{cases} \quad (13)
\]

where \( S_i \) is the specific score of each box, \( M \) is the box with the highest current score, \( b_i \) is any one of the remaining boxes, and \( N_i \) is the set threshold.

Equation (13) is not a continuous function. When the overlap IoU of a box and \( M \) exceeds the threshold \( N_i \), its score will jump, which will produce large fluctuations in the detection results. Therefore, the SNMS in the form of Gaussian function is selected in the CBAM-YOLOX network, as shown in Equation (14).

\[
S_i = S_i \exp\left(-\frac{(\text{IoU}(M, b_i))^2}{\sigma_1^2}\right), \forall b_i \not\in D \quad (14)
\]

where \( \sigma_1 \) is the smoothing parameter, \( \sigma_1 \) set to 0.5 \([29]\), \( S_i \) represents the specific score of each box, \( M \) is the box with the highest current score, \( b_i \) is any one of the remaining boxes, and \( D \) is the optimal bounding box set, so \( b_i \) does not belong to the optimal bounding box set.

It can be seen from Equations (12)–(14) that the basic principle of NMS and SNMS are the same, the difference is that the prediction box whose IoU values of \( M \) and \( b_i \) are larger than the suppression threshold in SNMS is not directly deleted, but their confidence is weighted according to the value of IoU. Finally, the prediction box that is smaller than the soft threshold is retained, the prediction box larger than the soft threshold is suppressed, and the confidence of the remaining prediction box is revised.

3. A Ship’s Small Target Detection Results and Analysis

3.1. Experimental Design

3.1.1. Dataset

The public dataset SeaShips is selected to verify the effectiveness of the CBAM-YOLOX algorithm. There are total of 7000 images in the SeaShips dataset, including 6 categories of ships (ore carrier, passenger ship, container ship, bulk cargo carrier, general cargo ship, fishing boat). All images are collected from real scenes with a resolution of 1920 × 1080, and part of the dataset is shown in Figure 5.

It can be seen from Figure 5 that the SeaShips dataset includes six types of ship targets: ore carrier, passenger ship, container ship, bulk cargo carrier, general cargo ship, and fishing boat. The obvious feature in the dataset is that some ships have small targets and occlusion problems, which makes it easy to affect the target detection work.

For small targets, Chen \([30]\) and others defined them as follows: the relative area of all target instances in the same category, that is, the median of the ratio of bounding box area
to image area, is between 0.08% and 0.58%. In this experiment, the dataset image size is \(1920 \times 1080\), so the bounding box with an image size between 1658.88–12,026.88 is a small target. The size statistics of the bounding box in the dataset are shown in the Figure 6.

Figure 5. Partial SeaShips Dataset.

Figure 6. Partial SeaShips Dataset.
It can be seen from Figure 6 that some data in the dataset still have the problem of small targets, which accounts for a large proportion. Therefore, it is meaningful to improve the network for small target problems.

3.1.2. Experimental Environment and Parameter Settings

The experimental environment in this paper is configured with a Tesla V100 graphics card (32 G video memory), the operating system is Ubuntu 18.04, the deep learning framework is pytorch 1.9.0, and the training time is about 10 h.

In the experiment, the dataset is first divided, 90% of which is used as the training set and 10% of which is used as the test set. When building the network, the backbone network, feature fusion network, and prediction network are firstly built based on the YOLOX_s network. On this basis, the CBAM module is added to the backbone network, and the loss function and post-processing algorithm are improved to complete the construction of the CBAM-YOLOX network. During training, the epoch is set to 100, and the frozen training method is adopted. The backbone network is frozen in the first 50 epochs, and the pre training weights are loaded. The BatchSize is set to 16, and the Adam optimizer is used. The initial learning rate is set to 0.001 and weight_decay is set to 0.0005. The last 50 epochs unfreeze the backbone network. The batchsize is set to 8, and the Adam optimizer is used. The initial learning rate is set to 0.0001 and weight_decay is set to 0.0005. There are four versions of the YOLOX network (s, m, l, x), and the difference between each version is that Depth_Multiple and Width_Multiple are different, and the parameters are shown in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Depth_Multiple</th>
<th>Width_Multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOX_s</td>
<td>0.33</td>
<td>0.50</td>
</tr>
<tr>
<td>YOLOX_m</td>
<td>0.67</td>
<td>0.75</td>
</tr>
<tr>
<td>YOLOX_l</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>YOLOX_x</td>
<td>1.33</td>
<td>1.25</td>
</tr>
</tbody>
</table>

It can be seen from Table 1, the Depth_Multiple refers to the number of layers of the neural network, and the Width_Multiple refers to the number of channels in each layer of the neural network. The specific parameters represent scaling factors, which are used to adjust the Depth_Multiple and Width_Multiple of the neural network. Due to the different parameters of different versions of YOLOX networks, the YOLOX_s network has the lowest detection accuracy whereas the network weight is the smallest, and the YOLOX_x network has the highest detection accuracy whereas the network weight is the largest [18]. To improve the detection accuracy of the YOLOX network for small targets while ensuring that the weight of the YOLOX network is small, YOLOX_s is chosen as the benchmark in this paper, and the CBAM module is added for comparative analysis.

After the network training, the testset is used for testing. The 700 images used for testing are input into the prediction network one by one, the trained weight file is loaded, and the detection results are output to check the network performance.

3.1.3. Evaluation Index

Mean average precision (mAP), Recall (R), Precision (P), F1 curve, and detection time (time) are used to evaluate the target detection capability of the CBAM-YOLOX network. The Equations are shown in Equations (15)–(18).

\[
mAP = \frac{1}{N} \sum_{i=0}^{N-1} \int_{0}^{1} P(R) dR
\]  

\[
R = \frac{TP}{TP + FN}
\]
In the field of target detection, accuracy and recall are two important indicators used to evaluate detection algorithms. Accuracy refers to how many of the candidate boxes output by the algorithm are correct, reflecting the algorithm’s ability to detect positive classes. Recall refers to how many of the pre-labeled boxes are detected, reflecting the algorithm’s ability to distinguish positive and negative samples. If precision remains high while recall increases, it indicates that the performance of the algorithm is good. Relatively, with the increase of recall value, precision suffers a serious loss, which indicates that the performance of the algorithm is poor. AP is the area enclosed by PR curve and coordinate axes; a larger AP value means better target detection for that category. mAP refers to the average value of each category of AP. To calculate mAP, first calculate and draw the PR curve of each category separately to obtain AP, and then calculate the mean value to obtain mAP. First, a low confidence threshold, usually 0.001, is specified to filter the prediction boxes of the network and retain as many prediction boxes as possible. The filtered prediction boxes are suppressed using the SNMS algorithm to remove highly overlapping boxes, and the mAP values are calculated based on the processed prediction boxes and the true boxes. The specific calculation steps are as follows: in the specific calculation steps, first calculate the AP value of each category, calculating AP50 as an example, iou_threshold = 0.5, and determine whether each prediction box belongs to TP or FP. Then, accumulate the total number of TP in each row, and calculate precision and recall for each row, and then calculate the average precision corresponding to recall. This paper follows the current mainstream COCO101 point method. The recall value is in the range of 0–1, starting from 0 every 0.01 until 1, for each point, and the maximum precision value is found on the PR graph to the right, finding 101 values in turn. The definite integral of these values is AP, which is approximated as the area under the PR curve. Finally, the AP values of all categories are averaged to obtain the mAP value. In the target detection task, the mAP value is an important indicator of target detection accuracy, and a larger mAP value means better target detection.

3.2. Ablation Experiment

3.2.1. Comparison of Network Training Results

In the YOLOX_s network, the target location regression and confidence regression use the IoU Loss function, and the classification of targets uses the BCE Loss function. In the CBAM-YOLOX network proposed in this paper, the CIoU Loss function is used for target location regression, the Focal Loss function is used for target confidence regression, and the BCE Loss function is also used for classification of target categories. In the experiment, under the condition that other network structures and training parameters are kept unchanged, the network training results using different loss functions are compared. The loss curves of the YOLOX_s network and the CBAM-YOLOX network are shown in Figure 7.

It can be seen from Figure 7a that the Loss curve of the YOLOX_s network converges at the 80th epoch in the process of network training, whereas the Loss curve of the CBAM-YOLOX network in Figure 7b converges at the 90th epoch. The Loss curve of YOLOX_s network converges faster, but the final Loss value is larger, and the train loss value is larger than the validation loss (val loss) value, which indicates that the generalization of the YOLOX_s network is poor. In contrast, although the CBAM-YOLOX network converges slower, it converges within 100 epochs, whereas the value of loss is slightly larger than the value of train loss, showing that CBAM-YOLOX network has a better fitting degree.
The effectiveness of using CIoU Loss and Focal Loss is preliminarily proven, which can better perform the location regression of the detected target and alleviate the problem of unbalanced positive and negative samples in the one-stage target detector.

Figure 7. Comparison of YOLOX_s network and CBAM-YOLOX network Loss curves. (a) YOLOX_s network Loss curve; (b) CBAM-YOLOX network Loss curve.

3.2.2. Comparison of The Effect of CBAM Modules in Different Locations

Under the condition that other experimental conditions and parameters remain unchanged, the CBAM attention module is placed in different locations (head, Dark5, Dark4, Dark3) of the YOLOX_s network for comparison [31]. Dark3, Dark4, and Dark5 represent the three feature map output modules of the YOLOX backbone network, respectively. The mAP values and detection speeds of the YOLOX_s network and the CBAM-YOLOX network are shown in Table 2.

Table 2. Detection accuracy and speed of different versions of CBAM-YOLOX network.

<table>
<thead>
<tr>
<th>Model</th>
<th>Head</th>
<th>Dark3</th>
<th>Dark4</th>
<th>Dark5</th>
<th>mAP@0.5%</th>
<th>Time/ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOX_s</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>94.66</td>
<td>14.26</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>98.59</td>
<td>16.10</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>98.61</td>
<td>15.77</td>
</tr>
<tr>
<td>CBAM-YOLOX</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>98.63</td>
<td>15.73</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>98.67</td>
<td>15.72</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>98.77</td>
<td>16.38</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>98.05</td>
<td>18.35</td>
</tr>
</tbody>
</table>

In Table 2, time/ms is the detection speed of each image, ✓ in Table 2 indicates that the structure has not been added. The CBAM module is placed at different locations of the YOLOX network, and the mAP of the network has different degrees of improvement, but the detection speed of the network also has different degrees of decrease. Placing the CBAM module behind both the head, Dark3, Dark4, and Dark5 structures, the mAP of the CBAM-YOLOX network improves by 3.39% and the detection time of the CBAM-YOLOX network is reduced by 4.09 ms relative to the YOLOX_s network. Placing the CBAM module behind both Dark3, Dark4, and Dark5 structures, the mAP of the CBAM-YOLOX network is improved by 4.11% and the detection time of the CBAM-YOLOX network is reduced by 2.12 ms relative to the YOLOX_s network. Compared with the YOLOX_s network, the mAP value of the CBAM-YOLOX network improves by 4.01% and detection time decreases by 1.46 ms when the CBAM module is placed behind the Dark3 structure. Compared with the YOLOX_s network, the mAP value of the CBAM-YOLOX network improves by 3.93% and detection speed decreases by 1.84 ms when the CBAM module is placed behind head structure alone. Therefore, in the
CBAM-YOLOX network, to balance the detection accuracy and detection speed, the CBAM module is placed behind the Dark3 structure of the backbone network.

Structurally, CBAM includes a spatial attention mechanism and a channel attention mechanism. Both abstract features turn into attention weights by using global average pooling, and then attach the weights to the original space or channel features. From the perspective of feature extraction position, the CBAM attention mechanism is used for feature extraction in the shallow network, and the RGB channel attention weight and the spatial attention weight of the whole image size are obtained, which is usually difficult in order to obtain effective information. In the deeper network, the number of channels is small, and the extracted channel weights are not concentrated on some specific features, which is more likely to cause negative effects. Therefore, combining the theoretical analysis and experimental results, adding the CBAM attention mechanism in the backbone network and YOLOhead has not achieved the best effect. Relatively speaking, only integrating the CBAM attention mechanism into the backbone network obtained more accurate detection results.

### 3.2.3. Comparison Experiments of Different Versions of the YOLOX Network

To verify the target detection performance of the CBAM-YOLOX network proposed in this paper, comparison tests were conducted with other versions of YOLOX networks in the same experimental environment and the same dataset. The experimental results are shown in Table 3.

#### Table 3. Detection accuracy and speed of different versions of YOLOX network.

<table>
<thead>
<tr>
<th>Model</th>
<th>R/%</th>
<th>P/%</th>
<th>mAP@0.5/%</th>
<th>Time/ms</th>
<th>Weights/KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOX_s</td>
<td>88.56</td>
<td>89.59</td>
<td>94.66</td>
<td>14.26</td>
<td>35,110</td>
</tr>
<tr>
<td>YOLOX_m</td>
<td>93.70</td>
<td>94.77</td>
<td>97.73</td>
<td>17.95</td>
<td>99,065</td>
</tr>
<tr>
<td>YOLOX_l</td>
<td>95.87</td>
<td>94.49</td>
<td>97.78</td>
<td>23.76</td>
<td>212,966</td>
</tr>
<tr>
<td>YOLOX_x</td>
<td>94.48</td>
<td>95.75</td>
<td>97.44</td>
<td>34.84</td>
<td>387,311</td>
</tr>
<tr>
<td>YOLOv5s</td>
<td>92.51</td>
<td>93.38</td>
<td>94.26</td>
<td>15.88</td>
<td>7372.8</td>
</tr>
<tr>
<td>YOLOv4</td>
<td>90.86</td>
<td>91.73</td>
<td>92.80</td>
<td>22.73</td>
<td>25,1801.6</td>
</tr>
<tr>
<td>CBAM-YOLOX</td>
<td>97.37</td>
<td>96.52</td>
<td>98.67</td>
<td>15.72</td>
<td>35,119</td>
</tr>
</tbody>
</table>

In Table 3, weights/KB is the network weight size, and time/ms is the detection speed per image. Among the four versions of YOLOX, the mAP value of YOLOX_s network is 94.66%, the detection speed is 14.26 ms, and the weight size is 35,110 KB. For the CBAM-YOLOX network, the mAP value is 98.67%, the detection speed is 15.72 ms, and the weight size is 35,119 KB. Relative to the YOLOX_s network, the weight size of the CBAM-YOLOX network increased by only 9 KB, the detection speed decreased by only 1.46 ms, the mAP value increased by 4.01%, and the recall rate increased by 8.81%, proving that the target detection capability of the CBAM-YOLOX network was effectively improved. In the YOLOX_x network, the mAP value is 97.44%, the detection speed is 34.84 ms, and the recall rate is 94.48%. In contrast, the CBAM-YOLOX network improves the mAP value by 1.23%, the recall rate by 2.89%, and the detection speed by 19.12 ms relative to the YOLOX_x network, and the weight size is 1/11 of the YOLOX_x network. From the above data, the CBAM-YOLOX network ensures both the detection accuracy and the detection speed of the network, and the CBAM-YOLOX network has smaller weights. For ship equipment, it can obtain higher accuracy detection results without increasing the hardware cost of the target detection platform.

To verify the effectiveness of the CBAM-YOLOX network, it is also compared with the YOLOv5s network and YOLOv4 network. Compared with the YOLOv5s network, the recall rate of CBAM-YOLOX network is increased by 4.86%, the mAP value is increased by 4.41%, and the detection speed is increased by 0.16 ms. Compared with the YOLOv4 network, the recall rate of the CBAM-YOLOX network is increased by 6.51%, the mAP value is increased by 5.87%, and the detection speed is increased by 7.01 ms. Therefore,
compared with other networks, the performance of the CBAM-YOLOX network has also been greatly improved, which proved the effectiveness of the CBAM-YOLOX network for target detection.

3.2.4. F1 Curve Comparison

To better illustrate the small target detection capability of the CBAM-YOLOX network, the F1 curves of the YOLOX_s network and CBAM-YOLOX network under different categories of targets were selected for comparison, and the experimental results are shown in Figure 8.

Figure 8. F1 curves of YOLOX_s network compared with CBAM-YOLOX network. (a) F1 curve for bulk cargo carrier; (b) F1 curve for container ship; (c) F1 curve for fishing boat; (d) F1 curve for general cargo ship; (e) F1 curve for ore carrier; (f) F1 curve for passenger ship.
As shown in Figure 8a–f, the F1 curves of the CBAM-YOLOX network outperformed the YOLOX_s network for all six categories of ships in the SeaShips dataset, indicating that the CBAM-YOLOX network achieved better results between accuracy and recall, proving the effectiveness of the CBAM-YOLOX network for target detection.

3.3. Comparison of Test Results

The confusion matrix can be used to visualize the performance of the algorithm. Each row of the matrix represents the real category, and each column represents the predicted category. To better show the performance of the CBAM-YOLOX network, the confusion matrix is shown in Figure 9.

As shown in Figure 9, the X-axis represents the real category, and the Y-axis represents the prediction category. According to the confusion matrix, the CBAM-YOLOX network has almost no false detections and missed detections, which proved the effectiveness of the CBAM-YALOX network for target detection.

Some of the images with small targets in the test set were selected for comparison of detection results shown in Figures 10–13, with YOLOX_s detection results on the left and CBAM-YOLOX detection results on the right.

In Figure 10, it can be seen from the first row that the detection result of YOLOX_s is a fishing boat, whereas the detection result of CBAM-YOLOX is a passenger ship. According to the images, it is known that the detection result of YOLOX_s is wrong, which means that CBAM-YOLOX is more accurate in detecting small targets with occlusion.

In the second row, the YOLOX_s has the problem of missing detection. In contrast, the CBAM-YOLOX can detect the target accurately, which proves the effectiveness of the CBAM-YOLOX network.

From Figure 11, the YOLOX_s network has the problem of missing detection when the light is dark and the target is small, whereas the CBAM-YOLOX network can effectively detect the target, which further proves the effectiveness of the CBAM-YOLOX network for small target detection.

From Figure 12, the YOLOX_s network shows a missed detection in the presence of both partial occlusion and dark light, whereas the CBAM-YOLOX network has a better detection effect, which proves the effectiveness of the CBAM-YOLOX network.
Figure 10. Comparison of small target detection results with occlusion conditions. (a) YOLOX_s detection results; (b) CBAM-YOLOX detection results.

Figure 11. Comparison of small target detection effect under dark light conditions. (a) YOLOX_s detection results; (b) CBAM-YOLOX detection results.

In Figure 13, the passenger ship is a typical small target. The YOLOX_s network can only detect the larger target ore carrier, and there is a missing detection problem for the smaller target passenger ship, whereas the CBAM-YOLOX network can detect the small target well, which proves the effectiveness of CBAM-YOLOX for small target detection.

From the detection results of each group in Figures 10–13, for small targets, the detection results of the YOLOX_s network have low accuracy, especially in the case of occlusion and low luminance, and there are missed detections. When using the CBAM-YOLOX network for target detection, the detection effect of small targets is improved. Even if the small target is occluded, the target can still be accurately detected, which proves the effectiveness of the CBAM-YOLOX network for a ship’s target detection.
Figure 12. Comparison of small target detection effects under dark and partially occluded conditions. (a) YOLOX_s detection results; (b) CBAM-YOLOX detection results.

Figure 13. Comparison of the detection effect of small targets around large targets. (a) YOLOX_s detection results; (b) CBAM-YOLOX detection results.

4. Conclusions

Aiming at the problem of the insufficient small target detection ability of the current target detection algorithm, a CBAM-YOLOX network integrating a hybrid domain attention mechanism is proposed. The advantages of convolutional neural network and attention mechanism are effectively combined to better focus on small targets from both spatial and channel dimensions in the CBAM module. The loss function is also improved, and the CIoU Loss function is used to regress the location of the target with full consideration of three factors: coverage area, centroid distance, and aspect ratio. The Focal Loss function is used to regress the confidence of the target, which effectively overcomes the problem of unbalanced positive and negative samples in the one-stage target detector. Finally, SNMS is used for post-processing to solve the problem of missing detection in the detection of close-range ship targets. Comprehensive analysis of the experimental results shows that the mAP value of the CBAM-YOLOX algorithm is increased by 4.01% compared with the YOLOX_s network, the recall rate is increased by 8.81%, and the detection speed and network weight changes are small. Compared with the YOLOX_x network, the mAP value of the CBAM-YOLOX network is increased by 1.23%, and the recall rate is increased by 2.89%. Moreover, the detection speed of the CBAM-YOLOX network is 2.22 times faster than the YOLOX_x network, and the weight is only 1/11 of the YOLOX_x network. From the above experimental results, the CBAM-YOLOX algorithm has better detection capability for small targets and a small model size, which can be better deployed on ship systems for real-time target detection task and has a certain application value.
Author Contributions: Conceptualization, Y.W.; methodology, Y.W. and J.L.; software, Y.W. and J.L.; validation, J.L., Z.C. and C.W.; writing—original draft preparation, J.L. and C.W.; writing—review and editing, J.L. and C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52271313; the Innovative Research Foundation of Ship General Performance, grant number 21822216; and Fundamental Research Funds for the Central Universities, grant number 3072022QBJ0406.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study.

References


