Journal of *Marine Science and Engineering*

MDPI

Article

StereoYOLO: A Stereo Vision-Based Method for Maritime Object Recognition and Localization

Yifan Shang, Wanneng Yu *, Guangmiao Zeng, Huihui Li and Yuegao Wu

School of Marine Engineering, Jimei University, Xiamen 361021, China; 202112855021@jmu.edu.cn (Y.S.); gm.zeng@foxmail.com (G.Z.); huihuili@jmu.edu.cn (H.L.); 202212855074@jmu.edu.cn (Y.W.)
* Correspondence: wnyu2007@jmu.edu.cn

Abstract: Image recognition is vital for intelligent ships' autonomous navigation. However, traditional methods often fail to accurately identify maritime objects' spatial positions, especially under electromagnetic silence. We introduce the StereoYOLO method, an enhanced stereo vision-based object recognition and localization approach that serves autonomous vessels using only image sensors. It is specifically refined for maritime object recognition and localization scenarios through the integration of convolutional and coordinated attention modules. The method uses stereo cameras to identify and locate maritime objects in images and calculate their relative positions using stereo vision algorithms. Experimental results indicate that the StereoYOLO algorithm boosts the mean Average Precision at IoU threshold of 0.5 (mAP50) in object recognition by 5.23%. Furthermore, the variation in range measurement due to target angle changes is reduced by 6.12%. Additionally, upon measuring the distance to targets at varying ranges, the algorithm achieves an average positioning error of 5.73%, meeting the accuracy and robustness criteria for maritime object collision avoidance on experimental platform ships.

Keywords: object detection; stereo vision; attention mechanism; deep neural network; YOLO

Citation: Shang, Y.; Yu, W.; Zeng, G.; Li, H.; Wu, Y. StereoYOLO: A Stereo Vision-Based Method for Maritime Object Recognition and Localization. *J. Mar. Sci. Eng.* **2024**, *12*, 197. https://doi.org/10.3390/jmse12010197

Academic Editor: Alessandro Ridolfi

Received: 28 December 2023
Revised: 14 January 2024
Accepted: 19 January 2024
Published: 22 January 2024

1. Introduction

With the deployment of innovative unmanned vessels such as the USV Mariner and LUSV Ranger by the U.S. Navy, a pivotal shift in ship autonomy, maritime navigation, and surveillance technology is marked. Despite these advancements, maritime target recognition still relies on active detection methods, such as radar, ESM, sonar systems, and Automatic Identification System (AIS). While radar and sonar are fundamental tools for identifying other vessels, they often lack rich texture information and typically require data fusion with other sources like AIS and ESM for effective target detection [1]. However, AIS and ESM are not equipped on all vessels, and their update frequency fails to meet the real-time requirements of autonomous navigation. Importantly, radar, as a radio emission source, is unusable in scenarios where unmanned vessels need to navigate autonomously while maintaining radio silence. Therefore, leveraging passive detection technologies, such as computer vision, is crucial for enhancing situational awareness under silent conditions, vital for improving the autonomy and safety of unmanned ships.

The shift from traditional computer vision methods to deep learning in image recognition has been decisive. Traditional computer vision methods, such as the Sobel [2] and Canny [3] edge detection algorithms, identify object contours and edges by analyzing changes in image brightness. The Sobel algorithm highlights edges by calculating the gradient of image brightness, while the Canny algorithm further optimizes edge detection through a multi-stage process, enhancing accuracy and robustness. Histogram of Oriented Gradients (HOG) is another significant traditional method, building feature descriptors by tallying the direction and magnitude of gradients in local image areas [4]. Scale-Invariant Feature Transform (SIFT) detects and describes local image features for feature matching

and object recognition. SIFT, with its scale invariance, effectively handles problems caused by image scaling, rotation, and partial changes in perspective [5]. These traditional methods have played a crucial role in initial feature identification and extraction in images. However, they typically require manual adjustments and are less effective in handling the complexity of high-dimensional image data encountered in modern applications [6,7].

Deep neural networks present unique advantages over traditional computer vision methods. With their deep structure, they can automatically learn and extract hierarchical feature representations from vast image data [8], a capability particularly crucial in the field of computer vision. In object detection, deep neural networks can effectively distinguish and identify different objects by analyzing various patterns and textures in images [9]. In image segmentation tasks, they segment images into multiple regions, identifying the attributes and boundaries of each [10], which is essential for understanding image content and context. Furthermore, in image classification tasks, deep neural networks accurately categorize images into predefined classes by analyzing both global and local features [11].

As autonomous driving technology continues to evolve, the importance of computer vision-based target localization algorithms has become increasingly prominent. Some automobile manufacturers have adopted pure vision detection methods that do not rely on high-precision maps, using multi-camera systems for autonomous driving, presenting a promising alternative [12]. In this context, researchers have proposed methods for target tracking and angle tracking using stereo images [13] and introduced the Stereo R-CNN algorithm for implementation [14]. Additionally, studies have focused on geometric information from stereo images, using Stereo Centernet to detect and locate targets in three-dimensional space, supporting the autonomous driving of vehicles [15]. As target detection networks have evolved, adjusting algorithms for the spatial and channel-wise distribution probability of detected features has become an important form of improvement in the image recognition field [16–18]. However, these algorithms are designed for terrestrial applications and have not been specifically optimized for marine target characteristics.

The uniqueness of marine target recognition tasks lies in large-scale variations, potential occlusions, overlaps, and blurriness of targets in images. Furthermore, marine targets in images are predominantly located near the sea–sky line. Thus, algorithms need to be specifically improved for marine target recognition tasks. Some studies have designed algorithms for maritime conditions and achieved good results [19], such as ISDet, which improved the accuracy of marine target recognition by enhancing the ShuffleNet network structure and applying the PD-NAML training method [20]; CLFR-Det, which enhanced recognition accuracy by using features of different levels and semantics, combined with cross-layer deformed convolution and a multi-scale feature refinement mechanism for enriching the semantic information of small vessels [21]; and methods that improved recognition accuracy by merging multiple visual features and segmenting sea-surface images after detecting the sea–sky line [22]. YOLO-based object detection algorithms, tailored for marine targets, have also demonstrated real-time target recognition capabilities, capable of recognizing ship targets in satellite [23–25], aerial [26], and horizontal perspective images [27], proving the feasibility of deep learning-based marine target recognition algorithms. However, they often require data fusion with spatial information obtained from radar to locate targets [28], hindering their use under radio silence conditions for maritime target localization.

An ideal approach to achieve marine spatial information perception based solely on images is to utilize stereoscopic vision algorithms. In maritime platform applications, binocular stereo vision, monocular vision, and point cloud semantic segmentation algorithms have achieved significant success, especially in 3D reconstruction performance. Binocular stereo vision algorithms estimate depth information by capturing images from two cameras at different angles and comparing the differences between these images [29]. Monocular stereo vision relies on a single camera, combining data from Inertial Measurement Units (IMU) and analyzing changes in image disparity to perceive three-dimensional space [30]. Point cloud semantic segmentation algorithms classify objects in three-dimensional space by analyzing point cloud data obtained from LiDAR or other 3D scanning devices [31,32].

The application of these technologies, particularly in complex maritime environments, has significantly improved the accuracy and efficiency of sea surface stereoscopic perception. However, these efforts have focused on spatial perception in maritime environments without integrating with target recognition methods to obtain distance and location information to support ship autonomy.

Therefore, to meet the requirements of autonomous ship navigation for marine target recognition and localization, combining deep learning-based marine target recognition with stereoscopic vision algorithms is both feasible and urgent. This paper proposes the StereoYOLO algorithm, which improves the YOLOv5 object recognition algorithm by incorporating stereo vision and attention mechanisms, thereby achieving recognition and localization of marine targets. By analyzing a large dataset of marine targets, a deep convolutional neural network-based method for recognizing marine targets is developed. The algorithm accurately identifies the category of marine targets based on their spatial and channel characteristics. Then, using stereo vision algorithms for depth perception of recognized marine targets and transforming the coordinates into world coordinates to obtain the location of marine targets. Additionally, to apply the algorithm to small, unmanned surface vessels, it is necessary to adapt the algorithm for use with the NVIDIA Jetson embedded development board in experiments.

2. StereoYOLO: A Maritime Target Recognition and Motion State Detection Algorithm

2.1. Target Recognition and Motion State Detection Process

The stereo camera system comprises left and right cameras that capture corresponding images independently. Initially, the left image is processed through an attention-integrated deep target recognition network, encompassing a Focus network, a main feature extraction network, and an enhanced feature extraction network. This sequential processing leads to the identification of target anchor boxes. In the subsequent stage within the enhanced feature extraction framework, these identified anchor boxes guide the precise selection of feature points from both left and right images of the stereo pair. For the purpose of matching these feature points, the Semi-Global Block Matching (SGBM) algorithm is employed [33]. SGBM is adept at identifying distinct and reproducible features in stereo imagery, achieved by aggregating matching costs across multiple directions and implementing a semi-global optimization strategy. This strategy is instrumental for ensuring reliable feature point matching, a critical factor for computing accurate disparity. Following the matching phase, a disparity-to-depth conversion is applied to the selected feature points to obtain their three-dimensional relative coordinates. The precision inherent in this feature point selection and matching procedure plays a crucial role in estimating distances accurately, which is fundamental to the robustness and precision of the three-dimensional localization process.

The target distance is calculated using the stereo ranging algorithm and further refined by applying the K-means clustering algorithm to the distance information of multiple feature points. The overall framework of the algorithm is shown in Figure 1.

2.2. Stereo Vision Model

The stereo vision measurement principle is depicted in Figure 2. O_1 and O_2 represent the optical centers of the left and right camera imaging planes, with pixel coordinates (u_{O1}, v_{O1}) and (u_{O2}, v_{O2}), respectively. $O_1x_1y_1z_1$ and $O_2x_2y_2z_2$ define the coordinate systems for the left and right cameras. The target feature point X_c, with coordinates (x, y, z) in the left camera coordinate system, projects onto the left camera imaging plane ($X_1O'_1Y_1$) at point p_1 and onto the right camera imaging plane($X_2O'_2Y_2$) at point p_2. $O_1O'_1$ and $O_2O'_2$ denote the focal lengths f of the left and right camera lenses.

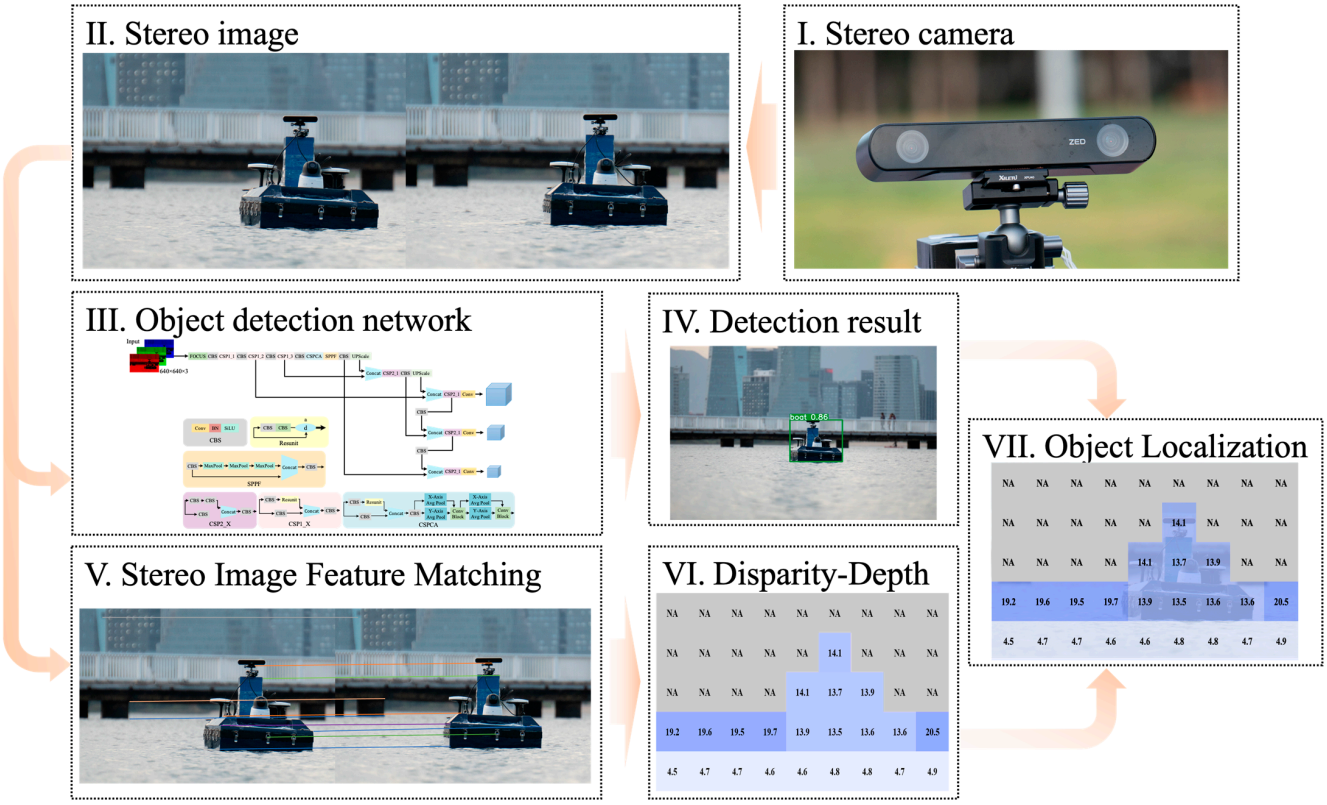

Figure 1. Ship motion state detection algorithm framework. Note: A larger version of Subfigure III is presented in Figure 4 for detailed view.

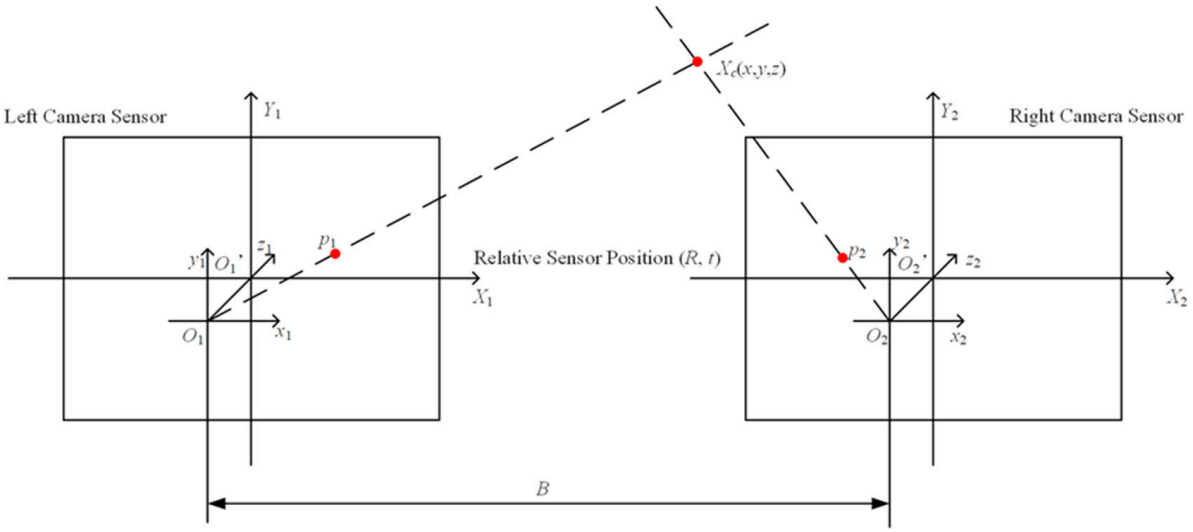

Figure 2. Stereo vision distance measurement principle.

In the process of stereo matching using the SGBM algorithm, the homogeneous pixel coordinates of the projection points p_1 and p_2 for the maritime target feature point X_c on the imaging planes of the left and right cameras are computed as $P_1 = (u_1, v_1)$ and $P_2 = (u_2, v_2)$. We should first convert the pixel coordinates of the feature points into homogeneous coordinates, i.e., $P'_1 = (u_1, v_1, 1)$ and $P'_2 = (u_2, v_2, 1)$. The homogeneous coordinates are back-projected to obtain the three-dimensional coordinates relative to each camera's coordinate system:

$$\begin{aligned} \widetilde{X}_{c1} &= M_1^{-1} \cdot P_1' \\ \widetilde{X}_{c2} &= M_2^{-1} \cdot P_2' \end{aligned} \tag{1}$$

In Equation (1), the matrices M_1 and M_2 represent the 3×3 intrinsic parameter matrices of the left and right cameras, respectively, which are obtained post-calibration and determined by the camera sensor's physical characteristics.

The point in the right camera's coordinate system is transformed to the left camera's coordinate system using the relative rotation matrix R and translation vector t:

$$\widetilde{X}_{c2}' = R \cdot \widetilde{X}_{c2} + t \tag{2}$$

With the point expressed in both camera coordinate systems, triangulation methods are employed to resolve the actual three-dimensional coordinates of X_c: The three-dimensional coordinates X_c are obtained by solving Equation (3):

$$\lambda_1 \cdot \widetilde{X}_{c1} = \lambda_2 \cdot \widetilde{X}_{c2} \tag{3}$$

Here, λ_1 and λ_2 are scale factors, determined by minimizing the disparity between the two expressions using Levenberg–Marquardt algorithm. The three-dimensional coordinates X_c (x, y, z) are computed by applying the scale factor λ_1 to the inverse of the left camera's intrinsic matrix M_1 and multiplying it with the homogenized image coordinates P'_1 as Equation (4):

$$X_c = \lambda_1 \cdot M_1^{-1} \cdot P_1' \tag{4}$$

This computational approach is anchored in the principles of triangulation inherent to stereo vision, leveraging the disparity in perspectives of the same point as observed by two distinct cameras to ascertain its position in three-dimensional space.

In maritime target detection tasks, we need a horizontal coordinate to localize the object. So, we should correct the coordinate system defined by X_c (x, y, z) which the Z-axis extends perpendicularly from the imaging plane.

To calculate the horizontal distance of an object on the water surface, taking into account the ship's roll, pitch, and yaw, we can use rotation matrices to adjust the object's position onto a 2D plane where the y-axis represents the ship's forward direction on the water's surface, and the x-axis represents the lateral direction. The process is as follows:

Define rotation matrices for roll, pitch, and yaw angles as $R_{roll}(\alpha)$, $R_{pitch}(\beta)$, and $R_{yaw}(\gamma)$ adjustments. These matrices are given in Equation (4):

$$\begin{aligned} R_{roll}(\alpha) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix} \\ R_{pitch}(\beta) &= \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \\ R_{yaw}(\gamma) &= \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{5}$$

Apply these rotations to X_c to adjust for the ship's orientation as Equation (6):

$$X_{horizontal} = R_{yaw}(\gamma) \cdot R_{pitch}(\beta) \cdot R_{roll}(\alpha) \cdot X_c \tag{6}$$

Here, the $X_{horizontal}[0]$ and $X_{horizontal}[1]$ distance of the horizontal plane can be extract from $X_{horizontal}$. And the distance can be calculated as Equation (7):

$$D_{horizontal} = \sqrt{X_{horizontal}[0]^2 + X_{horizontal}[1]^2} \quad (7)$$

The schematic diagram of the improved marine distance measurement algorithm is shown in Figure 3:

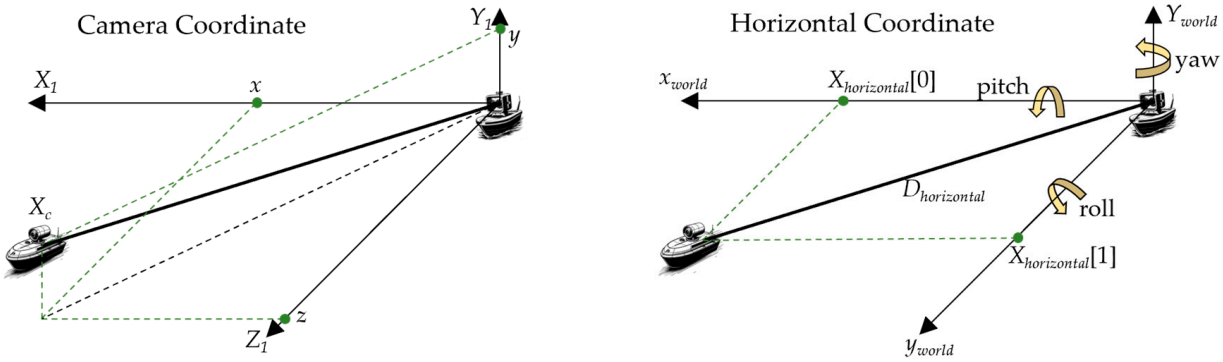

Figure 3. Representation of Distance in Different Coordinate Systems.

2.3. Improved Maritime Target Detection Algorithm

2.3.1. Algorithm Framework

The target recognition network algorithm in StereoYOLO is based on the improved YOLOv5 object detection algorithm, a leading algorithm in the field of object detection. The algorithm divides the image into grids and detects objects within these grids. Each grid cell is responsible for detecting targets within itself. Due to its excellent efficiency and accuracy, YOLO has become one of the most famous object detection algorithms.

Since its release, YOLOv5 has continuously improved its algorithm to enhance efficiency and accuracy. In this paper, we use the YOLOv5 v6.0 version code implementation. The model size can be divided into YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The weights of the five models are stacked sequentially. The algorithm is based on the improved YOLOv5s detection model.

The target recognition network framework in the StereoYOLO algorithm is shown in Figure 4. The main structure of the algorithm consists of the Backbone, Feature Pyramid Network (FPN), and Yolo Head. Here, Resunit represents the residual module, Conv represents the convolutional block, BN represents batch normalization, SiLU represents the activation function, Concat represents merging arrays, and MaxPool represents max pooling. Channel Attention represents the channel attention module, Spatial Attention represents the spatial attention module, and complex network structures are established by combining basic modules. This algorithm replaces the CSP1_X module in the backbone network with the improved CSPCBAM and CSPCA modules with attention mechanisms, using channel and spatial attention mechanisms to apply targeted weighting for maritime target recognition scenarios, thus optimizing the detection accuracy of maritime target detection tasks.

2.3.2. Backbone Network

The algorithm uses the CSPDarknet backbone network to extract features from the input images and output them as feature layers. This process is performed three times to obtain three feature layers, called effective feature layers.

At the input end, the backbone network employs the Focus network structure, extracting a value from every other pixel in each image to generate four independent feature layers. This expands the input channel count by four times, resulting in a better depth compared to the three-layer structure in other networks.

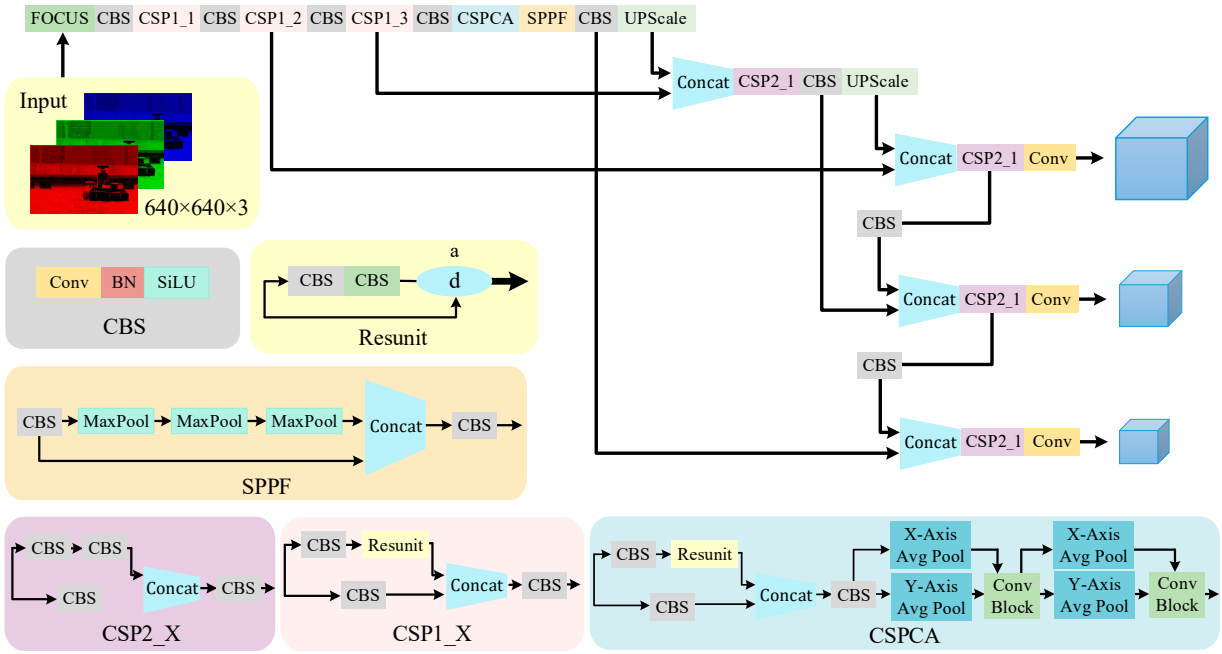

Figure 4. Object detection network structure.

The backbone network extensively employs residual networks (Resunit), where the residual edges are not processed, and the input and output of the backbone are directly combined to avoid gradient dispersion and network degradation problems. The use of residual networks alleviates the vanishing gradient problem caused by increased depth in the backbone network. At the same time, CSPnet splits the stacking of residual blocks, with the main part continuing the stacking of the original residual blocks and the remaining part directly connected to the end after minimal processing. Additionally, the backbone network employs SiLU as the activation function, which has characteristics such as unbounded upper, bounded lower, smooth, and non-monotonic. By smoothing the ReLU activation function, SiLU performs better than ReLU in deep models.

The final part of the backbone network is the spatial pyramid pooling (SPP) structure, which extracts features through max pooling with different kernel sizes, increasing the receptive field.

2.3.3. Feature Extraction Network

The main function of the enhanced feature extraction network is to perform feature fusion on the three effective feature layers output by the backbone network, combining features of different scales to enhance feature extraction. In the feature utilization part, the algorithm extracts three feature layers for target detection by extracting multiple feature layers.

The three feature layers are located at different positions in the backbone part of CSPdarknet, specifically in the middle layer, middle-lower layer, and bottom layer. These three effective feature layers will be used to build the FPN layer, allowing FPN to fuse feature layers of different sizes, thus promoting feature extraction and generating three enhanced features.

2.3.4. Detection Head

The main function of the detection head is to convolve the three enhanced features generated by FPN separately and determine whether there are objects corresponding to the feature points in the feature map.

By inputting the three enhanced features with dimensions (20,20,1024), (40,40,512), and (80,80,256) obtained from the FPN feature pyramid into the detection head, the prediction results are obtained.

2.3.5. Loss Function

The loss function of the model uses Generalized Intersection over Union (GIoU) as the loss function for bounding box regression, which is derived from the improvement of the Intersection over Union (IoU) loss function.

The calculation of the IoU loss function is shown in Equation (8):

$$\mathrm{IoU} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{8}$$

where A is the predicted box and B is the ground truth box. IoU calculates the overlap between the predicted box and the ground truth box, reflecting the detection result and performing backpropagation. However, when used as a loss function, if there is no intersection between the ground truth box and the predicted box, IoU = 0, and the gradient is zero. Without gradient backpropagation, backpropagation cannot continue. Therefore, an improved loss function is needed to ensure that the gradient is not zero when there is no overlap between the ground truth box and the predicted box, allowing for backpropagation.

The GIoU calculation formula is shown in Equation (9):

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|\mathrm{C} \backslash (\mathrm{A} \cup \mathrm{B})|}{|\mathrm{C}|} \tag{9}$$

The GIoUloss calculation formula is shown in Equation (10):

$$\mathrm{GIoUloss} = 1 - \mathrm{GIoU} \tag{10}$$

where C is the minimum convex closure containing both A and B. By taking the difference set of C, GIoU is non-zero and decreases as C.

2.3.6. Attention Mechanism-Based Maritime Target Detection Algorithm

In maritime target recognition tasks, camera-captured maritime targets are predominantly located near the sea–sky line, necessitating tailored attention mechanisms for enhanced detection. These targets, often represented as tensors, exhibit spatial correlations critical for recognition.

To capitalize on this, a spatial attention module is employed, assigning higher weight values to tensor regions where targets are more likely to emerge, typically the lower half of the image in maritime scenarios. Concurrently, the unique environmental backdrop of bluish skies and greenish seas in these tasks influences the pixel values in captured images. Specifically, in the red, green, and blue (RGB) channels, blue, and green values tend to be higher due to this backdrop, whereas pixels containing maritime targets display varied channel values, reflective of distinct target features. This variation underscores the necessity of channel-specific attention.

By introducing channel attention, the algorithm can differentially weigh the RGB channels based on their relevance to target features, enhancing detection accuracy. Such channel-specific adjustments, in synergy with spatial attention, forge a more nuanced and effective approach to maritime target recognition, optimizing the likelihood of accurate target identification.

Due to the specificity of maritime target recognition tasks, this paper uses the attention mechanism to improve the maritime target detection algorithm, expecting to achieve better target detection accuracy. In this paper, CBAM attention module and CA attention module are separately introduced into the backbone network to improve recognition accuracy by targeting the distribution characteristics of targets to be recognized in space and channels in maritime target recognition tasks. Attention modules are introduced after the CSP feature

extraction module to change the model's sensitivity to different regional features. The improved CSPCBAM and CSPCA network structures are shown in Figure 5:

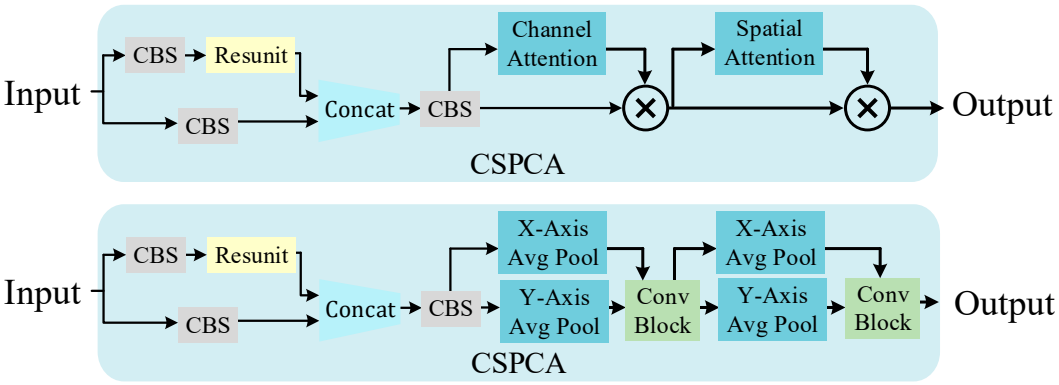

Figure 5. CSPCBAM (**top**) and CSPCA (**bottom**) attention module.

In the figure, Channel Attention represents channel attention, Spatial Attention represents spatial attention, X-Axis Avg Pool represents X-axis average pooling, Y-Axis Avg Pool represents Y-axis average pooling, X-Axis Attention represents X-axis attention, and Y-Axis Attention represents Y-axis attention.

CBAM combines channel attention and spatial attention mechanisms. Given an intermediate feature map, the module sequentially infers attention maps along the two independent dimensions of channels and space, and then multiplies the attention maps by the input feature map for adaptive feature refinement.

In the CBAM attention module, given an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$ as input, the module successively infers a 1D channel attention map $M_c \in \mathbb{R}^{C\times H\times W}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{C\times H\times W}$. The algorithm structure is represented by Equation (11):

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \tag{11}$$

where F represents the output tensor of the CSP module, F' represents the weight tensor after channel attention processing, and F'' represents the weight tensor after channel-spatial attention mechanism processing. $\otimes$ represents the tensor inner product.

The Convolutional Block Attention Module (CBAM) sequentially applies two distinct attention mechanisms to the input tensor: first, the channel attention module assesses the importance of each feature channel, then the spatial attention module evaluates the significance of different spatial regions. The tensor is progressively refined through these mechanisms, with the final output reflecting the enhanced feature representation after attention has been applied.

The properties of the CBAM attention module can take into account both channel attention and spatial attention, and affect the weights of related feature values, which meets the needs of maritime target recognition tasks. Therefore, this paper introduces it into the feature extraction backbone network for the improvement of maritime target recognition tasks.

CA attention utilizes the different probabilities of detecting targets in image width and height to optimize target detection tasks. It encodes precise target information in image width and height, inputs the feature map, and performs global average pooling operations in the h and w directions, respectively, to obtain feature maps in the h and w directions. The formula is shown in Equation (12):

$$\begin{aligned} z_c^h(h) &= \frac{1}{W} \sum_{0\le i\le w} x_c(h,i) \\ z_c^w(w) &= \frac{1}{H} \sum_{0\le j\le h} x_c(j,w) \end{aligned} \tag{12}$$

The feature maps along the h and w directions are concatenated, and then sent into a 1×1 convolution block with shared weights. The feature map F_1 is then subjected to batch normalization and sent into a Sigmoid activation function to obtain the feature map f, as shown in Equation (13):

$$f = \sigma\left(F_1\left[z_c^h, z_c^w\right]\right) \tag{13}$$

Following that, perform a 1×1 convolution on the feature map f according to the original height and width, obtaining new feature maps F_h and F_w. After applying the Sigmoid activation function, the attention maps for the h and w directions, g^h and g^w, can be obtained, as shown in Equation (14):

$$\begin{aligned} g^h &= \sigma\left(F_h\left(f^h\right)\right) \\ g^w &= \sigma(F_w(f^w)) \end{aligned} \tag{14}$$

Finally, compute the multiplication of the attention weights g^h and g^w with the original feature maps in the h and w directions, respectively, to obtain the feature maps with attention weights in the h and w directions, as shown in Equation (15):

$$y_c(i,j) = x_c(i,j) \times g^h(i) \times g^w(j) \tag{15}$$

The CA attention module primarily focuses on spatial attention and refines it into attention weights in the image height and width directions, generating attention-weighted feature maps. This provides a targeted improvement for the greater probability of targets appearing near the sea–sky line in maritime target recognition tasks.

2.3.7. Experimental Platform

Our research is based on the ZED stereo camera for development, as shown in Figure 6, with a resolution of 1920×1080 (single-lens) and capturing 30 frames per second. The development platform utilizes a Windows 10 21H2 operating system, 64 GB RAM, a CPU i7-10700F with a base frequency of 2.9 GHz, a GPU RTX A4000, and Python 3.9.7 as the experimental platform.

Figure 6. ZED stereo camera (**left**) and experimental vessel (**right**).

The experimental platform uses a 1.8 m test boat as the target. By performing target detection, ranging, and positioning experiments on the platform, the accuracy of the maritime target recognition algorithm and ranging errors can be verified, thereby validating the target positioning algorithm.

To meet the real-time and offline computing requirements of maritime target recognition and ship motion state monitoring tasks, the mobility and power requirements of the monitoring platform must be considered. Therefore, it is necessary to port the maritime target recognition algorithm to make it suitable for small, embedded devices.

The embedded platform in this research adopts the Ubuntu 20.04 operating system, with NVIDIA Jetson AGX Orin as the hardware platform [34]. It has 32 GB RAM, 5.32TFLOPS of Single-Precision floating-point performance, and supports CUDA 11.3.

The detailed parameters of the development and embedded platforms are shown in Table 1:

Table 1. Hardware platform parameters.

Platform Information	Development Platform	Embedded Platform
CPU	8 Core Intel i7-10700F@2.9GHz	12 Core Arm Cortex-A78AE@1.3GHz
RAM	64 GB	32 GB
GPU	NVIDIA RTX A4000	2048 Core NVIDIA Ampere GPU
FP32 Performance	19.2 TFLOPS	5.32 TFLOPS
Operation System	Windows10 21H2	Ubuntu20.04
Power Consumption	300 W	60 W

The detection results, including three-dimensional position and speed data, are transmitted from the embedded platform to the lower-level machine via the CAN bus interface. This transmission encompasses the target's identification number and its position and speed in three dimensions. Additionally, the lower-level machine is connected to an Inertial Measurement Unit (IMU), enabling it to calculate two-dimensional horizontal positioning information. The transmitted data are formatted in an extended frame, consisting of the target sequence number, category, three-dimensional positions (x, y, z), and velocities (x speed, y speed, z speed).

3. Model Training and Experiment

3.1. Model Training and Evaluation Metrics

According to the needs of maritime target recognition tasks, the public dataset Seaships7000 [35] is used for data augmentation and model training. The dataset classifies ship targets into six types: ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship. The dataset uses VOC format annotation, and VOC annotations are used to convert VOC format dataset annotations to COCO format dataset annotations. The dataset contains a total of 7000 images, divided into test, training, and validation datasets at a ratio of 1:9:0.9, resulting in 700 test images, 6300 training images, and 630 validation images. The ship target categories, positions, and the aspect ratio of the target's width and height in the images are shown in Figure 7:

As can be seen, in maritime target recognition tasks, the probability of targets appearing in the camera detection area in the y-direction between 0.3 and 0.6 is relatively high. Due to the specificity of maritime target recognition tasks, the algorithm should apply different confidence levels to different areas of the image. By using attention mechanisms, targeted performance optimization can be achieved for maritime target recognition tasks.

During model training, to accommodate the GPU memory capacity of the experimental platform, the actual loaded images are scaled and padded to obtain input images with a training resolution of 640 × 640. Model training uses the Adaptive Moment Estimation (Adam) optimizer, with an initial learning rate of lr0 = 1×10^{-3}, a batch size of 16, and 100 epochs. When loading training data, data augmentation techniques such as mosaic, image distortion, changing brightness, contrast, hue, adding noise, random scaling, random cropping, flipping, rotation, and random erasing are used.

3.2. Simulation Experiment Analysis of Ship Target Detection Based on CBAM Improvement

The maritime target recognition algorithm proposed in this paper is based on improvements to the YOLOv5 algorithm. Ablation experiments are conducted on the Seaships dataset for ship target detection to determine the optimal position of the attention module in the algorithm improvement. This experiment uses the development platform, and some test parameters are shown in Table 2:

Table 2. Hyperparameters of ablation experiment.

resolution	640 × 640
maximum epoch	100
optimizer	Adam
batch Size	16
Training data	6300
Val data	700
pretrained	No

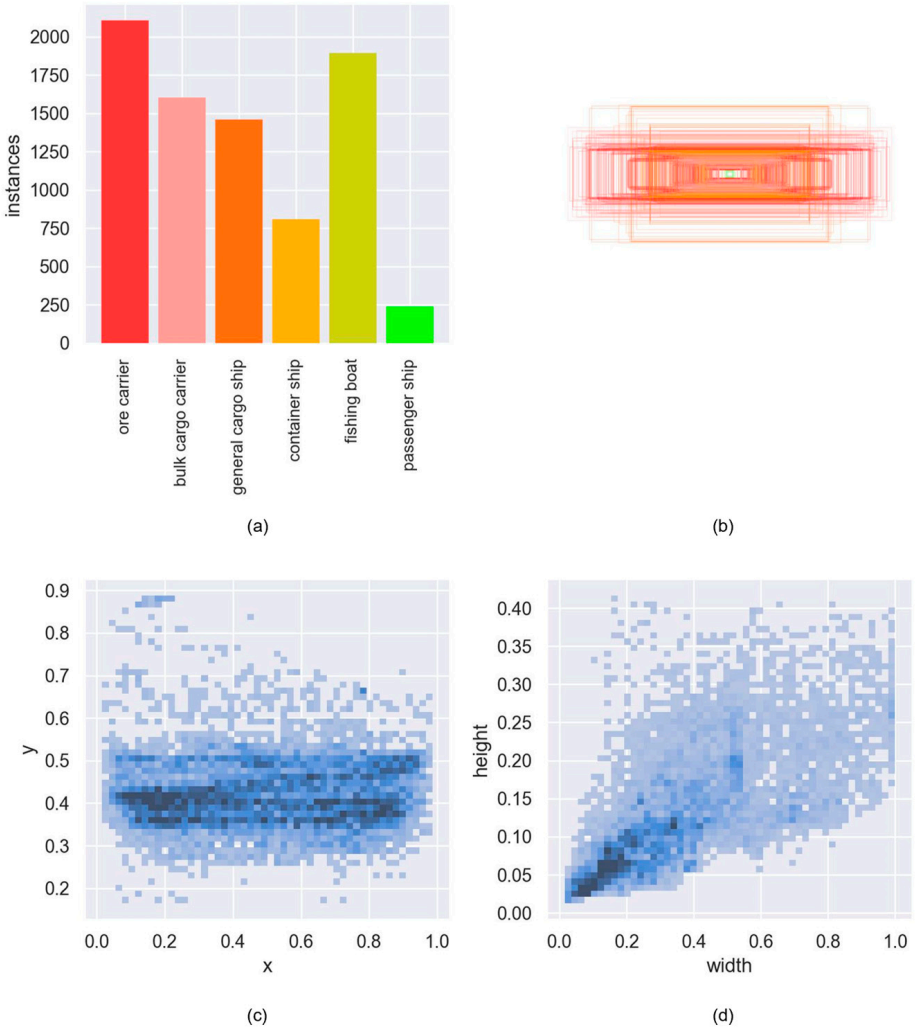

Figure 7. Dataset target category (**a**), Anchor Boxes Visualization (**b**), Normalized Coordinates (**c**), Normalized size (**d**).

To evaluate the detection performance of the ship target detection algorithm that incorporates the CBAM attention mechanism, the detection results of introducing the CBAM attention module after the CSP1_X layer and the CONV layer at different positions in the model are compared with the detection results of the YOLOv5s target detection algorithm for maritime targets. Figure 8 shows the loss function value curves obtained during the network training process when the CBAM attention mechanism is added after the CSP1_4 layer, after all CSP layers, and after the CONV_1 and CONV_2 layers. Figure 9 removes the improved algorithms with less improvement, retaining only the improved StereoYOLO network with the CBAM attention mechanism module added after the CSP1_4 layer and the YOLOv5s network loss function curve.

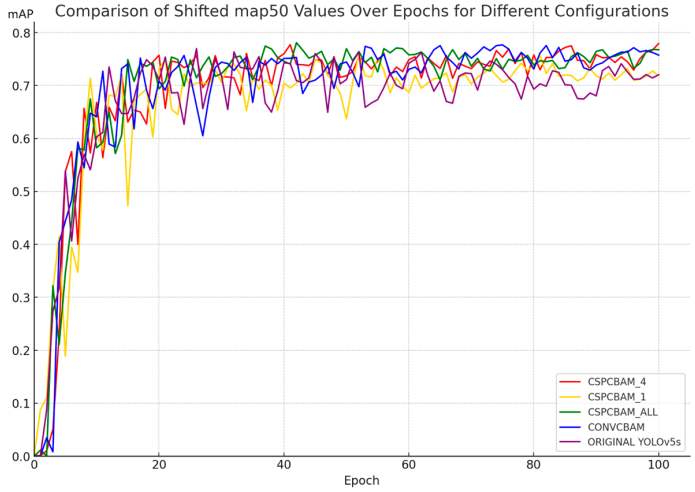

Figure 8. Results of the ablation experiment.

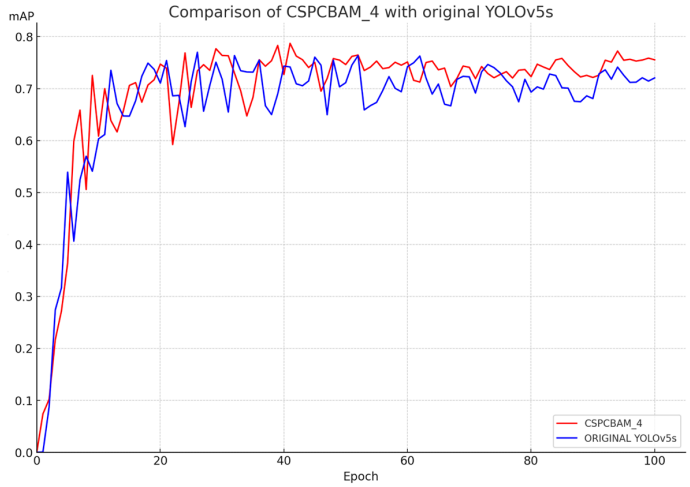

Figure 9. Comparison of our CSPCBAM_4 method with Original YOLOv5s.

By conducting ablation experiments on the CSPCBAM-improved backbone feature extraction network, replacing the CSP1_X module with the CBAM attention mechanism at different positions in the CSPDarknet backbone network, we observed changes in the training process of the StereoYOLO target detection network. Notably, introducing the attention mechanism after the last CSP layer resulted in a noticeable modification in the network's performance. In the absence of pre-training, the model achieved a final convergence mean Average Precision at IoU threshold of 0.5 (mAP50) of 78.70%. This represents an improvement over the YOLOv5s algorithm, suggesting the potential effectiveness of the algorithmic enhancements.

The optimal weights during the training process are taken when the CBAM attention mechanism is added at different positions in the backbone network, and comparative experiments are conducted on the Seaships test set. The test results are shown in Table 3:

Table 3. Results of the ablation experiment.

Modified Module	mAP50/%
CSP1_1	77.69
CSP1_4	78.70
CSP1_1/CSP1_2/CSP1_3/CSP1_4	78.09
YOLOv5s	76.99

By adding the CBAM attention mechanism after the fourth CSP layer in the backbone feature extraction network, the detection accuracy is increased from 76.99% to 78.70%. The maritime target detection algorithm with the added CBAM attention mechanism improved the mAP50 by 1.71% compared to the YOLOv5s target detection network in the Seaships dataset test. At the same time, the algorithm's performance requirements are the same as the original algorithm, and overfitting problems can be effectively avoided. The detection results of the improved StereoYOLO target detection algorithm are shown in Figure 10.

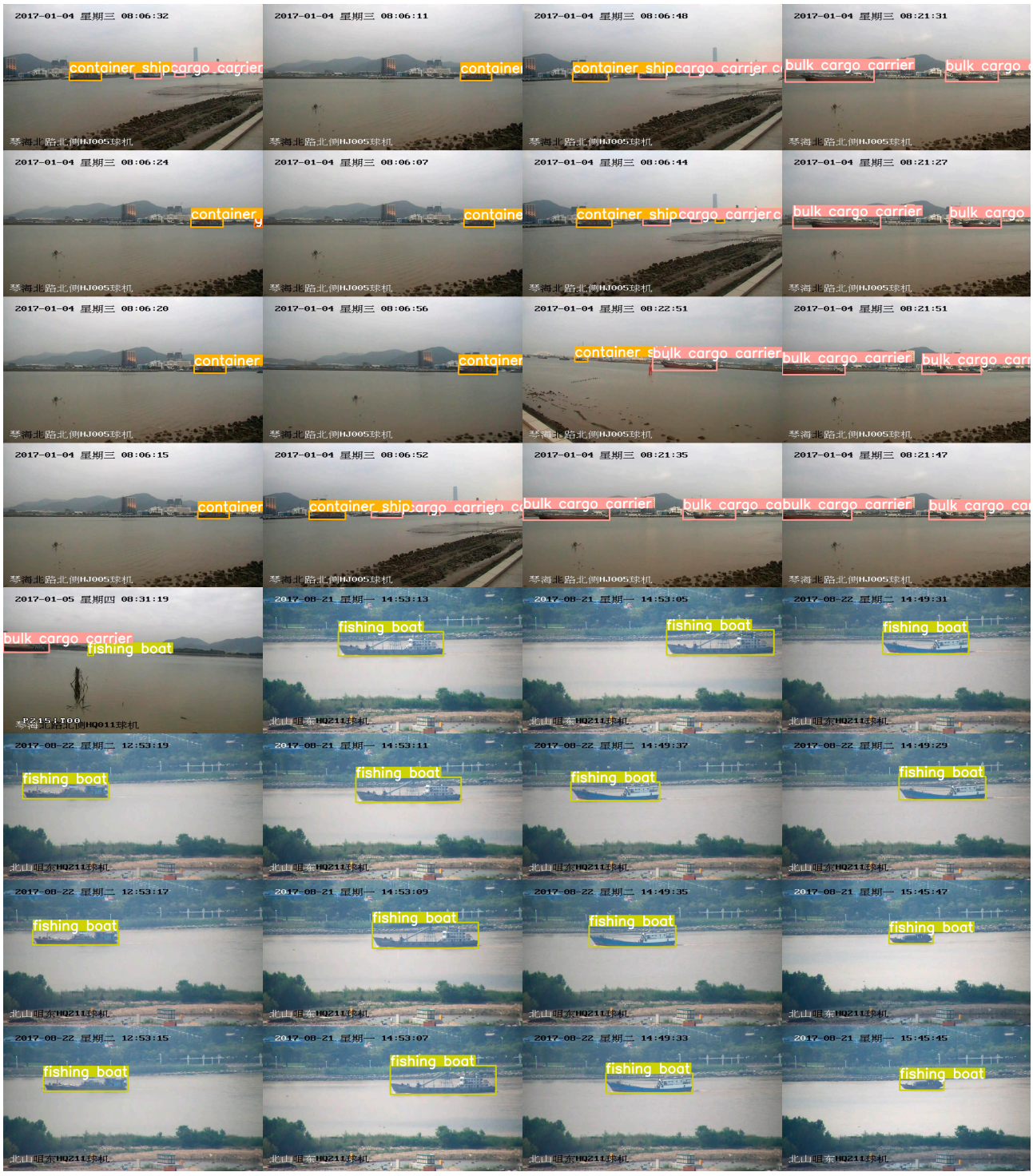

Figure 10. Results of the improved object detection algorithm.

It can be seen that the target detection algorithm improved based on the attention mechanism has a higher prediction box accuracy in the test dataset after mosaic image enhancement (left) and the standard maritime images without image enhancement (right) and can complete the image detection task of maritime ship targets.

3.3. Simulation Analysis of Ship Target Detection with Improved Maritime Target Detection Algorithm Based on Multiple Attention Mechanisms

According to the experimental results in Section 3.2, the introduction of the attention mechanism can significantly improve the accuracy of the maritime target algorithm. To explore the role of different attention mechanism modules in maritime target recognition tasks, we introduce different attention mechanism modules to improve the algorithm and determine the direction of attention mechanism improvement by calculating the arithmetic average of 5 repetitions to eliminate the errors caused by the randomness of neural networks.

After improving the algorithm by adding CA attention module and CBAM attention module behind CSP1_4 layer, the model accuracy mAP50 is obtained, and the influence of adding CA attention module and CBAM attention module on the mAP50 accuracy of maritime target recognition tasks in YOLOv5 algorithm is compared with the original algorithm mAP50 accuracy, as shown in Table 4:

Table 4. Results of the improved object detection algorithm.

Experiment ID	Attention Plugins	FLOPs	mAP50/%
1	None	15.8G	77.48
2	None	15.8G	75.35
3	None	15.8G	76.20
4	None	15.8G	76.10
5	None	15.8G	76.25
6	CA	15.9G	77.04
7	CA	15.9G	77.66
8	CA	15.9G	78.80
9	CA	15.9G	86.14
10	CA	15.9G	87.86
11	CBAM	15.9G	80.31
12	CBAM	15.9G	76.53
13	CBAM	15.9G	77.87
14	CBAM	15.9G	77.22
15	CBAM	15.9G	79.78

The mAP50 target recognition accuracy after adding CA attention mechanism or CBAM attention mechanism to the backbone network CSP1_4 of YOLOv5 algorithm is 81.50% and 78.34%, respectively, while the original YOLOv5 algorithm target recognition accuracy is 76.27%. That is, the accuracy of YOLOv5 algorithm after adding CA attention and CBAM attention improvements is increased by 5.23% and 2.07% compared to the original algorithm. At the same time, the model calculation cost only increases from 15.8GFLOP to 15.9GFLOP, with a computational performance loss of 0.6%. Using the CA attention mechanism, the accuracy of maritime target recognition tasks can be improved with minimal computational cost, providing precise reference for the fusion of stereo vision maritime target recognition algorithm.

3.4. Experiment on Ship Trajectory Detection Method Fused with Stereo Vision

To validate the effectiveness of our experimental platform, we established a controlled experimental environment in a closed water area. Here, we utilized the platform to measure the distance to a target ship, comparing these measurements with data from a laser rangefinder to ascertain the accuracy of our experimental data. It is important to note that the laser rangefinder provides single-point ranging data, while our stereo vision ranging

algorithm extracts edge feature points within the target recognition range for K-nearest calculation, leading to a systematic difference in results. In these experiments, the average error was found to be 5.73%, as detailed in Table 5. For all experiments, the same trained network was consistently used, and the experimental platform boat remained uniform in type and size. This consistency was maintained to ensure the reliability and comparability of our results across different measurements.

Table 5. Distance measurement results of the stereo vision algorithm.

Experiment ID	Detect Distance/m	Actual Distance/m
1	2.306	2.253
2	4.35	4.6
3	5.3	5.8
4	8.9	10.14
5	11.41	10.654
6	8.2	7.616
7	5.6	5.519
8	3.5	3.467

In the binocular stereo camera-based ranging experiment of the target ship, the standard error in the distance measurement results for the same target was initially 9.59%. This percentage reflects the variability in the distance measurements due to factors such as camera angle, environmental conditions, and inherent limitations of the pre-improvement algorithm. After the maritime positioning algorithm was enhanced, the coefficient of variation in the ranging results was significantly reduced to 3.47%. This reduction indicates a substantial increase in the consistency and reliability of the distance measurements under varying conditions. The detailed experimental setup involved controlled conditions where the target ship's distance and relative angle to the cameras were systematically varied. The data were then rigorously analyzed to assess the impact of these variables on the ranging accuracy. Figure 11 illustrates the relationship between the target's relative camera angle and the target ranging results, both before and after the algorithmic improvements, providing a visual representation of the enhanced accuracy and consistency achieved by the refined algorithm.

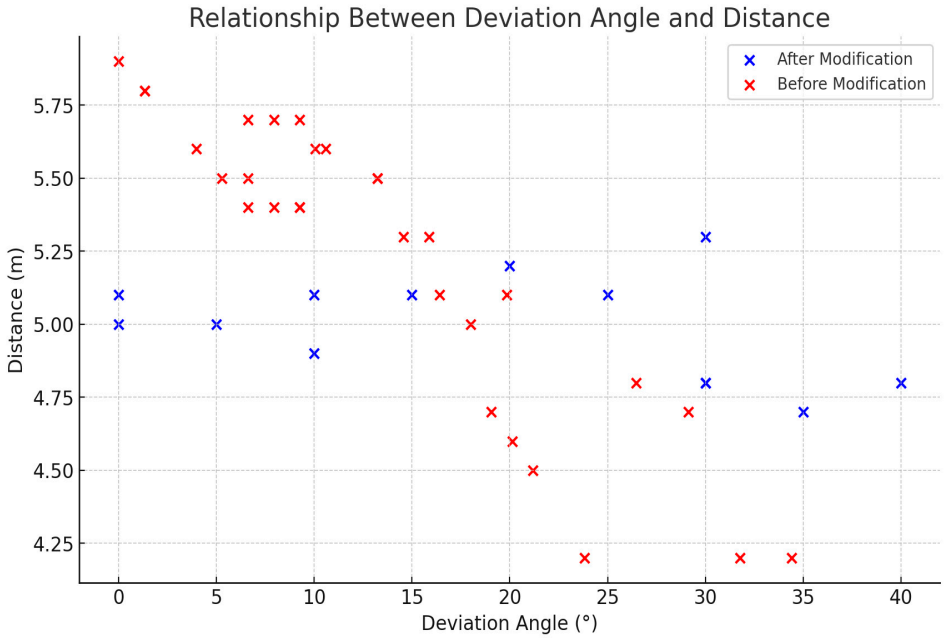

Figure 11. Comparison of target angle and distance measurement results.

It can be seen that the improved ranging results correct the ranging errors caused by the camera's orientation angle, thereby improving the positioning accuracy of maritime targets.

4. Conclusions

This paper proposes an improved maritime target detection and positioning method based on stereo vision. During the training phase, the Seaships ship dataset and maritime target images in the COCO general dataset are used for forward propagation of the ship through deep neural networks, and the weight gradient calculation and backpropagation in the deep neural networks are performed using the annotation information in the dataset, ultimately obtaining the improved maritime target detection algorithm for ships. In the recognition and positioning phase, target recognition is performed through deep neural networks, and spatial positioning of key points in the target area is performed in conjunction with stereo vision to obtain the relative position of the target. The algorithm has undergone several improvements based on the attention mechanism and has been compared with the YOLOv5 algorithm for target detection accuracy. The research results show:

1. The CBAM attention mechanism-improved StereoYOLO algorithm has increased target recognition accuracy compared to the original algorithm while keeping the computational requirements almost unchanged. Among them, adding the attention module after the CSP1_4 layer achieves the highest accuracy improvement for maritime target recognition tasks, reaching 1.71%. In subsequent multiple tests, the improved algorithm achieved a 2.07% increase in mAP50 performance compared to the original YOLOv5;
2. The CA attention-based improved StereoYOLO algorithm, which performs feature pooling operations separately for the h-direction and w-direction, has a higher detection accuracy compared to other attention algorithms, with a mAP50 accuracy improvement of 5.23% compared to the pre-improvement algorithm;
3. Enhancements to the distance measurement algorithm have significantly increased the robustness of positioning accuracy in maritime target recognition tasks, especially under the conditions of a ship's angular oscillations in three dimensions. This improvement becomes evident when detecting the same target at consistent distances, where the refined algorithm has notably reduced the coefficient of variation in data deviation, attributable to the ship's roll, pitch, and yaw movements, from 9.59% to 3.47%;
4. The SGBM feature point matching algorithm used in the algorithm has limitations. The use of stereo transformer and other depth neural network-based multi-view vision algorithms to improve target positioning accuracy will become the focus of subsequent research.

Author Contributions: Conceptualization, W.Y.; Data cruation, Y.S., G.Z. and H.L.; Funding acquisition, W.Y.; Investigation, Y.S., G.Z., H.L. and Y.W.; Methodology, Y.S.; Project administration, W.Y.; Resources, G.Z. and W.Y.; Software, Y.S. and Y.W.; Supervision, W.Y.; Validation, G.Z., H.L. and Y.W.; Visualization, Y.S. and Y.W.; Writing—original draft, Y.S.; Writing—review & editing, W.Y., G.Z., H.L. and Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: The financial support is provided by the National Natural Science Foundation of China (52171308), Key Project of Fujian Provincial Science and Technology Department (2021H0021) Natural Science Foundation of Fujian Province (2022J01333) and National Key Research and Development Program of Ministry of Science and Technology (2021YFB3901500).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Seaships dataset: https://github.com/jiaming-wang/SeaShips/.

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

Nomenclature

AI	Artificial Intelligence
AIS	Automatic Identification System
SVM	Support Vector Machine
NPU	Neural Processing Unit
CNN	Convolutional Neural Network
FPN	Feature Pyramid Network
CBAM	Convolutional Block Attention Module
CA	Coordinate Attention Module
YOLO	You Only Look Once
SGBM	Semi-Global Block Matching
mAP	mean Average Precision
IoU	Intersection over Union
GIoU	Generalized Intersection over Union
TFLOPS	Tera Floating Point Operations Per Second
GFLOP	Giga Floating Point Operations
$\otimes$	tensor inner product

References

1. Sun, S.; Lyu, H.; Dong, C. AIS Aided Marine Radar Target Tracking in a Detection Occluded Environment. *Ocean Eng.* **2023**, *288 Pt 2*, 116133. [CrossRef]
2. Nudd, G.; Nygaard, P. Demonstration of a C.C.D. Image Processor for Two-Dimensional Edge Detection. *Electron. Lett.* **1978**, *14*, 83–85. [CrossRef]
3. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]
4. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
5. Lowe, D. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
6. Dosovitskiy, A.; Brox, T. Inverting Visual Representations with Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4829–4837. [CrossRef]
7. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Yan, J.; Yan, K. A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition. *IEEE Trans. Multimed.* **2016**, *18*, 2528–2536. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1 (NIPS'12), Lake Tahoe, CA, USA, 3–6 December 2012; Curran Associates Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
9. Kalake, L.; Wan, W.; Hou, L. Analysis Based on Recent Deep Learning Approaches Applied in Real-Time Multi-Object Tracking: A Review. *IEEE Access* **2021**, *9*, 32650–32671. [CrossRef]
10. Liu, Z.; Waqas, M.; Yang, J.; Rashid, A.; Han, Z. A Multi-Task CNN for Maritime Target Detection. *IEEE Signal Process. Lett.* **2021**, *28*, 434–438. [CrossRef]
11. Liu, K.; Yu, S.; Liu, S. An Improved InceptionV3 Network for Obscured Ship Classification in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4738–4747. [CrossRef]
12. Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. Planning-Oriented Autonomous Driving. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 17853–17862. [CrossRef]
13. Li, P.; Qin, T. Stereo Vision-Based Semantic 3d Object and Ego-Motion Tracking for Autonomous Driving. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 646–661.
14. Li, P.; Chen, X.; Shen, S. Stereo R-CNN Based 3d Object Detection for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7644–7652.
15. Shi, Y.; Guo, Y.; Mi, Z.; Li, X. Stereo CenterNet-based 3D Object Detection for Autonomous Driving. *Neurocomputing* **2022**, *471*, 219–229. [CrossRef]
16. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
17. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
18. Wang, Y.; Zhang, J.; Kan, M.; Shan, S.; Chen, X. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12275–12284.

19. Li, B.; Xie, X.; Wei, X.; Tang, W. Ship Detection and Classification from Optical Remote Sensing Images: A Survey. *Chin. J. Aeronaut.* **2021**, *34*, 145–163. [CrossRef]
20. Wang, B.; Jiang, P.; Gao, J.; Huo, W.; Yang, Z.; Liao, Y. A Lightweight Few-Shot Marine Object Detection Network for Unmanned Surface Vehicles. *Ocean Eng.* **2023**, *277*, 114329. [CrossRef]
21. Liu, L.; Fu, L.; Zhang, Y.; Ni, W.; Wu, B.; Li, Y.; Shang, C.; Shen, Q. CLFR-Det: Cross-Level Feature Refinement Detector for Tiny-Ship Detection in SAR Images. *Knowl. Based Syst.* **2024**, *284*, 111284. [CrossRef]
22. Lin, C.; Wu, C.; Zhou, H. Multi-Visual Feature Saliency Detection for Sea-Surface Targets through Improved Sea-Sky-Line Detection. *J. Mar. Sci. Eng.* **2020**, *8*, 799. [CrossRef]
23. Patel, K.; Bhatt, C.; Mazzeo, P. Deep Learning-Based Automatic Detection of Ships: An Experimental Study Using Satellite Images. *J. Imaging* **2022**, *8*, 182. [CrossRef] [PubMed]
24. Xiong, B.; Sun, Z.; Wang, J.; Leng, X.; Ji, K. A Lightweight Model for Ship Detection and Recognition in Complex-Scene SAR Images. *Remote Sens.* **2022**, *14*, 6053. [CrossRef]
25. Kizilkaya, S.; Alganci, U.; Sertel, E. VHRShips: An Extensive Benchmark Dataset for Scalable Deep Learning-Based Ship Detection Applications. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 445. [CrossRef]
26. Cheng, S.; Zhu, Y.; Wu, S. Deep Learning Based Efficient Ship Detection from Drone-Captured Images for Maritime Surveillance. *Ocean. Eng.* **2023**, *285 Pt 2*, 115440. [CrossRef]
27. Zhang, Q.; Huang, Y.; Song, R. A Ship Detection Model Based on YOLOX with Lightweight Adaptive Channel Feature Fusion and Sparse Data Augmentation. In Proceedings of the 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Madrid, Spain, 29 November–2 December 2022; pp. 1–8. [CrossRef]
28. Thombre, S.; Zhao, Z.; Ramm-Schmidt, H.; Vallet Garcia, J.M.; Malkamaki, T.; Nikolskiy, S.; Hammarberg, T.; Nuortie, H.; Bhuiyan, M.Z.H.; Sarkka, S.; et al. Sensors and AI Techniques for Situational Awareness in Autonomous Ships: A Review. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 64–83. [CrossRef]
29. Xu, S.; Jiang, Y.; Li, Y.; Liu, S.; Ding, S.; Ma, D.; Qi, H.; Zhang, W. A Stereo Vision Localization Method for Autonomous Recovery of Autonomous Underwater Vehicle. *J. Harbin Eng. Univ.* **2022**, *43*, 1084–1090.
30. He, H.; Wa, N. Monocular Visual Servo-Based Stabilization Control of Underactuated Unmanned Surface Vehicle. *Chin. J. Ship Res.* **2022**, *17*, 166–174.
31. Zhu, S.; Li, C.; Change Loy, C.; Tang, X. Face Alignment by Coarse-to-Fine Shape Searching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4998–5006.
32. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. Shapenet: An Information-Rich 3D Model Repository. *arXiv* **2015**, arXiv:1512.03012.
33. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]
34. Barnell, M.; Raymond, C.; Smiley, S.; Isereau, D.; Brown, D. Ultra Low-Power Deep Learning Applications at the Edge with Jetson Orin AGX Hardware. In Proceedings of the 2022 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 23–27 September 2022; pp. 1–4. [CrossRef]
35. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A Large-Scale Precisely-Annotated Dataset for Ship Detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [CrossRef]