Journal of *Marine Science and Engineering*

MDPI

Article

GT-YOLO: Nearshore Infrared Ship Detection Based on Infrared Images

Yong Wang [1,*], Bairong Wang [1], Lile Huo [1] and Yunsheng Fan [1,2]

[1] Marine Electrical Engineering College, Dalian Maritime University, Dalian 116026, China; wbr1120221285@dlmu.edu.cn (B.W.); hll_123@dlmu.edu.cn (L.H.); yunsheng@dlmu.edu.cn (Y.F.)
[2] Key Laboratory of Technology and System, Intelligent Ships of Liaoning Province, Dalian 116026, China
* Correspondence: wy_521@dlmu.edu.cn

Abstract: Traditional visible light target detection is usually applied in scenes with good visibility, while the advantage of infrared target detection is that it can detect targets at nighttime and in harsh weather, thus being able to be applied to ship detection in complex sea conditions all day long. However, in coastal areas where the density of ships is high and there is a significant difference in target scale, this can lead to missed detection of some dense and small targets. To address this issue, this paper proposes an improved detection model based on YOLOv5s. Firstly, this article designs a feature fusion module based on a fusion attention mechanism to enhance the feature fusion of the network and introduces SPD-Conv to improve the detection accuracy of small targets and low-resolution images. Secondly, by introducing Soft-NMS, the detection accuracy is improved while also addressing the issue of missed detections in dense occlusion situations. Finally, the improved algorithm in this article increased $mAP_{0.5}$ by 1%, $mAP_{0.75}$ by 5.7%, and $mAP_{0.5:0.95}$ by 5% on the infrared ship dataset. A large number of comparative experiments have shown that the improved algorithm in this article is effective at improving detection capabilities.

Keywords: object detection; infrared ship; yolov5

Citation: Wang, Y.; Wang, B.; Huo, L.; Fan, Y. GT-YOLO: Nearshore Infrared Ship Detection Based on Infrared Images. *J. Mar. Sci. Eng.* **2024**, *12*, 213. https://doi.org/10.3390/jmse12020213

Academic Editor: Marco Cococcioni

Received: 8 December 2023
Revised: 24 December 2023
Accepted: 22 January 2024
Published: 25 January 2024

1. Introduction

With the advent of the information age, the rapid development of computers and related cutting-edge technologies has driven breakthroughs in machine vision technology. Machine vision, as a fusion technology that uses computers as tools and combines image processing with sensors, has emerged in multiple research directions such as object detection, semantic segmentation, motion tracking, 3D reconstruction, and action recognition. In many practical application scenarios, object detection has important application value, covering multiple fields such as video surveillance, autonomous driving, unmanned ship detection, and navigation [1]. The world has abundant ocean and inland river resources, so research on water target detection is increasingly receiving attention [2], especially in the field of ship detection at sea. Initially, ship detection mainly relied on visible light technology, but with the continuous development of infrared thermal imaging technology, infrared-based ship detection has gradually become a research hotspot. Compared to visible light, infrared detection technology has significant characteristics such as all-weather detection, long detection distance, and strong anti-interference ability, which enable it to effectively detect ships in navigation at night and in adverse weather conditions. Compared to other detection tasks, the nearshore ship detection task based on infrared images has its distinct characteristics:

(1) Based on its physical properties, the resolution of infrared images is lower compared to visible light images. This results in limited information for small target ships, and during the feature extraction process, crucial details are easily lost, thereby affecting the detection.

(2) From an ocean perspective, ships of the same category exhibit different scales in the images, which can easily result in the loss of small targets. This necessitates that detection algorithms have a stronger capability for multi-scale target detection.
(3) Coastal areas and ports often harbor a substantial number of ships. The occurrence of mutual occlusion between ships poses a significant challenge for the accurate positioning and classification of ships in nearshore environments.

Deep learning-based object detection methods have gradually become mainstream, giving rise to numerous algorithm frameworks. These encompass two-stage algorithms based on anchor boxes (such as RCNN [3] and Faster RCNN [4]), single-stage detection algorithms (such as YOLO [5] and SSD [6]), and detection algorithms without anchor boxes (such as CenterNet [7] and FCOS [8]). With continuous and in-depth research in detection technology, the focus of detection tasks has shifted from large and medium-sized targets to enhancing the detection performance of small targets while preserving the effectiveness of larger ones.

In response to this challenge, researchers have introduced various methods, including multi-scale feature fusion and context learning. These approaches aim to further refine the detection performance of small targets without compromising the accuracy of larger and medium-sized targets. This trend underscores the pursuit of a more comprehensive performance in object detection algorithms, enabling them to excel in handling diverse target sizes and scenes.

Shi et al. [9] proposed a strategy for fusing deep and shallow features to enhance detection probability. This strategy captures low-level structures and texture features of small targets, along with high-level semantic information, to prevent missed detections. Ye et al. [10] added a small object detection layer to the original network model to solve the problem of small object detection and adopted a new connection method based on the introduced BIFPN to address the issue caused by multi-scale changes in ships. With the continuous research on and development of attention mechanisms, they have been widely applied in various fields of deep learning in recent years, including object detection. Si et al. [11] proposed an improved YOLO-RSSD algorithm, where an enhanced bidirectional feature pyramid network structure is embedded in the feature fusion section. This enables cross-layer multi-scale weighted feature fusion, and a channel attention mechanism is introduced in the convolutional units to further enhance the detection effectiveness of small ship targets in infrared images. Guo et al. [12] proposed a nearshore ship detection method based on the FCOS network, incorporating a bidirectional attention feature pyramid network. This approach enhances the detection accuracy of small targets. Although the current improved methods have shown some enhancement in the detection of small targets, there is still room for further improvement.

Wang et al. [13] designed a CNeB2 module to enhance the spatial correlation in encoding, reducing interference from redundant information and improving the model's capability to recognize dense targets. With the increasing improvement in feature extraction and fusion in network models, some scholars have made enhancements in the post-processing stage of network models. Shi et al. [14] proposed an improved YOLOv5s_SE to address the issue of insufficient performance of existing algorithms in detecting small targets. It is achieved by integrating Soft_NMS and EIOU_Loss, replacing the non-maximum suppression function (NMS) in the original network, and thereby improving the detection ability of occluded objects.

Infrared ships near the coast exhibit distinctive characteristics. Firstly, due to equipment limitations, infrared images have lower resolution, appear more blurred compared to visible light, and are often accompanied by significant noise. Secondly, the coastal perspective of the infrared images, facing the ocean, results in a broader field of view for detection scenes. This leads to uneven distribution of target sizes, and in coastal areas, the density of ships is relatively high, making occlusion incidents more likely.

Considering these characteristics of infrared ship detection near the coast, and through a comparison of various network models, a GT_YOLO infrared ship detection algorithm

based on YOLOv5s is proposed. YOLO, as a single-stage detection algorithm, has significant advantages in terms of speed, accuracy, and deployment, making it more adaptable to diverse and complex scenes than other detection algorithms. This algorithm not only performs well in small target detection but also effectively addresses the challenges introduced by dense scenes. Furthermore, the $mAP_{0.5}$ of infrared ships has been improved by 1%, $mAP_{0.75}$ by 5.7%, and $mAP_{0.5:0.95}$ by 5%. The main contributions of the research can be summarized as follows.

To address prominent challenges in the detection of infrared ships near the coast, this paper proposes a GT_YOLO algorithm for infrared ship detection based on YOLOv5s. The algorithm introduces an attention mechanism to allow the network model to focus more on crucial features to improve performance in specific scenarios near the coast.

(1) To capture distant contextual information of the targets, this paper introduces a feature fusion module based on a fused attention mechanism. This module enhances feature fusion while suppressing noise introduced by shallow feature layers.
(2) To suppress the unclear detail information caused by low resolution and its impact on small object detection, the SPD-Conv module was introduced to improve the detection accuracy of small objects.
(3) To address the issue of dense occlusion, which is prone to occur near the coast, this paper introduces Soft-NMS to ensure that the detection model still has excellent detection performance in dense occlusion scenes.

2. GT_YOLO

2.1. YOLOv5

YOLO [3] (you only look once) is a classic single-stage detection algorithm, and in this paper, the mature version YOLOv5 from the YOLO series is selected. The network structure is divided into three parts: the backbone, neck, and head. The backbone uses the BottleneckCSP structure for feature extraction, the neck introduces a feature pyramid network (FPN) and PANet to enhance feature fusion and extraction, and the head classifies and predicts targets based on the learned features. YOLOv5 achieves rapid and accurate detection of targets in various scenarios through the collaborative work of these three parts, demonstrating outstanding performance.

YOLOv5 adopts the input method of mosaic data augmentation, synthesizing new images through random cropping, scaling, and composition methods. This strategy not only enriches the experimental data but also helps improve the inference speed. In terms of loss calculation, YOLOv5 utilizes CIOU_Loss, which aids in more accurately assessing the error of the target detection results. Finally, the output is optimized through NMS to obtain the final detection results. This entire set of data augmentation and loss calculation mechanisms allows YOLOv5 to achieve more robust target detection in complex scenarios and produce more accurate results. The YOLOv5 network architecture is illustrated in Figure 1.

2.2. Feature Fusion Enhancement

In the original YOLOv5 network, the FPN structure is employed, but there are some issues with the simple fusion of shallow-level features with other multi-scale features in the channel dimension. This approach struggles to accurately reflect the importance of different channel features and may lead to the diffusion of noise throughout the entire network model, impacting the fusion effectiveness. To address this problem, this paper introduces a feature fusion module GT based on a fusion-style attention mechanism, as illustrated in Figure 2.

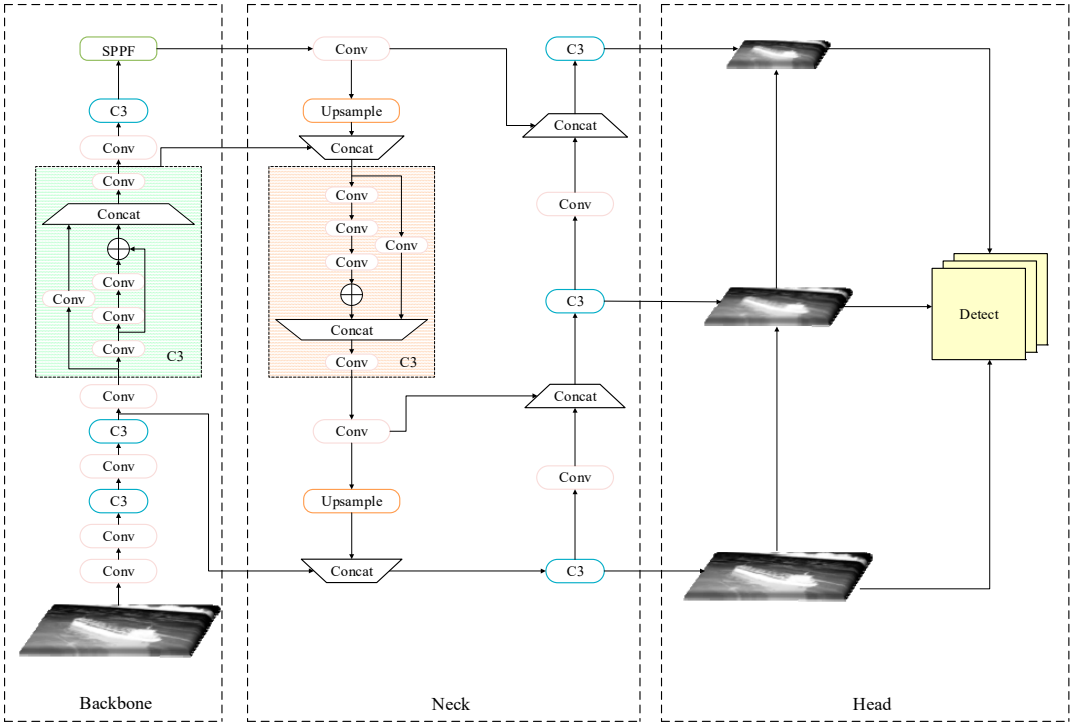


Figure 1. YOLOv5 network structure diagram.

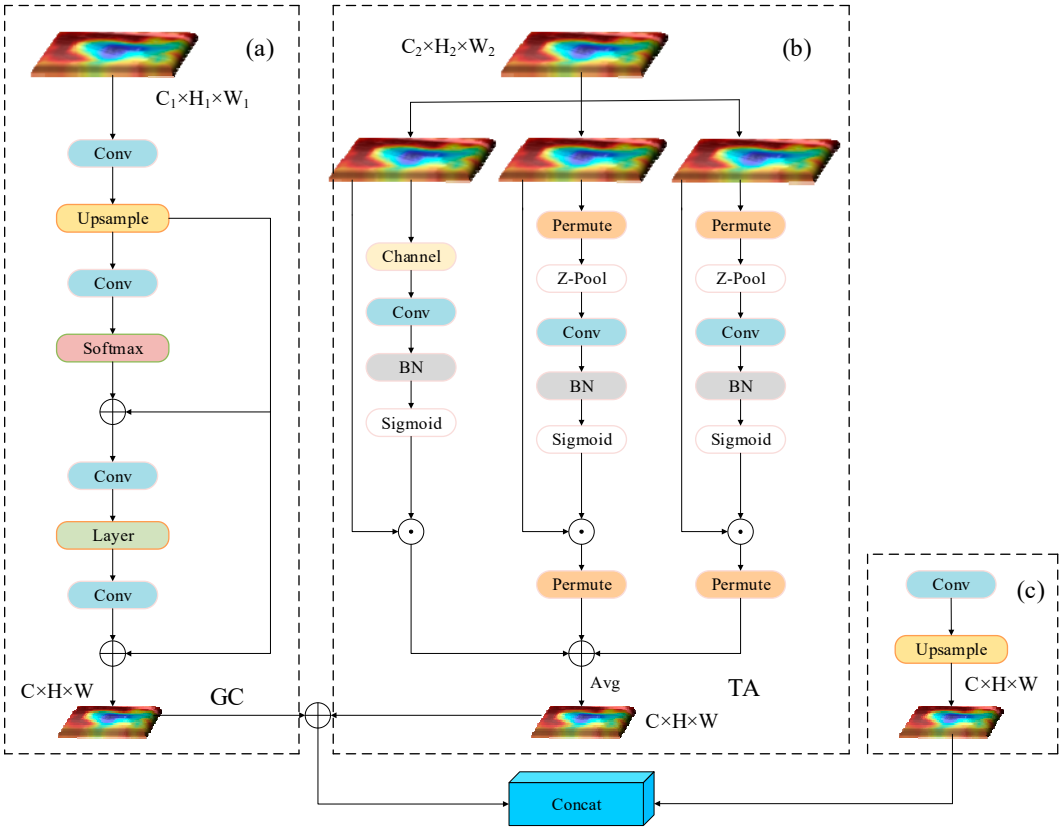


Figure 2. GT feature fusion module. (**a**) High-level features utilize the GCNet attention mechanism; (**b**) detailed features employ the triplet attention mechanism; (**c**) feature information propagated by the FPN.

This module effectively enhances the capabilities of global context modeling and local feature extraction, not only providing better attention to small targets but also suppressing the propagation of noise. The features of the GT module come from the high-level features of the backbone network, features at the same level, and features from the previous layer, generating new features through clever fusion. This feature fusion module based on the fusion-style attention mechanism contributes to improving the network's perception of global and local information, thereby more effectively addressing the challenges in small target detection.

For the high-level features of the backbone network, this paper employs the GCNet [15] for effective global context modeling, capturing long-range dependencies while extracting a global understanding of the visual scene. The attention mechanism module integrates the advantages of simplified non-local (SNL) blocks and SENet, providing not only modeling capabilities for long-range dependencies but also lightweight characteristics. The input feature tensor $C_1 \times H_1 \times W_1$ undergoes processing through SNL blocks and SE blocks, transforming into $C \times H \times W$, thereby capturing more useful features.

For features at the same level, the module adopts triplet attention [16] to make full use of spatial and channel information in the features. This attention mechanism consists of three branches, each processing the input feature tensor $C_2 \times H_2 \times W_2$. The first branch is the channel attention branch, where the input undergoes pooling and a 7×7 convolution, and finally, a Sigmoid function generates spatial attention weights. The second branch establishes interaction between channels C_2 and space H_2 by rotating the feature tensor along the W_2 dimension, transforming it into $H_2 \times C_2 \times W_2$. After processing with Z-Pool and a 7×7 convolution, it undergoes a Sigmoid function to generate attention weights for the spatial dimension H_2 and channel dimension C_2. The third branch is similar to the second branch, establishing interaction between channels C_2 and space W_2, generating attention weights between channel dimension C_2 and spatial dimension W_2. Finally, the outputs from the three branches are added and averaged to transform into $C \times H \times W$. This module, through the triplet attention mechanism, not only utilizes the rich features from the shallow layers more comprehensively but also suppresses a significant amount of noise within them. This enables the network to better focus on infrared ship targets, enhancing the performance of infrared ship detection.

The feature fusion module designed in this paper processes the high-level and low-level features transmitted by the backbone network through attention mechanisms. The fusion is performed by addition, and the resulting features are combined with the features propagated by the FPN, ultimately producing the fused feature map.

2.3. SPD-Conv

Due to hardware limitations, infrared target images exhibit lower resolution and pixel blurring compared to visible light images, making it challenging for networks to extract detailed features, especially in the case of small infrared targets (less than 32×32 pixels) that may be overlooked in multi-scale object detection scenarios. To address this issue, researchers have designed BIFPN based on the FPN+PAN structure, introducing a shorter path to improve feature fusion. On the other hand, some scholars have introduced the scale-invariant subspace (SAN) [17] to map multi-scale features, aiming to enhance detection performance. However, these improvement methods often involve stride convolutions and maximum pooling, which may lead to information loss in infrared target detection, particularly being unfriendly to small targets.

To address this issue, this article introduces SPD-Conv [18] into the YOLOv5 network architecture. By replacing the stride convolution and pooling layers of the original network model, the detection performance for small targets and low-resolution images has been improved. Assuming a feature map U with a size of $W \times W \times V_1$, the feature map is divided into several sub-series feature maps at each stride:

$$f_{0,0} = U[0:W:X, 0:W:X], f_{1,0} = U[1:W:X, 0:W:X], \ldots, f_{X-1,0} = U[X-1:W:X, 0:W:X]$$
$$f_{0,1} = U[0:W:X, 1:W:X], f_{1,1} = U[1:W:X, 1:W:X], \ldots, f_{X-1,1} = U[X-1:W:X, 1:W:X]$$
$$f_{0,1} = U[0:W:X, 1:W:X], f_{1,1} = U[1:W:X, 1:W:X], \ldots, f_{X-1,1} = U[X-1:W:X, 1:W:X]$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$f_{0,X-1} = U[0:W:X, X-1:W:X], f_{1,X-1}, f_{1,X-2}, \ldots, f_{X-1,X-1} = U[X-1:W:X, X-1:W:X]$$

Each sub-mapping feature $f_{x,y}$ of the feature map U is composed of feature vectors $U_{(i,j)}$, where $x+i$ and $y+j$ can be proportionally divided. In this way, each sub-mapping is downsampled from U according to the scaling factor. When the scaling factor X = 2, as shown in Figure 3, four sub-mapping feature maps can be obtained.

$$f_{0,0} = U[0:W:X, 0:W:X], f_{1,0} = U[1:W:X, 0:W:X]$$
$$f_{0,1} = U[0:W:X, 1:W:X], f_{1,1} = U[1:W:X, 1:W:X]$$

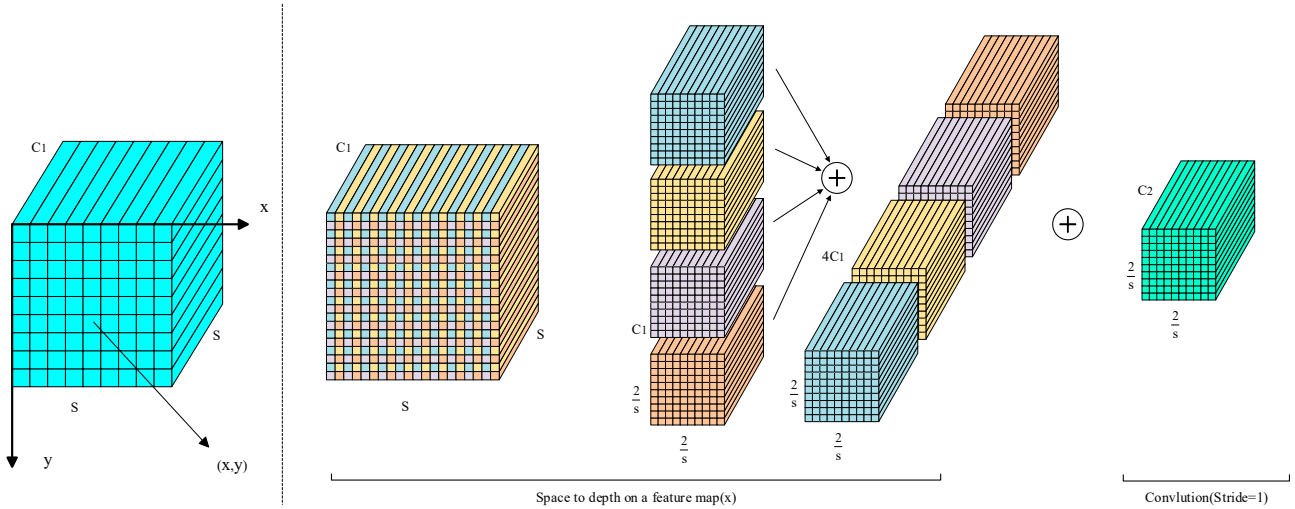


Figure 3. SPD-Conv with X = 2.

The size of each sub-feature map is $\left(\frac{W}{2}, \frac{W}{2}, X_1\right)$, with a downsampling factor of 2.

Then, the sub-feature maps are merged along the channel dimension to form a new feature map U'. It undergoes downsizing in the spatial domain by a scaling factor of X and an expansion in the channel dimension by X^2. In other words, the original feature map $U(W \times W \times V_1)$ is transformed into $U'\left(\frac{W}{X}, \frac{W}{X}, X^2 V_1\right)$ through this process.

After the SPD module, non-strided convolution filters are applied for further transformation, resulting in $U'\left(\frac{W}{X}, \frac{W}{X}, X^2 V_1\right) \rightarrow U''\left(\frac{W}{X}, \frac{W}{X}, V_2\right)$. The use of non-strided convolution helps better preserve feature information. Otherwise, when using filters with odd strides, such as a stride of 3, the feature map will be downsized proportionally, but each pixel will only be sampled once. If the filter has an even stride, such as a stride of 2, it leads to uneven sampling, causing inconsistency in the sampling between even and odd rows (columns).

2.4. Soft-NMS

In detection tasks, the same object may be detected multiple times, resulting in the generation of numerous overlapping candidate boxes. To address this issue, NMS is an effective method. The operation of NMS involves generating all candidate boxes in each round and organizing them in descending order based on their confidence scores, with

higher scoring candidates placed at the front. In each round, the candidate box with the highest confidence score is selected, and attention is focused on the highly overlapping portions with all remaining candidate boxes. These highly overlapping boxes are suppressed in that round, while the selected candidate box is retained and not considered in the next round. By repeating this operation, the final result is the candidate box with the highest confidence score, suppressing highly overlapping candidate boxes.

However, NMS may encounter issues of missed detections when dealing with dense situations of infrared ships near the coast. This is because NMS directly excludes candidate boxes when their overlap exceeds a certain threshold, and this more aggressive approach may lead to the incorrect exclusion of some important targets. Faced with this challenge, there is a need to seek a more flexible and adaptive approach to enhance the robustness of detection.

This paper addresses this issue by introducing Soft-NMS [19], which significantly improves the algorithm's performance in dense scenarios of infrared ships. Soft-NMS still follows the idea of NMS, suppressing candidate boxes that overlap with the highest scoring candidate box. However, for densely occluded scenes, Soft-NMS gradually shifts its focus to those candidate boxes with greater overlap, attenuating their scores more. The evolved pruning step rules are as follows:

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ 0, & iou(M, b_i) \geq N_t \end{cases} \rightarrow s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)), & iou(M, b_i) \geq N_t \end{cases} \quad (1)$$

Here, M represents the candidate box with the highest score, b_i and s_i represent the i-th candidate box and its score, and N_t represents the threshold.

Based on the rules mentioned above, the function adjusts the overlap above the threshold with a linear decay function relative to M. In this way, candidate boxes that are farther from M will not be affected, while those closer will receive a larger penalty. However, the overlap is not continuous, and when the set threshold is reached, a penalty is suddenly applied. The ideal scenario is a continuous penalty function that imposes no penalty when there is no overlap and a very heavy penalty when the overlap is high. At the same time, M should not affect the scores of boxes with low overlap, and the penalty should gradually increase when the overlap is low. Building on this idea, Soft-NMS introduces a Gaussian penalty function as follows:

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i e^{-\frac{iou(M, b_i)^2}{\sigma}}, & iou(M, b_i) \geq N_t \end{cases} \quad (2)$$

Here, σ is a hyperparameter. In this way, when the overlap between two boxes is high, the score will be smaller. Compared to traditional NMS, Soft-NMS assigns a very small score instead of directly removing the box. This can significantly improve the detection performance of infrared ships in dense occlusion situations.

This paper presents an improved GT-YOLO based on YOLOv5s, as shown in Figure 4. By incorporating the SPD-Conv module and the designed feature fusion module, the model not only achieves excellent detection results for small infrared ships but also enhances the fusion of multi-scale ships. It performs well in detecting multi-scale targets in the same scene. Finally, Soft-NMS is applied at the end of the network to address the issue of dense occlusion in infrared ships near the coast. The integration of these improvement points greatly enhances the detection performance of the network.

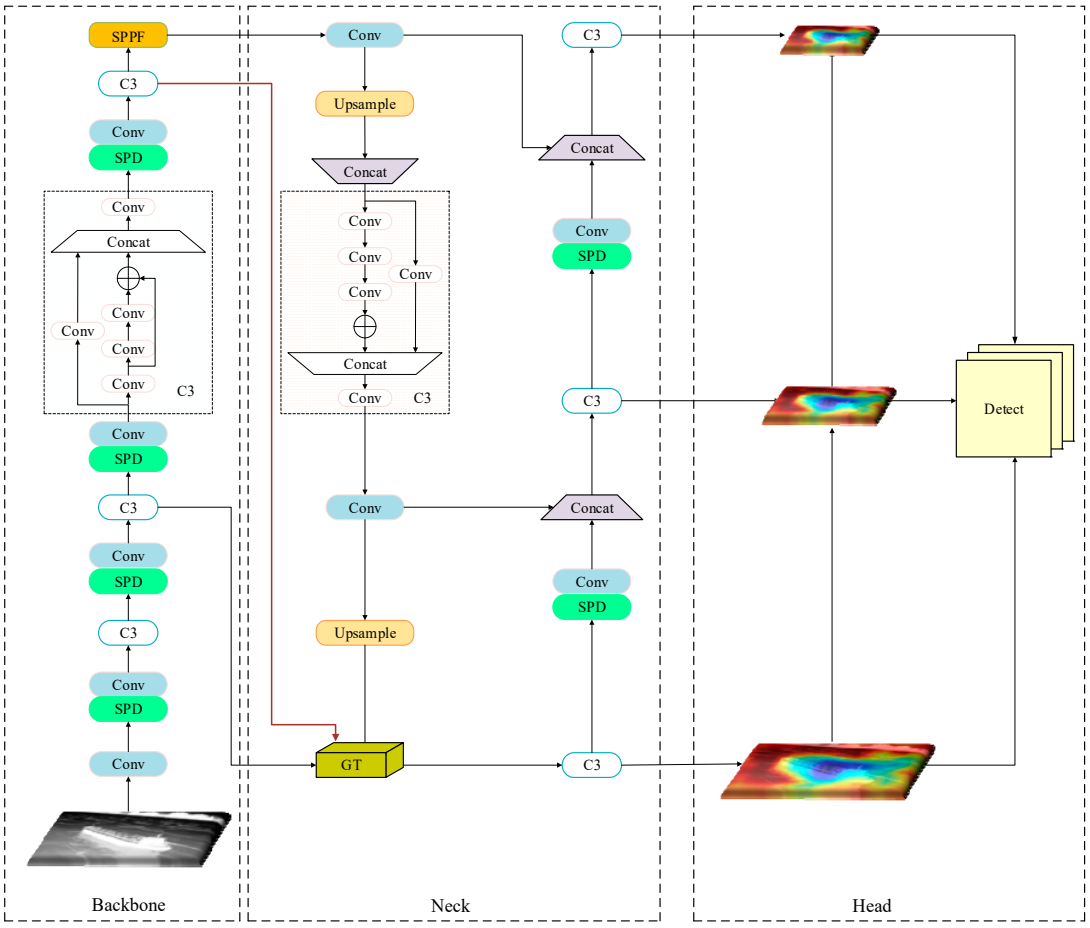


Figure 4. The network diagram of GT-YOLO.

3. Experiments

This section introduces the experimental dataset, relevant details, and evaluation metrics. Additionally, a substantial number of experiments are conducted to demonstrate the effectiveness of the proposed improved algorithm in this paper. To ensure the fairness of the experiments, no pre-trained weights are used, and parameters are kept consistent across all experiments.

3.1. Experimental Settings

In the experiments of this paper, the operating system used is Windows 10, with an Intel i7-13700k CPU (Santa Clara, CA, USA) and an NVIDIA RTX3080 (10G) GPU (Santa Clara, CA, USA). The experiments are conducted using the PyTorch framework (version 2.0.1), and the parameter settings are based on the default parameters of YOLOv5, including adaptive anchor boxes and mosaic data augmentation. The overall training parameter settings for the experiments are shown in Table 1.

Table 1. Training parameters.

Name	Configuration
Learning rate	0.01
Momentum	0.937
Data enhancement	MOSAIC
Epochs	150
Batch size	8

3.2. Infrared Ship Dataset

The dataset used in this study is obtained from the InfiRay Infrared Open Platform [20], which is an open dataset dedicated to infrared ships. The creation of an infrared dataset requires accurate camera calibration and proper positioning to obtain high-quality and precise infrared images [21]. This dataset is generated through on-site testing and adjustments, employing a fixed position and stable lighting conditions within the same scene. These measures are implemented to enhance the effectiveness of object detection. The dataset consists of 9402 images of various types of infrared ships along with their corresponding labels. The details of the dataset are shown in Figure 5. The dataset comprises 9402 images of different types of infrared ships, each accompanied by its corresponding labels. The infrared ship categories in the dataset include seven types: liner, bulk carrier, warship, sailboat, canoe, container ship, and fishing boat. Many images in the dataset have low resolutions, with a high count of 4818 images at 384 × 288 resolution and 3472 images at 1280 × 1024 resolution. The resolutions of the remaining images are also all below 640 × 640. During model training, a fixed input size of 640 × 640 will be used. The dataset predominantly captures scenes in coastal ports and docks, where the presence of sailboats and canoes is common on docks, while fishing boats are prevalent in coastal port areas. The dataset contains a significant number of small objects, with fishing boats and canoes being the most frequently occurring types. This also adds a certain level of difficulty to the detection task. In this study, to enhance the reliability of the dataset, corrections were made to some of the labels with errors.

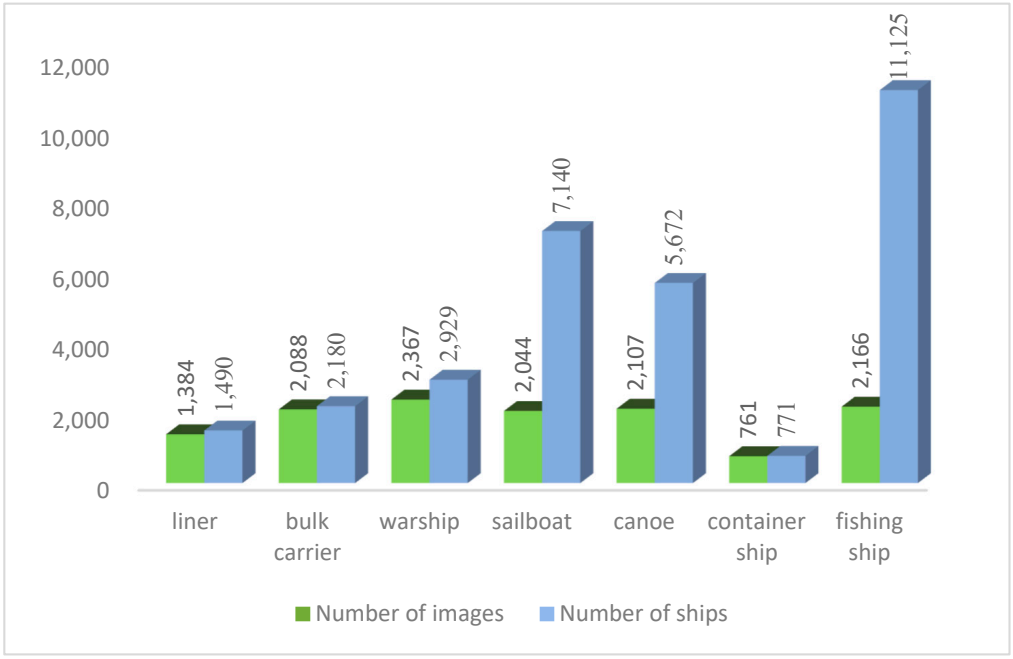


Figure 5. Dataset overview.

3.3. Evaluation Metrics

In this article, for a more comprehensive evaluation of the model's performance, classic metrics such as precision (P), recall (R), $mAP_{0.5}$ (average precision), $mAP_{0.75}$ and $mAP_{0.50:0.95}$, GFLOPS, parameters, and FPS are evaluated. The formulas for these metrics are as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$AP = \int_0^1 PdR \tag{5}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{6}$$

where TP, FP, and FN represent true positive samples, false positive samples, and false negative samples, respectively. mAP is a metric that measures the accuracy of object detection, and a higher mAP value indicates a better detection performance of the model. GFLOPS refers to the number of floating-point operations executed per second and serves as a metric for measuring the computational performance of a model. Parameters refers to the weights and biases of the model, determining the computational complexity of the model. FPS refers to the number of frames a model can process per second when handling images, serving as a crucial metric for measuring real-time performance and processing speed of the model.

3.4. Results and Discussion

This section discusses the impact of the feature fusion module on detection performance. Through a large number of experiments and comparisons with the latest experimental methods, it is demonstrated that the improvements proposed in this paper exhibit excellent detection performance.

3.4.1. The Impact of the Feature Fusion Module

The attention fusion module in this paper selects GCNet over other attention mechanisms for integration. Due to the imaging characteristics of infrared thermal images, infrared images possess a global nature, and utilizing GCNet allows for better capturing of global contextual information. The model can gain a more comprehensive understanding of the overall infrared scene, thereby enhancing the accuracy of ship detection. GCNet incorporates bottleneck transformations to reduce redundancy in global contextual features, resulting in a slight increase in computational complexity during ship detection tasks.

To demonstrate the effectiveness of the feature fusion module designed in this paper, various attention mechanisms were employed for fusion. The same parameter settings were used under this dataset, and no pre-trained weights were utilized to ensure experimental fairness. Considering factors such as model parameters, GFLOPS, FPS, and mAP, comparative experiments were conducted by fusing SENet [22], GAM Attention [23], and SKNet [24] with the GCNet [15]. The attention fusion module proposed in this paper exhibits a notable improvement in detection average precision, with a limited increase in parameters and computational load. The results are presented in Table 2. A visualization of the results is presented in Figure 6.

Table 2. Experimental results of the feature fusion module.

Model	$mAP_{0.5}$ (%)	$mAP_{0.75}$ (%)	$mAP_{0.5:0.95}$ (%)	GFLOPS	Parameters (M)	FPS
YOLOv5s	0.944	0.777	0.695	15.8	7	244
5s + GC + SE	0.947	0.774	0.695	17.0	7.2	243
5s + GC + GAM	0.948	0.779	0.699	18.4	7.3	217
5s + GC + SK	0.947	0.776	0.698	21.4	7.5	210
5s + GT	0.948	0.783	0.704	17.1	7.2	238

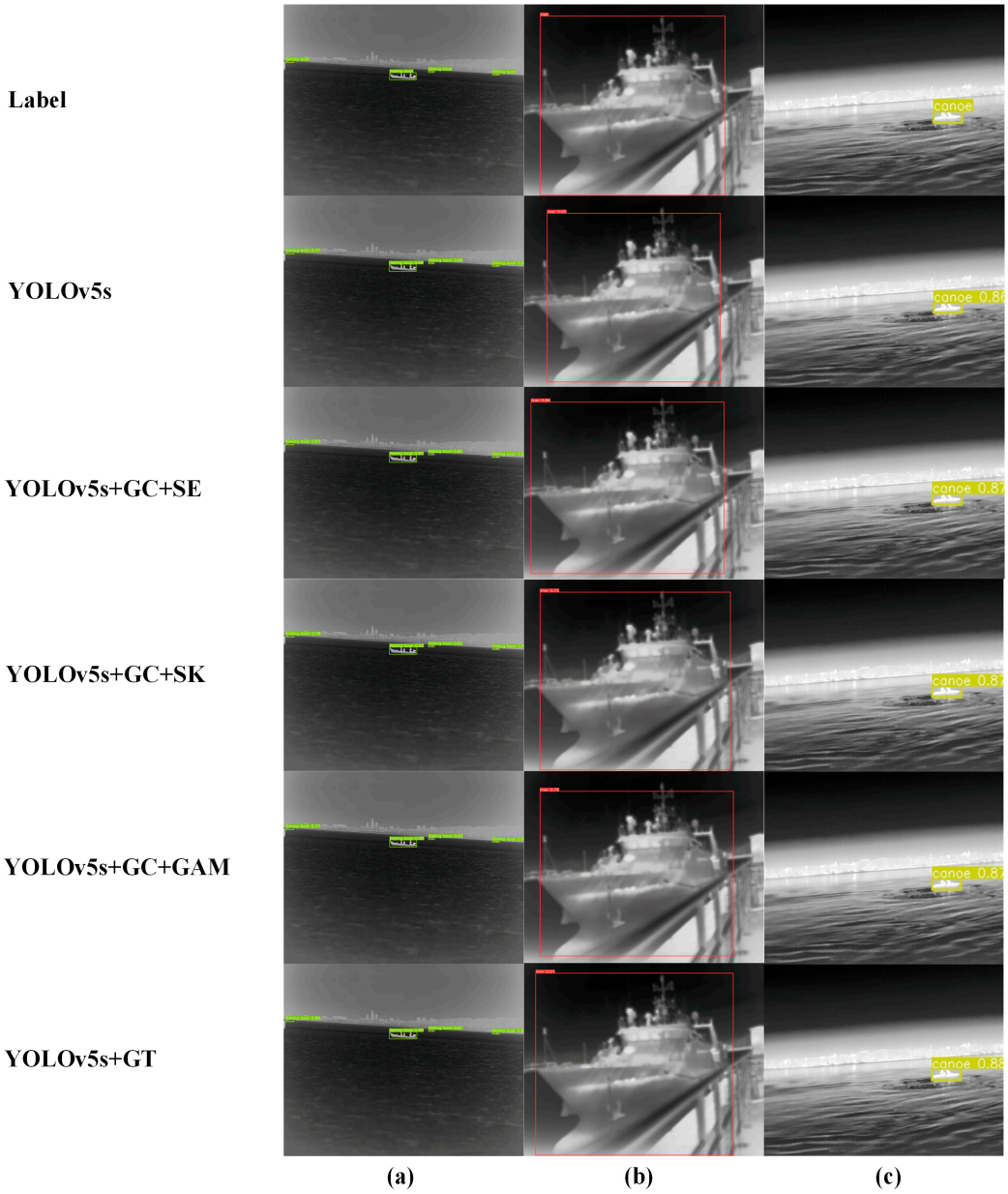


Figure 6. The experimental results of the feature fusion module. (**a**) The test result of the fishing boat; (**b**) The detection result of liner; (**c**) The detection result of canoe.

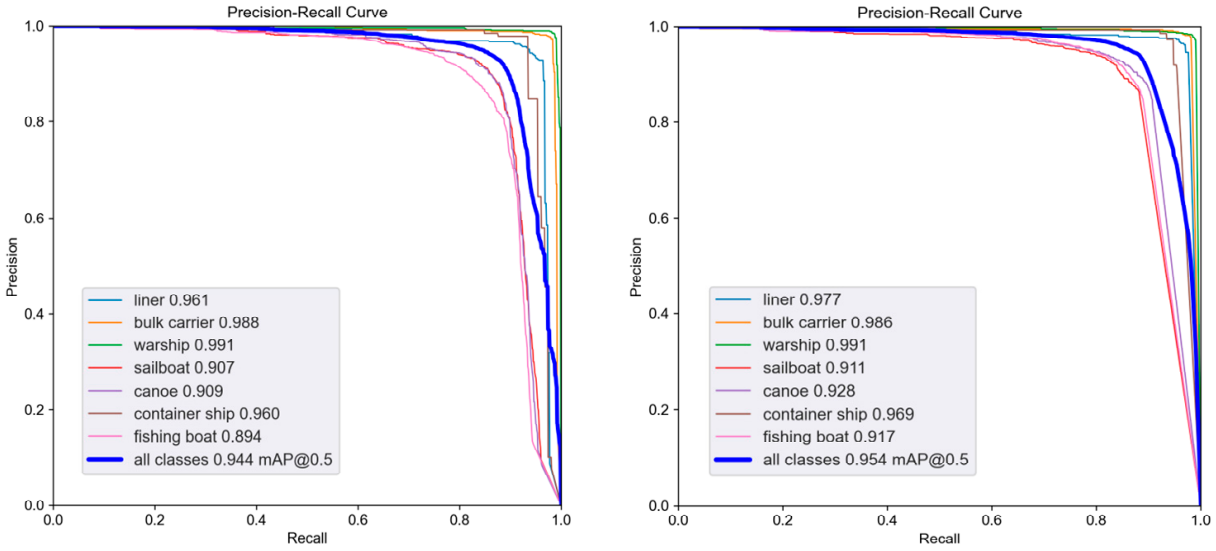
3.4.2. Ablation Experiment

The proposed GT-YOLO model in this paper enhances the detection performance for each category in the dataset. It not only demonstrates notable improvements for small targets within multi-scale objectives but also addresses detection challenges more effectively in situations involving dense occlusion. In comparison to the original baseline algorithm, the improved model in this paper shows a 5% increase in mAP, with an additional 1.7 M parameters, as shown in Table 3.

The detection results of the trained model are illustrated in the figure. The "Label" indicates annotated boxes from the dataset. From top to bottom, on the left side, we have the original model and various improved models (a–c), respectively. The detection results obtained by different models are displayed. It can be observed from the figure that the proposed module in this paper exhibits favorable performance in detecting objects on multiple scales and small-sized targets. The proposed algorithm exhibits higher accuracy, as illustrated in Figure 7.

Table 3. The experimental results of different improvements.

Model	$mAP_{0.75}$ (%)	$mAP_{0.5:0.95}$ (%)	GFLOPS	Parameters (M)
YOLOv5s	77.7	69.5	15.8	7
YOLOv5s + GT	78.3	70.4	17.1	7.2
YOLOv5s + GT + SPD	79.2	71.3	34.4	8.7
YOLOv5s + GT + SPD + Soft-NMS	83.4	74.5	34.4	8.7



(**a**) (**b**)

Figure 7. PR curve. (**a**) YOLOv5; (**b**) GT-YOLO.

The average precision for each category of YOLOv5s and GT-YOLO is presented in Table 4. Among them, sailboat, canoe, and fishing boat have relatively higher quantities and pose a challenge due to the presence of numerous small targets. Additionally, the detection difficulty is increased as sailboats and canoes are often situated in dock scenes with complex backgrounds, introducing significant interference. Consequently, these three types exhibit lower average precision compared to other ship categories. The initial three types have lower resolutions, mostly at 384 × 288, but with fewer ships in the images and most targets being larger, which are relatively easier to detect.

Table 4. The comparative results of average precision for each category.

Class	YOLOv5s		GT-YOLO	
	$AP_{0.75}$ (%)	$AP_{0.5:0.95}$ (%)	$AP_{0.75}$ (%)	$AP_{0.5:0.95}$ (%)
Liner	78.3	68	85.8	74.4
Bulk carrier	94.9	82	96.4	85.7
Warship	95.7	85.2	96.9	87.5
Sailboat	59.8	56.7	69	62.6
Canoe	67.3	59.9	75.8	65.4
Container ship	91.6	80.7	93.1	84.6
Fishing boat	56.5	54.3	66.6	60.8

In the improvement process of GT-YOLO, efforts were made not only to enhance the detection performance of small targets but also to address the issue of low resolution in images within the dataset, which impacts detection accuracy. Additionally, improvements were implemented to alleviate densely occluded situations between targets. Through

experiments, GT-YOLO has achieved a significant improvement in detection accuracy for each category.

3.4.3. Comparison with Other Algorithms

This paper compares the proposed GT-YOLO with other algorithms in the same series without using pre-trained weights. To demonstrate the performance of the improved algorithm, it is compared with YOLOv5m, showing better detection results with fewer parameters and computations. Additionally, a comparison is made with the latest YOLOv8s, demonstrating superior detection performance with a relatively smaller number of parameters and a modest increase in computational load. The results are shown in Table 5.

Table 5. Comparison with algorithms in the same series.

Model	$mAP_{0.5}$ (%)	$mAP_{0.5:0.95}$ (%)	GFLOPS	Parameters (M)
YOLOv5s	94.4	69.5	15.8	7.0
YOLOv5m	95.2	71.1	47.9	20.9
YOLOv7-Tiny [25]	92.9	65.6	13.2	6.0
YOLOv8s	93.2	71	28.5	11.1
GT-YOLO	95.4	74.5	34.4	8.7

To further validate the effectiveness of the proposed algorithm, it is compared with SSD, CenterNet, and EfficientDet-D0 [26] on the same dataset. The results are shown in Table 6. The first two algorithms use Resnet50 as the backbone network for feature extraction. CenterNet, however, differs from other algorithms in that it does not require anchor boxes to predict the bounding boxes of targets. Instead, it directly predicts the center point of the target and generates the target box through regression, eliminating the need for anchor boxes.

Table 6. Comparison with other algorithms.

Model	$mAP_{0.5}$ (%)	$mAP_{0.5:0.95}$ (%)	GFLOPS	Parameters (M)	FPS
SSD	78	46.6	39	12.50	149
EfficientDet-D0	62.2	39.5	4.8	3.83	53
CenterNet	78.5	48.6	70	32	125
GT-YOLO	95.4	74.5	34.4	8.7	152

From the results, it can be observed that the EfficientDet-D0 algorithm tends to lose target information, especially in the detection of small targets and dense scenes. In some images that are blurry or have lower resolutions, both SSD and EfficientDet-D0 may lead to the loss of some targets. Moreover, in dense scenes, recognition errors may occur. In comparison, GT-YOLO demonstrates not only accurate localization and recognition across various scenarios but also exhibits higher detection accuracy than other algorithms, as shown in Figure 8.

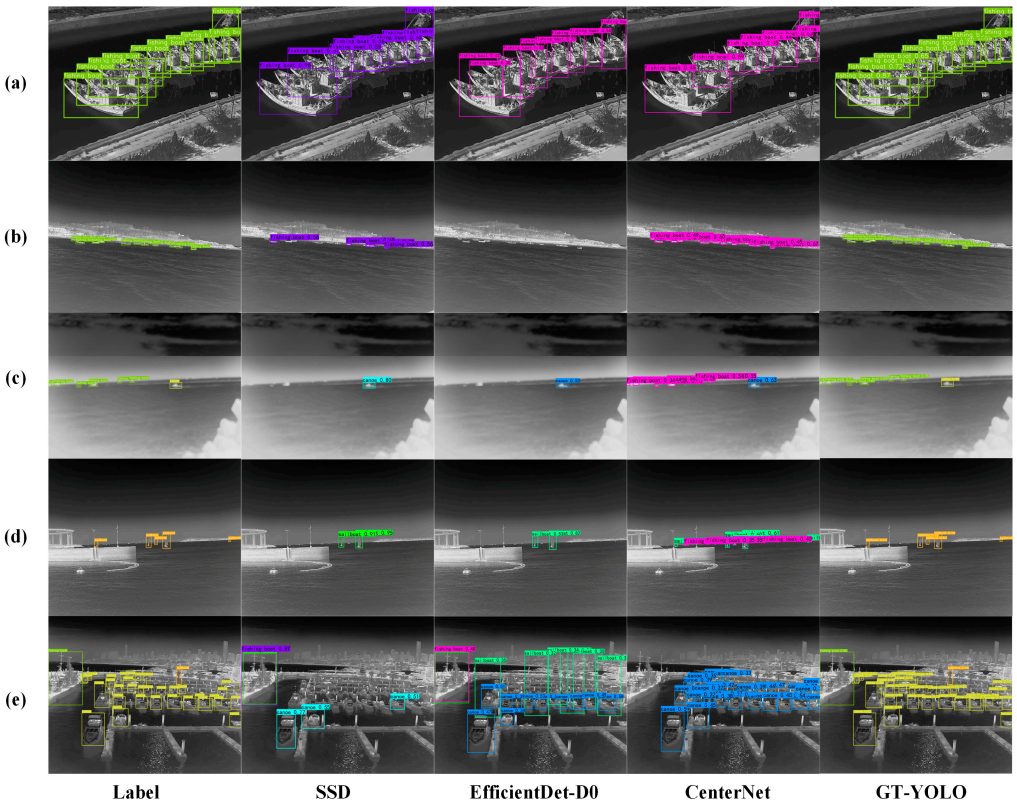


Figure 8. Comparative experiment images; Label represents the original labels; (**a**–**e**) depict different scenes for detection. (**a**) The detection result of fishing boats in close range along the coast; (**b**) The detection result of the fishing boat in the distant view; (**c**) The result of ship detection in blurred images; (**d**) The result of ship inspection at the port exit; (**e**) The ship detection result in the dock scene.

4. Conclusions

This paper introduces a feature fusion module to facilitate better integration of high-level and low-level features in neural network models while suppressing noise interference in the input. The dataset comprises a diverse range of infrared ships. By incorporating SPD-Conv into the network architecture, the model is optimized, resulting in improved accuracy for low-resolution and small target detection.

In scenes with dense occlusion, the original network model is prone to interference, affecting the localization and detection accuracy of ships. To address this, this paper replaces the traditional NMS with Soft-NMS, significantly enhancing the model's performance in dense occlusion scenarios.

The improved algorithm demonstrates a 1% increase in $mAP_{0.5}$ compared to the original algorithm and a 5% improvement in $mAP_{0.5:0.95}$. Although there is a slight decrease in FPS, it still achieves a frame rate of over 150 f/s. The proposed algorithm in this paper exhibits higher detection accuracy, enabling better monitoring of ships in nearshore ship detection tasks. Through extensive experiments and comparisons with other benchmark algorithms, GT-YOLO demonstrates a notable advantage in detection performance with a considerably smaller parameter count. Nevertheless, there are still limitations. The algorithm proposed in this paper improves the original algorithm by introducing SPD-Conv and simultaneously enhancing NMS. However, this comes at the cost of an increased parameter count by 1.7 million and a slight reduction in detection speed, requiring the network to handle larger computational loads and impacting real-time performance. Future work will focus on further optimizing the network to reduce parameters and computational overhead. Additionally, efforts will be directed toward refining Soft-NMS to mitigate its impact on the network's detection speed.

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

Funding: This research was funded by the National Key Research and Development Program of China (Grant number 2022YFB4301401), the Pilot Base Construction and Pilot Verification Plan Program of Liaoning Province of China (Grant numbers 2022JH24/10200029), and the Fundamental Research Projects of the Educational Department of Liaoning Province (Grant number LJKMZ20220362).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Zhu, J.; Yang, Y.; Cheng, Y. A Millimeter-Wave Radar-Aided Vision Detection Method for Water Surface Small Object Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 1794. [CrossRef]
2. Li, Y.; Wang, R.; Gao, D.; Liu, Z. A Floating-Waste-Detection Method for Unmanned Surface Vehicle Based on Feature Fusion and Enhancement. *J. Mar. Sci. Eng.* **2023**, *11*, 2234. [CrossRef]
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
4. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
5. Redmon, J.; Divvala, S.; Girshick et, al. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision–ECCV 2016, In Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I 14*; Lecture notes in computer science; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
7. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
8. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933. [CrossRef] [PubMed]
9. Shi, Q.; Zhang, C.; Chen, Z.; Lu, F.; Ge, L.; Wei, S. An infrared small target detection method using coordinate attention and feature fusion. *Infrared Phys. Technol.* **2023**, *131*, 104614. [CrossRef]
10. Ye, J.; Yuan, Z.; Qian, C.; Li, X. Caa-yolo: Combined-attention-augmented yolo for infrared ocean ships detection. *Sensors* **2022**, *22*, 3782. [CrossRef] [PubMed]
11. Si, J.; Song, B.; Wu, J.; Lin, W.; Huang, W.; Chen, S. Maritime Ship Detection Method for Satellite Images Based on Multiscale Feature Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6642–6655. [CrossRef]
12. Guo, H.; Gu, D. Closely arranged inshore ship detection using a bi-directional attention feature pyramid network. *Int. J. Remote Sens.* **2023**, *44*, 7106–7125. [CrossRef]
13. Wang, J.; Pan, Q.; Lu, D.; Zhang, Y. An Efficient Ship-Detection Algorithm Based on the Improved YOLOv5. *Electronics* **2023**, *12*, 3600. [CrossRef]
14. Shi, T.; Ding, Y.; Zhu, W. YOLOv5s_2E: Improved YOLOv5s for Aerial Small Target Detection. *IEEE Access* **2023**, *11*, 80479–80490. [CrossRef]
15. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980.
16. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3139–3148.
17. Kim, Y.; Kang, B.N.; Kim, D. San: Learning relationship between convolutional features for multi-scale object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 316–331.
18. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer Nature: Cham, Switzerland, 2022; pp. 443–459.
19. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS--improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.

20. InfiRay Dataset [OL]. Available online: http://openai.iraytek.com/apply/Sea_shipping.html/ (accessed on 15 March 2023).
21. Konovalenko, I.; Maruschak, P.; Kozbur, H.; Brezinová, J.; Brezina, J.; Nazarevich, B.; Shkira, Y. Influence of uneven lighting on quantitative indicators of surface defects. *Machines* **2022**, *10*, 194. [CrossRef]
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
23. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
24. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
25. Wang, C.Y.; Bochkovskiy, A.; Liao HY, M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
26. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.