electronics

MDPI

Article

BFE-Net: Object Detection with Bidirectional Feature Enhancement

Rong Zhang, Zhongjie Zhu *, Long Li, Yongqiang Bai and Jiong Shi

Ningbo Key Lab of DSP, Zhejiang Wanli University, Ningbo 315000, China
* Correspondence: zhuzhongjie@zwu.edu.cn

Abstract: In realistic scenarios, existing object detection models still face challenges in resisting interference and detecting small objects due to complex environmental factors such as light and noise. For this reason, a novel scheme termed BFE-Net based on bidirectional feature enhancement is proposed. Firstly, a new multi-scale feature extraction module is constructed, which uses a self-attention mechanism to simulate human visual perception. It is used to capture global information and long-range dependencies between pixels, thereby optimizing the extraction of multi-scale features from input images. Secondly, a feature enhancement and denoising module is designed, based on bidirectional information flow. In the top-down, the impact of noise on the feature map is weakened to further enhance the feature extraction. In the bottom-up, multi-scale features are fused to improve the accuracy of small object feature extraction. Lastly, a generalized intersection over union regression loss function is employed to optimize the movement direction of predicted bounding boxes, improving the efficiency and accuracy of object localization. Experimental results using the public dataset PASCAL VOC2007test show that our scheme achieves a mean average precision (mAP) of 85% for object detection, which is 2.3% to 8.6% higher than classical methods such as RetinaNet and YOLOv5. Particularly, the anti-interference capability and the performance in detecting small objects show a significant enhancement.

Keywords: object detection; bidirectional feature enhancement; anti-interference; small object detection

Citation: Zhang, R.; Zhu, Z.; Li, L.; Bai, Y.; Shi, J. BFE-Net: Object Detection with Bidirectional Feature Enhancement. *Electronics* **2023**, *12*, 4531. https://doi.org/10.3390/electronics12214531

Academic Editor: George A. Papakostas

Received: 16 October 2023
Revised: 31 October 2023
Accepted: 2 November 2023
Published: 3 November 2023

1. Introduction

Object detection, as a research hotspot in computer vision, holds extensive applications in domains including autonomous driving [1] and remote sensing monitoring [2]. Nevertheless, object detection faces persistent challenges arising from complex environments and the diversity of object characteristics, encompassing scale variations, background interference, and noise.

Presently, object detection predominantly leverages deep learning methodologies, which are categorized into two-stage and one-stage detection algorithms. Two-stage object detection algorithms involve the generation of candidate regions, which are subsequently classified and localized. Among them, Girshick et al. [3] proposed Fast R-CNN, which employs ROI pooling to extract features from regions of interest (ROI). However, its dependency on Selective Search for candidate box generation leads to slower detection speeds. In the same year, Ren et al. [4] proposed Faster R-CNN, which enhances the efficiency and detection performance of candidate box generation by introducing the Region Proposal Network (RPN) for automatic ROI generation. However, in scenarios involving small objects, RPN may produce fewer candidate regions, potentially leading to omissions in small object detection. To address this concern, Hu et al. [5] proposed SFGNet, which leverages the integration of high-resolution fine-grained features with low-resolution high-level semantic features to obtain spatially refined features. These features are incorporated into an enhanced RPN to enhance the detection performance for small objects. Nonetheless, this algorithm demonstrates limited robustness when faced

with complex scenes, variations in lighting, noise, and other forms of interference. Overall, two-stage object detection algorithms are plagued by redundant computations and slower detection speeds. Additionally, their detection performance may degrade in scenarios involving small objects and complex scenes.

To tackle these challenges, one-stage object detection algorithms that abandon candidate boxes are proposed. Instead, they rely on predicting key points to obtain bounding boxes, which has the advantages of fewer network parameters and fast detection. Notable algorithms in this category include YOLO [6], SSD [7], and CenterNet [8]. YOLO [6] divides the image into grids for detection, which can increase speed. However, its detection accuracy is low, due to the single-scale approach. To address this, YOLOv3 [9] introduced multi-scale prediction, which not only enhances detection accuracy but also handles objects of different scales more effectively. YOLOv5, the improvement of YOLOv3, adopts Feature Pyramid Network (FPN) [10] to enhance the detection accuracy of large objects through fully utilizing low-level features with high resolution and high-level features with semantic information. However, there are still challenges relating to small objects. Zhang et al. [11] proposed an object detection method that combines MobileNet v2, YOLOv4, and attention mechanisms, effectively enhancing the speed and accuracy of underwater object detection. However, it does not fully address issues like noise and blurriness in underwater images.

SSD [7] achieved better detection speed via the Anchor mechanism and multi-scale detection method. The Anchor mechanism, as a vital component of SDD, proficiently identifies objects of varying sizes and proportions by means of predefined rectangular boxes distributed at various positions and dimensions within the input image. However, the underlying feature mapping in SSD lacks semantic information, resulting in suboptimal detection performance. In response, RetinaNet [12] leverages FPN to combine low-level features with detail information and high-level features with semantic information, improving the detection capability for objects of different scales. Deng et al. [13] introduced FPN into SSD and modified the structure of effective feature layers, enriching the semantic information in shallow feature mappings. Nonetheless, it may still be susceptible to information loss during FPN propagation, potentially compromising the preservation of detailed information and the detection performance for small objects.

The core idea of CenterNet is to redefine object detection as the task of locating object center points [8]. While CenterNet has certain advantages in object detection, it still has limitations in terms of detecting small objects and handling interferences. On the one hand, centroid localization may be affected by pixel-level errors in small objects, resulting in inaccurate detection frames. On the other hand, centroid localization is also affected by complex backgrounds, leading to inaccurate detection. As an improvement upon CenterNet, DC-CenterNet [14] is designed to improve convergence speed, stability, and prediction accuracy by regressing the diagonal half-length and the central angle. And it still relies on center point prediction, which results in weaker anti-interference ability. Hence, one-stage algorithms are afflicted by insufficient feature extraction capabilities, constrained receptive fields, and suboptimal feature fusion, ultimately resulting in diminished performance in the detection of small and noisy objects.

To improve the detection accuracy for small objects and the ability to handle interferences, a novel scheme termed BFE-Net is proposed based on bidirectional feature enhancement. Firstly, in terms of feature extraction, a self-attention mechanism is introduced to optimize the extraction of multi-scale features by increasing the receptive field, thereby improving overall detection performance. Secondly, in terms of information propagation, a bidirectional feature pyramid structure is constructed to preserve fine-grained details in the top-down and fully utilize features at different levels and scales in the bottom-up, thus improving the detection performance for small objects. Furthermore, in feature fusion, the diversity of features is increased through multi-scale feature fusion, enhancing the detection capability for objects of different scales. Lastly, a generalized intersection over union (GIoU) regression loss function [15] is employed to optimize the position loss, improving both the accuracy of object localization and convergence speed. Moreover, to

enhance the anti-interference capability in the top-down, bicubic interpolation is used to perform up-sampling on feature maps, preserving image details and smoothness while mitigating the influence of noise to some extent. Our contributions are in these aspects:

(1) In order to effectively improve the detection capability for small objects and anti-interference, we propose an object detection algorithm named BFE-Net based on bidirectional feature enhancement, which consists of a perceptually optimized multi-scale feature extraction module, feature enhancement and denoising module with bidirectional information flow, and a classification regression network.

(2) To enhance multi-scale feature extraction and improve the overall performance of BFE-Net, the perceptually optimized multi-scale feature extraction module utilizes a pyramid hierarchical self-attention mechanism, which can capture global information and long-range dependencies between pixels, thereby facilitating the optimization and extraction of multi-scale features from the input image.

(3) To mitigate the influence of noise on various objects, particularly small objects, we design a feature pyramid structure with bidirectional information flow and gradually obtain high-resolution images through bicubic interpolation up-sampling, which maintains the details and smoothness of the images. This approach is beneficial to reduce the impact of noise on the feature extraction across different-scale objects, thereby enhancing the detection capability, especially for small-scale objects.

2. BFE-Net

The proposed algorithm BFE-Net, depicted in Figure 1, consists of three main components: a perceptually optimized multi-scale feature extraction module, a feature enhancement and denoising module with bidirectional information flow, and a classification regression network. To begin with, the feature extraction module uses a pyramid hierarchical approach to reorganize the input image chunks and calculate self-attention mechanisms, which optimize the extraction of multi-scale features. Next, the feature enhancement and denoising module applies bicubic interpolation up-sampling, weighted feature fusion, and cross-scale connections to reduce noise interference and improve feature completeness. Lastly, the classification regression network fuses the multi-scale features and utilizes the GIoU regression loss function to optimize the movement direction of predicted bounding boxes, which enhances the classification and localization accuracy for small objects.

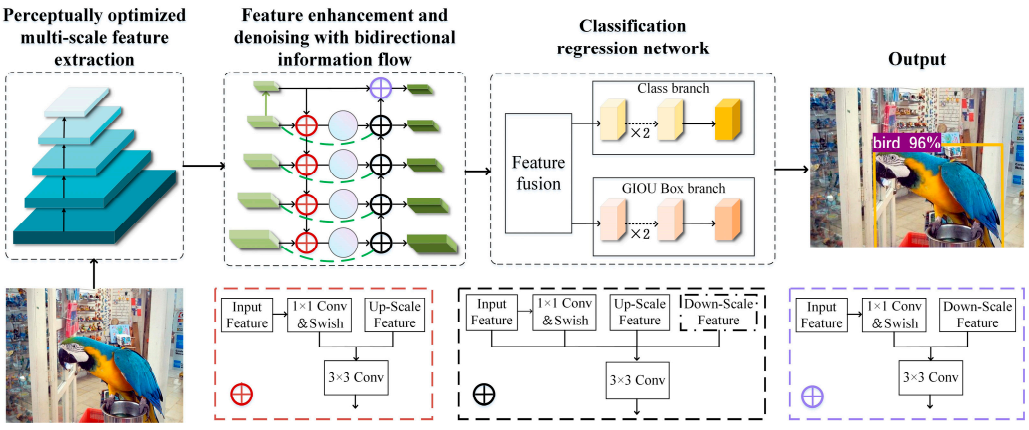

Figure 1. Overall architecture of BFE-Net.

2.1. Perceptually Optimized Multi-Scale Feature Extraction

The traditional FPN in computer vision typically builds upon different levels of feature maps extracted by a bottom-up network. Each level contains information with varying resolutions and semantics. However, due to the limited local receptive field of the network, the bottom-level feature maps can only capture local details of the image, making it difficult to obtain global contextual information, thus impacting the final detection results. The Swin Transformer [16], serving as an enhancement of the conventional Transformer

architecture [17], boasts significant advantages in the realm of image data manipulation. It amplifies its capability to handle high-resolution images and capture local details by incorporating techniques such as chunking, window-based attention, and hierarchical attention. Consequently, it has emerged as a pivotal instrument within the domain of computer vision. Therefore, BFE-Net draws inspiration from the Swin Transformer and employs self-attention mechanisms to simulate human visual perception and then captures global information and long-range dependencies between pixels, optimizing the extraction of multi-scale features.

In the perceptually optimized multiscale feature extraction module, an improved version of the Swin Transformer serves as the bottom-up component of the feature pyramid, extracting features at different scales. Specifically, the correlation between features is modelled by the Swin Transformer's self-attention mechanism to obtain both local details and global context, thereby improving feature expressiveness and detection performance, especially for small objects. To apply the pyramid structure to the output features of the bottom-up component, these are divided into five stages, representing different scales of feature outputs. Stage 1 corresponds to the lowest level, emphasizing fine details, while Stage 5 represents the highest level, emphasizing semantic information. By leveraging the improved Swin Transformer, BFE-Net can enhance perception and comprehension across different scales, contributing to an overall performance boost. Figure 2 illustrates the extraction of multi-scale features.

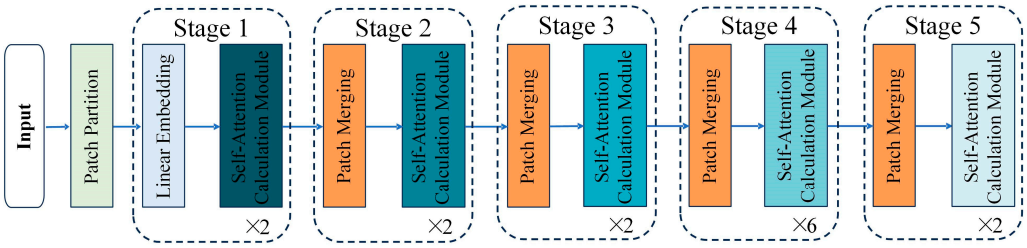

Figure 2. Extraction of multi-scale features.

Initially, the input raw image undergoes patch partitioning and linear embedding. Subsequently, in Stage 1, the Self-attention Calculation Module computes attention weights between patches, effectively downsizing the output feature map to 1/4 of the input. Moving to Stage 2, Patch Merging combines patches post self-attention calculations into an overarching feature representation. This representation then goes through a two-layer Self-attention Calculation Module to determine attention weights for this stage, further reducing the output feature map to 1/8 of the input. Stages 3, 4, and 5 replicate the operations of Stage 2, progressively yielding feature map sizes of 1/16, 1/32, and 1/64 of the input, respectively. These output feature maps from the five stages are used in the feature pyramid structure, allowing for a comprehensive and rich representation of multi-scale features, as well as global and local contextual awareness. This improves the understanding and analysis of different regions and objects in the image. Notably, the Self-attention Calculation Module employs a windowed self-attention mechanism, effectively reducing computational complexity and enhancing the efficiency and speed.

As depicted in Figure 3, the Self-attention Calculation Module employs a window-based approach for computation [16]. Specifically, W-MSA stands for multi-head self-attention module with regular windowing configuration, confining attention calculations within a fixed-size window. On the other hand, SW-MSA stands for shifted window multi-head self-attention module, which moves the window through translation operation to enhance the interaction of feature information at different locations. This approach not only reduces computational complexity but also captures long-range dependencies within the image. Apart from W-MSA and SW-MSA, the Self-attention Calculation Module

also contains Layer Normalization (LN), Residual Connection, and Multi-layer Perceptron (MLP). The formula for the Self-attention Calculation Module is provided below.

$$\hat{z}^m = W - MSA(LN(z^{m-1})) + z^{m-1} \quad (1)$$

$$z^m = MLP(LN(\hat{z}^m)) + \hat{z}^m \quad (2)$$

$$\hat{z}^{m+1} = SW - MSA(LN(z^m)) + z^m \quad (3)$$

$$z^{m+1} = MLP(LN(\hat{z}^{m+1})) + \hat{z}^{m+1} \quad (4)$$

where z^{m-1}, z^m, and z^{m+1}, respectively, denote the input of the previous layer, the output of the current layer, and the output of the next layer within the Self-attention Calculation Module of each stage. LN contributes to regularization and model training optimization, ultimately enhancing performance and stability. Residual connections are utilized to address the issues of gradient vanishing and information loss. MLP enhances the expressive power of the network through nonlinear transformation.

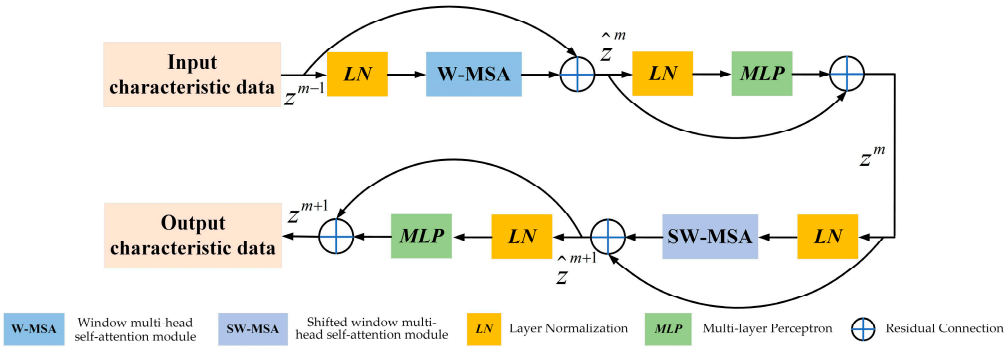

Figure 3. Self-attention Calculation Module.

By employing the perceptually optimized multi-scale feature extraction module, BFE-Net enhances the semantic representation and multi-scale information of low-level features, which proves advantageous in improving the performance of object detection tasks, particularly for small objects and scenes with complex backgrounds. Subsequently, the feature representations of different scales are outputted to the next module.

2.2. Feature Enhancement and Denoising with Bidirectional Information Flow

Although the perceptually optimized multi-scale feature extraction module enhances multi-scale feature extraction using the self-attention mechanism, there are still factors that negatively affect the accuracy of object detection in feature extraction. On the one hand, the resolution of the feature map gradually decreases, resulting in subtle edges, textures and local features not being preserved completely. On the other hand, in the low-resolution region of the feature map, noise is amplified, which adversely affects the detection effect. Noise may come from sensor noise of image acquisition devices, artifacts caused by image compression, and error propagation during feature extraction.

Figure 4 shows the object detection results of RetinaNet and YOLOv5 before and after Gaussian noise processing. Both algorithms have reduced detection accuracy for dogs, as large objects, after introducing noise. For small objects, RetinaNet fails to detect the airplane, while YOLOv5 mistakenly detects the airplane as a bird. In summary, the noise adversely affects the accuracy and robustness of object detection, especially for small objects. Therefore, it is crucial to address the problem of feature loss and noise, improving the performance and reliability of the algorithms.

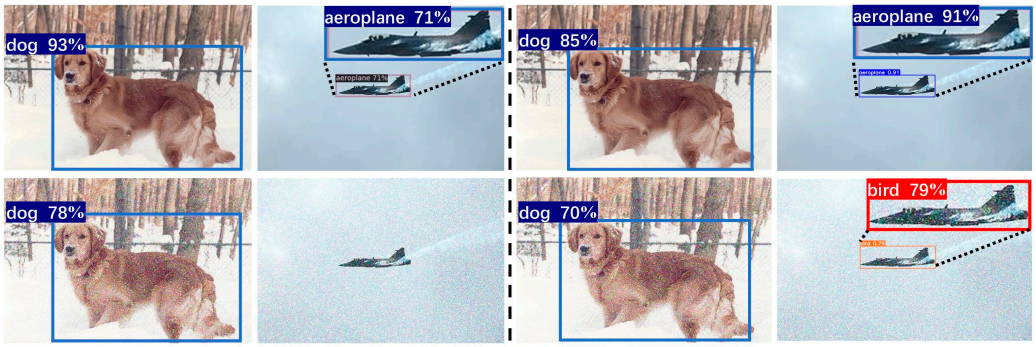

Figure 4. Comparative results before and after introducing gaussian noise. (**a**,**b**) represent the detection results of RetinaNet and YOLOv5, respectively.

To overcome these challenges, an improved feature pyramid structure (Figure 5) has been devised. This structure establishes bidirectional information flow through both a top-down and bottom-up, enabling the network to maximize the utilization of input data and create stronger connections across different layers. In the top-down, to weaken the noise present in the input data, the interpolation method is used, which draws on the idea of filters to amplify the feature maps, maintaining the detail and smoothness of the image and weakening the effect of noise to a certain extent. In the bottom-up, to further enhance the feature extraction, features from different layers and scales are fully utilized through skip connections, cross-scale connections, and weighted feature fusion. This can improve the perception and expression ability of the network, which in turn improves the performance of small object detection. The specific operation of the feature pyramid with bidirectional information flow is as follows.

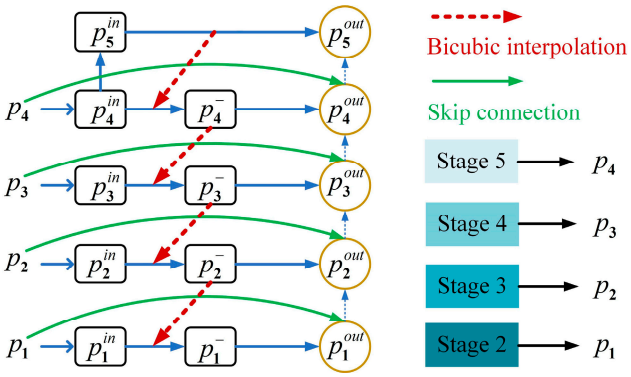

Figure 5. Feature pyramid with bidirectional information flow.

In Figure 5, p_1, p_2, p_3, and p_4 denote the features of different-scale output from Stage 2~5 in the perceptually optimized multi-scale feature extraction module. Firstly, p_1~p_4 are the initial input features, which are uniformly convolved with 1×1 and Swish activation function to obtain the corresponding input feature maps: ${p_1}^{in}$, ${p_2}^{in}$, ${p_3}^{in}$, ${p_4}^{in}$. Then, ${p_5}^{in}$ is obtained from ${p_4}^{in}$ through 1×1 convolution and 3×3 maximum pooling with a step size of 2. Finally, ${p_1}^{in} \sim {p_5}^{in}$ are regarded as the final features that are inputted into the feature pyramid structure with bidirectional information flow.

In the top-down, we employ bicubic interpolation as a smooth and accurate interpolation method, which has been widely used by researchers [18,19]. Therefore, we employ bicubic interpolation to perform quadruple up-sampling, leveraging the continuity and connectivity between pixels to enhance image resolution and mitigate the impact of noise. Specifically, bicubic interpolation calculates interpolation weights based on the distance between the interpolation point and its 16 neighboring pixels, reflecting the contribution of these nearby pixels to the interpolation point. By taking a weighted average of the

neighboring pixels, an estimated value for the interpolation point can be obtained, thus achieving image up-sampling. Through the gradual restoration of image details and clarity via quadruple up-sampling, BFE-Net generates higher-quality, high-resolution images. Specifically, we up-sample ${p_5}^{in}$ through bicubic interpolation, and then splice it with ${p_4}^{in}$ in the channel dimension to obtain ${p_4}^{-}$, which is a richer feature representation. Next, we up-sample ${p_4}^{-}$ through bicubic interpolation, and then splice it with ${p_3}^{in}$ in the channel dimension to obtain ${p_3}^{-}$. Similarly, we can obtain ${p_2}^{-}$ and ${p_1}^{-}$ in turn. In the above process, BFE-Net realizes feature fusion, which fuses higher-level with lower-level features to improve perception and anti-interference capability.

In the bottom-up, skip connections are added to the paths where p_1, p_2, p_3 and p_4 are located to avoid the feature extraction network missing information of the small objects through multiple-feature propagation, so as to enhance the feature extraction at different resolutions and then enhance the feature expression ability of the small objects. Moreover, cross-scale connections are used in the bottom-up to realize the feature propagation from shallow to deep and fuse features at different levels, thus obtaining a more comprehensive and enriched feature representation. To be more specific, p_1 is spliced with ${p_1}^{-}$ in the channel dimension to obtain ${p_1}^{out}$ through skip connection. Next, ${p_1}^{out}$ is maximum-pooled by 3×3 with a step size of 2 and then spliced with p_2 and ${p_2}^{-}$ in the channel dimension to obtain ${p_2}^{out}$. Similarly, we can obtain ${p_3}^{out}$ and ${p_4}^{out}$ in turn. Finally, ${p_5}^{out}$ is obtained by splicing ${p_5}^{in}$ with ${p_4}^{out}$ in the channel dimension via a cross-scale connection. Taking ${p_4}^{-}$ and ${p_4}^{out}$ as examples, the specific calculations are expressed in Formulas (5) and (6).

$$P_4^- = conv[\frac{w_1 \cdot p_4^{in} + w_2 \cdot BI(p_5^{in})}{w_1 + w_2 + \beta}] \tag{5}$$

$$p_4^{out} = conv[\frac{w_1{}' \cdot p_4 + w_2{}' \cdot p_4^- + w_3{}' \cdot Resize(p_3^{out})}{w_1{}' + w_2{}' + w_3{}' + \beta}] \tag{6}$$

where p_i and p_i^{in} represent the initial and final input features, respectively, for the feature enhancement and denoising module with bidirectional information flow. p_i^- denotes the fusion features in the top-down, and p_i^{out} represents the output features. w_1 and w_2 represent the learnable weights of the top-down, while $w_1{}'$, $w_2{}'$, and $w_3{}'$ represent the learnable weights of the bottom-up. β is a value significantly smaller than 1. The up-sampling operation, represented by the function BI, is performed using bicubic interpolation to enlarge the feature maps while preserving image details and smoothness, thereby reducing the impact of noise. The function Resize denotes 3×3 max pooling with a step size of 2, which adjusts the size of the feature maps to achieve consistent dimensions across different hierarchical levels, thus facilitating the fusion of features.

2.3. Classification Regression Network

The classification regression network, as an important component of BFE-Net, is used to fuse multi-scale feature maps and perform object classification and regression prediction. For the diverse-scale feature maps, p_i^{out}, output from previous modules, a multi-scale feature fusion is initially conducted. Subsequently, the fused feature maps are fed into the classification prediction branch and the regression prediction branch, enabling the classification regression network to differentiate between different objects and predict their locations.

Within the classification prediction branch, the fused feature maps are initially processed through four 3×3 convolutional layers, and nonlinearities are introduced using the ReLU activation function to extract features and map them to the classes of the object. Then, an additional 3×3 convolutional layer is used to map the features to the scores corresponding to the object classes, facilitating the classification prediction of objects. Similarly, within the regression prediction branch, object location is performed through regression prediction of the bounding box. To optimize the movement direction of the bounding box

and then enhance the accuracy of object localization, the GIoU regression loss function (L_{GIoU}) is employed to train the regression branch.

In contrast to the conventional intersection over union regression loss function (L_{IoU}), the L_{GIoU} demonstrates greater sensitivity to change in the size of predicted bounding boxes. Specifically, L_{IoU} only considers the area of the intersection and union during loss computation. In contrast, L_{GIoU} introduces the concept of Minimum Bounding Box (MBB), which accounts for both the relative position and size of predicted and ground truth bounding boxes, thereby providing more accurate location for both small and large objects.

For L_{GIoU}, we start by calculating the areas of the intersection and union regions between the predicted bounding boxes and the ground truth boxes. Afterwards, we compute the area of MBB. Finally, we calculate the value of L_{GIoU} using the following formula.

$$L_{GIoU} = 1 - IoU + \frac{(MBB - A \cup B)}{MBB} \tag{7}$$

where IoU represents the calculation result of the intersection over union, that is, the area of the intersection region divided by the area of the union region. A and B represent the area of the ground truth box and predicted bounding box, respectively. $A \cup B$ represents the area of the union region between A and B. The value of L_{GIoU} falls within the range of $[-1, 1]$. When the predicted bounding box perfectly overlaps with the ground truth box, IoU is 1, and as L_{GIoU} approaches 0, it indicates the best localization accuracy. Conversely, when the predicted bounding box has no overlap with the ground truth box, IoU is 0, and as L_{GIoU} approaches 1, it represents the worst localization accuracy.

L_{GIoU} not only focuses on the overlapping region but also considers other non-overlapping areas, providing a better reflection of the intersection between the predicted and ground truth boxes within MBB. Moreover, L_{GIoU} helps optimize the movement direction of the predicted box and provides more accurate localization information, thereby improving the performance of the object detection.

3. Results and Analysis

3.1. Experimental Datasets and Environment Settings

To validate the effectiveness of BFE-Net, we trained and tested on the PASCAL VOC2007+2012 dataset [20], which comprises 20 common object classes, including boat, bird, people, etc. During the model training, we set the input image size to 640 × 640, the epoch to 18,000, the batch size to 15, and the initial learning rate to 0.0001, and chose Adam, a stochastic gradient descent method, as the optimization algorithm.

The Ubuntu 16.04 operating system was used as the hardware platform for the experiment. The GPU part of the experiment used three NVIDIA TITAN RTX, each GPU's memory size was 24 G, and the processor used an Intel Xeon(R) Silver 4214 CPU. In terms of software, we utilized PyTorch (1.10.1) as the deep learning framework. The versions of torchvision, torchaudio, and cudatoolkit were 0.11.2, 0.10.1, and 10.2, respectively.

3.2. Comparison of Experimental Results

3.2.1. Overall Performance Analysis

To evaluate the performance of BFE-Net in object detection, we applied the trained model weights to the PASCAL VOC2007test dataset and calculated the average precision (AP) for each object class, then computed mean average precision (mAP). In addition, we compared BFE-Net to other leading object detection algorithms, presenting the results in Table 1. This table shows the detection accuracy of different algorithms for 20 object categories in the PASCAL VOC2007test dataset in detail. Comparing the results below, the effectiveness and superiority of BFE-Net in object detection are proven.

As shown in Table 1, BFE-Net exhibits a high level of detection accuracy for most object classes in the PASCAL VOC2007test dataset, with mAP reaching 85%. Compared to the typical algorithms RetinaNet and YOLOv5, BFE-Net demonstrates 4.2% and 2.3% increases in mAP, respectively. When compared to the latest algorithms SFGNet and DC-

CenterNet, the mAP of BFE-Net is an improvement by 3.7% and 3.4%, respectively. For a visual representation of BFE-Net's object detection capability, we randomly selected several images from the PASCAL VOC2007test dataset and conducted a comparative analysis with RetinaNet and YOLOv5 to highlight the differences in detection, as illustrated in Figure 6.

Table 1. Object detection results for PASCAL VOC2007test dataset (%).

Methods	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
Faster R-CNN [4]	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	56.7	87.8
SSD [7]	77.3	78.8	85.3	75.7	71.5	49.1	85.7	86.4	87.8	60.6	82.7
YOLOv3 [9]	78.3	79.0	85.6	77.2	72.3	55.6	86.8	87.3	88.0	62.3	84.2
CenterNet [8]	80.3	82.1	89.8	80.6	66.7	58.7	90.5	91.2	90.7	65.0	83.8
RetinaNet [12]	80.8	87.4	85.4	83.4	71.3	72.5	86.4	88.6	87.8	65.9	85.4
SLMS-SSD [21]	81.2	88.5	87.1	83.2	76.4	59.2	88.3	88.4	89.0	66.6	86.9
SFGNet [5]	81.3	82.2	83.9	80.3	71.5	78.2	89.6	86.9	90.0	65.7	87.9
Zhang et al. [11]	81.6	88.5	87.5	83.1	75.2	67.1	85.3	90.2	88.9	60.9	89.7
DC-CenterNet [14]	81.6	84.8	90.9	83.5	70.6	64.9	90.8	91.3	91.3	64.9	80.6
YOLOv5	82.7	95.2	85.6	65.7	50.5	89.0	65.4	92.6	92.0	82.3	78.6
BFE-Net	85.0	89.6	89.0	86.6	80.0	78.5	88.8	89.6	89.0	73.5	88.6
Methods	**mAP**	**Table**	**Dog**	**Horse**	**Mbike**	**Person**	**Plant**	**Sheep**	**Sofa**	**Train**	**TV**
Faster R-CNN [4]	76.4	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72.0
SSD [7]	77.3	76.5	84.9	86.7	84.0	79.2	51.3	77.5	78.7	86.7	76.3
YOLOv3 [9]	78.3	76.4	85.0	87.2	84.3	80.3	51.9	77.2	78.9	86.7	78.9
CenterNet [8]	80.3	75.1	88.5	89.9	89.9	86.5	42.0	82.0	80.2	89.4	83.8
RetinaNet [12]	80.8	75.5	86.6	87.0	86.0	85.3	57.0	82.8	75.7	85.7	80.6
SLMS-SSD [21]	81.2	74.6	87.3	88.6	86.5	82.2	54.8	85.5	80.9	87.9	81.0
SFGNet [5]	81.3	72.4	90.3	89.9	83.5	82.5	67.8	79.0	81.6	86.7	75.7
Zhang et al. [11]	81.6	78.4	89.5	89.5	84.9	84.8	55.1	86.9	74.3	90.8	82.0
DC-CenterNet [14]	81.6	73.8	87.5	90.6	90.8	86.4	53.3	82.9	79.0	87.7	85.6
YOLOv5	82.7	81.7	88.2	66.3	99.5	91.9	83.8	87.5	89.8	99.5	68.3
BFE-Net	85.0	81.0	89.0	90.0	89.4	87.1	66.6	88.6	80.5	88.3	87.1

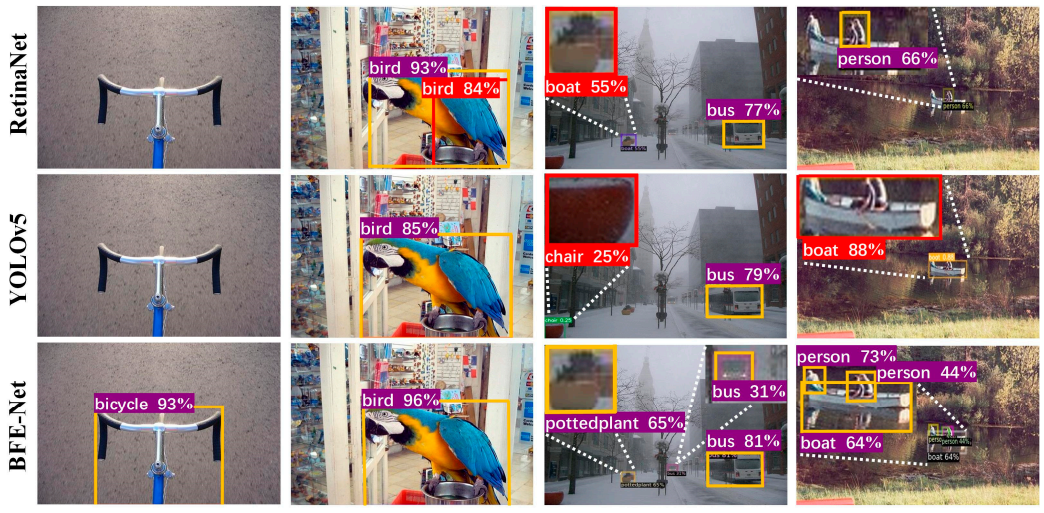

Figure 6. Visual results of BFE-Net with typical algorithms.

As can be seen from Figure 6, the detection accuracy of BFE-Net for bicycle, bird, bus, and boat is significantly better than that of the comparison algorithms, RetinaNet and YOLOv5. In terms of overall network, compared with other algorithms, BFE-Net draws on the idea of the Swin Transformer in feature extraction, which effectively captures the global and local information in the image through the window self-attention mechanism, thus improving the object detection performance. Additionally, BFE-Net constructs a feature pyramid with bidirectional information flow, facilitating the fusion of multi-scale features

and further boosting detection accuracy. Lastly, L_{GIoU} is used in the classification regression network to optimize the movement direction of predicted bounding boxes, leading to improved localization accuracy. In summary, during object detection visualization, BFE-Net consistently outperforms typical algorithms, proving its effectiveness in delivering superior accuracy and efficiency in a wide range of object detection tasks.

3.2.2. Performance Analysis of Small Object Detection

Many algorithms have poor detection accuracy for small objects, and small objects may even be missed and wrongly detected due to the small-sized object samples in the PASCAL VOC2017+2012 dataset, such as bird, boat, bottle, and television. To address this, BFE-Net improves its algorithmic structure to enhance the detection accuracy of small objects. For feature extraction, BFE-Net uses a window-based self-attention mechanism that adaptively focuses on important regions in the image, especially the local regions where small objects are located, improving the perception of small objects. For feature fusion, the bidirectional information flow within the feature pyramid structure allows the fusion of features from different levels, enabling the network to better capture the feature information of multi-scale objects. This helps BFE-Net handle small objects and improve the detection accuracy. Moreover, the adoption of L_{GIoU} optimizes the movement directions of predicted boxes, ultimately enhancing the detection performance for small objects. BFE-Net specifically considers bird, boat, bottle, and television as small object categories, and the mAP is 83.1%, surpassing other comparative algorithms. This demonstrates BFE-Net's effectiveness in detecting small objects and addressing the challenges presented by their presence in the dataset. As a result, BFE-Net is a reliable choice for object detection tasks where small object detection is of critical importance. This performance is illustrated in Figure 7 (created using Origin 2018).

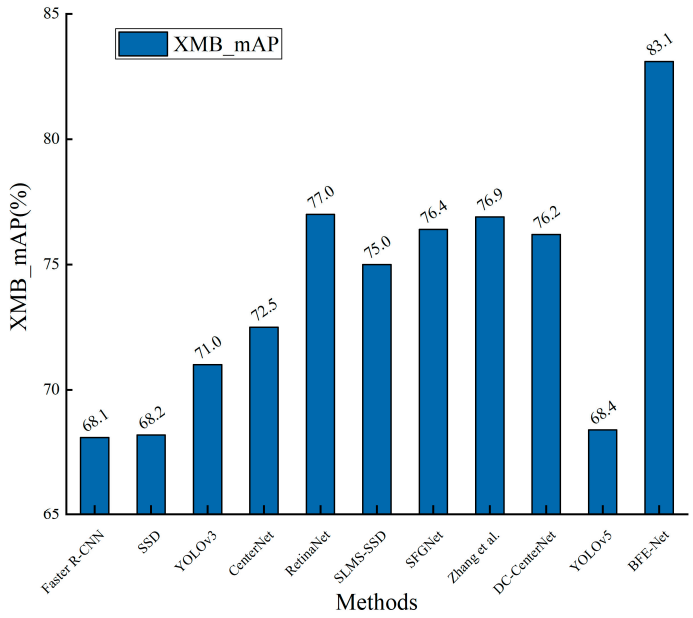

Figure 7. Comparison of mAP for four small objects (%).

Furthermore, we have visualized the detection results for small objects in Figure 8, showing the performance of BFE-Net compared to the typical algorithms, RetinaNet and YOLOv5. BFE-Net shows superior detection performance for small objects in the four categories of bird, boat, bottle, and television. In scenes 1 and 2, BFE-Net detects small objects such as bird and boat effectively. In scene 3, where the background is more complex, the comparative algorithms miss some small objects, while BFE-Net can detect the bottles better. In scene 4, RetinaNet detects a false object, while BFE-Net achieves a high detection accuracy of 99% for the television. Through the comparison of mAP

and visual analysis, these confirm the performance of BFE-Net in small object detection, highlighting its effectiveness and reliability in tackling the challenges posed by small objects in various scenarios.

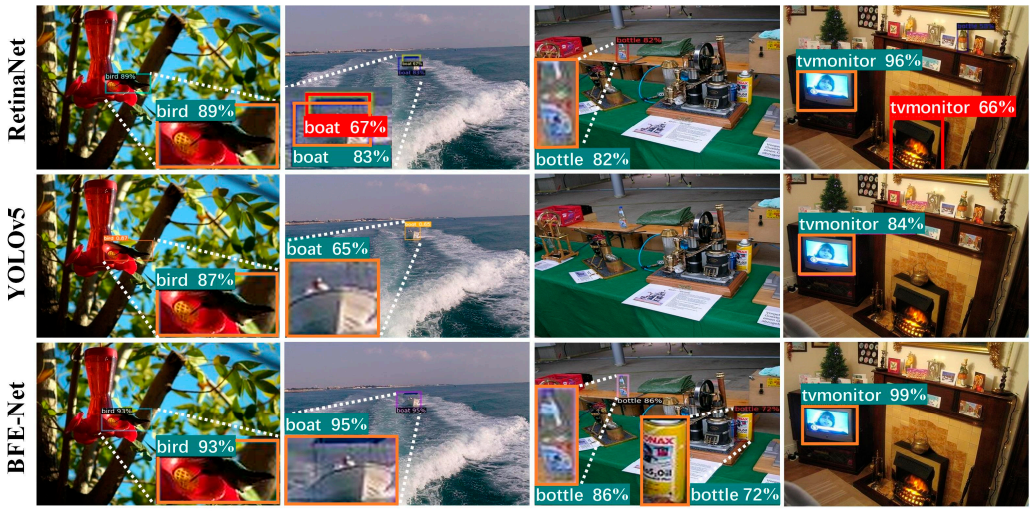

Figure 8. Visual results of different algorithms for four small objects.

3.2.3. Analysis of Anti-Interference Performance

In complex real scenarios, object detection algorithms are always plagued by noise, light, and so on. To mitigate the impact of various interfering factors on object detection, a bicubic interpolation for noise filtering is employed by BFE-Net. This approach effectively preserves image details and smoothness while mitigating the impact of noise to a certain extent. Simultaneously, through a quadruple up-sampling, image details and clarity are progressively restored by BFE-Net, yielding higher-quality, high-resolution images and further enhancing the extraction of fine-grained features.

Next, to verify the anti-interference performance of BFE-Net, we experimented with four different kinds of common interferences using the PASCAL VOC2007test dataset: Gaussian noise (GN), Gaussian blur (GB), motion blur (MB), and encoding distortion (ED). Specifically, the four interference datasets were tested separately using the original training weights to obtain the mAP, as shown in Table 2. As can be seen from Table 2, on the four interference datasets, the detection accuracy of BFE-Net is higher than that of the typical RetinaNet and YOLOv5. To further evaluate the anti-interference performance of BFE-Net, we randomly selected one image, respectively, from four interference datasets for visual comparison. The result is shown in Figure 9.

Table 2. Comparison of mAP for PASCAL VOC2007test and four interference datasets (%).

Methods	mAP				
	VOC2007test	+GN	+GB	+MB	+ED
RetinaNet	80.80	71.89	78.25	74.39	75.55
YOLOv5	82.70	71.50	78.30	76.20	77.30
BFE-Net	85.00	80.24	83.88	82.06	82.35

Figure 9 shows that the visual detection results of BFE-Net for the four interference datasets are significantly better than the comparative algorithms. In the PASCAL VOC2007test dataset with GN introduced, for scenario 1, RetinaNet failed to detect the plane and YOLOv5 incorrectly detected the plane as a bird, while BFE-Net could detect the plane with a high accuracy of 95%. In the PASCAL VOC2007test dataset with GB introduced, for scenario 2, RetinaNet mistakenly detected two cats instead of one, while BFE-Net avoided this. In the PASCAL VOC2007test dataset with MB and ED introduced,

for scenes 3 and 4, RetinaNet and YOLOv5 showed object omission and low accuracy, while BFE-Net could detect interference object with high accuracy. These tests and visualization effect comparison confirm that BFE-Net can effectively reduce the noise interference and has a certain anti-interference ability.

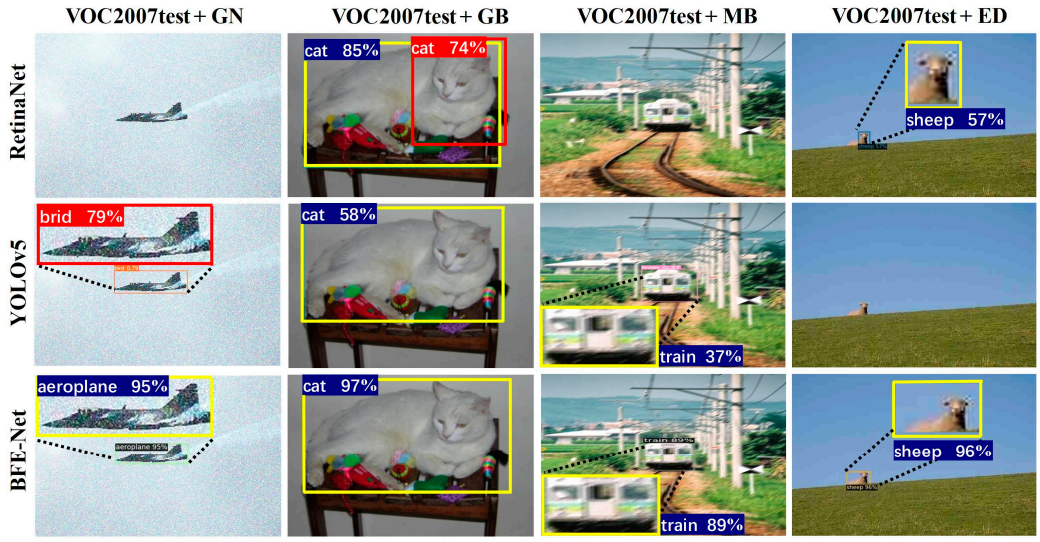

Figure 9. Visual results of different algorithms for interference datasets.

3.2.4. Analysis of Regression Loss Function

In classification regression networks, the performance can be improved by utilizing L_{GIoU} to optimize the predicted bounding box locations. This loss function provides more accurate positioning information. As shown in Figure 10, compared to RetinaNet, the optimized regression loss function of BFE-Net converges faster and has smaller oscillations. This indicates the effectiveness of L_{GIoU}, which enhances the efficiency and accuracy of object localization, resulting in better detection performance and a higher level of precision in predicting bounding box locations for objects in various scenarios.

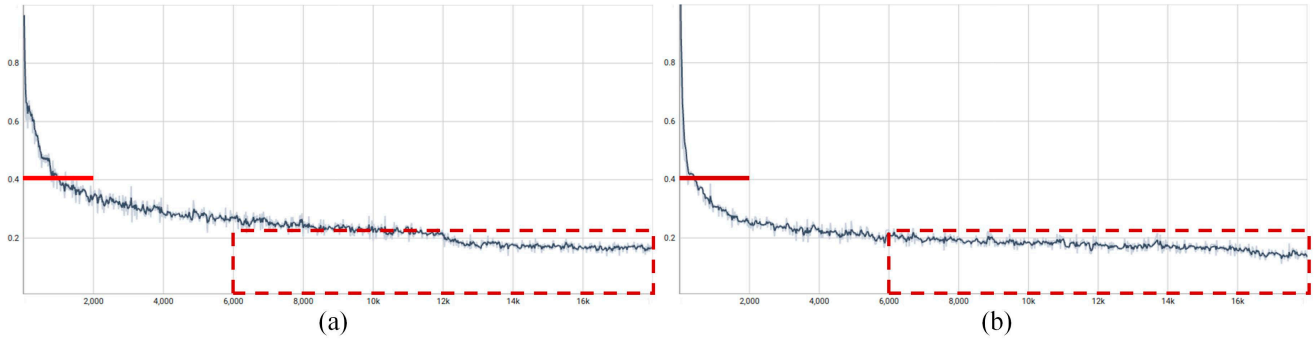

Figure 10. Comparison of regression loss between RetinaNet and BFE-Net. (**a**,**b**) illustrate the regression loss functions of RetinaNet and BFE-Net, respectively.

3.3. Ablation Experiment

This paper validated the impact of different modules on the detection performance using the PASCAL VOC2007+2012 dataset. The experimental results are detailed in Table 3.

Experiment 1 presents the original RetinaNet algorithm with ResNet-50 as the backbone feature extraction network and FPN for constructing the feature pyramid. The mAP achieved is 80.8%.

Experiment 2 replaces the backbone feature extraction network of RetinaNet with the perceptually optimized multi-scale feature extraction module, which uses the improved

Swin Transformer as the bottom-up part of the backbone network to obtain multi-scale feature representations. This improvement leads to a 1.5% mAP increase.

Table 3. Effect of different modules on detection performance.

/	ResNet-50	FPN	Perceptually Optimized Multi-Scale Feature Extraction	Feature Enhancement and Denoising with Bidirectional Information Flow	L_{GIoU}	mAP
1	√	√	×	×	×	80.8
2	×	√	√	×	×	82.3
3	√	×	×	√	×	83.6
4	×	×	√	√	×	84.5
5	×	×	√	√	√	85.0

Experiment 3 replaces the original FPN in Experiment 1 with a bidirectional feature enhancement and denoising module based on bidirectional information flow, which establishes two paths, top-down and bottom-up, for bidirectional information flow. It improves the detailed feature extraction and reduces the noise influence. The mAP is increased by 2.8% compared to Experiment 1.

Experiment 4 combines the improvements of Experiments 2 and 3, which involves feeding the multi-scale feature maps extracted by the perceptually optimized multi-scale feature extraction module into the bidirectional feature enhancement and denoising module. It mitigates the noise interference in feature extraction and enhances the feature completeness through bicubic interpolation up-sampling, weighted feature fusion, and cross-scale connections. It achieves a mAP of 84.5%.

Finally, in Experiment 5, the positional loss function is improved by using L_{GIoU} to optimize the predicted bounding box locations, which enhances the object localization accuracy, resulting in the best mAP of the network.

4. Conclusions

In this paper, we propose a one-stage object detection algorithm based on bidirectional feature enhancement. It consists of a self-attention mechanism module, which effectively captures the long-range dependencies and global information between objects, enhancing the overall performance. To further address the challenges of detecting small objects and missing detections, we employ a novel interpolation up-sampling method, which reduces the impact of noise. Additionally, we introduce skip connections and cross-scale connections to enhance feature fusion across various scales. Moreover, the position loss function is optimized to address inaccurate localization that leads to missed detections and false alarms. The experimental results show that the proposed algorithm can effectively enhance the detection accuracy of different-scale objects, especially small objects. It also maintains good detection ability in complex and noisy real environments.

However, a single technique for object detection is not enough to accurately identify objects. In the next research work, the semantic information of the object can be employed in the object detection pipeline to adjust the accuracy of the detection frame, which is potentially advantageous for small object detection and anti-interference.

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

Funding: This research was funded by the National Natural Science Foundation of China (61671412), Zhejiang Provincial Natural Science Foundation of China (LY21F010014), Ningbo Municipal Major Project of Science and Technology Innovation 2025 (2022Z076), and Zhejiang Provincial Public Welfare Program under Grant (LGF21F020021).

Data Availability Statement: Not applicable.

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

References

1. Jiang, S.; Hong, Z. Unexpected Dynamic Obstacle Monocular Detection in the Driver View. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 68–81. [CrossRef]
2. Jiang, S.; Yao, W.; Wong, M.S.; Li, G.; Hong, Z.; Kuc, T.; Tong, X. An Optimized Deep Neural Network Detecting Small and Narrow Rectangular Objects in Google Earth Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1068–1081. [CrossRef]
3. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
5. Hu, J.; Wang, Y.; Cheng, S.; Liu, J.; Kang, J.; Yang, W. SFGNet detecting objects via spatial fine-grained feature and enhanced RPN with spatial context. *Syst. Sci. Control Eng.* **2022**, *10*, 388–406. [CrossRef]
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
7. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
8. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
9. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
10. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
11. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight underwater object detection based on yolo v4 and multi-scale attentional feature fusion. *Remote Sens.* **2021**, *13*, 4706. [CrossRef]
12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
13. Deng, X.; Li, S. An Improved SSD Object Detection Algorithm Based on Attention Mechanism and Feature Fusion. *J. Phys. Conf. Ser.* **2023**, *2450*, 012088. [CrossRef]
14. Xia, Y.; Kou, X.; Jia, W.; Lu, S.; Wang, L.; Li, L. CenterNet Based on Diagonal Half-length and Center Angle Regression for Object Detection. *KSII Trans. Internet Inf. Syst.* **2023**, *17*, 1841–1857.
15. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
18. Saeed, F.; Ahmed, M.J.; Gul, M.J.; Hong, K.J.; Paul, A.; Kavitha, M.S. A robust approach for industrial small-object detection using an improved faster regional convolutional neural network. *Sci. Rep.* **2021**, *11*, 23390. [CrossRef] [PubMed]
19. Han, G.; Chen, Y.; Wu, T.; Li, H.; Luo, J. Adaptive AFM imaging based on object detection using compressive sensing. *Micron* **2022**, *154*, 103197. [CrossRef] [PubMed]
20. Everingham, M.; Eslami, S.M.; Gool, L.V.; Williams, C.K.; Winn, J.M.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2014**, *111*, 98–136. [CrossRef]
21. Wang, K.; Wang, Y.; Zhang, S.; Tian, Y.; Li, D. SLMS-SSD: Improving the balance of semantic and spatial information in object detection. *Expert Syst. Appl.* **2022**, *206*, 117682. [CrossRef]