land

MDPI

Article

YOLOV4_CSPBi: Enhanced Land Target Detection Model

Lirong Yin [1], Lei Wang [1], Jianqiang Li [2], Siyu Lu [2], Jiawei Tian [2], Zhengtong Yin [3], Shan Liu [2] and Wenfeng Zheng [2,*]

[1] Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA; lyin5@lsu.edu (L.Y.)
[2] School of Automation, University of Electronic Science and Technology of China, Chengdu 610054, China
[3] College of Resource and Environment Engineering, Guizhou University, Guiyang 550025, China
* Correspondence: winfirms@uestc.edu.cn

Abstract: The identification of small land targets in remote sensing imagery has emerged as a significant research objective. Despite significant advancements in object detection strategies based on deep learning for visible remote sensing images, the performance of detecting a small and densely distributed number of small targets remains suboptimal. To address this issue, this study introduces an improved model named YOLOV4_CPSBi, based on the YOLOV4 architecture, specifically designed to enhance the detection capability of small land targets in remote sensing imagery. The proposed model enhances the traditional CSPNet by redefining its channel partitioning and integrating this enhanced structure into the neck part of the YOLO network model. Additionally, the conventional pyramid fusion structure used in the traditional BiFPN is removed. By integrating a weight-based bidirectional multi-scale mechanism for feature fusion, the model is capable of effectively reasoning about objects of various sizes, with a particular focus on detecting small land targets, without introducing a significant increase in computational costs. Using the DOTA dataset as research data, this study quantifies the object detection performance of the proposed model. Compared with various baseline models, for the detection of small targets, its AP performance has been improved by nearly 8% compared with YOLOV4. By combining these modifications, the proposed model demonstrates promising results in identifying small land targets in visible remote sensing images.

Keywords: remote sensing; multi-scale feature fusion; land target detecting; deep learning; CNN; reasoning ability; YOLO network; BiFPN

Citation: Yin, L.; Wang, L.; Li, J.; Lu, S.; Tian, J.; Yin, Z.; Liu, S.; Zheng, W. YOLOV4_CSPBi: Enhanced Land Target Detection Model. *Land* **2023**, *12*, 1813. https://doi.org/10.3390/land12091813

Academic Editor: Deodato Tapete

Received: 8 August 2023
Revised: 7 September 2023
Accepted: 19 September 2023
Published: 21 September 2023

1. Introduction

Optical remote sensing image has the advantages of high resolution and rich feature information and being intuitive and easy to understand [1]. The primary objective of target detection is to identify interesting targets from massive data and extract their location information. The task of target detection in optical images involves the automated analysis of details about the image characteristics using relevant algorithms, followed by the classification of targets and extraction of their positional characteristics. During the early stages of exploration, target detection in optical images primarily relies on manual classification and positioning, which proves to be both time-consuming and labor-intensive while also falling short of meeting real-time requirements.

Over the course of several years, the landscape of automatic recognition and detection methods for optical images has undergone a progressive transformation. This evolution has encompassed various techniques, such as template matching [2] and image analysis [3]. However, these methodologies necessitate prior manual design and calibration of feature information. This dependence on expert-engineered feature information introduces a reliance on a substantial cadre of experts and is characterized by a tailored focus, often lacking in robust generalization capability. In tandem with the advancement of artificial intelligence technology, object detection approaches grounded in deep learning have

emerged as prominent contenders. These methods have garnered widespread adoption and dissemination across numerous traditional disciplines [4–6]. This method uses the theoretical basis that neural networks can fit any function and studies a neural network that can automatically complete feature information learning and task reasoning, which significantly enhances target detection tasks' speed and precision in natural scenes [7]. While this deep learning-based network has demonstrated remarkable performance in natural scenes, its application in optical remote sensing images remains challenging due to significant disparities between the two types of imagery.

As deep learning-based target detection continues to be extensively investigated, an increasing number of one-stage object-detecting algorithms are being employed in optical remote sensing images. A mainstream approach is to perform cluster analysis on the target dataset through an automatic clustering algorithm, then design an adaptive distance calculation formula to obtain a more meaningful intersection ratio, and finally learn from the model architecture of the YOLO series to complete the adaptability. For example, the idea of the Densely Connected Convolutional Networks (Densenet) [8] is applied, which combines the dense connection layer in the Dense network and the residual block to improve the network's capacity to extract feature information. Moreover, within the Neck network that performs the feature fusion, diverse structures of the characteristic pyramid are extensively employed, and these prevailing optimization approaches exhibit superior performance compared to conventional approaches.

In 2019, Ghorbani et al. introduced a novel approach utilizing the PIIFD characterization operator [9] to address differentiated samples and their background changes. The study demonstrated the superior performance of this method in optical remote sensing target detection compared to traditional approaches; Cao C et al. introduced a ship detection algorithm based on YOLO in 2020 [10], which actually adopts the above mainstream method. The author deviated from the conventional YOLO approach of utilizing three anchor boxes and instead recalculated the anchor box parameters, specifically using a clustering algorithm, and incorporated the detection scale into the output layer of the network extracting characteristics. The receptive field method enhances the detection accuracy of smaller objects like ships, reducing the network further. Xu et al. adopted an improved method by selecting the feature fusion structure of the network [11]. It is difficult to disseminate low-level semantic information when the target is small.

In the feature extraction network, Yang et al. [12] increased the low-level feature information, which is beneficial to the classification and localization of small targets. Meanwhile, the author changes the connection mode of the network to dense connection, which reduces the loss of the propagation of the underlying features in the network. Wang et al. [13] also tried to integrate low-level characteristic details into the network and amplified the significance of smaller target samples by assigning them higher weights within the loss function, thus increasing the accuracy of detecting small targets. The deconvolution layer was employed to integrate shallow characteristics with deep characteristics, thereby augmenting the detection capability of small targets in Li et al.'s study [14]. To mitigate the influence of background information on the detection task. Fu et al. [15] performed weight distinction before the fusion of low-level characteristic details and characteristic details. The weight details are implemented by a balance operator, but the robustness is poor. Zhang et al. [16] utilized the two-stage object detection network Faster RCNN, up-sampling all candidate objects obtained in the first stage. This feature upsampling operation is usually done by deconvolution calculation so that a larger-scale feature map can be obtained. Using a similar idea, Schilling et al. [17] studied scale improvement by adding high-level feature maps and also used deconvolution layers to achieve this scale expansion and fused low-level characteristic details with the expanded high-level characteristics to achieve the final target detection task. Liu et al. [18] replaced the deconvolution calculation with the atrous convolution calculation to reduce the computational cost. While this enhancement does yield a reduction in network parameters to some extent, it leads to the loss of certain features while maintaining the same receptive field. To this end, Ying et al. [19] studied the

problem caused by atrous convolution by completing partial information fusion through the attention mechanism.

Optical distant sensing images are not only unique from images in common natural scenes but also have huge differences between various targets in their own sample images. This phenomenon of huge differences in appearance and shape makes it difficult to preset the anchor frame of the network, resulting in the result that the target is missed [20–22].

The current mainstream solution is to increase the number of anchor boxes to try to cover more possible targets. In addition, the network's robustness about the appearance and shape of the object can be enhanced through variable neural networks and key feature detection. Currently, numerous studies have focused on augmenting the network's generalization capabilities toward target appearance by explicitly increasing the quantity and diversity of anchor boxes [23–25]. Since the target angle is relatively random in optical remote sensing images, it is difficult to fundamentally solve such problems with poor robustness, and increasing the number of anchor boxes also causes the overall computational cost of the network with greater pressure. It can be seen that the solution at the current stage is essentially to alleviate the occurrence of this problem through brute force calculation.

The primary focus of this study is on enhancing the detection capability of small-scale objects in optical remote sensing imagery without significantly increasing computational complexity. By improving the channel division method of the Cross Stage Partial (CSP) structure [26] and simultaneously applying it to the Neck component of the YOLO network, the reusability of features is enhanced. The proposed enhancements are employed in both the Neck and Backbone structures of the YOLOv4 network, resulting in the introduction of the CSPX_1 structure. Given the necessity for specific adaptive improvements to the CSP structure in the Neck component, ResBlock structures are removed, and a stack of Cross-body Link (CBL) structures is added. A feature fusion structure named CSPX_2 is designed to acquire fused features with stronger semantic information. Building upon the Bidirectional Feature Pyramid Network (BiFPN), the feature fusion network within the model is improved, introducing bidirectional feature fusion mechanisms into the Neck network structure of YOLOv4. Adaptive improvements are tailored to the structure of the YOLO network and the characteristics of remote sensing imagery. The DOTA dataset is used for experimentation, and during data preprocessing, various data augmentation schemes are employed to enhance features related to small objects.

The experimental phase involves a performance comparison between the proposed model and baseline models. The proposed model showcases approximately a 3.2% improvement in mAP (mean Average Precision) compared to the traditional YOLOv4 model. This validates the effectiveness of the proposed model improvements. Additionally, a comparison of small object detection performance with multiple detection models further demonstrates that the improvements made in this study to the YOLOv4 network's feature fusion aspect successfully enhance the model's detection performance and robustness concerning medium to small-scale objects. It proves the feasibility of the weighted bidirectional multi-scale feature fusion mechanism on the YOLO network architecture and provides certain improvement ideas for models with similar structures.

The main innovations and contributions of this article are as follows:

1. The main contribution of this research is the improvement in feature fusion of the YOLOv4 network, which successfully enhances the model's detection performance and robustness against small land targets.
2. This study introduces improvements to the CSP structure, which enhances the reusability of functions. Through the introduction of the CSPX_1 structure, the CSP enhancement function is integrated into the Neck and Backbone structures of the YOLOv4 network, which is an innovation aimed at improving the overall detection performance.
3. Research on introducing the CSPX_2 structure and bidirectional feature fusion mechanism into the Neck network structure. This helps capture stronger semantic infor-

mation and improve object detection accuracy. This research verifies the feasibility of the weighted bidirectional multi-scale feature fusion mechanism within the YOLO network architecture. This innovation has the potential to benefit models with similar structures in various fields.

2. Data

2.1. DOTA Dataset

The DOTA dataset [27] is a comprehensive dataset specially designed for identifying visible light remote sensing images produced and maintained by Wuhan University. The dataset comprises a wide range of image sizes, varying from 800 × 800 pixels to large-scale images measuring 20,000 × 20,000 pixels. The data comes from multiple satellite data and different platforms, including data from GF-2 and JL-1 satellites, as well as data from Google Earth and Optical remote sensing imagery. The DOTA dataset includes 15 categories, which are marked in a total of 2806 image files of different sizes, and a total of 188,282 instances of real object detection tasks are marked.

2.2. Improvements to the DOTA Dataset

This section aims to improve the quality and usefulness of the DOTA dataset in optical remote sensing image detection tasks. It does this by segmenting images, augmenting images with pixel-level augmentation, and applying region-based augmentation methods, ultimately generating rich datasets for more efficient training and evaluation of network models. The specific method is as follows.

2.2.1. Sample Cutting Method Based on Sliding Overlapping Area

In this study, the original image files of the DOTA dataset are cut to 832 × 832 size. This paper tries the method of overlapping area cutting, which adopts the idea of sliding window to overlap the image with the same step distance to solve the loss of the labeling box information in the original image after cutting. Finally, it is a more suitable solution to segment the primary image with a step size of 50%. After cutting, 55,992 optical remote sensing images, each with dimensions of 832 × 832, were acquired.

2.2.2. Analysis of Pixel-Level Enhancement Methods

In consideration of the inherent properties of remote sensing images, this study employed various image augmentation techniques, including random rotation, random zoom, horizontal flip, contrast adjustment, saturation adjustment, translation transformation, brightness adjustment, noise transformation, vertical flip, sharpness modification, and random cropping, on the original images. Pixel-level enhancement combination, these image enhancement methods can expand the size of the primary dataset according to the actual characteristics of the samples, without bringing irrelevant image feature information into the new dataset.

The noise transformation method is Gaussian noise transformation. This data enhancement method brings more random data disturbance, which can guide the detection model to learn more meaningful characteristics and enhance the robustness of network prediction. In addition, this paper does not perform indiscriminate enhancement on all the original samples, but randomly selects about 60% of the original samples and applies 3 random enhancement methods to the samples.

2.2.3. Analysis of Area Random Erase Method

A variety of pixel-level data enhancement methods, including the GridMask [28] method based on the idea of regional random erasure and the Mosaic method, are applied to the DOTA dataset through different thresholds and scales, and a new dataset applied to the improved network model described later is produced. Finally, the self-made dataset contains 87,382 sample images with a size of 832 × 832 and a total of 334,585 target frames are marked. The annotation information adopts the Pascal VOC label format, and the file

is in xml format and follows the xml syntax specification. The annotation information includes Filename, Aize, Object_Name, Pose, Truncated, and Difficult.

3. Methods

3.1. The Structure of Cross-Stage Partial

The main purpose of the CSPNet [29] is to enhance the network's structural level in order to mitigate the computational burden of the network. CSPNet partitions the input into two distinct components: short-connected edge and convolutional edge.

The convolutional edge extracts feature information through the operation of the traditional convolution structure, and the short-connected edge is directly processed with the output of the CSPNet structure after a small amount of processing. Feature maps are connected, as shown in Figure 1.

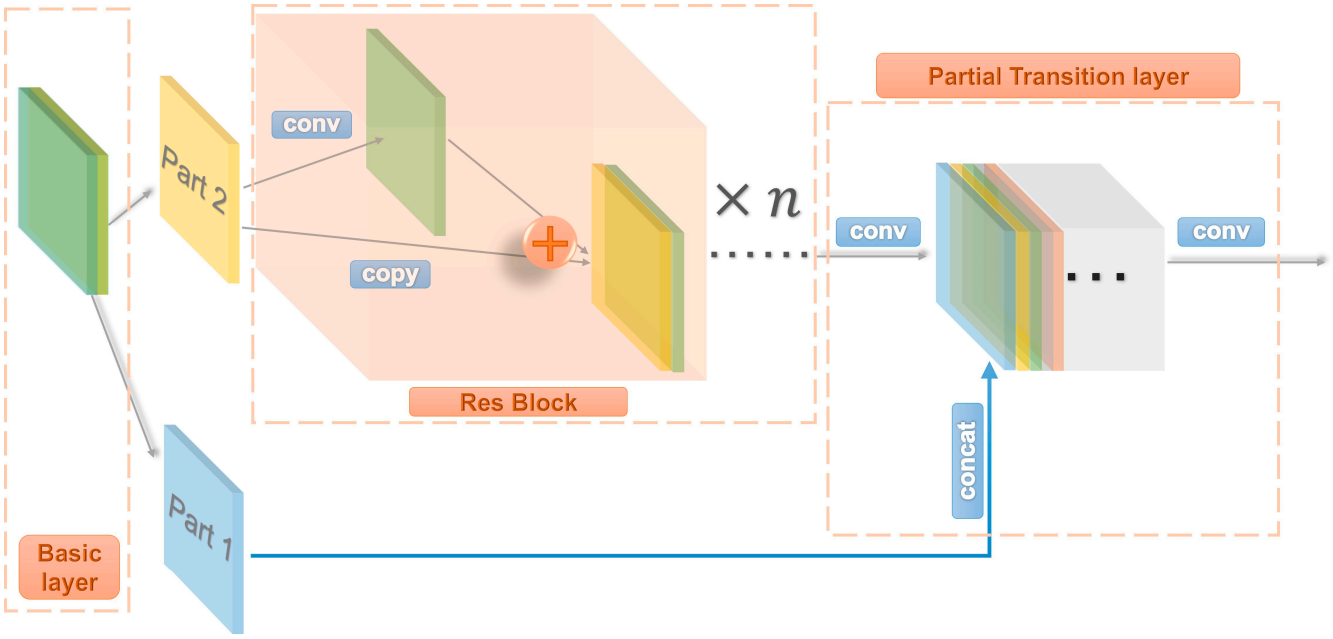


Figure 1. The structure of CSPNet.

Through this channel division method, CSPNet reduces the memory cost required by the network in the operation process, and on this basis, improves the learning ability and maintains the performance of the model.

The main reason why CSPNet can accelerate network processing is that by intercepting the gradient flow, it prevents the network from continuously calculating repeated content when updating gradient information, as shown in Equations (1) and (2) for the forward propagation of ordinary DenseNet and backpropagation [30].

$$\begin{aligned} x_1 &= w_1 \otimes x_0 \\ x_2 &= w_2 \otimes [x_0, x_1] \\ &\cdots\cdots \\ x_k &= w_k \otimes [x_0, x_1, \ldots, x_{k-1}] \end{aligned} \tag{1}$$

$$\begin{aligned} w'_1 &= f(w_1, g_0) \\ w'_2 &= f(w_2, g_0, g_1) \\ w'_3 &= f(w_3, g_0, g_1, g_2) \\ &\cdots\cdots \\ w'_k &= f(w_k, g_0, g_1, g_2, \ldots\ldots, g_{k-1}) \end{aligned} \tag{2}$$

where $\otimes$ represents the convolution operator, $[x_0, x_1, \ldots, x_k]$ represents feature information splicing, f is the update method of the weight parameters, g_i is the i-th layer's gradient

information, w_i is the i-th layer's weight information, while x_i is the changed layer's feature output.

It can be seen from this formula that a substantial quantity of gradient information is reused in backpropagation, which causes different hierarchical structures in the network to learn the same feature information. Therefore, the CSPNet structure proposes to divide the input feature information, denoted as $x_0 = \left[x_0{}', x_0{}''\right]$. Among them, $x_0^{'}$ represents the shorted edge, $x_0{}''$ represents the convolution edge, and the output x_τ is obtained after the calculation of the convolution edge of $x_0{}''$, and, finally, the final output result x_U can be obtained by splicing x_τ and x_0. Its forward propagation and backpropagation weight update rules are shown as Equations (3) and (4) [31].

$$\begin{aligned} x_k &= w_k \otimes [x_0'', x_1, \ldots, x_{k-1}] \\ x_\tau &= w_\tau \otimes [x_0'', x_1, \ldots, x_k] \quad \cdots \\ x_U &= w_U \otimes [x_0', x_\tau] \end{aligned} \tag{3}$$

$$\begin{aligned} w_k' &= f(w_k, g_0'', g_1, g_2, \ldots, g_{k-1}) \\ w_\tau' &= f(w_\tau, g_0'', g_1, g_2, \ldots, g_{k-1}) \\ w_U' &= f(w_U, g_0', g_\tau) \\ &\cdots\cdots \\ w_k' &= f(w_k, g_0, g_1, g_2, \ldots, g_{k-1}) \end{aligned} \tag{4}$$

From the provided equation, it is evident that each side's gradient update channel operates independently. As a result, these channels do not contain redundant gradient information with respect to each other. This approach prevents the flow of gradients and thereby mitigates the computation of a substantial portion of redundant gradient information. While it is true that certain feature information within an individual CSP structure may not undergo processing by predetermined feature extraction networks, the CSP structure itself functions as a fundamental component of the network. It incorporates established computational units from the original network. Consequently, in the practical application of the CSP structure, the feature extraction segment of the network is formed by stacking multiple CSP structures consecutively. This approach effectively prevents the loss of feature information.

YOLOV4 replaces all residual structures of DarkNet with CSP structures in the backbone part of the network [27]. This replacement not only substantially enhances the characteristic extraction ability of the backbone network in YOLOV4, but also reduces the inference calculation to a certain extent quantity. On this basis, this study explores the possibility of applying the CSP structure to the Neck part of the YOLO network.

To reduce the amount of computation during inference, this paper uses the CSPX_1 structure shown in Figure 2a. Unless otherwise specified, all CSPX_1 structures in this article are of this type. For the Neck part, because the task of the backbone network is to extract features, a large number of residual networks are designed to improve the learning ability of features. But in the Neck part of the network, its main task is to fuse features, so specific adaptive improvements are required for the CSP structure of the Neck part. This article makes modifications on the basis of CSPX_1, deletes the ResBlock structure, and increases the stacking of the CBL structure. This improvement is to obtain stronger fusion features of semantic information, and its structure diagram is shown in Figure 2b.

The enhanced CSP architecture introduced in this article also involves the division of feature channels after the input of features. However, the data flow from these two divisions is managed by predetermined computational units. The resulting outputs from these units do not require dimension adjustment before being directly concatenated. This approach not only enhances feature reusability compared to the traditional CSP structure but also adheres to the core principles of the CSP architecture. Moreover, it reduces computational complexity in comparison to standard convolutional structures by truncating gradient information.

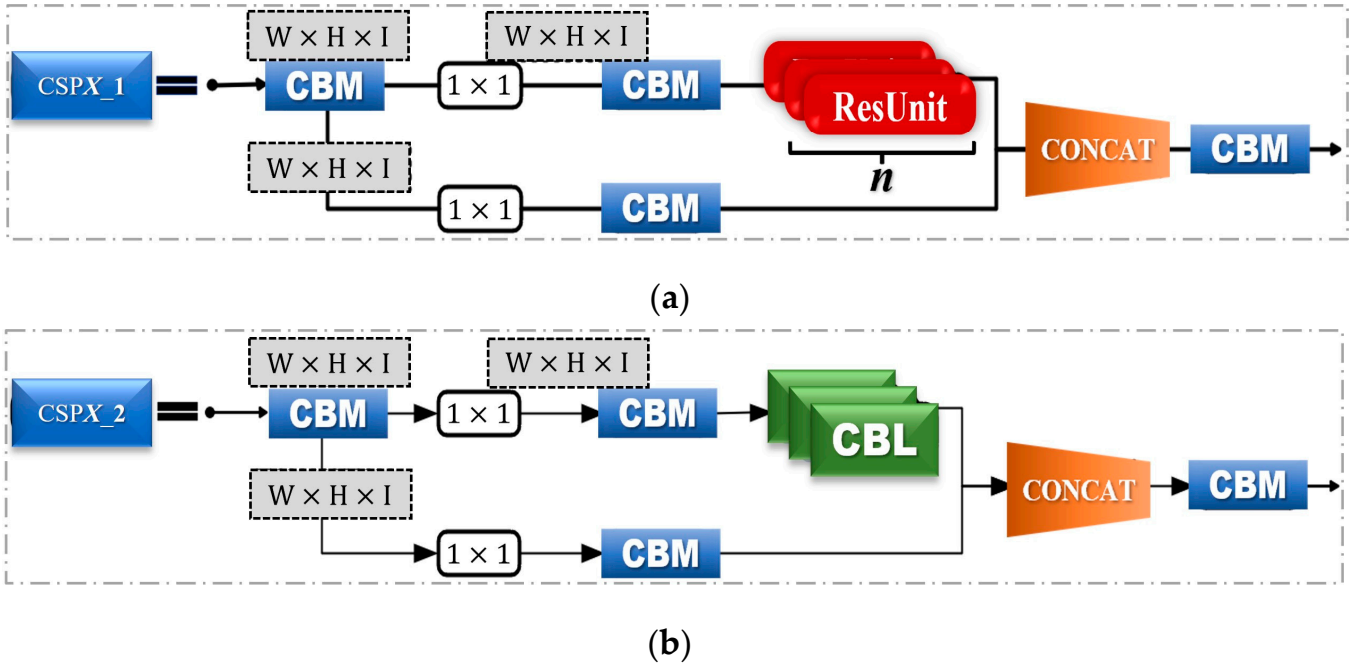


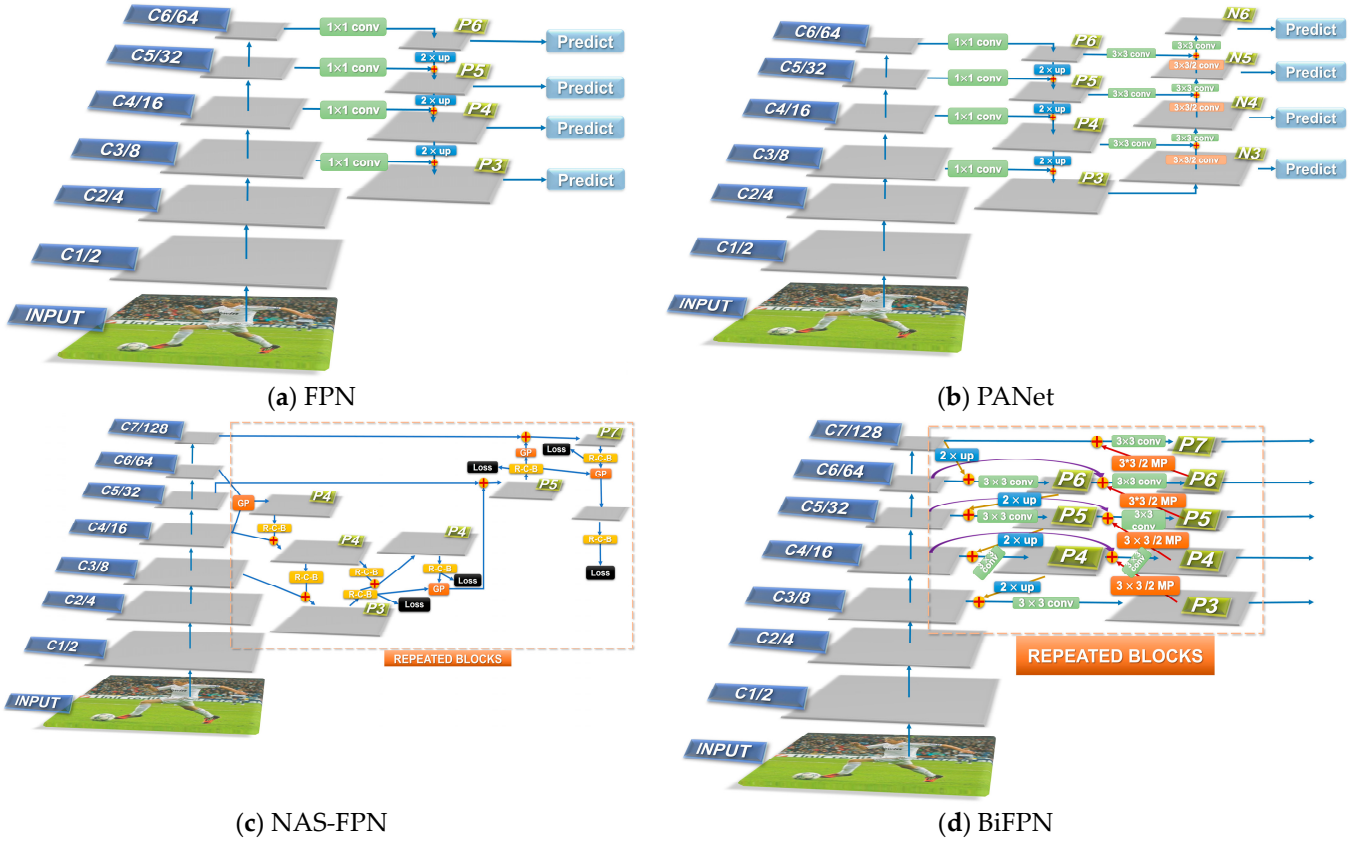
Figure 2. Improved CSPX_1 and CSPX_2 structures. (**a**) The basic unit of the backbone feature extraction network; (**b**) the picture shows the basic unit of the feature fusion structure.

3.2. The Structure of BiFPN

YOLO network's first characteristic fusing method was influenced by the Feature Pyramid Network (FPN) structure and introduced into YOLOV3's Neck network. Although many cross-scale feature fusion network structures have been developed since FPN, such as Path Aggregation Network and Neural Architecture Search-feature Pyramid Networks used in the YOLO series, these structures are used to fuse feature information extracted from different network levels. However, these input feature information from different network levels often has different resolutions, so the input feature information of different scales also has unequal contributions to the output feature information after fusion. Therefore, it is a reasonable solution to weigh the feature information of different scales when fusing features, allowing the network adaptively to learn the weight during training.

This is one of the main design ideas of BIFPN [32], which introduces weights that can be adaptively learned by the network to distinguish the importance of characteristic details at different scales for effective feature layers. In addition, the feature fusion mechanism, rooted in the FPN concept, clarifies the importance of feature fusion between different levels, but these methods all use simple upper- and lower-layer connections as a fusion method and do not consider the problem of excessive abstraction of feature information caused by such repeated fusion, so BiFPN also proposes a bidirectional cross-scale connection feature fusion mechanism. By adding a skip connection to the network at the same level, the underlying semantic characteristic is enhanced, and since the connected characteristic is at the same network level, it does not introduce too much computational cost. Figure 3 illustrates FPN's structures and its variants.

The feature fusion network described in this paper fully absorbs the idea of BiFPN, introduces the bidirectional feature fusion mechanism into the Neck network structure of YOLOv4, and makes adaptive improvements to the structure of the YOLO network.



(**a**) FPN (**b**) PANet

(**c**) NAS-FPN (**d**) BiFPN

Figure 3. Comparison of BiFPN network and other feature fusion methods.

3.3. Improved Remote Sensing Image Object Detection Model YOLOV4_Bi

After the improvement of the CSP and the Neck network structure, YOLOV4_Bi shown in Figure 4 is obtained.

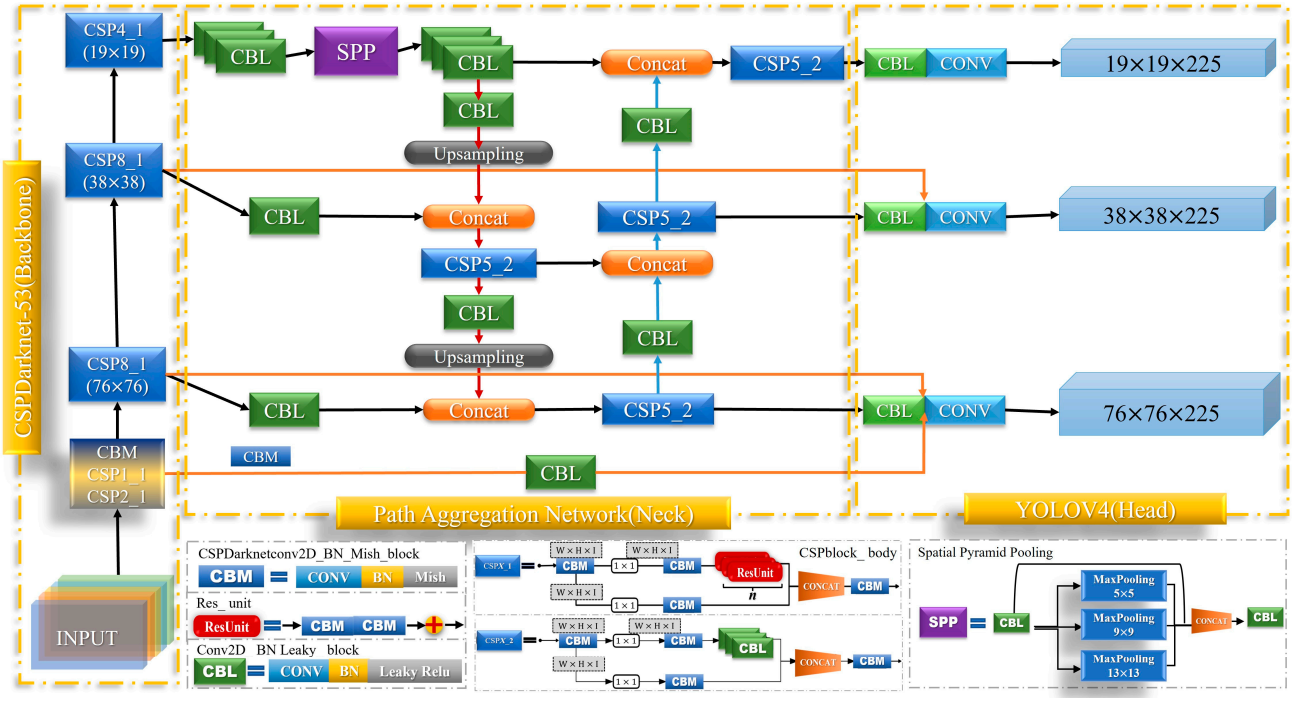


Figure 4. Improved YOLOV4_CPSBi network structure.

The orange data flow in the figure represents the horizontal skip connection, and the red and blue data flow represents the bidirectional cross-scale connection. As mentioned above, considering the fine-grained and densely distributed features exhibited in optical remote sensing images, the YOLOV4_CPSBi network proposed in this paper removes the horizontal skip connection at the top of the traditional YOLOV4 network and transfers the computation of the connection to the bottom of the network. It provides more abundant underlying feature information for the small target detector. This improvement effectively improves the detection ability of small-scale objects in optical remote sensing images without escalating computational complexity.

4. Experiment

4.1. Performance Index

Table 1 provides a comprehensive overview of the common performance indicators employed in this study for the target detection model [33]. Specifically, the main performance metric utilized in this experiment is mAP@0.5, which is computed based on the precision and recall rates. The recall rate (Recall) signifies the proportion of accurately detected targets in relation to total similar targets present in the test dataset, and its calculation is defined in Equation (5). On the other hand, precision indicates the ratio of correctly identified objects to all the detected objects, and its calculation is described by Equation (6).

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Table 1. Common performance metrics.

True Label	Prediction Results	Common Performance Metrics
True	Positive	TP
True	Negative	TN
False	Negative	FN
False	Positive	FP

The data of *Recall* and *Precision* are formed into two tuples. Under the determined IoU threshold, the area enclosed by all the two tuples above the two-dimensional coordinate axis is calculated as each category. The AP are averaged together to derive the mAP index value, which is defined in Equations (7) and (8).

$$AP = \int_0^1 P(y)dy \tag{7}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{8}$$

4.2. Software and Hardware Environment

All the experimental data mentioned in this study were generated in the environment shown in Tables 2 and 3.

Table 2. Software environment.

Software Environment	Version
Python	V 3.6+
TensorFlow	V 2.7.0
CUDA	V 11.2
CuDNN	V 8101
Matplotlib	V 3.5.0
TensorBoard	V 2.7.0
Operating System	Ubuntu 20.04.2 LTS

Table 3. Hardware environment.

Hardware	Model	Performance Parameters
GPU	RTX 3090 × 1	25.4 GB
CPU	Xeon Gold 6142 × 6	2.6~3.7 GHz
RAM	DDR5 6000 16 GB × 4	64 GB
ROM	SAMSUNG 980 PRO	400 GB

4.3. Pre-Training Parameter Settings

This experiment employed a pre-training approach, where the main portion of the network underwent parameter updates on the VOC dataset. Despite the VOC dataset consisting of natural images, the universality of image features allows for increased training efficiency. This approach also serves to prevent significant oscillations in model performance during training. The training of the pretrained model, including the number of initially frozen epochs and other parameter configurations during network training, is outlined in Table 4.

Table 4. Network training parameters.

Parameter	Setting	Related Parameter Setting
Freeze_Epoch	50	YOLOV4_CSPBi triggers EarlyStop at 41 YOLOV4 triggers EarlyStop at 25 BatchSize: 12 Freeze_lr is set to 0.001
UnFreeze_Epoch	150	Turn off the EarlyStop mechanism BatchSize: 6 UnFreeze_lr set to 0.0001
Optimizers	Adam	CosineDecayRestarts, initialize Lr to 0.0001

5. Results

5.1. Quantitative Analysis of the Detection Performance of YOLOV4_CSPBi

This section uses the parameters provided in Section 4.3 for quantitative analysis of the detection performance of YOLOV4_CSPBi. Figure 5 shows the Loss curves of the training process of the proposed model and the baseline model.

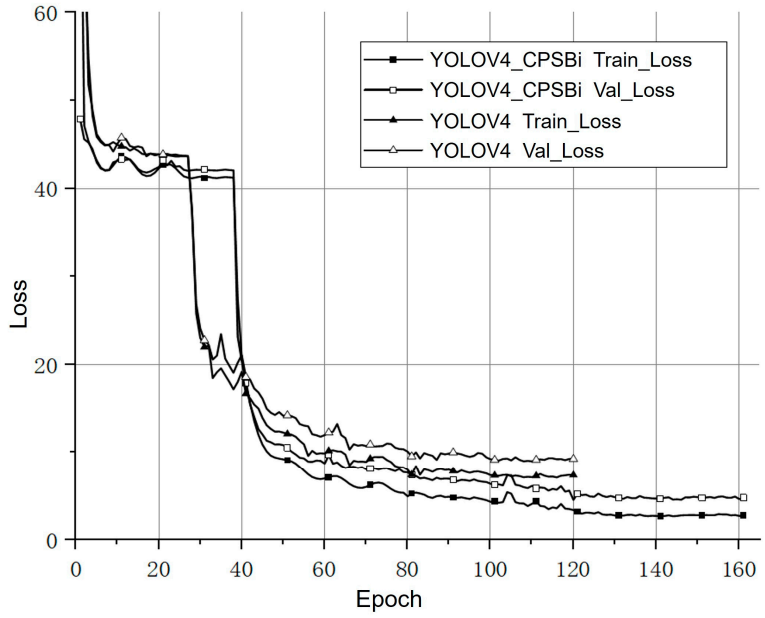


Figure 5. Loss curves of two network training processes.

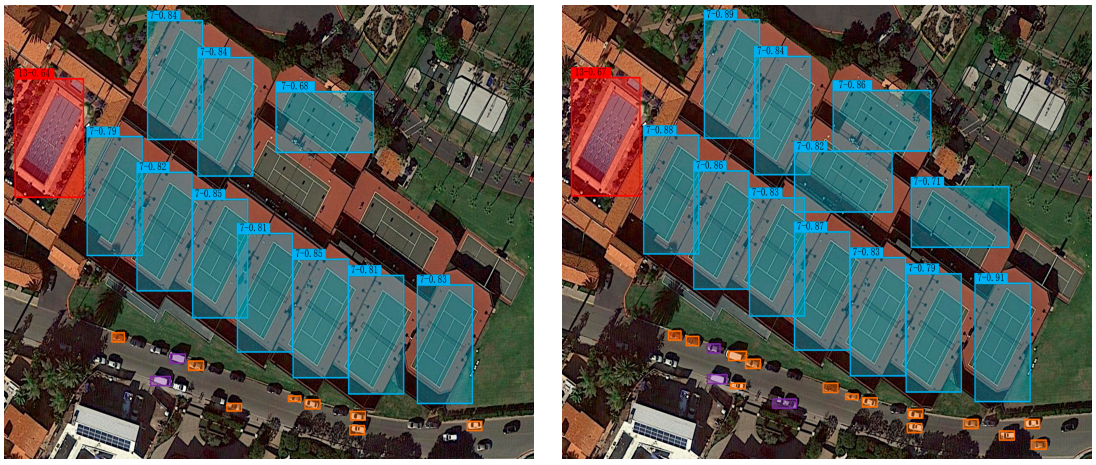
Figure 5 presents that during the phase of freezing the backbone characteristic extraction network due to the low degree of feature fusion of the traditional YOLOV4 network, the convergence of this training stage is entered earlier. The proposed YOLOV4_CPSBi model, by enhancing the reusability of feature details within the feature fusion network, triggers the Early_Stop mechanism later in the freezing stage of the backbone network and has a stronger learning ability.

From the experimental findings presented in Table 5, it is evident that YOLOV4_CPSBi exhibits significantly enhanced target detection capabilities compared to YOLOV4.

Table 5. Comparison of mAP and FPS indicators of the two networks.

Target Category	YOLOV4	YOLOV4_CPSBi
Plane	84.92	88.47
Baseball Diamond	79.58	80.17
Bridge	46.62	48.73
Ground Track Field	71.78	75.51
Small Vehicle	70.67	73.38
Large Vehicle	63.29	69.57
Ship	77.37	79.42
Tennis Court	86.71	87.96
Basketball Court	81.69	81.2
Storage Tank	70.73	72.88
Soccer Ballfield	61.94	62.39
Roundabout	60.11	63.03
Harbor	71.08	77.71
Swimming Pool	68.27	76.25
Helicopter	48.81	55.44
TOTAL_mAP@0.5	69.57	72.8
FPS	45	38

The mAP metric has demonstrated an improvement of approximately 3.2% compared to the traditional YOLOv4 model. Moreover, across the majority of object categories, the proposed model exhibits notably superior detection performance. In the case of four specific object categories—large vehicles, harbors, ships, and helicopters—the detection AP has achieved enhancements of 6.3%, 6.6%, 8.0%, and 6.6%, respectively. These results vividly highlight the efficacy of the various enhancements introduced in YOLOv4_CPSBi. Figure 6 presents a visual illustration comparing the final detection outcomes.



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

Figure 6. Visual display of two network detection results (Red box: swimming pool; Blue box: Tennis court). (**a**) YOLOV4 detection effect; (**b**) YOLOV4_CPSBi detection effect.

Figure 6a showcases the detection effect obtained using the conventional YOLOV4 network, while Figure 6b portrays the detection effect achieved by the YOLOV4_CPSBi network. The detection confidence of the traditional YOLOV4 network is generally lower than that of YOLOV4_CPSBi, and three tennis courts and a large number of car targets are missed. Among them, the reason for the missed inspection of the tennis courts is that the three tennis courts that were missed are all because their directions have changed greatly. However, the dataset used in this experiment has been enhanced and expanded on the features of rotation in similar directions, so it can be shown that the YOLOV4_CPSBi model has a stronger learning ability for this rotation difference feature than the traditional YOLOV4 network.

5.2. Ablation Experiment

This section primarily focuses on validating the impact of the Focal Loss [34] and the two optimizers on both the baseline model and the proposed YOLOV4_CPSBi in terms of performance. Focal Loss is a specific loss function based on the target detection model, which addresses the uneven distribution of positive and negative samples while detecting the one-stage target. The loss function is a relatively simple sample image with a small loss weight. The method improves the detection success rate of difficult samples. Its formal expression is shown in Equation (9).

$$\begin{aligned} FL(p_t) &= -(1-p_t)^{\gamma} log(p_t) \\ p_t &= \begin{cases} p, & y=1 \\ 1-p, & \text{else} \end{cases} \end{aligned} \tag{9}$$

As shown in Table 6. Adam and SGD optimizers have little effect on model performance, but the addition of Focal Loss greatly diminishes the model's overall efficiency. This may happen because the Focal Loss incorrectly marks the correct samples of lower quality as Difficult, which instead makes the model pay more attention to some False Positives, resulting in an increase in the false positive rate. However, in view of the successful application of Focal Loss in the RetinaNet network, after fully studying its positive and negative sample calibration principle, it is theoretically possible to improve the difficult sample mining ability of the YOLO series network, which is also one of the directions that should be continued in the future.

Table 6. Ablation experiment results.

Focal Loss	Optimizer	YOLOV4 Map@0.5	YOLOV4_CPSBi Map@0.5	Change
\	Adam	69.57	72.80	—
√	Adam	61.64	66.26	↓
√	SGD	62.38	65.77	↓
\	SGD	70.44	72.45	~

In Table 6, \ represents that Focal Loss is not used, and √ represents that Focal Loss is used. — represents no change; ~ represents the same level and little change; ↓ represents performance degradation.

5.3. Quantitative Analysis Detection Capability for Small and Medium-Sized Objects

Since YOLOV4_CPSBi has made adaptive improvements for remote sensing images with small and dense targets, this paper focuses on the comparison of the detection performance in the DOTA dataset specifically for small and medium-scale objects, employing the same type of network. The scales are divided into small-scale objects (S, image size less than 56 × 56 pixels), medium-scale objects (M, image size less than 126 × 126 pixels), and large-scale objects (L, image size larger than 126 × 126 pixels).

In this experiment, YOLOV3, YOLOV4, SSD, and RetinaNet of the same type as YOLOV4_CPSBi were selected as comparison models. These models are all representative

networks of the one-stage target detection model. The main advantage is that the network performs the inference task of classification and localization, and the speed of object detection is higher than that of other types of networks.

Table 7 presents the comparative data of this experimental study, wherein AP_L, M, and S denote the average precision values for large, medium, and small scales, respectively.

Table 7. Comparison of the same type of network for medium and small-scale target detection.

Detection Model	Backbone Network	Total_AP	AP_L	AP_M	AP_S
YOLOV3	Darknet-53	61.16	69.72	63.44	50.32
YOLOV4	CSPDarknet-53	69.57	77.58	70.42	60.71
SSD	VGG-16	63.71	71.93	64.77	54.43
YOLOV4_CPSBi	CSPDarknet-53	72.8	77.47	72.42	68.51
S2ANet	ResNet-152	73.97	80.41	71.66	69.84

It is evident from the results that YOLOV4_CPSBi, which eliminates the horizontal feature fusion connection in the large-size target detector, exhibits a slight decrease in AP_L. However, it demonstrates a notable enhancement in the detection capability of AP_M and AP_S, particularly in detecting small targets, where its AP performance shows an improvement of nearly 8% compared to YOLOV4.

S2ANet is specifically designed for rotating object detection in aerial images, so it may perform better in the presence of some oriented objects. However, the method proposed in this article is designed for small and dense objects rather than directional objects, so there are certain differences from S2ANet on the DOTA dataset.

Based on the favorable performance indicated in the table, this paper selects YOLOV4 and YOLOV4_CPSBi for further visual comparison when faced with a substantial amount of small and densely distributed detection samples, as depicted in Figure 7.

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

Figure 7. Visual display of detection results of YOLOV4 and YOLOV4_CPSBi for small targets (Orange boxes indicate Large-Vehicle (LV) targets, purple boxes represent Small-Vehicle (SV) targets). (**a**) YOLOV4; (**b**) YOLOV4_CPSBi.

Figure 7a represents the detection outcomes of the YOLOV4 network, while Figure 7b corresponds to the detection results of the YOLOV4_CPSBi network. Notably, the orange boxes indicate Large-Vehicle (LV) targets, while the purple boxes represent Small-Vehicle (SV) targets. It is evident from the visual analysis that YOLOV4_CPSBi exhibits higher

detection accuracy for small targets. YOLOV4 missed 6 cases of SV targets and 1 case of LV targets, while YOLOV4_CPSBi only missed 2 cases of SV targets.

Moreover, in Figure 7, a noticeable observation can be made regarding the stronger robustness of the YOLOv4_CPSBi network. The YOLOv4 network demonstrates a significant number of false positives in the detection image, inaccurately classifying 13 instances of negative samples as Small-Vehicle targets and 1 instance of a negative sample as a Large-Vehicle target. In contrast, the detection results of YOLOv4_CPSBi do not exhibit any misclassifications. This further validates that the improvements made in this study to the YOLOv4 network's feature fusion aspect have successfully enhanced the model's detection performance and robustness concerning medium to small-scale objects.

6. Discussion

This paper preliminarily discusses the application and improvement of YOLO series networks in target detection in optical remote sensing images. This study introduces a novel network architecture tailored for the task of small object detection in remote sensing images, thereby improving the accuracy of the traditional YOLOv4 model in this field.

The key aspects encompass:

(1) Enhanced Dataset Construction: Leveraging the DOTA dataset as the foundational dataset, several strategic approaches were employed to enhance and refine optical remote sensing images. These strategies encompassed techniques such as sample segmentation utilizing a sliding window concept, pixel-level data augmentation, mosaic-based enhancements, and the implementation of GridMask with the concept of area random erasing. These efforts culminated in a dataset containing 87,382 sample images with dimensions of 832 × 832, encompassing a total data size of 87.2 GB and comprising 334,585 target boxes. Comparative experiments substantiated the efficacy of the improved dataset in enhancing the learning capacity of pertinent deep learning neural networks.

(2) Augmented Feature Multiplicity: To bolster feature multiplexing, the study advanced the channel division methodology within the CSP structure. The enhanced CSP structure was incorporated within the backbone and neck components of YOLOV4 networks. Through the fusion of bidirectional multi-scale connections and a weighted feature fusion technique, the study effectively reallocated computational resources from large-scale target detection to small-target detection. This refinement significantly fortified the network's capacity to detect small targets, without incurring substantial computational overhead. The resultant enhanced model, denoted as YOLOV4_CPSBi, demonstrated a noteworthy 3.2% improvement in mean Average Precision (mAP) compared to the conventional YOLOV4 model. Particularly notable was the approximately 8% enhancement in Average Precision (AP) performance pertaining to small target detection when juxtaposed with YOLOV4.

However, certain limitations persist, as outlined below:

(1) Potential Expansion of Rotation Angle Index: To further optimize densely arranged target detection, it is conceivable to augment the rotation angle index of the target, akin to methodologies employed in detectors such as SCRDet [35] and IENet [36], with the aim of achieving superior outcomes;

(2) Holistic Contextual Comprehension: Most existing target detection techniques predominantly rely on visual characteristics derived solely from the target, thereby neglecting the pivotal process of holistic image comprehension and contextual interpretation. While some endeavors have incorporated contextual and global information, they predominantly concentrate on visual attributes and lack the incorporation of high-level semantic knowledge, thus yielding diminished interpretability.

7. Conclusions

The detection of targets within optical remote sensing images holds significant implications across both civilian and military contexts. This study centers on the intricate

challenge presented by the identification of small and densely clustered land targets within such images. To address this challenge, we propose an enhanced target detection network, denoted as YOLOV4_CPSBi, which builds upon the conventional YOLOV4 network. This novel network architecture enhances convolution computations and augments the technique of feature fusion. Furthermore, it fosters improved feature information utilization by employing a bidirectional cross-scale weighted connection approach.

The efficacy of the proposed model approach is substantiated through an exhaustive array of comparative experiments, establishing its prowess in target detection. In particular, the mAP metric of the proposed model when applied to the DOTA dataset surpasses that of the conventional YOLOV4 model by a margin of 3.2%. This augmentation in performance is consistently pronounced across numerous target categories, with a notable boost observed in the detection accuracy for objects such as carts, ports, ships, and helicopters. Impressively, the detection AP for these target categories demonstrates improvements of 6.3%, 6.6%, 8.0%, and 6.6%, respectively.

Additionally, the YOLOV4_CPSBi model showcases remarkable advancements in its detection capabilities, particularly in the context of AP_M and AP_S, attributes that are especially pertinent to small-scale target identification. Comparative evaluation indicates a nearly 8% enhancement in AP performance when juxtaposed with YOLOV4. Collectively, these results validate the substantial enhancement that the proposed model brings to land object detection tasks in optical remote sensing images. Notably, its heightened robustness concerning medium and small-scale target recognition reinforces its utility and efficacy in this domain.

Author Contributions: Conceptualization, W.Z. and L.W.; methodology, J.L., J.T. and L.Y.; software, J.L. and J.T.; formal analysis, Z.Y., S.L. (Siyu Lu) and L.Y.; data curation, J.L. and J.T.; writing—original draft preparation, L.Y., S.L. (Siyu Lu) and W.Z.; writing—review and editing, L.Y., L.W. and W.Z.; visualization, J.T. and S.L. (Siyu Lu); resources, S.L. (Shan Liu) and Z.Y.; supervision, S.L. (Shan Liu) and L.W.; funding acquisition, W.Z. and S.L. (Shan Liu). All authors have read and agreed to the published version of the manuscript.

Funding: Supported by the Sichuan Science and Technology Program (2023YFSY0026, 2023YFH0004).

Data Availability Statement: The DOTA dataset, which serves as a valuable resource for validating the results obtained in this study, is publicly accessible at the following location: https://captain-whu.github.io/DOTA/dataset.html (accessed on 1 September 2022).

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

References

1. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
2. Weber, J.; Lefevre, S. A multivariate hit-or-miss transform for conjoint spatial and spectral template matching. In *Image and Signal Processing, Proceedings of the 3rd International Conference, ICISP 2008, Cherbourg-Octeville, France, 1–3 July 2008*; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2008; pp. 226–235.
3. Van der Meer, F. Remote-sensing image analysis and geostatistics. *Int. J. Remote Sens.* **2012**, *33*, 5644–5676. [CrossRef]
4. Wang, L.; Tang, J.; Liao, Q. A Study on Radar Target Detection Based on Deep Neural Networks. *IEEE Sens. Lett.* **2019**, *3*, 7000504. [CrossRef]
5. Zhou, L.; Liu, J.; Chen, L. Vehicle detection based on remote sensing image of Yolov3. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 468–472. [CrossRef]
6. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J. Sig-NMS-based faster R-CNN combining transfer learning for small target detection in VHR optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8534–8545. [CrossRef]
7. Zhang, D.; Zhan, J.; Tan, L.; Gao, Y.; Župan, R. Comparison of two deep learning methods for ship target recognition with optical remotely sensed data. *Neural Comput. Appl.* **2021**, *33*, 4639–4649. [CrossRef]
8. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
9. Ghorbani, F.; Ebadi, H.; Sedaghat, A. Geospatial Target Detection from High-Resolution Remote-Sensing Images Based on PIIFD Descriptor and Salient Regions. *J. Indian Soc. Remote Sens.* **2019**, *47*, 879–891. [CrossRef]

10. Cao, C.; Wu, J.; Zeng, X.; Feng, Z.; Wang, T.; Yan, X.; Wu, Z.; Wu, Q.; Huang, Z. Research on Airplane and Ship Detection of Aerial Remote Sensing Images Based on Convolutional Neural Network. *Sensors* **2020**, *20*, 4696. [CrossRef]
11. Xu, Y.; Zhu, M.; Xin, P.; Li, S.; Qi, M.; Ma, S. Rapid Airplane Detection in Remote Sensing Images Based on Multilayer Feature Fusion in Fully Convolutional Neural Networks. *Sensors* **2018**, *18*, 2335. [CrossRef]
12. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position Detection and Direction Prediction for Arbitrary-Oriented Ships via Multitask Rotation Region Convolutional Neural Network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
13. Wang, P.; Sun, X.; Diao, W.; Fu, K. FMSSD: Feature-Merged Single-Shot Detection for Multi-scale Objects in Large-Scale Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3377–3390. [CrossRef]
14. Li, M.; Guo, W.; Zhang, Z.; Yu, W.; Zhang, T. Rotated Region Based Fully Convolutional Network for Ship Detection. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 673–676. [CrossRef]
15. Fu, Y.; Wu, F.; Zhao, J. Context-Aware and Depthwise-based Detection on Orbit for Remote Sensing Image. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1725–1730. [CrossRef]
16. Zhang, W.; Wang, S.; Thachan, S.; Chen, J.; Qian, Y. Deconv R-CNN for Small Object Detection on Remote Sensing Images. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2483–2486. [CrossRef]
17. Schilling, H.; Bulatov, D.; Niessner, R.; Middelmann, W.; Soergel, U. Detection of Vehicles in Multisensor Data via Multibranch Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4299–4316. [CrossRef]
18. Liu, W.; Ma, L.; Wang, J.; xsChen, H. Detection of Multiclass Objects in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 791–795. [CrossRef]
19. Ying, X.; Wang, Q.; Li, X.; Yu, M.; Jiang, H.; Gao, J.; Liu, Z.; Yu, R. Multi-Attention Object Detection Model in Remote Sensing Images Based on Multi-Scale. *IEEE Access* **2019**, *7*, 94508–94519. [CrossRef]
20. Long, H.; Chung, Y.; Liu, Z.; Bu, S. Object Detection in Aerial Images Using Feature Fusion Deep Networks. *IEEE Access* **2019**, *7*, 30980–30990. [CrossRef]
21. Mastrorosa, S.; Crespi, M.; Congedo, L.; Munafò, M. Land Consumption Classification Using Sentinel 1 Data: A Systematic Review. *Land* **2023**, *12*, 932. [CrossRef]
22. Liu, Y.; Pan, X.; Liu, Q.; Li, G. Establishing a Reliable Assessment of the Green View Index Based on Image Classification Techniques, Estimation, and a Hypothesis Testing Route. *Land* **2023**, *12*, 1030. [CrossRef]
23. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Körner, M.; Reinartz, P. Towards multi-class object detection in unconstrained remote sensing imagery. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 150–165. [CrossRef]
24. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multi-scale Rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
25. Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
26. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13029–13038.
27. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
28. Yang, J. Gridmask based data augmentation for bengali handwritten grapheme classification. In Proceedings of the 2020 2nd International Conference on Intelligent Medicine and Image Processing, Tianjin, China, 23–26 April 2020; pp. 98–102. [CrossRef]
29. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 390–391.
30. Zhang, J.; Lu, C.; Li, X.; Kim, H.-J.; Wang, J. A full convolutional network based on DenseNet for remote sensing scene classification. *Math. Biosci. Eng.* **2019**, *16*, 3345–3367. [CrossRef]
31. Ju, C.; Guan, C. Tensor-cspnet: A novel geometric deep learning framework for motor imagery classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–15. [CrossRef]
32. 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.
33. Basha, S.M.; Rajput, D.S. Survey on evaluating the performance of machine learning algorithms: Past contributions and future roadmap. In *Deep Learning and Parallel Computing Environment for Bioengineering Systems*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 153–164.
34. 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; pp. 2980–2988.

35. Li, C.; Luo, B.; Hong, H.; Su, X.; Wang, Y.; Liu, J.; Wang, C.; Zhang, J.; Wei, L. Object Detection Based on Global-Local Saliency Constraint in Aerial Images. *Remote Sens.* **2020**, *12*, 1435. [CrossRef]
36. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459. [CrossRef] [PubMed]