electronics

Article

Small Object Detection Method Based on Weighted Feature Fusion and CSMA Attention Module

Chao Peng [1], Meng Zhu [2,3], Honge Ren [1,3,*] and Mahmoud Emam [4,*]

[1] College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China
[2] College of Information Engineering, Harbin University, Harbin 150086, China
[3] Heilongjiang Forestry Intelligent Equipment Engineering Research Center, Harbin 150040, China
[4] Faculty of Artificial Intelligence, Menoufia University, Shebin El-Koom 32511, Egypt
* Correspondence: rhe@nefu.edu.cn (H.R.); memam@ai.menofia.edu.eg (M.E.); Tel.: +86-13766805518 (H.R.); +20-1093210321 (M.E.)

Abstract: Small object detection is one of the challenging tasks in computer vision. Most of the existing small object detection models cannot fully extract the characteristics of small objects within an image, due to the small coverage area, low resolution and unclear detailed information of small objects in the image; hence, the effect of these models is not ideal. To solve this problem, a simple and efficient reinforce feature pyramid network R-FPN is proposed for the YOLOv5 algorithm. The learnable weight is introduced to show the importance of different input features, make full use of the useful information of different feature layers and strengthen the extraction of small object features. At the same time, a channel space mixed attention CSMA module is proposed to extract the detailed information of small objects combined with spaces and channels, suppress other useless information and further improve the accuracy of small object detection. The experimental results show that the proposed method improves the average accuracy AP, AP50 and AR100 of the original algorithm by 2.11%, 2.86% and 1.94%, respectively, and the detection effect is better than the existing small object detection algorithms, which proves the effectiveness of the proposed method.

Keywords: small object detection; deep learning; feature fusion; attention mechanism; YOLOv5 algorithm; CSMA module; R-FPN

Citation: Peng, C.; Zhu, M.; Ren, H.; Emam, M. Small Object Detection Method Based on Weighted Feature Fusion and CSMA Attention Module. *Electronics* **2022**, *11*, 2546. https://doi.org/10.3390/electronics11162546

Academic Editor: Youngbae Hwang

Received: 7 July 2022
Accepted: 11 August 2022
Published: 15 August 2022

1. Introduction

Object detection is one of the basic tasks of computer vision. Its main task is to identify and locate the objects that people are interested in within an image. Object detection has been widely used in many applications, such as remote sensing, autonomous driving, visual search, virtual reality (VR) and augmented reality (AR), etc. [1]. Before the rise of deep learning methods, object detection methods relied on artificially designed features and classifiers designed based on the way humans understand objects. With the development of convolutional neural networks (CNN) [2] and deep learning, deep learning methods based on convolutional neural networks are gradually replacing traditional object detection methods. Currently, object detection algorithms based on convolutional neural network are mainly divided into two types. The first type is the single-stage algorithm, such as YOLO and SSD (single shot detectors), which directly predicts the categories and positions of different objects in the input images through a convolutional neural network. The second type of algorithms are based on the proposal region, which is divided into two stages. Firstly, proposal regions are generated through the region proposal network (RPN), and then objects are classified, and regression based on the proposal region. After these two stages, detection results are obtained [3], such as Fast-RCNN (fast-region based convolutional neural network) and Faster-RCNN. While two-stage detectors tend to be more flexible and accurate, one-stage detectors are generally considered simpler and more effective by using predefined anchors. One-stage detectors attract great attention due to

their efficiency and simplicity. In this paper, we mainly follow the design of single-stage detector and improve it on its basis. Although accurate detection of medium and large objects in images was achieved in many applications, it is still challenging to accurately detect small objects. There are two definitions of small object: the first one states that when the object size is one tenth of the original image size, it can be defined as small object; the second method defines objects with pixels less than 32×32 as small objects. Small objects are difficult to detect due to their indistinguishable characteristics, low resolution, complex background, limited context information and other reasons.

At present, many scholars propose some effective methods to improve the performance of small object detection. Niu Haoqing et al. [4] proposed an improved YOLOv3 algorithm combining EGCA (efficient gating channel attention) and an adaptive upsampling module. The upsampling module of the original network structure is replaced by an adaptive upsampling module, and an EGCA attentional mechanism is added before the output of prediction results. The improved algorithm has good robustness and a strong ability to resist environmental interference. Zhao Pengfei et al. [5] proposed a focus mechanism and deep level of small object detection algorithm fusion method: this method replaces the residual connection structure in the backbone network with the packet residual connection structure, improves the output of the receptive field size in the multi-scale detection stage, and feature enhancement module and channel attention mechanism are used to fuse different feature layers solved the problem of the lack of characteristics of shallow semantic information. Qiu Nanhao et al. [6] removed large-scale object detection, and added the small-scale detection. At the same time, the method of intersection union ratio loss function is used to improve the effect of small object detection.

Inspired by the above-mentioned literature, this paper aims at solving the problem that the detection effect of the YOLOv5 algorithm is not ideal for small objects on the VisDrone-DET2021 dataset [7]. To improve the detection results of small objects based on the structure of the YOLOv5 algorithm, the main contribution of this paper can be categorized as follows:

(1) A reinforce feature pyramid network (R-FPN) is proposed for the YOLOv5 algorithm, aiming at the intrinsic problems of small objects' features being difficult to distinguish and low resolution in images. R-FPN learns different input features by introducing learnable weights. Multiple features are fused to enhance the algorithm's learning of small object features and perform more accurate position regression for small objects.
(2) A channel space mixed attention (CSMA) protocol is proposed for the small objects which contains few information factors and limited context information. This protocol is built behind CSPDarknet53 [8], in which the feature information of the small object is fully extracted to enhance the effective feature.

This paper improves the YOLOv5 model through the above two methods, in order to solve the problem of insufficient feature extraction of input image in small object detection, so as to improve the accuracy of small object detection results. Experimental results show that the proposed model can improve the accuracy of small object detection and obtain a better detection effect.

2. Reinforced Feature Pyramid Network R-FPN

The main aim of feature fusion is to combine the features extracted from the image into a more discriminant feature than the input feature. The earliest feature fusion is to directly extract the features of the pyramid for prediction, without the fusion of multiple features [9], but the accuracy is relatively low in this case. Therefore, feature pyramid networks (FPN) [10] were proposed, as shown in Figure 1. A top-down connection is set up in FPN for feature fusion, and the fused feature graph with higher semantic information is used for prediction, which can improve the detection accuracy.

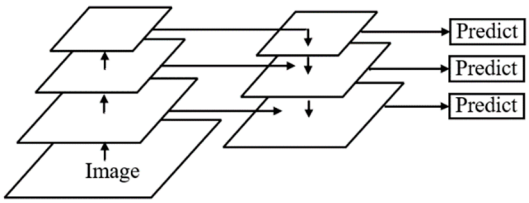

Figure 1. The structure of FPN.

In recent years, PANet [11] is the most widely used model for object detection. It adds a bottom-up pathway on the basis of FPN. The feature maps at the top of the pyramid have richer semantic information (conducive to object classification), while the feature maps at the bottom have stronger location information (conducive to object positioning). So, the FPN top-down structure makes the forecast figure improve the semantic information, but in theory, loses a lot of location information. Therefore, PANet creates a new bottom-up path based on the original FPN, the location information is also transmitted to the prediction feature map, which makes the prediction feature map have high semantic information and location information at the same time. The structure of PANet is shown in Figure 2.

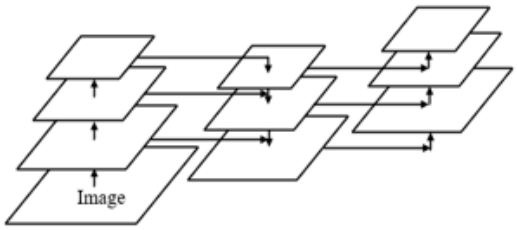

Figure 2. The structure of PANet.

Based on the development of feature fusion network, a new feature fusion network structure, reinforce feature pyramid network (R-FPN), is proposed in this paper. Based on PANet, the R-FPN builds an additional path between the original input node and the output node to fuse more features without increasing too much cost. The traditional feature fusion is often just a simple superposition or addition of feature maps, such as using concat or shortcut connection, treating all input features equally without distinguishing between them. However, different input feature graphs have different resolutions, and each feature graph contains different information, so their contribution to the fusion input feature graph is also different and simple addition or superposition is not the best operation. To sum up, this paper proposes a simple and efficient weighted special fusion mechanism. The specific structure is shown in Figure 3.

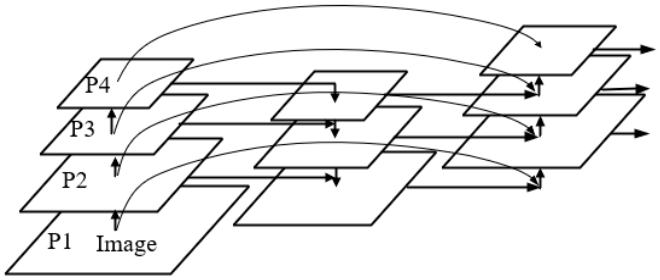

Figure 3. The structure of weighted feature fusion mechanism.

The calculation formula of the weighted feature fusion can be expressed as:

$$O = \sum_i \frac{w_i \times I_i}{\varepsilon + \sum_j w_j} \tag{1}$$

where O represents the output, w represents the learnable weight and I represents the characteristics of the input, ε represents the offset value. For example, the feature fusion of P_4 layer can be expressed as:

$$P_4^{td} = Conv\left(\frac{w_1 \cdot P_4^{in}}{w_1 + \varepsilon}\right) \tag{2}$$

$$P_4^{out} = Conv\left(\frac{w_1' \cdot P_4^{in} + w_2' \cdot P_4^{td} + w_3' \cdot P_3^{out}}{w_1' + w_2' + w_3' + \varepsilon}\right) \tag{3}$$

P_4^{td} is the intermediate feature of layer 4 in the top-down path, P_4^{out} is the output feature of layer 4 in the bottom-up path and P_3^{out} is the output feature of layer 3. The functional implementations of other layers are constructed in a similar way. In order to further improve efficiency, we use the feature fusion depth separable convolution. Deep separable convolution is an algorithm obtained by improving the standard convolution calculation in a convolution neural network, compared with the ordinary convolution, which not only reduces the amount of training required for participation in the process but implements the channel and regional separation and improves the training speed model. We add normalization and activation functions after each convolution to improve the stability of the network.

In addition, we found that in the feature extraction network, due to the stacking of convolutional layers, the receptive field would gradually increase, and the resolution of the feature map would decrease, with insufficient detail information. This makes it impossible to obtain the best results of feature map resolution and receptive field at the same time. To get rid of this dilemma, this paper proposes the context information extraction module (CEM). After obtaining the feature maps of the first few layers through the feature extraction network, we input them into CEM. CEM uses dense connections to extract the feature maps of different receptive fields by using the dilated convolution [12] of 3, 6, 12, 18 and 24 for the input features, and introduces deformable convolution for each connection, and deformable convolution is introduced for each connection to ensure that CEM can learn and transform different features from given data [13]. In addition, in order to fine-merge multi-scale information, we use a densely linked approach in CEM, where the output of each extension layer is connected to the input feature map, and then the input to the next extension layer. By means of dense linking, we obtain better multi-scale features. The CEM structure is shown in Figure 4.

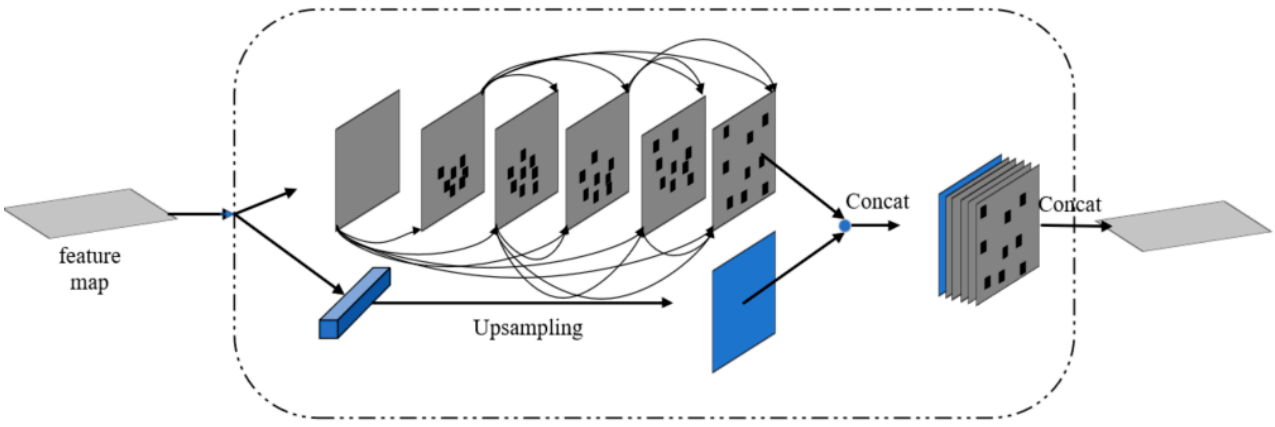

Figure 4. The structure of CEM.

We combine the CEM module with the weighted feature mechanism and put the CEM module after the high-level feature map of the pyramid to obtain the enhanced feature fusion network structure R-FPN. The structure of R-FPN is shown in Figure 5.

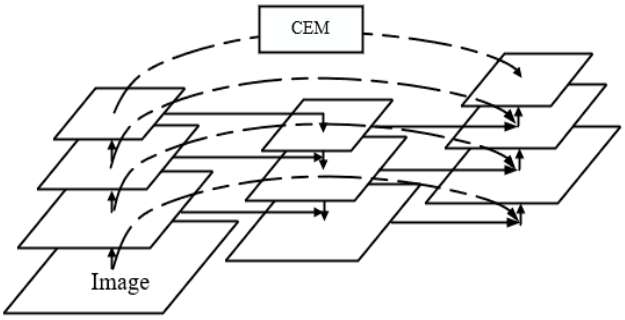

Figure 5. The structure of R-FPN.

R-FPN, by adding an extra weight for each input, makes the network learn the importance of each input characteristic in the process of training and makes the information in the network make full use of each input characteristic. It also solves the problem that the resolution and receptive field cannot, achieves the best results at the same time in the process of feature extraction and enhances model learning ability, so as to improve the accuracy of small object detection.

3. CSMA Attention Module

It has become an important means to improve the performance of deep neural networks by adding attention mechanism to neural networks, which can accurately focus on all relevant elements input by the outside world [14]. There are two kinds of attention mechanisms widely used in computer vision research: spatial attention and channel attention. Channel attention focuses on "what" is the meaningful input image, while spatial attention focuses on "where" the most informative part is. Spatial attention is a supplement to channel attention [15]. Spatial attention and channel attention complement each other to improve feature extraction performance. Although fusing them together may yield better results than they would individually, this inevitably increases computational overhead. In this paper, we propose a CSMA attention module to solve this problem that uses a hybrid unit to effectively combine the two types of attention mechanisms.

For a given input feature map $X \in \mathbb{R}^{C\times H\times W}$, where C, H and W represent the channel number, the height and width of the feature map, respectively, CSMA first divides X into G groups along the channel dimension, $X = [X_1, \cdots, X_G]$, $X_k \in \mathbb{R}^{C/G\times H\times W}$, in which each sub feature X_k gradually captures specific semantic information in the training process. Then, we generate the corresponding importance coefficient for each sub feature through the attention module. At the beginning, the input of X_k in each attention unit is divided into two branches along the channel dimension, $X_{k1}, X_{k2} \in \mathbb{R}^{C/2G\times H\times W}$. As shown in Figure 6, the first branch generates a channel attention diagram using the relationship among channels; the second branch uses the spatial relationship between features to generate spatial attention map. Therefore, it is meaningful for the model to focus on "what" and "where".

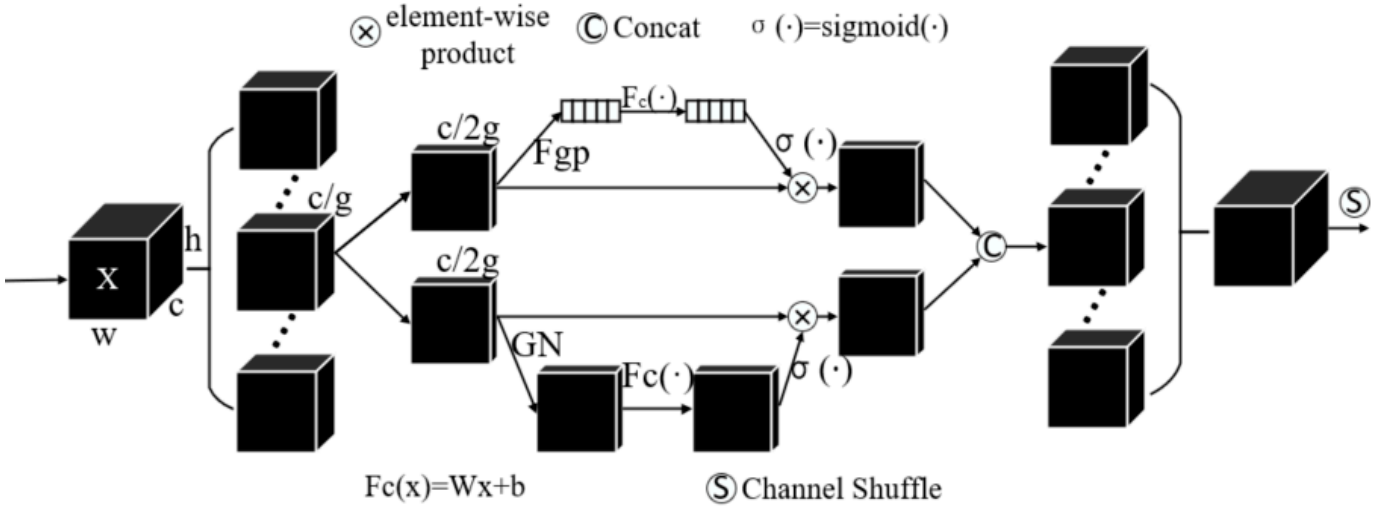

Figure 6. The structure of CSMA.

Detailed steps are described below. First, we use global average pooling (GAP) to embed global information and generate the channel feature of $s \in \mathbb{R}^{C/2G \times 1 \times 1}$. This is calculated by shrinking X_{k1} in the direction of height H and width W:

$$s = F_{gp}(X_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{k1}(i, j) \tag{4}$$

A compact feature is also created through a simple fully connected layer with sigmoid activation functions. The final output of channel attention can be expressed as:

$$X'_{k1} = \sigma(F_c(s)) \cdot X_{k1} = \sigma(W_1 s + b_1) \cdot X_{k1} \tag{5}$$

where $W_1 \in \mathbb{R}^{C/2G \times 1 \times 1}$ and $b_1 \in \mathbb{R}^{C/2G \times 1 \times 1}$ are used for scaling and moving s.

Different from channel attention, spatial attention focuses on "where" and is the supplement of channel attention. First, we use the Group Norm (GN) to obtain the spatial-wise statistics. Then, we use $F_c(\cdot)$ to enhance feature representation in X_{k2}. The final output of spatial attention can be expressed as:

$$X'_{k2} = \sigma(W_2 \cdot GN(X_{k2}) + b_2) \cdot X_{k2} \tag{6}$$

$W_2 \in \mathbb{R}^{C/2G \times 1 \times 1}$, $b_2 \in \mathbb{R}^{C/2G \times 1 \times 1}$. The two branches are then connected so that the number of channels is the same as the number of entered channels.

After that, all the sub-features are aggregated, and we use a "channel mixing" operation, which scrambles the feature map channel order so that cross-group information flows along channel dimensions. The final output of the CSMA module is the same size as X, making CSMA easy to integrate with modern architectures. Six hyperparameters W_1, b_1, W_2, b_2 and GN (which has two hyperparameters) are introduced in the CSMA module, and the number of channels in each branch of a single CSMA module is $C/2G$. As a result, the total number of parameters is $3C/G$ (G usually uses 32 or 64), which is trivial compared to the millions of parameters in the entire network. In this paper, the CSMA module is built on the backbone network structure of CSPDarknet53, and the model structure is shown in Figure 7.

The CSMA module enables the network to pay more attention to the small objects area and extract the feature information of the small objects from the spatial and channel dimensions, respectively. The feature covers more parts of the objects to be recognized, making the final probability of identifying the object higher. At the same time, effective features in the low-level feature map are strengthened, invalid features and noise are

suppressed, and more details of small objects are obtained. Furthermore, the detection accuracy of the proposed algorithm for the small objects' detection is improved. In order to compare the performance of the algorithm with different attention, Table 1 lists the specific detection results.

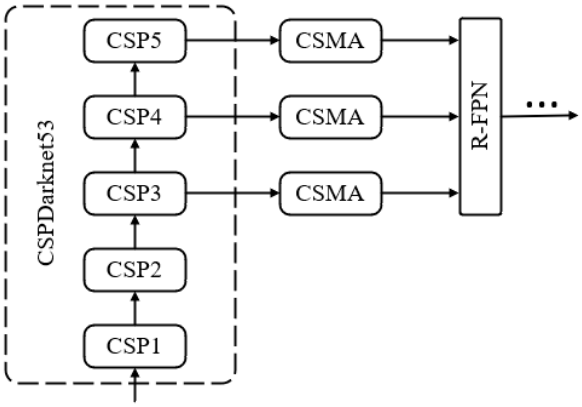

Figure 7. The position of CSMA.

Table 1. Object Detection Result of Different Attention Method in VisDrone-DET2021 Dataset.

Method	Params	GFLOPs
YOLOv5	7.325 M	17.312
+SE	9.848 M	17.320
+CBAM	9.858 M	17.329
+SGE	7.327 M	17.317
+SK	7.923 M	17.377
+CSMA(Ours)	7.325 M	17.315

4. Experimental Results

4.1. Dataset

The Visdron-DET2021 dataset was selected for experimental results in this paper, which is a dataset of 8599 images taken at different locations and at different heights using drones. In addition, more than 540 K object bounding boxes were annotated into 10 predefined categories. The dataset was divided into training, validation and testing sets (6471 for training, 548 for validation and 1580 for testing), with subsets from different sites but similar environments. Since the object is usually very small in the *uav* scene, we use it as a dataset for small object detection.

4.2. Model Training

The experiment was carried out under the Ubuntu16.04.12 operating system with Pytorch as the deep learning framework and Python as the development language. Hardware configuration includes Intel (R) Xeon (R) Gold 5218R CPU, 2.10 GHz, 64 GB memory. GPU is two NVIDIA RTX 2080 Ti, 11 GB display memory.

The total number of iterations in this experiment was 200, the iteration batch size was set to 32, and the optimizer selected SGD. During model training, Warmup was used for the learning rate to slow down the over-fitting phenomenon caused by a too-small data amount in the initial stage of the model and avoid model oscillation to ensure the deep stability of the model. After Warmup, cosine annealing learning algorithm is used to update the learning rate.

Detailed steps are described below. First, we use global average pooling (GAP) to embed global information and generate the channel feature of $s \in \mathbb{R}^{C/2G \times 1 \times 1}$. This is calculated by shrinking X_{k1} in the direction of height H and width W:

$$s = F_{gp}(X_{k1}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{k1}(i, j) \tag{4}$$

4.3. Performance Evaluation

In this paper, a number of experiments were conducted on the VisDrone-DET2021 dataset to verify the detection performance of the proposed small object detection algorithm. In order to compare and evaluate the detection accuracy of YOLOv5 model and the three models trained in this paper for this dataset, average precision (AP), AP50 and AR were used as an evaluation metrics.

AP refers to the area enclosed by the precision–recall (P–R) curve and the first quadrant of coordinates. Generally, AP is not directly calculated for the P–R curve in practical application, but is smoothed for P–R curve; that is, the maximum accuracy value on the right of each point on P–R curve is selected. Then, the smoothed accuracy value is used for AP calculation, and the calculation formula can be expressed as:

$$\mathrm{AP} = \int_0^1 P_{smooth}(r) dr \tag{7}$$

$$P_{smooth}(r) = \max_{r\prime \geq r} P(r\prime) \tag{8}$$

AP50 is the AP value when the Intersection-over-Union (IoU) value is 0.5. IoU is the overlap rate of candidate bound and ground truth bound in object detection, and the ratio of their intersection to union: whereas AR refers to the maximum recall in a given fixed number of test results in each picture.

The structure of the improved model is shown in Figure 8. The experimental data are shown in Tables 2 and 3. The YOLOv5 algorithm of the Pytorch version was selected as the comparison baseline model. It can be seen that, compared with the object detection accuracy of the YOLOv5 algorithm on the VisDrone-DET2021 dataset, the average accuracy of the model with R-FPN improved by 1.15% AP and 1.28% AP50 in this paper. The average accuracy of the model with the CSMA module AP increased by 1.96% and AP50 increased by 2.54%. After the comprehensive improvement, the average accuracy of the model AP is increased by 2.11% and AP50 is increased by 2.86%. The results show that the proposed method effectively extracts the feature information of small objects and improves the small object detection accuracy of the YOLOv5 algorithm.

Table 2. Detection Precision of Different Models.

Algorithm	AP (%)	AP50 (%)	AR1 (%)	AR10 (%)	AR100 (%)	AR500 (%)
YOLOv5 (Pytorch)	28.88	49.33	0.48	3.19	11.01	45.32
YOLOv5 + R-FPN	30.03	50.61	0.35	3.03	12.17	43.53
YOLOv5 + CSMA	30.84	51.87	0.48	3.22	11.57	43.76
Our method	**30.99**	**52.19**	**0.51**	**3.25**	**12.95**	44.02

Table 3. Performance for Different Input Combinations of Refinement Network.

Method	R-FPN	CEM	CSMA	AP (%)	AP50 (%)	AR1 (%)	AR10 (%)	AR100 (%)	AR500 (%)
YOLOv5 (Pytorch)				28.88	49.33	0.48	3.19	11.01	45.32
Method1	√	×	×	29.13	50.73	0.42	3.38	10.94	44.02
Method2	×	√	×	28.59	50.07	0.50	3.36	9.46	42.52
Method3	×	×	√	30.84	51.87	0.48	3.22	11.57	43.76
Method4	√	√	×	30.03	50.61	0.35	3.03	12.17	43.53
Method5	√	×	√	29.62	49.56	0.47	2.96	10.23	42.28
Method6	√	√	√	**30.99**	**52.19**	**0.51**	**3.25**	**12.95**	44.02

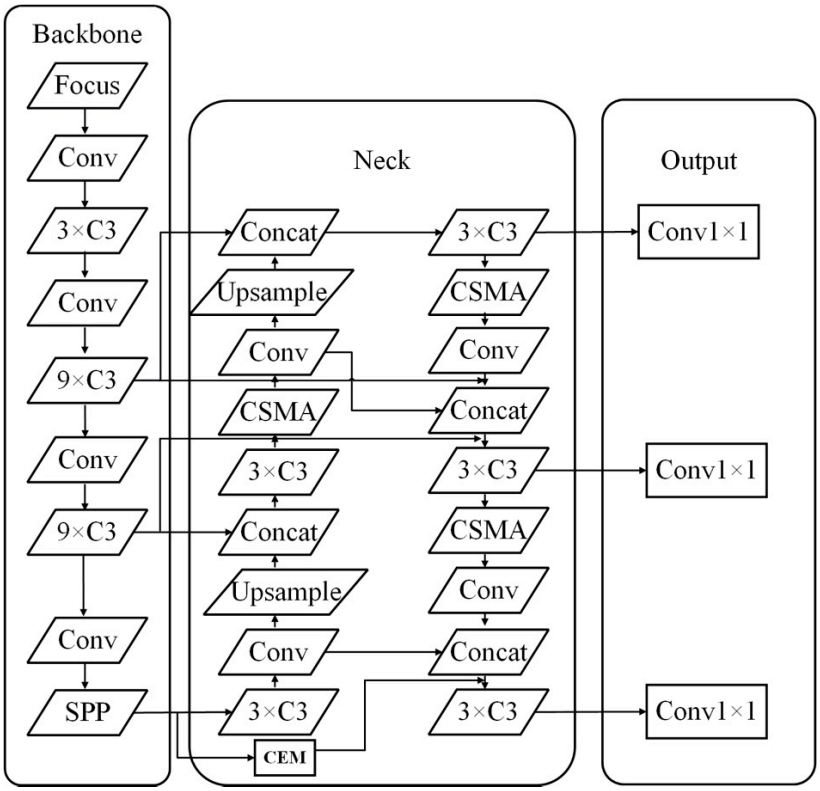

Figure 8. The Structure of the improved model.

In order to compare the performance of each algorithm, Table 4 lists the specific detection results of the latest small object detection algorithms, such as: RetinaNet [16], RefineNet [17], DetNet [18], CornerNet [19], IENet [20], DMNET [21], CascadeNet [22], GLSAN [23] and CenterNeSt [24] on the VisDrone-Det2021 dataset. Table 4 lists the AP scores of each object category on the VisDrone-DET2021 test set.

Table 4. Analysis of New Small Object Detection Algorithms in VisDrone-DET2021 database.

Algorithm	AP (%)	AP50 (%)	AR1 (%)	AR10 (%)	AR100 (%)	AR500 (%)
RetinaNet [16]	11.81	21.37	0.21	1.21	5.31	19.29
RefineNet [17]	14.90	28.76	0.24	2.41	18.13	25.69
DetNet [18]	15.26	29.23	0.26	2.57	20.87	22.28
CornerNet [19]	17.41	34.12	0.39	3.32	24.37	26.11

Table 4. *Cont.*

Algorithm	AP (%)	AP50 (%)	AR1 (%)	AR10 (%)	AR100 (%)	AR500 (%)
IENet [20]	29.13	51.33	0.50	3.38	15.59	42.72
DMNET [21]	28.20	47.60	0.43	3.01	10.23	42.88
CascadeNet [22]	30.12	51.02	0.41	2.96	7.78	46.81
GLSAN [23]	30.70	50.40	0.47	3.25	14.91	38.53
CenterNeSt	30.03	51.69	0.57	3.91	21.40	43.14
YOLOv5 (Pytorch)	28.88	49.33	0.48	3.19	11.01	45.32
Our method	**30.99**	**52.19**	**0.51**	3.25	12.95	44.02

In view of all precision data in Table 4, we can notice that the accuracy of the proposed algorithm is superior to the other state-of-the-art algorithms. In terms of average accuracy, the AP value of the proposed algorithm is improved to 30.99% compared with other existing algorithms. The AP50 value also recorded at 52.19%. Table 5 shows the analysis of The AP Scores of each object category in the VisDrone-DET2021 dataset.

Table 5. Analysis of The AP Scores of Each Object Category in VisDrone-DET2021 dataset.

Algorithm	Pedestrian	Person	Bicycle	Car	Van	Truck	Tricycle	Awning-Tricycle	Bus	Motor
RetinaNet	9.91	2.92	1.32	28.99	17.82	11.35	10.93	8.02	22.21	7.03
RefineNet	14.90	3.67	2.02	30.14	16.33	18.13	9.03	10.25	21.93	8.38
DetNet	15.26	4.07	3.13	36.12	17.29	20.87	13.52	10.45	26.01	10.92
CornerNet	20.43	6.55	4.56	40.94	20.23	20.54	14.03	9.25	24.39	12.10
IENet	27.76	11.72	7.97	50.72	36.38	28.04	19.61	18.50	35.32	21.28
DMNET	23.94	12.80	10.03	43.89	32.43	29.02	28.45	20.30	42.56	21.90
CascadeNet	17.75	5.08	3.54	42.01	26.50	22.58	15.96	12.71	33.28	12.31
GLSAN	17.49	6.81	2.59	39.17	25.17	14.41	11.02	8.64	18.11	11.84
CenterNeSt	27.99	11.61	9.02	51.03	36.52	27.88	20.09	19.88	37.71	20.09
YOLOv5 (Pytorch)	28.44	14.71	7.20	49.38	36.18	27.79	23.06	18.12	38.97	22.08
Our method	29.00	13.51	8.44	51.82	38.00	29.83	25.49	20.67	39.15	22.04

4.4. Test Results

The detection results of the improved YOLOv5 algorithm compared with the original YOLOv5 algorithm is shown in Figure 9. Through the comparison, it can be concluded that the traditional algorithm missed the detection of small objects in the image due to the insufficient extracted features of small objects and the limited context information of small objects. Compared with the original algorithm, the proposed algorithm has significantly improved the detection accuracy of small objects, and the phenomenon of false detection and missing detection are significantly reduced.

The proposed small object detection model has indeed made some achievements in detecting small objects, but there is still much room for improvement in accuracy. With the increasing complexity of the computer vision system, if you want to deploy the system in real-life application scenarios, we need to continue to optimize the performance of the model in terms of small object detection accuracy. In the future, we will refer to the characteristics of small objects to build an optimal network model in terms of high-resolution technology for small object detection in images and videos, so as to further improve the performance.

Figure 9. Test results comparison of the YOLOv5 algorithm: (**a**) The detection results of the original YOLOv5 algorithm; (**b**) The detection results of the proposed YOLOv5 algorithm.

5. Conclusions

Recently, the object detection method based on deep learning technology has become mainstream, and the algorithm model has become more and more effective but also more

complex. Small object detection has always been a difficult problem and focus in this field because of the characteristics of small objects. The main objective of this paper is to improve YOLOv5's unsatisfactory small object detection in the VisDrone-DET2021 dataset. According to the characteristics of small objects, such as low resolution, little information carried and easily affected by background factors, an enhanced feature fusion network R-FPN is proposed in this paper. The context extraction module CEM is used to enhance the extraction of small object feature information and perform more accurate position regression. At the same time, the CSMA attention module is proposed to fully extract the detailed features information of small objects, strengthen the effective feature information in the low-level feature map, suppress the invalid feature, and further improve the detection accuracy of small object. Experimental results show that the improved YOLOv5 model improves the average accuracy of object detection AP by 2.11%, AP50 by 2.86% and AR100 by 1.94% on the VisDrone-Det2021 data set, which proves that the method proposed in this paper can effectively improve the detection effect of small objects. Compared with other latest small object detection algorithms, the proposed algorithm achieves better detection performance, but there is still room for improvement.

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

Funding: This research was funded by Natural Science Foundation of Heilongjiang Province (LH2020F040). The work is also supported by Fundamental Research Funds for The Central Universities (2572017PZ10).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data were prepared and analyzed in this study.

Acknowledgments: The authors would like to thank all anonymous reviewers for their helpful comments and suggestions.

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

References

1. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
2. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
3. Liu, Y.; Liu, H.Y.; Fan, J.L.; Gong, Y.C.; Li, Y.H.; Wang, F.P.; Lu, J. A survey of research and application of small object detection based on deep learning. *Acta Electonica Sin.* **2020**, *48*, 590–601.
4. Niu, H.; Ou, O.; Rao, S.; Ma, W. Small Object Detection Method Based on Improved YOLOv3 in Remote Sensing Image. *Comput. Eng. Appl.* **2022**, *58*, 241–248. [CrossRef]
5. Zhao, P.; Xie, L.; Peng, L. Deep Small Object Detection Algorithm Integrating Attention Mechanism. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 927–937. [CrossRef]
6. Qiu, N.; Cao, J.; Ma, J.; Gong, Y. An improved UAV Ground small object detection method. *Electron. Des. Eng.* **2020**, *28*, 79–84. [CrossRef]
7. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The Vision Meets Drone Object detection Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2847–2854.
8. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
9. Zhao, S.; Zhao, L. Image recognition method based on bidirectional feature pyramid and deep learning. *J. Harbin Univ. Sci. Technol.* **2021**, *26*, 44–50. [CrossRef]
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; pp. 2117–2125.

11. Qu, J.; Su, C.; Zhang, Z.; Razi, A. Dilated convolution and feature fusion SSD network for small object detection in remote sensing images. *IEEE Access* **2020**, *8*, 82832–82843. [CrossRef]
12. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9197–9206.
13. Cao, J.; Chen, Q.; Guo, J.; Shi, R. Attention-guided context feature pyramid network for object detection. *arXiv* **2020**, arXiv:2005.11475.
14. Lin, R.; Huang, R.; Dong, A. Few-shot object detection based on attention mechanism and secondary reweighting of meta-features. *J. Comput. Appl.* **2022**, 1–7. [CrossRef]
15. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
16. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. *Remote Sens.* **2019**, *11*, 531. [CrossRef]
17. Rajaram, R.N.; Ohn-Bar, E.; Trivedi, M.M. Refinenet: Refining object detectors for autonomous driving. *IEEE Trans. Intell. Veh.* **2016**, *1*, 358–368. [CrossRef]
18. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Detnet: A backbone network for object detection. *arXiv* **2018**, arXiv:1804.06215.
19. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
20. Leng, J.; Ren, Y.; Jiang, W.; Sun, X.; Wang, Y. Realize your surroundings: Exploiting context information for small object detection. *Neurocomputing* **2021**, *433*, 287–299. [CrossRef]
21. Changlin, L.; Taojiannan, Y.; Sijie, Z.; Chen, C.; Shanyue, G. Density map guided object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 190–191.
22. Zhang, X.; Ebroul, I.; Krishna, C. Dense and small object detection in UAV vision based on cascade network. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Long Beach, CA, USA, 16–17 June 2019.
23. Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; Qin, H. A global-local self-adaptive network for drone-view object detection. *IEEE Trans. Image Process.* **2020**, *30*, 1556–1569. [CrossRef] [PubMed]
24. Yu, F.; Zhang, X.; Zhang, M.; Zhao, T. Automatic driving small target detection based on improved CenterNet. *Electr. Meas. Technol.* **2022**, 1–7. Available online: http://kns.cnki.net/kcms/detail/11.2175.TN.20220719.1838.026.html (accessed on 10 August 2022).