aerospace

MDPI

Article

Enhancing Small Object Detection in Aerial Images: A Novel Approach with PCSG Model

Kang An, Huiping Duanmu, Zhiyang Wu, Yuqiang Liu, Jingzhen Qiao, Qianqian Shangguan, Yaqing Song * and Xiaonong Xu *

The College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China; ankang526@foxmail.com (K.A.); 1000448148@smail.shnu.edu.cn (H.D.); 100512056@smail.shnu.edu.cn (Z.W.); 1000516142@smail.shnu.edu.cn (Y.L.); 3160602044@smail.shnu.edu.cn (J.Q.); shangguan@shnu.edu.cn (Q.S.)

* Correspondence: yqsong@shnu.edu.cn (Y.S.); seuxxn@shnu.edu.cn (X.X.)

Abstract: Generalized target detection algorithms perform well for large- and medium-sized targets but struggle with small ones. However, with the growing importance of aerial images in urban transportation and environmental monitoring, detecting small targets in such imagery has been a promising research hotspot. The challenge in small object detection lies in the limited pixel proportion and the complexity of feature extraction. Moreover, current mainstream detection algorithms tend to be overly complex, leading to structural redundancy for small objects. To cope with these challenges, this paper recommends the PCSG model based on yolov5, which optimizes both the detection head and backbone networks. (1) An enhanced detection header is introduced, featuring a new structure that enhances the feature pyramid network and the path aggregation network. This enhancement bolsters the model's shallow feature reuse capability and introduces a dedicated detection layer for smaller objects. Additionally, redundant structures in the network are pruned, and the lightweight and versatile upsampling operator CARAFE is used to optimize the upsampling algorithm. (2) The paper proposes the module named SPD-Conv to replace the strided convolution operation and pooling structures in yolov5, thereby enhancing the backbone's feature extraction capability. Furthermore, Ghost convolution is utilized to optimize the parameter count, ensuring that the backbone meets the real-time needs of aerial image detection. The experimental results from the RSOD dataset show that the PCSG model exhibits superior detection performance. The value of mAP increases from 97.1% to 97.8%, while the number of model parameters decreases by 22.3%, from 1,761,871 to 1,368,823. These findings unequivocally highlight the effectiveness of this approach.

Keywords: small target detection; model pruning; feature extraction; fine-grained information

Citation: An, K.; Duanmu, H.; Wu, Z.; Liu, Y.; Qiao, J.; Shangguan, Q.; Song, Y.; Xu, X. Enhancing Small Object Detection in Aerial Images: A Novel Approach with PCSG Model. *Aerospace* **2024**, *11*, 392. https://doi.org/10.3390/aerospace11050392

Academic Editor: Hailong Huang

Received: 3 April 2024
Revised: 10 May 2024
Accepted: 13 May 2024
Published: 14 May 2024

1. Introduction

UAV aerial image detection aims to locate and identify perceptual targets in the image, which belongs to small target detection. A small object refers to the following: (1) Based on relative scales: the width and height of an object is a target that is less than 1/10 of the image or a target whose bounding box to the area of the image is under a value (usually 0.03). (2) Based on absolute scales: objects with a pixel count of less than 32×32. Much meaningful progress has been made in object detection algorithms, which has also injected confidence into small target detection and led to the development of UAV technology. The increasing prevalence of UAV aerial images across diverse applications has provided a fertile ground for the integration of target detection and aerial image fusion. This integration has proven particularly advantageous in fields such as urban transportation [1], urban planning [2], and environmental monitoring [3].

Consequently, the study of small target detection assumes paramount importance, particularly in the context of aerial imagery where images often contain small targets due to the high altitude of aerial photography [4]. Additionally, variable angles and environmental conditions further complicate the detection process by exacerbating target–background mixing [5]. While mainstream object detection algorithms are widely used across various domains, their effectiveness in detecting small targets is often limited, primarily due to their reliance on mechanisms such as anchor frames [6]. Small object detection based on the anchor frame mechanism uses different methods to set anchor frames of different sizes on the feature map and predicts whether each frame contains target objects based on the features in each anchor frame. This reliance can lead to suboptimal results when applied directly to small target detection tasks as mechanisms like anchor frames exacerbate the challenge of localizing small targets. Furthermore, noise filtering during convolution operations can reduce image resolution, causing the loss of critical features essential for effectively learning from small targets [7]. Hence, addressing the challenge of small object detection, characterized by limited pixel proportions and complex feature extraction, is essential for advancing the capabilities of object detection algorithms in aerial imagery analysis [8,9].

The urgent demand for small target detection spans various fields, including environmental monitoring, urban development, and transportation, among others. Research breakthroughs in this area are becoming increasingly imperative as aerial image detection becomes an integral component of daily life, posing a significant real-world challenge. However, numerous obstacles persist, rendering small target detection a formidable task.

To tackle these challenges, this paper introduces PCSG, which optimizes computational performance while enhancing detection capabilities. This enhancement entails modifications to both the backbone and detection head. Regarding the detection head, we propose an innovative framework and conduct further analysis through pruning and upsampling optimization. Concerning the backbone, we optimize feature extraction performance by leveraging the Ghost framework [10] and SPD-Conv [11]. The specific contributions are outlined below:

- By enhancing shallow feature reuse, the model retains richer position information and bolsters feature extraction for small targets while also optimizing network structures and introducing additional prediction branches.
- Leveraging the lightweight and versatile CARAFE structure for upsampling mitigates the issue of local information loss associated with traditional nearest-neighbor upsampling methods.
- The adoption of the SPD-Conv (space-to-depth) module in place of strided convolution and pooling maximizes the retention of discriminative feature information.
- The integration of Ghost convolution as the backbone reduces parameters and computational complexity compared to the original CSPNet backbone while also enhancing real-time performance.

The structure of our article is as follows: Section 2 provides an overview of related works, focusing on anchor-free object detection methods, data augmentation techniques, and multi-scale learning approaches. Section 3 elaborates on our proposed method, which comprises a PCHead model and a PCSG model. In Section 4, we show the experiment results, where we apply our method to aerial images and compare its performance with that of YOLOv5. Finally, we list our findings and outline directions for future work.

2. Related Works

This paper first discusses the importance of resolution on small target detection. Then, we delve into the study of small object detection algorithms, which can be split into three main parts: anchor-free mechanisms, data augmentation technology, and multi-scale learning.

In object detection, image resolution plays a crucial role. Images with high resolution can provide more abundant information such as outlines, textures, and feature points. This information helps to improve the accuracy of the network. Low-resolution images contain less detail, and the image becomes blurry and contains more noise, which interferes with network detection. But, higher resolution increases computational complexity and slows down the network's recognition speed. For small object detection, higher resolutions can increase the size of the target in the image and reduce the difficulty of detection. However, when the resolution is too large, the proportion of the network's receptive field in the image decreases, making the network unable to predict objects at all scales.

The anchor box mechanism remains the prevailing approach in object detection, which applies different anchor frames to the feature map, predicts whether the target object is included based on the features in each box, and fits the ground truth box to obtain the target positioning. However, the detection of small objects poses unique challenges due to their limited pixel occupancy within images. Predicting the boundary box offset for small objects often leads to increased errors, compounded by the smaller number of anchor boxes available for their detection. To mitigate these challenges, various innovative solutions have emerged. One such solution [12] involves per-pixel prediction, while another approach [13,14] utilizes key points to replace anchor boxes, with enhancements made through the incorporation of a central point. Additionally, some methods [15] leverage global context information between detected instances and images to eliminate the reliance on anchor boxes and non-maximum suppression (NMS). Moreover, employing attention mechanisms to focus on the surrounding environment of detected instances [16,17] has also yielded promising results in small object detection.

Utilizing data augmentation techniques to enhance the quantity and quality of images in datasets can significantly improve the generalization ability of models. Small targets often suffer from low resolution, difficulty in feature extraction, and limited sample size. The problem of fewer small targets in the image can be solved by duplicating the small targets in the image through the oversampling strategy [18]. Crop out the small target by mask, paste it anywhere in the image, and generate a new ground truth and the pasted target can be randomly transformed (scale, fold, rotate, etc.). This increases small targets in each image. However, simplistic replication strategies may lead to problems such as scale and background mismatches. Addressing these challenges requires the consideration of contextual information during data replication and the implementation of adaptive resampling augmentation [19]. For the problem of feature extraction and limited sample size, image processing can be used, such as increasing the contrast of the image, histogram equalization, spatial filtering, spatial scale transformation, etc. These operations can enhance some of the features in the image, and the number of samples can also be increased by training the transformed image and the original image as a whole dataset.

Multi-scale learning enhances the feature learning capability of models by fusing deep abstract information with shallow detail information [20]. Small targets often face challenges in feature adulteration with background information during convolution. Shallow networks extract pixel-level information, including color, edges, textures, and corners, while deep networks capture semantic information. Integrating features from different levels of the feature pyramid [21,22] and employing adaptive multi-scale networks [23] are effective strategies to resolve this issue. This paper extends the application of multi-scale information to even shallower feature maps, prunes redundant network structures, and employs the CARAFE module to minimize feature information loss during upsampling, along with the SPD-Conv module to preserve fine-grained feature map information. Furthermore, processing the context region of targets instead of simple pixel-by-pixel processing during training [24] yields an efficient multi-scale training approach. Another study [25] has demonstrated improved detection performance through the utilization of relevant information across different feature maps.

3. Improvement on Detection Head and Backbone Networks

In object detection, optimizing both the detection head and the backbone network is crucial for achieving superior performance. The detection head relies on the features extracted by the backbone network to generate detection boxes and class confidences. Conversely, the backbone network extracts image features to facilitate accurate detection. By strategically optimizing these components, we can exploit their complementary strengths to enhance detection accuracy, robustness, and efficiency. This study explores novel optimization strategies for the detection head and backbone network to achieve state-of-the-art performance. Through this investigation, we aim to provide insights into the fundamental principles of effective object detection systems.

3.1. PCHead Model

In this paper, we address the challenge of small targets occupying a small proportion of the image. The changes make it difficult to extract useful information while also leading to redundant network structures. To tackle this issue, we propose several enhancements. Firstly, we deepen the feature pyramid and the path aggregation network. Secondly, we introduce a new detection layer specifically designed for smaller targets. Additionally, we trim the redundant parts of the original network structure, focusing on features relevant to larger targets. Finally, we utilize the lightweight and versatile upsampling operator CARAFE to construct the PCHead model.

3.1.1. A Novel Detection Head with Enhanced Feature Pyramid

The YOLO-v5 algorithm, classified as a one-stage target detection algorithm, directly predicts the object's category probability and positional coordinate value. Its speed surpasses that of two-stage target detection algorithms, and it exhibits higher efficiency in detecting small targets. Therefore, it is particularly fit for real-time target detection in high-quality aerial imagery.

To leverage this feature of the YOLO-v5 algorithm and improve the performance, this paper enhances the FPN + PAN structure by transforming the original two-layer feature pyramid structure into a three-layer structure. The addition of an extra layer of the feature pyramid facilitates feature fusion with the feature map obtained through $4\times$ downsampling. These modifications aim to capitalize on the combination of the FPN and PAN structures while enhancing the capability of the network.

In addition, the original YOLOv5 network uses a three-layer output method, corresponding to $8\times$ downsampling, $16\times$ downsampling, and $32\times$ downsampling, respectively. The calculation of the target resolution from the original resolution to the feature layer is shown in Formula (1), where $(h, w)_L$ represents the size of the pixel in L-th layer feature map, (H, W) represents original resolution of target, and $Stride_L$ represents the downsampling stride of L-th layer feature map.

$$(h, w)_L = \frac{(H, W)}{Stride_L} \tag{1}$$

On the basis of the above modifications, this paper adds a detection head that downsamples by a factor of 4, which is used to detect even smaller targets. By deepening the feature pyramid, the new detection head model can use shallower feature maps to obtain richer feature and position information. The optimized network is shown in Figure 1. The left side shows the feature maps corresponding to the downsampling factor obtained by the feature extraction structure from the original input. The middle shows the feature maps obtained through feature fusion by the FPN structure. The right side shows the PAN structure corresponding to the FPN, which finally obtains the detection head for different-sized targets.

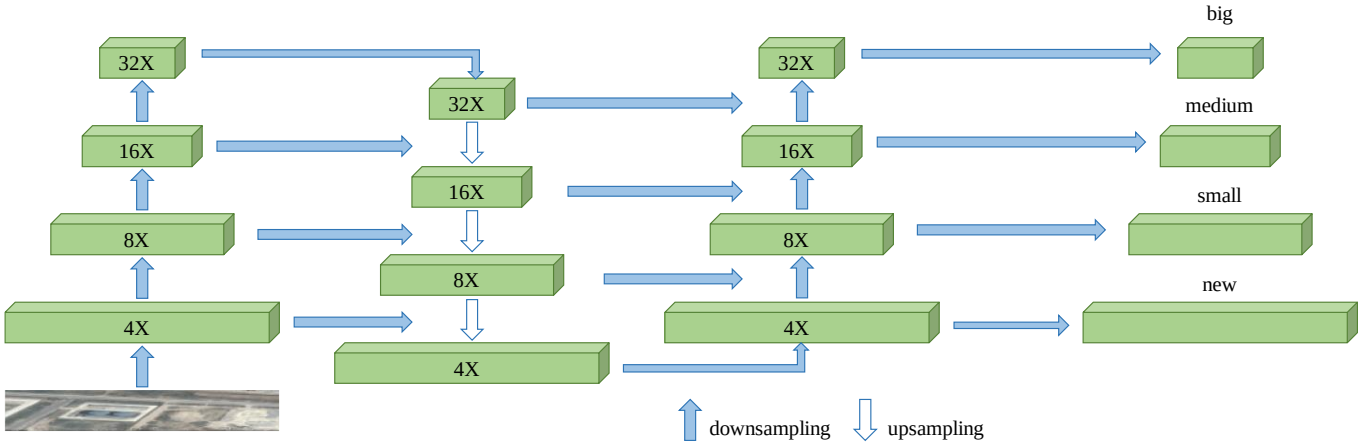

Figure 1. Added head network structure.

3.1.2. Network Structure Pruning for Complexity Reduction

Although the newly introduced detection head model has enhanced the accuracy, it has also contributed to an increase in the complexity of the network structure. The addition of the feature pyramid network (FPN) structure, path aggregation network (PAN) structure, and 4× downsampling detection head has significantly augmented the intricacy of the network. In aerial image datasets, large objects are scarce, rendering the 32-fold downsampling correlation module inapplicable to the primary objectives of detecting small objects effectively. Consequently, modifying the existing network structure becomes imperative to address the issue of redundancy.

Moreover, the feature maps obtained through 32× downsampling possess a large receptive field, making them less sensitive to small objects with low pixel ratios, thus offering minimal contribution to the detection of such objects. Therefore, after thorough deliberation, this paper proposes pruning the modules associated with 32× downsampling in the network architecture. Specifically, this involves removing the 32× downsampling feature extraction module from the backbone network, the 32× downsampling feature fusion module from the neck section, and the 32× downsampling detection head from the prediction branch, as illustrated in Figure 2. Following these pruning operations, adjustments are required for the FPN and PAN structures as the feature map size input into feature fusion network structure has been altered. Consequently, certain layers of the FPN structure and their corresponding PAN structures are pruned accordingly. Ultimately, the detection head designed for large objects is eliminated.

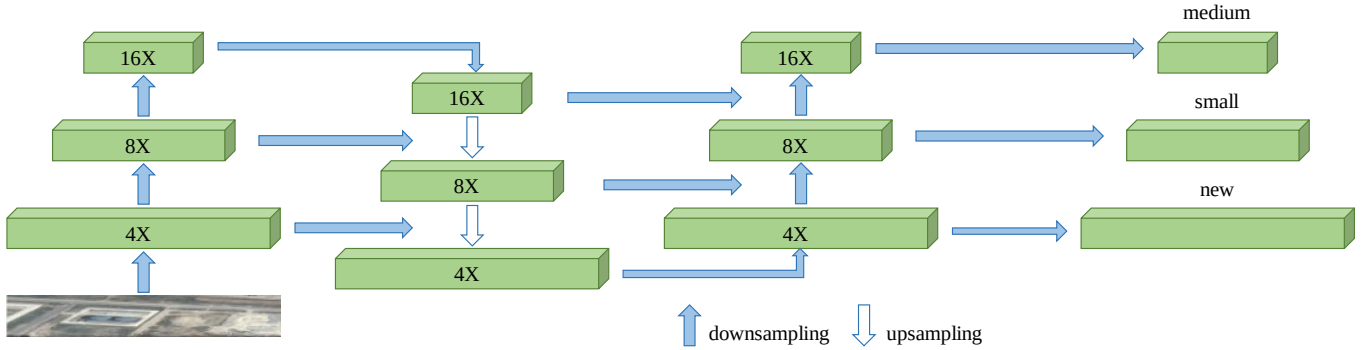

Figure 2. Prune network structure.

The increased complexity of the network structure poses significant challenges, particularly in terms of computational efficiency, memory consumption, and training time. Addressing these challenges is crucial for ensuring the scalability and practical applicability of the proposed model.

The overall network structure after pruning retains the advantages of using shallow feature maps and optimizes the problem of network complexity it brings. By removing the 32× downsampling feature extraction module in the feature extraction network, the subsequent feature fusion becomes more convenient without the need to add new upsampling structures. This also reduces unnecessary computations, decreases compute parameters, and increases the detection speed.

3.1.3. CARAFE-Based Feature Reassemble

The upsampling algorithm used in YOLOv5 is nearest-neighbor interpolation, which directly uses the closest existing color to generate missing pixel values. This approach of copying neighboring pixel values can create obvious aliasing artifacts. This strategy of copying samples is prone to emphasizing individual samples too much, leading to overfitting problems. In addition, this algorithm only considers the distance between instances and does not utilize the feature information of instances, which can also affect the detection performance to some extent.

In this paper, a lightweight and general-purpose upsampling algorithm called CARAFE (Content-Aware ReAssembly of Features) [26] is used to supersede the nearest-neighbor upsampling algorithm. The CARAFE algorithm is a content-aware and feature reassembly upsampling method that has a large receptive field during the feature reassembly process.

Compared to traditional upsampling algorithms, the CARAFE algorithm uses convolutional layers to transform the input feature channels, effectively solving the checkerboard effect. It enhances detail information during upsampling, which is helpful for feature learning. Moreover, the algorithm is more lightweight and has higher running efficiency. CARAFE contains two key components: a kernel prediction module and a content-aware reassembly module. The former is used to generate weights for convolution kernels, which can be adjusted based on the pixel values in the image. The latter is used to combine feature maps of different sizes to attain the final feature map.

CARAFE contains two steps: first, it predicts a reassembly kernel for each target location based on its content, as shown in Formula (2); the second step uses the predicted kernel to guide the feature reassembly process, as shown in Formula (3). For feature map X with size $C \times H \times W$ and upsampling rate σ (assuming σ is an integer), CARAFE generates a new feature map X' with size $C \times \sigma H \times \sigma W$. Any position ($l' = (i', j')$) in the new feature map (X') is associated with a corresponding original position ($l = (i, j)$) in the input feature map (X), where $i = [i'/\sigma]$ and $j = [j'/\sigma]$. $N(X_l, k)$ represents the $k \times k$ subregion of X centered at position l, which is the neighborhood of X_l. The kernel prediction module ψ predicts the position kernel $W_{l'}$ for each l' position based on the $k \times k$ subregion of X_l, while the content-aware reassembly module ϕ reassembles the $k \times k$ subregion of X_l and the kernel $W_{l'}$.

$$W_{l'} = \psi(N(X_l, k_{encoder})), \tag{2}$$

$$X'_{l'} = \phi\left(N\left(X_l, k_{up}\right), W_{l'}\right), \tag{3}$$

The CARAFE algorithm replaces the nearest-neighbor upsampling algorithm and applies it to the original network, which is modified by the previous work, and the final network structure is displayed in Figure 3. In this paper, we refer to it as the PCHead model.

3.2. PCSG Model

In this section, we delve into the detailed modifications made to the backbone network. With the aim of improving both fine-grained feature extraction capability and computational performance, this paper incorporates a novel CNN module, SPD-Conv, while also substituting the CSPDarknet53 network structure with Ghost convolution. These alterations culminate in the development of the PCSG model.

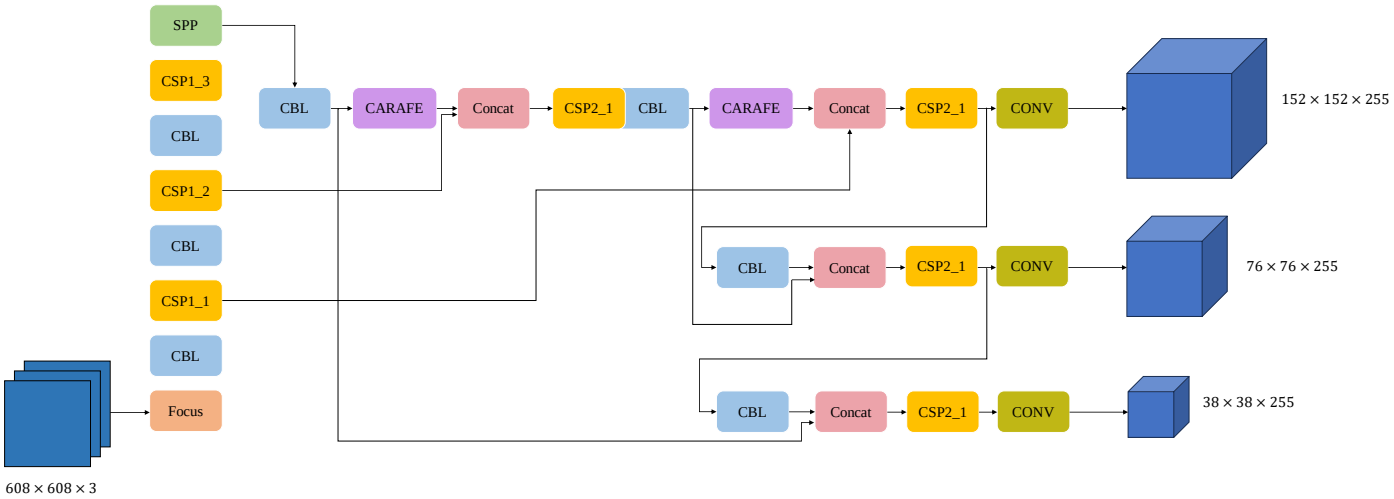

Figure 3. Schematic diagram of PCHead network structure.

3.2.1. Fine-Grained Information Network Structure

The underlying reason why mainstream target detection algorithms perform well in detecting regular-sized targets but poorly in detecting small targets is the flaw of existing network architectures, especially hierarchical convolution and pooling layers. These two layers cause the loss of fine-grained information and low efficiency. Limited by the low resolution of the input image and the small size of the target, existing CNN architectures are therefore not suitable for small object detection.

The SPD-Conv [11] efficiently solves the above-mentioned problem. As a new type of CNN module, SPD-Conv consists of an SPD layer (space-to-depth) and a Conv (non-strided convolution) layer. By replacing each strided convolution and pooling layer with the SPD-Conv, it achieves the preservation of fine-grained information and effective feature learning. The SPD module performs image transformation on the internal CNN and the entire feature map in order to perform downsampling.

For a middle feature map X with a size of $S \times S \times C_1$, it can be divided into a series of sub-feature maps by cutting, as shown in the Formula (4). For a feature map (X), each sub-map $f_{x,y}$ composes all entries $X(i+y)$ that are evenly divided by the scaling factors $i+x$, and $i+y$. Therefore, each sub-map downsamples the original feature map according to the scaling factors. Then, feature sub-maps are connected along channel dimension to gain a new feature map X', which has one less scaling factor in the spatial dimension and one more scaling factor in the channel dimension than the original feature map (X). Therefore, the dimension of feature map X changes from the original $S \times S \times C_1$ to $\frac{S}{scale} \times \frac{S}{scale} \times scale^2 C_1$.

$$\begin{aligned} f_{scale-1,0} &= X[scale-1:S:scale, 0:S:scale], \\ f_{scale-1,1} &= X[scale:S:scale, 1:S:scale], \\ f_{0,scale-1} &= X[0:S:scale, scale-1:S:scale], \\ f_{scale-1,scale-1} &= X[scale-1:S:scale, scale-1:S:scale], \end{aligned} \tag{4}$$

Regarding the non-strided convolutional layer, after the dimension transformation of the feature map is completed by the SPD layer, a non-strided convolutional layer with C_2 filters is added. C_2 is less than $scale^2 C_1$. Therefore, the dimension of the feature map is further transformed from $\frac{S}{scale} \times \frac{S}{scale} \times scale^2 C_1$ to $\frac{S}{scale} \times \frac{S}{scale} \times C_2$. By using non-strided convolution, the original network can preserve much feature discriminative information.

This article modifies the original YOLOv5 network structure by embedding SPD-Conv modules into the stride convolutional layers in the backbone and neck parts. Specifically, SPD-Conv is added to each stride convolutional layer and its subsequent connection layer, i.e., between the Conv module and the C3 module. In total, there are six replacements, with

four in the backbone part and two in the neck part because the backbone contains four stride convolutions and the neck contains two.

3.2.2. Lightweight Network Structure

Although the previous work can enhance the accuracy, it tends to make the network structure more complex, with larger parameters and computations. This leads to lower efficiency in small object detection and cannot meet the real-time needs of aerial image detection. In this paper, we use the Ghost convolution module [10] to lighten the weight of the network.

The Ghost module contains three steps: conventional convolution, Ghost generation, and feature map concatenation. In this paper, the Conv and C3 modules in the original network structure were replaced with Ghost convolutions, which greatly reduced the parameters of the network and improved the detection speed.

Combining the previous work, this paper has some enhancements to the original network's feature pyramid structure. By applying shallow feature maps, the ability to extract features for small objects is strengthened. The modules related to large objects in the original network's feature extraction network, feature fusion network, and prediction branch are pruned to enhance the model's accuracy and speed. The upsampling algorithm of the original network structure is modified, and the design flaws of the CNN module in the original network model are reduced to preserve fine-grained information. The entire model is lightweight to meet practical scenario requirements. The final network, called the PCSG model, is displayed in Figure 4. It can be described as follows: Firstly, the original input image elapses the Focus structure, as well as the GhostConv, SPD, and GhostC3 modules, to obtain the corresponding downsampling feature maps. Then, after passing through the SPP module and being upsampled through the CARAFE module, it undergoes feature fusion with the corresponding scale feature map obtained previously via the Concat module. Afterward, the PAN structure is utilized bottom-up with corresponding feature maps obtained previously by the FPN structure. Finally, detection heads are obtained for different-sized targets.

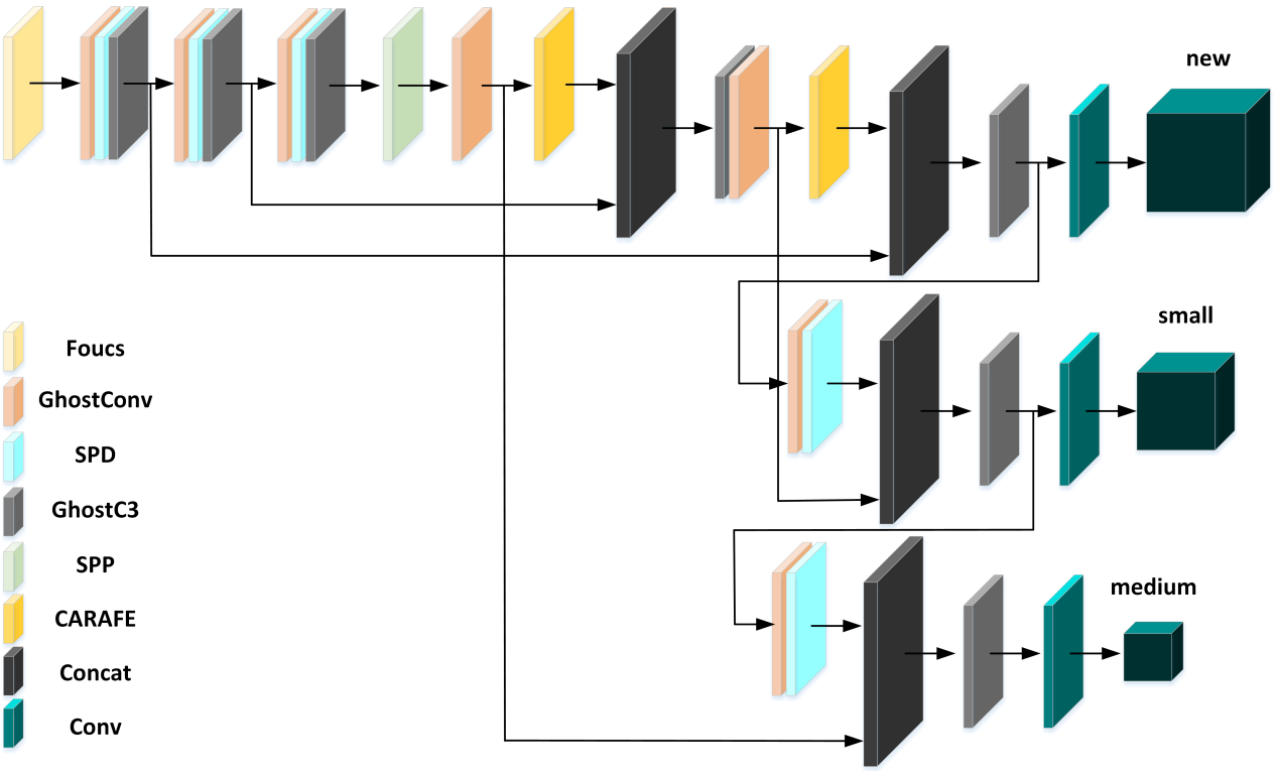

Figure 4. Schematic diagram of the PCSG network structure.

4. Experimental Results and Analysis

4.1. Datasets and Experimental Setting

We use the open source RSOD aerial object detection dataset [27] to select four types of targets: airplanes, playgrounds, overpasses, and oil drums. There are 446 images of airplanes with 4993 airplanes, 189 images of playgrounds with 191 playgrounds, 176 images of overpasses with 180 overpasses, and 165 images of oil drums with 1586 oil drums. In this dataset, the image resolution of airplanes is 0.5~2/pixel, the image resolution of playgrounds is 0.4~1/pixel, the image resolution of overpasses is 1.25~3/pixel, and the image resolution of oil drums is 0.3~1/pixel. Since the image resolution of the overpass is relatively large among them, and there is a partial lack of the playground image, this paper chooses the categories of "airport" and "oil tank" as the training and testing samples in the dataset. Training and testing sets are split into 9:1, and training and testing data are strictly independent.

The operating system that we used in this experiment is Windows 10 (64-bit), the CPU model is 11th Gen Intel(R) Core (TM) i7-11700 @2.50 GHz, the graphics card is NVIDIA GeForce RTX 3070 Ti (Santa Clara, CA, USA), the memory size is 32 GB, the Python version is 3.7.4, and the PyTorch version is 1.9.0. The parameter settings used in the training process of this article are shown in Table 1.

Table 1. Parameters setting.

Parameters	Values
Momentum	0.937
Batch size	16
Image size	640
Weight decay	0.0005
Learning rate	0.01
Epochs	300

4.2. Assessment of Indicators

We use the most commonly used mean average precision (mAP) as the evaluation index, and its value can comprehensively reflect the relationship between accuracy and recall, i.e., mAP is related to precision and recall. The closer to 1 the value, the better the algorithm performance. The confusion matrix of the evaluation indexes of mAP is shown in Table 2, where TP denotes true cases, FP denotes false positive cases, TN denotes true negative cases, and FN denotes false negative cases.

Table 2. Bicategory confusion matrix.

Real/Predicted	Positive Sample	Negative Sample
Positive Sample	True Positive (TP)	False Negative (FN)
Negative Sample	False Positive (FP)	True Negative (TN)

GFLOPS (giga floating-point operations per second) is defined as 1 billion floating-point operations per second.

4.3. Comparative Experimental Analysis

To confirm the validity of the improvements made in our article, corresponding experiments were managed on the aforementioned improved network structure, and the experiment results are described in Table 3.

Table 3 illustrates the efficacy of various model enhancements compared to the original network structure. The Add Head model improves feature fusion, deepens the FPN structure, and enhances shallow-level feature reuse, resulting in improved detection performance. The precision value increases from 96.8% to 98.4%, the recall value from 92.6% to 93.5%, and the mAP value from 97.1% to 97.3%. Conversely, the Prune model yields

even better results, with precision increasing to 97.8%, recall to 94.4%, and mAP to 98.0%. This suggests that excessive convolutional calculations in deeper network layers can lead to feature loss for small objects. The Prune model mitigates this by preserving feature information, with a reduction in parameters to 1,021,327, indicating reduced complexity.

Table 3. Performance comparison of different algorithms.

	Parameters	GFLOPS	P/%	R/%	mAP/%
YOLOv5	1,761,871	4.2	96.8	92.6	97.1
Detection Head	1,762,095	7.2	98.4	93.5	97.4
Module Pruning	1,021,327	7.5	97.8	94.4	98.0
CARAFE	1,803,335	4.3	95.2	94.6	97.4
PCHead	1,062,791	8.0	97.3	94.5	98.2

Similarly, the CARAFE upsampling algorithm enhances the mAP value to 97.4%, albeit with a slight increase in parameters to 1,803,335. The integrated PCHead model outperforms the original network, with P, R, and mAP values increasing to 97.3%, 94.5%, and 98.2%, respectively. This underscores the effectiveness of pruning, deepening the FPN structure, and replacing the upsampling algorithm with CARAFE in improving the detection results. Moreover, the PCHead model's reduction in parameters to 1,062,791 simplifies the model while increasing GFLOPS to 8.0, thus enhancing computation speed.

Since the size of convolutional kernels in the feature extraction network structure of the YOLOv5 algorithm increases with the network's depth, the pruning strategy reduces the number of covariates in the model significantly, which makes the SPD-Conv module only partially useful. Therefore, the shallow network in the feature extraction network of the model is supplemented, i.e., the corresponding network parameters are modified to ensure the number of convolutional kernels and the feature extraction ability of the model, so that the experimental setup is more rigorous. As we can see from Table 4, in SPD-Conv model, the P-value is increased to 96.9%, the R-value is increased from 92.6% to 95.8%, and the mAP value is increased from 97.1% to 98.6%, which is a big improvement in the detection accuracy. However, parameters increase from 1,761,871 to 2,442,503, which increases the model complexity, and the number of GFLOPS increases from 4.2 to 26.1, which increases the computation speed/required computation speed significantly, which illustrates the complexity of the model. The Ghost module is applied to models such as pruning and PCHead, and similarly, the parameters of the shallow layer of the feature extraction network are changed to maintain the rigor of the experiment. The P-value increases from 96.8% to 97.0%, the R-value decreases from 92.6% to 91.4%, and the mAP value is 96.6%, which is lower than the 97.1% of the original algorithm. The mAP value is 96.6%, which is lower than the 97.1% of the original algorithm. However, the parameter value is reduced from 1,761,871 to 987,895, which is more in line with the expected effect of the lightweighting of the network.

Table 4. Comparison of results of ablation experiments with overall improvement.

	Parameters	GFLOPs	P/%	R/%	mAP/%
YOLOv5	1,761,871	4.2	96.8	92.6	97.1
SPD-Conv	2,442,503	26.1	96.9	95.8	98.6
Ghost	987,895	7.2	97.0	91.4	96.6
SPD + Ghost	1,323,275	4.9	96.5	94.7	97.2

Since the YOLOv5 algorithm's number of convolutional kernels in the feature extraction network structure increases with the network's depth, pruning strategies significantly reduce the parameter count in the model and make the SPD-Conv modules only partially effective. Therefore, in this work, the shallow network in the model's feature extraction network was supplemented by modifying the corresponding network parameters to ensure the

model's convolutional kernel count and feature extraction ability, making the experimental setup more rigorous. As the data show in Table 5, compared to the YOLOv5 algorithm, the PCSG model had a precision value of 96.9%, and the recall value increased from 92.6% to 95.8%, with a mAP value increase from 97.1% to 97.8%, improving the model's detection accuracy. Additionally, the parameters decreased from 1,761,871 to 1,368,823, reducing the model's complexity and verifying the effectiveness of the light-weighting strategy of replacing the Conv and C3 modules in original network with Ghost modules. Overall, this work achieved good improvements in both accuracy and model complexity for small object detection and had ideal experimental results.

Table 5. Comparison of experimental results for the final model.

	Parameters	GFLOPS	P/%	R/%	mAP/%
YOLOv5	1,761,871	4.2	96.8	92.6	97.1
PCHead	1,062,791	8.0	97.3	94.5	98.2
SPD + Ghost	1,323,275	4.9	96.5	94.7	97.2
PCSG	1,368,823	14.9	96.9	95.8	97.8

4.4. Visualization of Test Results

To better illustrate the feasibility of the PCHead model, this paper selects some test set images for the test illustration, and the results are shown in Figure 5. Among them, the left picture is the truth value information, the middle picture is the YOLOv5 detection effect, and the right picture is the PCHead model detection effect.

The image can be seen from Figure 5. In (a) and (c), the original model has the phenomenon of missed detection, and it can be seen that the objects in green boxes are marked in the ground truth, but the original model is not detected. In (b), there is a double detection phenomenon with only two objects in the white box but three boxes in the source model. In (d), there is a misdetection: the object in the orange box is not marked as aircraft in the ground truth, but it is detected as aircraft in the original model. The PCHead model improves the above errors to a certain extent, which indicates that the new detection layer and the CARAFE upsampling algorithm can bring beneficial improvement to the detection of small targets and verifies the validity of the work in this paper. Specifically, the pruning model enhances the feature extraction capability by utilizing a shallower feature map, which brings richer detail information for the subsequent feature fusion network, while the CARAFE upsampling algorithm reorganizes the features based on the predicted upsampled features in a way that reduces the information loss caused by the original nearest-neighbor upsampling algorithm, which enhances the detection effect.

To better illustrate the feasibility of our model, we conducted tests on selected images from the test set. The results are shown in Figure 6. The left image shows the ground truth information, the middle image shows the results of YOLOv5, and the right image shows the results of PCSG. The green boxes in the figure indicate the detection errors of YOLOv5.

As can be seen from (a) to (c) in Figure 6, the original model has some limitations in small target detection. There are the phenomena of false detections. The orange box does not have an aircraft in ground truth, but the original model detects an aircraft. The small target detection results shown in Figure 6b also show that the PCSG model has better detection performance and exhibits a certain improvement on false detection. This further demonstrates the effectiveness and superiority of our work, which validates the effectiveness of the added detection layer, CARAFE upsampling algorithm, SPD-Conv module, and Ghost module in small object detection in aerial images.

Ground truth YOLOv5 PCHead

(**a**) Misdiagnosis

Ground truth YOLOv5 PCHead

(**b**) Reexamination

Ground truth YOLOv5 PCHead

(**c**) Misdiagnosis

Ground truth YOLOv5 PCHead

(**d**) Misdirection

Figure 5. Comparison of the experimental results of YOLOv5 and PCHead.

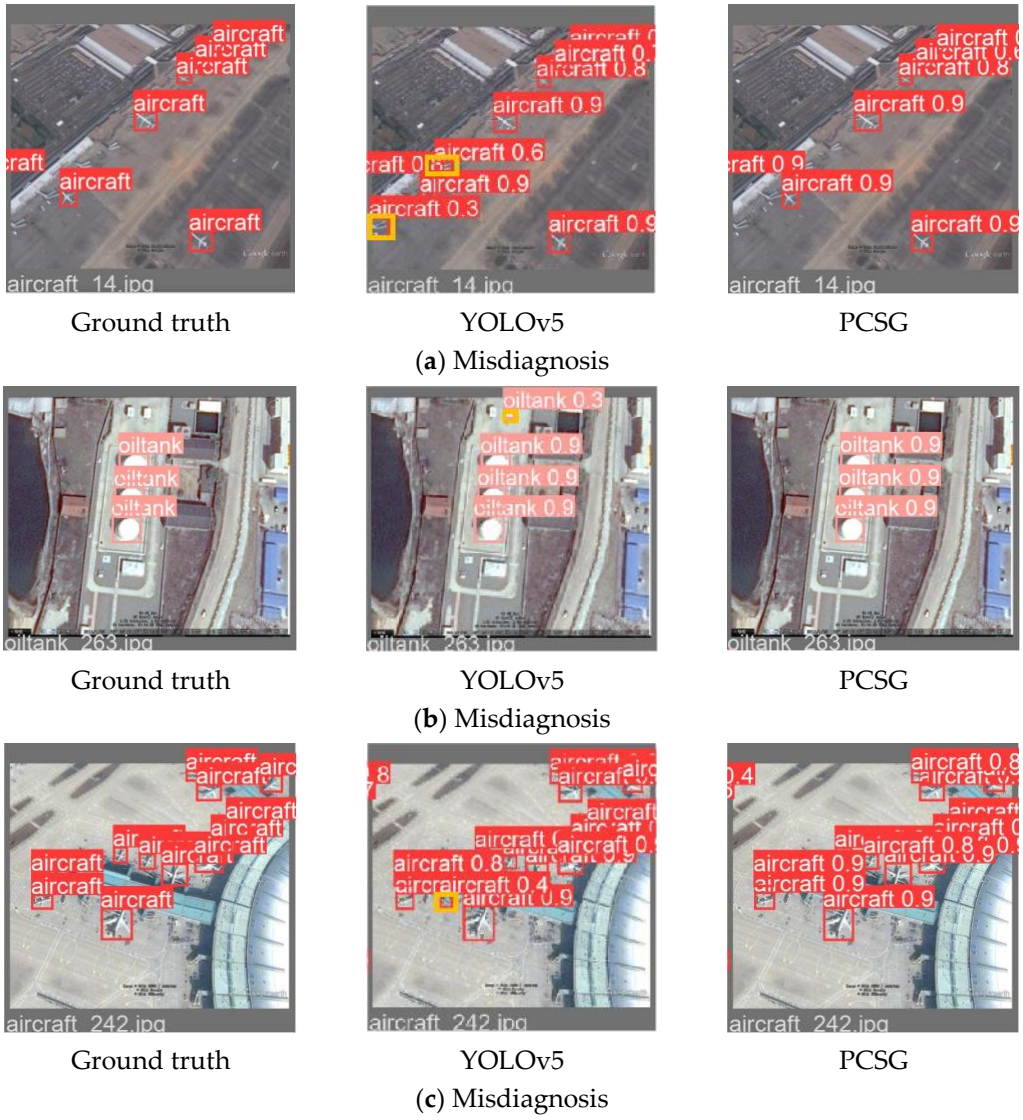

Figure 6. Comparison of the experimental results of YOLOv5 and PCSG.

5. Conclusions

While many object detection algorithms have shown effectiveness in detecting objects of regular sizes, small object detection remains a significant challenge. The main challenge is the limited pixel proportion and the complexity of feature extraction. Moreover, current mainstream detection algorithms tend to be overly complex, leading to structural redundancy for small objects. In this study, we delve into the domain of aerial imagery, where detecting small objects poses particular challenges due to their limited pixel representation and complex feature extraction. To address these challenges, we propose modifications to the feature pyramid structure, focusing on fusing shallow feature maps and enhancing their reusability. Additionally, tackling the issue of redundancy in existing network structures for small object detection, we prune the correlation structure and introduce a new detection head. Furthermore, adopting the lightweight and versatile upsampling operator CARAFE we address the problem of local feature information loss through content-aware feature recombination during upsampling. Striving to retain all discriminative feature information, the strided convolutions and pooling of the existing network are replaced with the SPD-Conv module. Leveraging Ghost convolution to reduce model complexity and enhance real-time performance, the resultant PCSG model achieves an impressive mAP value of 97.8% on the RSOD dataset, demonstrating superior detection capabilities. The research in this paper is for small target detection in aerial images, and the method adopted provides a

certain reference in this field. Moving forward, the focus will be on enhancing the model's generalizability and exploring further improvements from diverse perspectives, as well as incorporating YOLO v9 or other latest versions of the model into the comparison.

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

Funding: This research has recevied no external funding.

Data Availability Statement: Data are contained within the article.

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

References

1. Naranjo, M.; Fuentes, D.; Muelas, E.; Díez, E.; Ciruelo, L.; Alonso, C.; Abenza, E.; Gómez-Espinosa, R.; Luengo, I. Object Detection-Based System for Traffic Signs on Drone-Captured Images. *Drones* **2023**, *7*, 112. [CrossRef]
2. Zebedin, L.; Bauer, J.; Karner, K.; Bischof, H. Fusion of Feature- and Area-Based Information for Urban Buildings Modeling from Aerial Imagery. In *Computer Vision—ECCV 2008, Marseille, France, 12–18 October 2008. Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5305.
3. Watts, A.C.; Ambrosia, V.G.; Hinkley, E.A. Unmanned Aircraft Systems in Remote Sensing and Scientific Research: Classification and Considerations of Use. *Remote Sens.* **2012**, *4*, 1671–1692. [CrossRef]
4. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier Transformation-Based Histograms of Oriented Gradients for Rotationally Invariant Object Detection in Remote-Sensing Images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [CrossRef]
5. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [CrossRef]
6. 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.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
7. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
8. Carranza-García, M.; Lara-Benítez, P.; García-Gutiérrez, J.; Riquelme, J.C. Enhancing Object Detection for Autonomous Driving by Optimizing Anchor Generation and Addressing Class Imbalance. *Neurocomputing* **2021**, *449*, 229–244. [CrossRef]
9. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
10. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
11. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In *Machine Learning and Knowledge Discovery in Databases, Proceedings of the European Conference, ECML PKDD 2022, Grenoble, France, 19 –23 September 2022, Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2023; Volume 13715.
12. Sivapriya, M.S.; Suresh, S. ViT-DexiNet: A Vision Transformer-Based Edge Detection Operator for Small Object Detection in SAR Images. *Int. J. Remote Sens.* **2023**, *44*, 7057–7084. [CrossRef]
13. Xu, X.; Zhang, H.; Ma, Y.; Liu, K.; Bao, H.; Qian, X. TranSDet: Toward Effective Transfer Learning for Small-Object Detection. *Remote Sens.* **2023**, *15*, 3525. [CrossRef]
14. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended Feature Pyramid Network for Small Object Detection. *IEEE Trans. Multimed.* **2022**, *24*, 1968–1979. [CrossRef]
15. Kaur, R.; Singh, S. A Comprehensive Review of Object Detection with Deep Learning. *Digit. Signal Process.* **2022**, *132*, 103812. [CrossRef]
16. Cao, X.; Zhang, Y.; Lang, S.; Gong, Y. Swin-Transformer-Based YOLOv5 for Small-Object Detection in Remote Sensing Images. *Sensors* **2023**, *23*, 3634. [CrossRef] [PubMed]
17. Zhang, T.; Li, L.; Cao, S.; Pu, T.; Peng, Z. Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target under Complex Background. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [CrossRef]
18. Lu, S.; Ding, Y.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. Multiscale Feature Extraction and Fusion of Image and Text in VQA. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 54. [CrossRef]
19. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Wang, X.; Sun, X.; Dong, J. RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019.

20. Liu, H.; Duan, X.; Chen, H.; Lou, H.; Deng, L. DBF-YOLO: UAV Small Targets Detection Based on Shallow Feature Fusion. *IEEJ Trans. Electr. Electron. Eng.* **2023**, *18*, 605–612. [CrossRef]
21. Yi, H.; Liu, B.; Zhao, B.; Liu, E. Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 1734–1747. [CrossRef]
22. Zhao, H.; Zhang, H.; Zhao, Y. YOLOv7-Sea: Object Detection of Maritime UAV Images Based on Improved YOLOv7. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2–7 January 2023.
23. Liu, Y.; Li, W.; Tan, L.; Huang, X.; Zhang, H.; Jiang, X. DB-YOLOv5: A UAV Object Detection Model Based on Dual Backbone Network for Security Surveillance. *Electronics* **2023**, *12*, 3296. [CrossRef]
24. Li, S.; Yang, X.; Lin, X.; Zhang, Y.; Wu, J. Real-Time Vehicle Detection from UAV Aerial Images Based on Improved YOLOv5. *Sensors* **2023**, *23*, 5634. [CrossRef] [PubMed]
25. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef] [PubMed]
26. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware Reassembly of Features. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
27. Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]