



Article

A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios

Jianjun Ni ^{1,*}, Shengjie Zhu ¹, Guangyi Tang ¹, Chunyan Ke ¹ and Tingting Wang ²

¹ College of Artificial Intelligence and Automation, Hohai University, Changzhou 213200, China; zhushj@hhu.edu.cn (S.Z.); tang_gy@hhu.edu.cn (G.T.); chunyanke@hhu.edu.cn (C.K.)

² School of Mechanical and Electrical Engineering, Hohai University, Changzhou 213200, China; 20121894@hhu.edu.cn

* Correspondence: njjhhuc@gmail.com

Abstract: Small object detection for unmanned aerial vehicle (UAV) image scenarios is a challenging task in the computer vision field. Some problems should be further studied, such as the dense small objects and background noise in high-altitude aerial photography images. To address these issues, an enhanced YOLOv8s-based model for detecting small objects is presented. The proposed model incorporates a parallel multi-scale feature extraction module (PMSE), which enhances the feature extraction capability for small objects by generating adaptive weights with different receptive fields through parallel dilated convolution and deformable convolution, and integrating the generated weight information into shallow feature maps. Then, a scale compensation feature pyramid network (SCFPN) is designed to integrate the spatial feature information derived from the shallow neural network layers with the semantic data extracted from the higher layers of the network, thereby enhancing the network's capacity for representing features. Furthermore, the largest-object detection layer is removed from the original detection layers, and an ultra-small-object detection layer is applied, with the objective of improving the network's detection performance for small objects. Finally, the WIOU loss function is employed to balance high- and low-quality samples in the dataset. The results of the experiments conducted on the two public datasets illustrate that the proposed model can enhance the object detection accuracy in UAV image scenarios.

Keywords: deep learning; small object detection; feature enhancement; unmanned aerial vehicles



Citation: Ni, J.; Zhu, S.; Tang, G.; Ke, C.; Wang, T. A Small-Object Detection Model Based on Improved YOLOv8s for UAV Image Scenarios. *Remote Sens.* **2024**, *16*, 2465. <https://doi.org/10.3390/rs16132465>

Academic Editor: Riccardo Roncella

Received: 6 June 2024

Revised: 1 July 2024

Accepted: 3 July 2024

Published: 5 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The development of unmanned aerial vehicle (UAV) technology has led to the rapid application of camera-equipped UAVs, or general purpose drones, in a variety of fields [1–3]. These include agriculture [4,5], aerial photography [6], public safety [7], ecological protection [8], and military [9,10]. Therefore, it becomes particularly important to understand and analyze the information collected from these domains.

Deep learning-based object detection techniques are increasingly used in image scenarios [11–13]. Previous researchers have proposed various improved deep learning-based models in this field. For instance, Liu et al. [14] introduced a multi-scale feature fusion and representation module that was intended to facilitate the more effective reconstruction of residual images in the decoder stage. Qiu et al. [15] designed a multi-scale sliding window to rapidly extract candidate objects for better enhancing the objects and background suppression. Zhou et al. [16] proposed a new network, which introduced a correlation transformation module to fully utilize the positional information and correlations between objects. Fang et al. [17] proposed transforming small object detection into predicting residual images.

Despite the numerous advancements in deep learning-based models for small object detection, there are still several challenges in applying deep learning techniques to UAVs

small object detection. The main problems include the following: (1) The high flying altitude of UAVs and diverse angles during shooting pose interference to object detection. In addition, UAVs capture a wide variety of objects, leading to an imbalance in the number of samples of different types and sizes. (2) A higher proportion of small-sized objects in images captured by UAVs raises greater demands on the performance of detectors. (3) The presence of intricate variations in background noise levels between densely clustered objects results in a pronounced decline in the accuracy of detection.

In order to improve the precision and address the challenges posed by varying shooting angles and background noise, an enhanced model based on YOLOv8s is presented in this study. The main contributions of this paper are summarized as follows: (1) A parallel multi-scale feature extraction (PMSE) module is designed, which feeds the input features into two parallel dilated convolutions and deformable convolutions, generating adaptive weights for different receptive fields. This module can facilitate the perception of small objects adaptively. (2) A scale compensation feature pyramid network (SCFPN) module is designed, which compensates for the shallow feature maps of the backbone network by extracting high-dimensional feature maps at the neck, further enhancing the perception capability of coordinate information and size information. (3) A WIOU loss function is used to optimize the dynamic focusing of the relationship between different samples in UAV images, thus improving the model's detection accuracy for objects at long distances.

The key differences between the existing small object detectors for UAVs and the proposed model are as follows: Most small object detectors for UAVs have focused on the redesign of anchors, the augmentation of data, or the cropping of images. There have been few attempts to shift the focus of algorithmic self-learning towards the learning of small objects as a means of improving detection accuracy. However, in the proposed model, the PMSE module can improve the detection effect of occluded objects through the captured object context information. The SCFPN module has a higher receptive field and stronger feature representation capability. Meanwhile, a weighting term for small objects is introduced into the loss function to balance the training sample imbalance between small and large objects, thus further improving the performance of small object detection.

This paper is organized as follows: Section 2 introduces the related work of UAV object detection. Section 3 provides the details of the proposed YOLOv8s-based model for UAV object detection. In Section 4, the proposed model is compared with the state-of-the-art object detection model on two public datasets, and the results are analyzed. Section 5 carries out some discussions on the proposed model. Section 6 is the conclusion.

2. Related Work

2.1. Advances in Object Detection

Object detection models based on convolutional neural networks (CNNs) are mainly classified into two types: two-stage-based and one-stage-based models. The two-stage-based models are often referred to as candidate region-based methods, and the representative models include RCNN, Fast RCNN [18], and Faster RCNN [19,20]. Subsequently, to further optimize the detection performance, a series of improved models, such as FPN [21,22] and Mask RCNN [23], have emerged to enrich the model's components and improve the detection ability. Nevertheless, the two-stage models, despite exhibiting superior detection accuracy, are characterized by a considerable number of model parameters, which consequently results in a prolonged computation time. So the two-stage-based models have many limitations. In comparison to two-stage-based methods, one-stage object detection models accomplish the object position estimation and category classification by one computation, which is characterized by fast detection speed. Representative algorithms include YOLO [24,25] and its improved versions, such as YOLOv2, YOLOv3, YOLOv4, etc. [26,27].

There are some object detection models based on transformers, such as Vision Transformer (ViT) and Detection Transformer (DETR). The ViT model is a pioneer in applying transformers in the field of computer vision, with a larger receptive field, more flexible

weight-setting methods, and global modeling ability in feature learning [28]. Recently, lots of improved models based on Transformer have been presented. For example, Liu et al. [29] proposed a Swin Transformer model using sliding windows, constructed a hierarchical transformer, and introduced localization ideas to limit attention computation within the window. Wang et al. [30] proposed a transformer based on a convolutional backbone network and designed a novel pyramid structure that reduces the length of input sequences through hierarchical attenuation of image resolution. Carion et al. [31] designed a DETR model that treats the object detection task as a set prediction problem, treating the bounding boxes on each image as a set, and using set prediction for object detection. Although the above methods achieve end-to-end prediction and simplify the prediction process, the ViT model generates a significant amount of computation when processing high-resolution images, resulting in poor detection performance for small target objects. The DETR model uses a fixed length target query vector for global interaction with image features. This approach requires long attention weight training to focus on sparse and critical positions on the feature map, resulting in long model convergence time and poor detection performance on small objects.

The object detection algorithm for UAVs has high practical requirements for real-time and detection accuracy [32]. Thus, the transformer-based object detection models and the two-stage object detection models are not suitable for the object detection tasks in this study. It is necessary to study a network that can achieve high detection accuracy while maintaining fast inference time for aerial images. The one-stage YOLOv8 model represents one of the latest YOLO models, which is divided into five models with different parameter quantities. Among them, YOLOv8s has a low parameter count and high detection accuracy [33], which is used as the baseline network of the proposed model in this study.

2.2. UAV Object Detection

Compared to traditional object detection, there are a number of additional challenges in UAV object detection, including the prevalence of small targets, the potential for image occlusion, and the risk of information loss. A portion of the research pertains to the integration of information from a single feature layer. For example, Ye et al. [34] combined characteristics from a single layer to develop an innovative quantized feature extraction module, with the aim of enhancing the detection precision of small objects. Xu et al. [35] designed a BiFPN to achieve the bidirectional fusion of bottom-up and top-down shallow and deep feature maps, and fused more features through the feature information flow of the same layer to alleviate the false detection due to occlusion conditions.

While the above structure effectively enhances the expressive power of the network, it ignores the fact that different levels of the network have different sensitivities to small objects. For small objects, the detailed information in the shallow feature maps facilitates accurate localization. There are limitations in information extraction by only streaming information from the features at the same layer. Therefore, this paper presents a multi-scale feature compensation pyramid network that integrates multi-layer contextual information of varying scales to generate more comprehensive features.

In the field of small object detection for UAV images, there is also some research that seeks to enhance the attention of small objects by introducing an attention mechanism module. For example, Yang et al. [36] redesigned the network's neck by incorporating a triple attention mechanism module, which enhanced the feature extraction capabilities so as to detect objects in object-dense scenes and images covering a large area. Liu et al. [37] demonstrated an improvement in detection accuracy through the introduction of a SimAM attention module, which effectively fuses features for particularly small objects. Ye et al. [38] designed a network based on convolution and transformers, where an attention-enhanced transformer module is designed to improve the ability of feature extraction. Shen et al. [39] strengthened the perception of small target location information by introducing the coordinate attention mechanism.

Regarding attention mechanisms, the above literature emphasizes that the application of attention mechanisms can improve attention to small objects. However, the inclusion of a substantial quantity of background information in the shallow feature layer can impede the process of attending to minute objects by the attention mechanism, which results in suboptimal performance in the event of limited features. Some attention mechanisms focusing only on the position channel will be affected by the background noise, resulting in the low detection accuracy of the model. In addition, the multiple-attention mechanism requires multiple attention calculation matrices, which is bound to increase a large amount of computation compared with single-channel attention, reducing the real-time performance of the model.

Therefore, a feature enhancement module is designed in the proposed model, which adaptively extracts small target features from space and channels through three convolutional layers after splitting the input features, reducing the amount of computation by splitting the features. It uses deformation convolution and dilated convolution to help the neural network filter out a part of the noise information through the receptive fields of different sizes, and enhances the anti-interference ability of background noise [22,40]. Based on the proposed feature enhancement module, different receptive fields of different scales can extract different details and structures in the image, and these features can complement and enhance each other, improving the feature richness and expressive ability of the object detection model.

3. Proposed UAV Object Detection Model

In order to enhance the detection precision of small objects for UAVs, an improved YOLOv8s-based model is proposed. The network architecture of the proposed model is shown in Figure 1, which includes three main improved parts, namely, the PMSE module, the SCFPN module, and the WIOU loss function.

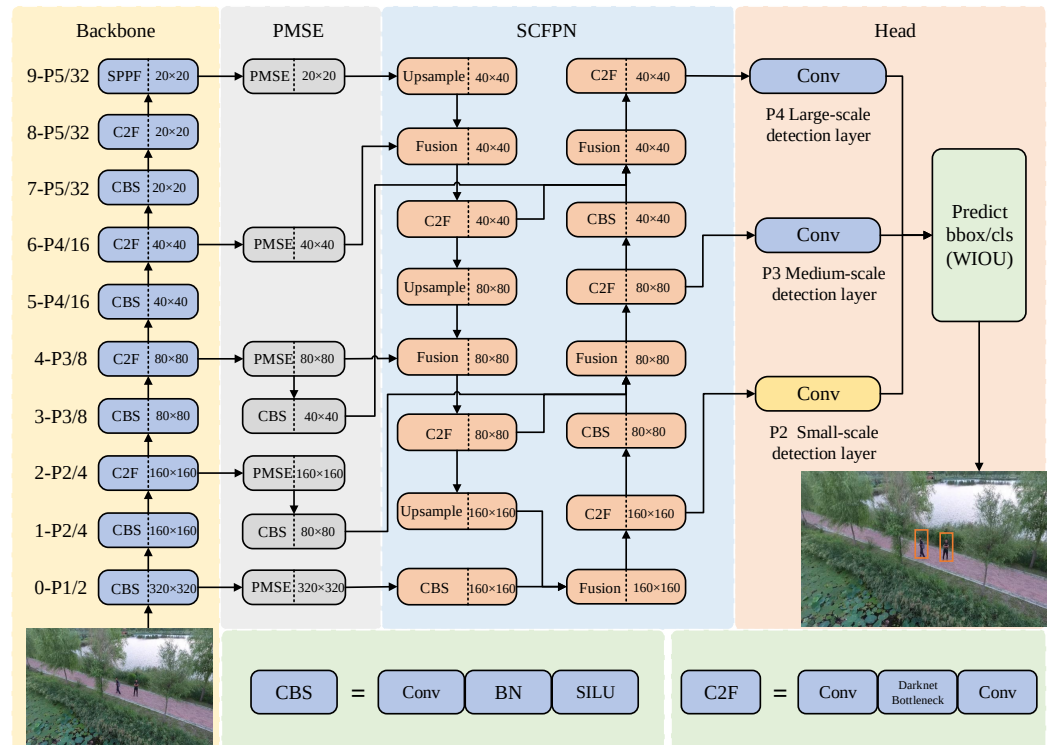


Figure 1. The whole structure of the proposed UAV object detection model. In the proposed model, PMSE denotes the parallel multi-scale feature extraction module, which expands the receptive field of

feature extraction and improves the correlation between features by introducing dilated convolution and variational convolution; SCFPN denotes the scale compensation feature pyramid module, which compensates for the shallow feature maps of the backbone network with high-dimensional feature maps extracted from the neck; and WIOU loss function is used to optimize the dynamic focusing of the small objects. Conv is the convolutional layer; BN is the normalization layer; SILU is the activation function; and Darknet Bottleneck is the feature extraction structure of the general YOLOV8s. CBS = Conv + BN + SILU. C2F = Conv + Darknet Bottleneck + Conv. In this model, we add a CBS layer behind the PMSE module to change the size of the feature map and maintain a consistent feature map size during feature fusion (see the size of the feature map in each feature layer).

3.1. The Proposed PMSE Module

The semantic information in the shallow feature maps of the YOLOv8s backbone network is disordered, and the receptive field is small. During the bidirectional context feature fusion process, it is easy to cause feature information loss, and the large amount of background information contained in the shallow feature layers will cause interference in the fusion process. To address these issues, a feature enhancement module PMSE is proposed to improve the feature fusion performance of the backbone network. The network structure and parameters structure of the proposed PMSE module are shown in Figure 2 and Figure 3, respectively.

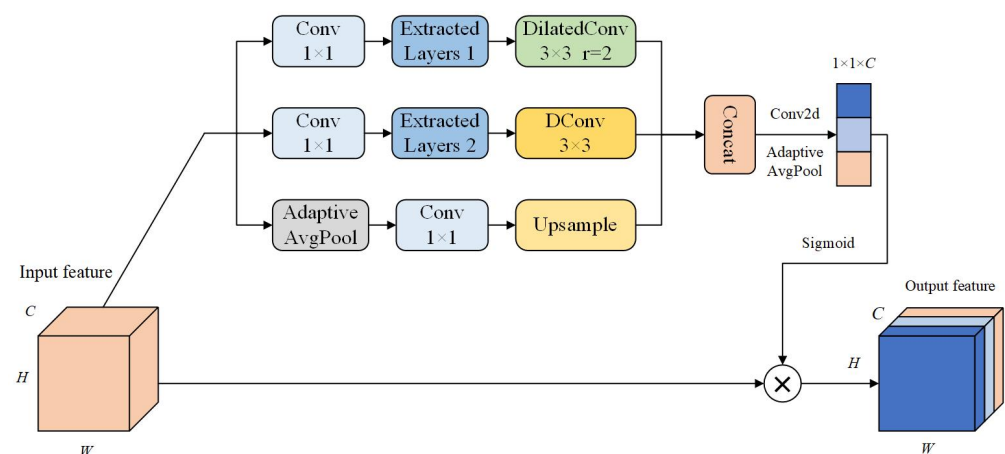


Figure 2. The structure of the parallel multi-scale feature extraction module (PMSE).

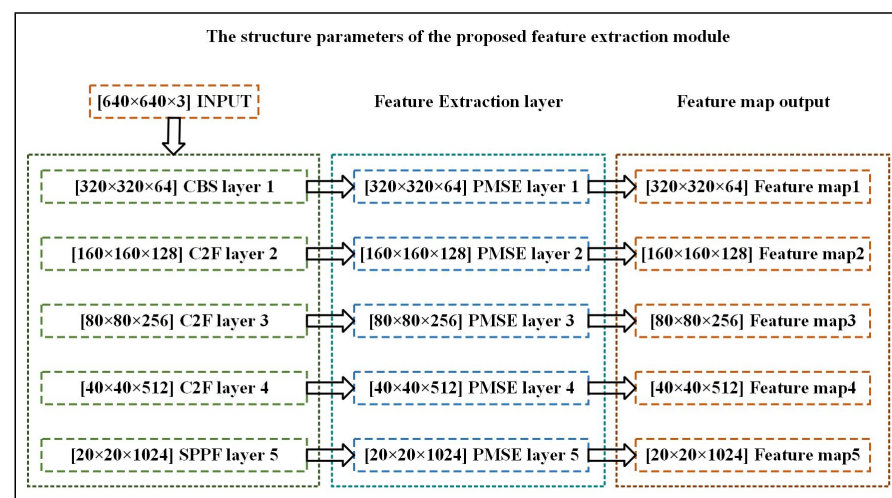


Figure 3. The structure parameters of the proposed PMSE module.

As shown in Figure 2, the feature map first passes through three parallel branches. One branch involves a convolutional kernel of size 1×1 to adaptively adjust the input channels, thereby reducing network complexity. Then, through the extracted layer convolutional layer, the feature information is further extracted: one path consists of 3×3 convolution combined with dilated convolution of rate 2, while the other path consists of 3×3 convolution combined with deformable convolution. The third branch preserves the original feature information through global average pooling, then reorganizes the space channel through a convolutional kernel of size 1×1 to reduce parameters, and restores the size information of the feature map through upsampling. The feature maps generated by the three branches are concatenated in the channel dimension. Then, the features are compressed along the spatial dimension to 1×1 scalar, and output as $1 \times 1 \times C$ feature map through global average pooling. Finally, through sigmoid non-linear feature mapping, the weight of each channel is multiplied by the weight of the original feature map, adaptively tuning algorithms for the target of interest.

The calculation process of the proposed PMSE module is as follows:

$$F' = S(F) + G(F) + T(F) \quad (1)$$

$$F'' = M(F') \quad (2)$$

$$F''' = F \otimes F'' \quad (3)$$

where F is the original feature map; F' is the feature map after splicing through three branches; $S(F)$ denotes the local features output from the first branch; $G(F)$ denotes the local features output from the second branch; $T(F)$ denotes the local features output from the third branch; F'' is the feature map activated by a function after compression and reorganization of the spliced feature map along the spatial dimensions; $M(F)$ is the globally average-pooled local features; and F''' is the final output feature map.

There are two extracted layers (ELs) in the proposed PMSE module (see Figure 2). The EL layer introduces a residual block structure, which helps to address the issues of gradient explosion and vanishing gradients. It can not only extract feature information at a deeper level but also ensures the robustness of the model. By using two 1×1 convolution layers to adjust the model's channel numbers, it reduces the computational load while maintaining the same number of output channels for the model. The structure of the EL is illustrated in Figure 4, and its calculation process is as follows:

$$M_o(F') = H(F) + W(F) \quad (4)$$

where $M_o(F')$ represents the output feature map obtained after processing with the residual block, $H(F)$ represents the local features obtained from the residual branch, and $W(F)$ represents the local features output by the convolutional branch.

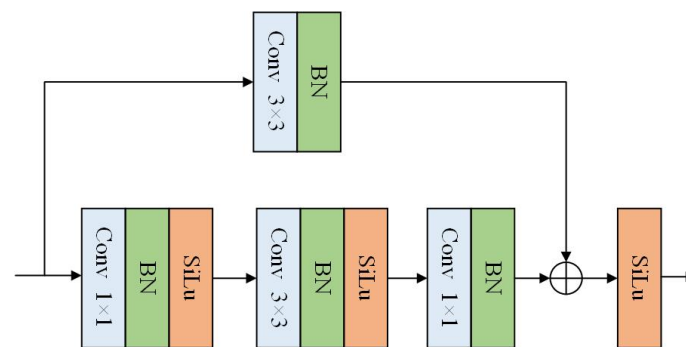


Figure 4. The structure of the EL layer in the proposed PMSE module.

Remark 1. By using convolutional kernels of different sizes, feature information can be extracted at different scales. The 3×3 convolutional kernel can capture local detailed information, while deformable convolution can capture a larger range of contextual information through adaptive receptive fields. Using dilated convolution and deformable convolution can flexibly adjust the sampling positions of convolutional kernels to adapt to the shapes and sizes of different objects, better capturing the deformations and geometric transformations of objects, thus improving the model's perceptual ability toward the objects. Thus, the proposed PMSE module can adaptively adjust the distribution of weights, enhancing useful features and suppressing irrelevant information. When inputting feature maps of varying sizes, the model is capable of adapting in such a way as to capture the receptive field of small objects. This adaptation is intended to enhance the detection performance of these objects.

3.2. The Proposed SCFPN Module

In the process of deep learning computation, as the network deepens gradually, the model complexity increases. In most cases, each layer of the network will cause a certain feature loss. In a neural network model, shallow networks have higher resolution than deep networks, providing more accurate location information for object detection. On the other hand, deep networks have a larger receptive field compared to shallow networks, containing more semantic information beneficial for object classification. Therefore, adopting a multi-scale feature fusion method to integrate feature information at different scales can reduce the feature loss caused by deepening the network.

The neck structure used in the general YOLOv8s model is PA-FPN [41], which establishes a bottom-up channel on top of the feature pyramid network (FPN). The structure of PA-FPN is shown in Figure 5a. To further enhance the detection performance of small objects, a new scale compensation pyramid network (SCFPN) is proposed in this study. The structure of the proposed SCFPN module is shown in Figure 5b.

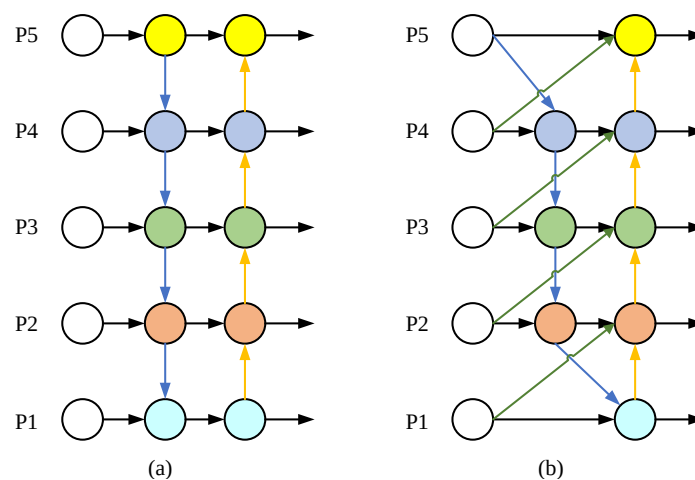


Figure 5. Two types of feature pyramid structures. (a) PA-FPN. (b) SCFPN.

In the proposed SCFPN module, firstly, nodes with only one feature input are removed to reduce the computational load further. The main reason is that when a node has only one feature for fusion, its contribution to feature extraction in the network is minimal. After considering the network's computational complexity and detection performance, these nodes are deleted.

Subsequently, the feature information in the backbone network is input into the bottom-up channel with the same dimensions, allowing more comprehensive interaction between low-level and high-level features to capture more positional and semantic information.

Finally, the feature maps from the backbone network are input across dimensions into the top-down pyramid. The low-level feature maps contain more positional information, and they are re-inputted to extract positional information at a higher-dimensional level,

enabling the entire feature pyramid structure to extract more positional information of small targets.

In addition, the general YOLOv8s model faces challenges in handling small objects due to the feature maps being obtained through simple downsampling, leading to the loss of relevant information when the downsampling multiplier is too large. The main reason is that high-level feature maps undergo multiple convolution operations, such as P4 and P5. At this time, the resolution of the feature map is very small, which only contains a large amount of semantic information of the object. However, shallow feature maps have a higher resolution and contain a large amount of position information for small objects, such as P1 and P2. To enhance the model's positioning accuracy for objects while maintaining good semantic information, high-level feature maps can be fused with the position information of shallow networks by performing multiple upsampling operations over longer paths [21,42]. Thus, to address these problems and enhance the model's accuracy and robustness, in the proposed SCFPN module, we use the input of the five feature layers P1, P2, P3, P4, and P5, and add an ultra-small detection layer P2.

After the original detection layer of the YOLOv8s model, to further extend the feature map, we incorporate convolution and upsampling operations because in the general YOLOv8s model, for small objects, the feature at P5 is downsampled to a 1×1 feature map, which leads to the problem of insufficient features. Therefore, in our proposed model, P2, P3 and P4 are used to output feature maps (with sizes of 160×160 , 80×80 , and 40×40 respectively) to achieve small object detection, and optimize the network structure of P5 layer (20×20). By obtaining lower spatial features and fusing them with higher-level semantic features to generate a P2 detection layer, it will result in an 8×8 feature map, which can better improve the model's ability to detect smaller objects. The two types of prediction heads are shown in Figure 6.

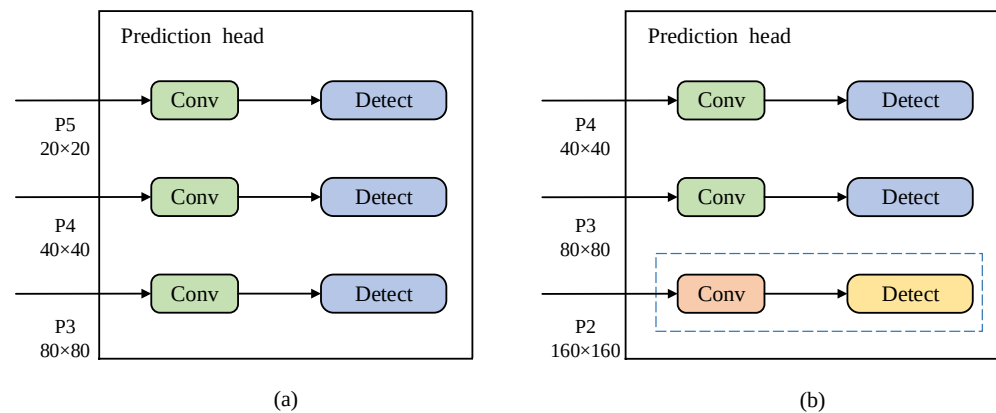


Figure 6. Two types of prediction heads. (a) The prediction head of the general YOLOv8s. (b) The prediction head of our proposed model.

The conventional approach to feature fusion often entails the overlaying or addition of feature maps, such as the utilization of Concat or Shortcut connections, without the appropriate allocation of weights to the concurrently added feature maps. In fact, the information content of different input feature maps varies, and their contributions to the fusion of input feature maps are also distinct. The mere addition or overlaying of these feature maps is insufficient to fully extract features. Thus, a weighted feature fusion mechanism is employed in the proposed SCFPN module [43]:

$$O = \sum_i \frac{\omega_i}{\epsilon + \sum_j \omega_j} \cdot I_i \quad (5)$$

where i and j represent the number of input layers for feature fusion nodes, with $i = j$; ω_i denotes the weights of respective input layers; and I is the input feature map of the node.

The use of the Relu activation function for non-linear mapping ensures $\omega_i \geq 0$, setting $\epsilon = 0.0001$ to prevent instability during training. This calculation method scales the weight values to be within $[0, 1]$, leading to fast and efficient training.

Remark 2. In the proposed SCFPN module, the improvements result in better performance and accuracy of the network in handling small object detection. An extra edge is added between the original input node and the output node in the proposed SCFPN module for location information extraction (see Figure 5b), which does not introduce new nodes nor increase the number of additional parameters. Adding an ultra-small-object detection P2 layer to the proposed SCFPN network improves the detection performance of small target objects (see Figure 6b). The novel weighted feature fusion mechanism in the fusion part allows adaptive adjustment of the contribution to the object detection based on input feature maps, facilitating more comprehensive feature extraction.

3.3. The Loss Function

The accuracy of the model in locating objects, particularly in UAV scenarios where objects are often small, can be improved by using an optimized loss function. The original YOLOv8 uses DFL and CIOU for bounding box regression. However, CIOU does not consider the balance of the dataset between high and low-quality samples. Therefore, in this paper, an improved WIOUv3 loss function instead of CIOU is adopted to ensure better bounding box prediction regression for small objects [44,45]. WIOUv1 is computed using a two-layer attention mechanism, which is as follows:

$$L_{IOU} = 1 - IOU \quad (6)$$

$$L_{WIOUv1} = R_{WIOU} \times L_{IOU} \quad (7)$$

$$R_{WIOU} = \exp\left(\frac{(b_x - b_x^{gt}) + (b_y - b_y^{gt})}{(b_w^2 + b_h^2)}\right) \quad (8)$$

where the coordinates of the center point of the actual box are represented by b_x and b_y , while the coordinates of the center point of the prediction box are represented by b_x^{gt} and b_y^{gt} . The width and height of the minimum perimeter rectangle between the prediction box and the actual box are represented by b_w and b_h .

WIOUv3 defines the anomaly degree coefficient β to measure the quality of anchor frames, uses β to construct the non-monotonic focusing coefficient γ and applies it to WIOUv1, with α and δ as hyperparameters to regulate the quality cut-off criterion. WIOUv3 adopts a dynamic gradient gain assignment strategy to reasonably allocate the high and low weights of quality anchor frames. The specific computation of WIOUv3 is as follows:

$$L_{WIOUv3} = \gamma \times L_{WIOUv1} \quad (9)$$

$$\gamma = \frac{\beta}{\delta \alpha^{\beta-\delta}} \quad (10)$$

$$\beta = \frac{L_{IOU}^*}{L_{IOU}} \quad (11)$$

Remark 3. The proposed WIOUv3 increases the focusing mechanism by calculating the focusing coefficients, achieves the boundary frame regression loss for dynamic focusing, and optimizes the dynamic focusing on small objects in UAV images. Furthermore, the introduction of WIOUv3 does not lead to an increase in additional parameters, thereby maintaining the lightness.

4. Experiments

4.1. Datasets and Metrics

Two public datasets are used for the purposes of training and testing the proposed model, namely, the VisDrone2019 dataset and the DOTA dataset. The two datasets are introduced as follows:

(1) VisDrone2019 [46]: This dataset contains 10,209 still images from different scenes and environments. To facilitate the validation of the algorithm for detecting small objects, this paper further divides the dataset through image preprocessing. In the completed dataset, 6471 images are used as the training set, and 548 images are used as the validation set. The VisDrone2019 dataset was captured using drones, covering common objects such as pedestrians, cars, bicycles, tricycles, and 10 other classes.

(2) DOTA [47]: This dataset contains 2806 high-resolution aerial images from different scenes and environments, with a total of 188,282 annotated object instances. To facilitate the validation of the algorithm for detecting small objects, this paper further divides the dataset through image preprocessing. In the completed dataset, 15,749 images are used as the training set, and 5297 images are used as the validation set. Fourteen common objects are selected, including large vehicles, small vehicles, ships, bridges, and airplanes.

In the metrics of object detection, most of the references in the small object detection for UAVs use the mean average precision (mAP) as the evaluation metrics [39,48,49]. Here, the average precision (AP) is a numerical measure that summarizes the shape of the Precision–Recall curve, which is used to determine the optimal balance between precision and recall in practice [13]. So mAP is used as the primary metric in the comparison experiments, which can comprehensively evaluate the performance of the models. Additionally, parameters and floating point operations (FLOPs) are utilized as supplementary evaluation criteria to assess the calculation efficiency. The definition of mAP is as follows:

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (12)$$

where n represents the number of object categories that have been identified, and AP_i represents the area under the Precision–Recall curve for the i -th object category [50].

4.2. Experimental Settings and Implementation Details

In this paper, we utilize the Ubuntu 20.04 system as the experimental platform. The experimental environments are PyTorch 1.10.0, CUDA 11.3 and Python 3.8. All models were trained, tested, validated under identical hyperparameters. The dimensions of the input feature maps for both the VisDrone2019 and DOTA datasets were set to 640×640 , the maximum number of detections was fixed to 100, and the number of training rounds for model training was 100. The NMS threshold was set to 0.65; the specific parameter settings are shown in Table 1.

Table 1. Experimental parameter configurations.

Network parameters	Learning rate	0.0005
	Weight decay	0.0005
	Batch size	8
	Workers	8
	Momentum	0.9
	Image size	640×640
Computer configurations	Operating System	Ubuntu 20.04
	Cpu	2.10GHZ
	Gpu	GeForce RTX 3090
	RAM	24.0 GB

4.3. Comparison Experiments

In order to valuate the efficacy of the proposed model in the context of small object detection in UAV image scenarios, a series of comparative experiments were conducted on the Visdrone2019 and DOTA datasets. Some models were selected for comparison, including some classical two-stage models and one-stage models and some state-of-the-art models proposed recently. The main reasons for these models being selected in this study are as follows: (1) the code for these models is available, or (2) the experiments conducted on these models utilize the same public datasets employed in this paper.

(1) *VisDrone2019*: The results of the comparison experiment on the VisDrone2019 dataset are shown in Table 2. In this experiment, the comparison models include SSD [51], RetinaNet [52], Faster-RCNN [19], YOLOv5s [37], YOLOv8s [53], YOLOv9 [54], GCGE-YOLO [55], TPH-YOLO [48], ATSS [56], LV-YOLOv5 [57], Swin Transformer [29], C3TB-YOLOv5 [49], and DTSSNet [58].

Table 2 indicates that our detection model has a mAP_{50} of 47.1%, which is an improvement of 11.7% in comparison to the model YOLOv8s, and mAP improved by 7.1%. The parameters are decreased by 0.9 M. In comparison to the suboptimal model LV-YOLOv5, the mAP_{50} is improved by 5.4% and 3.1%. The parameters are decreased threefold. Although GCGE-YOLO has the smallest parameters and FLOPs, the mAP_{50} of our model is improved by 13% compared to GCGE-YOLO. Compared to the transformer-based C3TB-YOLOv5 model, the mAP_{50} of our model is improved by 8.8% and mAP by 6.7%. Compared to the model YOLOv9, the mAP_{50} of our model is improved by 3.1% and mAP is improved by 2.2%, and our model parameter counts and FLOPs are much smaller than YOLOv9. Comprehensively, it reveals that the improved model greatly improves the small object detection accuracy, making it more efficient and feasible in detecting UAV aerial images.

Table 2. Detection performance (%) of different methods on Visdrone2019 validation dataset.

Models	mAP_{50}	mAP	Parameters (M)	FLOPs (G)
SSD [51]	23.9	13.1	24.5	87.9
RetinaNet [52]	27.3	15.5	19.8	93.7
ATSS [56]	31.7	18.6	10.3	57.0
Faster-RCNN [19]	33.2	17.0	41.2	207.1
GCGE-YOLO [55]	34.1	19.2	4.5	10.8
YOLOv5s [37]	35.4	20.5	9.1	23.8
Swin Transformer [29]	35.6	20.6	34.2	44.5
YOLOv8s [53]	36.4	21.6	11.1	28.5
C3TB-YOLOv5 [49]	38.3	22.0	8.0	19.7
TPH-YOLO [48]	39.3	23.6	51.5	138.1
DTSSNet [58]	39.9	24.2	10.1	50.4
YOLOv9 [54]	43.4	26.5	51.0	239.0
LV-YOLOv5 [57]	41.7	25.6	36.6	38.8
Ours	47.1	28.7	10.2	64.9

(2) *DOTA*: The results of the comparison experiment on the DOTA dataset are shown in Table 3. In this experiment, the comparison models include SSD [51], RetinaNet [52], Faster-RCNN [19], YOLOv5s [37], YOLOv8s [53], SDS-Det [59], A2-Net [60], FSoD-Net [61], RoI-Transformer [62], Swin-YOLOv5 [63], and SCA-YOLO [64].

Table 3 indicates that our detection model mAP_{50} achieved a score of 74.2%, which represents a 3.5% improvement in comparison to the model YOLOv8s. Additionally, mAP demonstrates a 3.4% increase. The parameters are reduced by 0.9 M. In comparison to SCA-YOLO, which has the highest detection accuracy, although mAP_{50} is reduced by 1.6%, our model is more superior in terms of parameters and FLOPs. The parameters are only one-quarter in SCA-YOLO, and the FLOPs is only one-half in SCA-YOLO. Despite SDS-Det

having the smallest parameters and FLOPs, our model achieves a mAP_{50} that is 6.7% higher than that of SDS-Det. The performance of the two Transformer-based models are not good either on this dataset. Overall, our model demonstrates a balance between lightweight and detection accuracy, meeting the real-time requirements for UAV detection.

Table 3. Detection performance (%) of different methods on DOTA validation dataset.

Models	mAP_{50}	mAP	Parameters (M)	FLOPs (G)
Faster-RCNN [19]	42.0	26.2	41.2	207.1
SSD [51]	52.4	32.3	24.5	87.9
RetinaNet [52]	61.6	38.7	19.8	93.7
SDS-Det [59]	67.5	43.0	4.9	11.3
A2-Net [60]	69.5	/	9.6	50.3
RoI-Transformer [62]	69.6	/	/	/
YOLOv5s [37]	70.2	46.5	9.1	23.8
YOLOv8s [53]	70.7	46.7	11.1	28.5
Swin-YOLOv5 [63]	74.7	/	/	/
FSOD-Net [61]	75.3	/	232.2	165.0
SCA-YOLO [64]	75.8	50.7	47.9	126.4
Ours	74.2	49.7	10.2	64.9

Note: “/” means that the original paper does not provide the corresponding results.

5. Discussion

In Section 4, the superiority of the proposed model is proven by comparison experiments. In this section, the performance of the key improved module (SCFPN) is discussed by some comparison experiments first. Then, to further discuss the effectiveness of the improvements in the proposed model, some ablation experiments are conducted. At last, the qualitative result analyses are provided to further discuss the effectiveness of the proposed model for UAV small object detection.

5.1. About the SCFPN Structure

To show the efficiency of the proposed SCFPN structure, some comparison experiments are conducted on VisDrone2019, where the baseline model is YOLOv8s and the feature pyramid structure used for comparison includes PA-FPN [41], BiFPN [43], and AFPN [65]. The results of these experiments are shown in Table 4.

Table 4. Comparison of different feature pyramid structures on VisDrone2019.

Models	mAP_{50}	mAP	Parameters (M)
PA-FPN [41]	36.4	21.6	11.1
BiFPN [43]	36.7	21.9	7.4
AFPN [65]	34.6	20.3	7.1
SCFPN (Ours)	41.9	25.3	7.6

Table 4 reveals that the mAP_{50} is enhanced by 5.5% when the SCFPN feature structure is employed, while parameters are reduced to 3/4 of the base model. Furthermore, the mAP_{50} of SCFPN is augmented by 5.2% in comparison to the BiFPN model. Comparing to the AFPN model, the mAP_{50} of the SCFPN model is observed to increase by 6.3%. The final experimental results demonstrate that the proposed SCFPN module is effective in enhancing the detection of small target objects.

5.2. About Different Detection Layers

To evaluate the impact of detection layers of different sizes on the accuracy of small object detection, some ablation experiments are conducted on the VisDrone2019 dataset, where YOLOv8s is selected as the baseline model in these experiments. The performance comparison between different modules is shown in Table 5.

Table 5. Comparison of different detection layers on VisDrone2019.

Baseline	P2	P3	P4	P5	mAP_{50}	mAP	Parameters (M)	FLOPs (G)
✓		✓			36.5	21.6	9.4	25.6
✓		✓	✓		36.5	21.7	10.0	27.6
✓		✓	✓	✓	36.4	21.6	11.1	28.5
✓	✓	✓			41.1	24.6	9.5	34.9
✓	✓	✓	✓		41.3	24.8	9.7	36.5
✓	✓	✓	✓	✓	41.3	24.8	10.6	37.0

According to the results in Table 5, it can be seen that when only the P3 detection layer is used, the mAP_{50} is 36.5%, and the number of model parameters is 9.4 M. When the detection layer P4 is added, the mAP rises by 0.1%, and the number of model parameters rises by 0.6 M. When layers P4 and P5 are added, the mAP_{50} is 36.4%, which decreases by 0.1%, and the number of parameters rises by 1.7 M. When P2, P4, P5 layers are added, the mAP_{50} is 41.3%, and the number of parameters rises by 1.2 M. When P2 and P3 layers are used, the mAP_{50} is 41.1%, and the number of parameters of the model is 9.5 M. When the P2, P3, and P4 layers are used, the mAP_{50} is 41.3%, and the number of parameters of the model is 9.7 M. We can see that the addition of the P2 layer greatly enhances the model's performance in detecting small objects, whereas the performance of the P5 layer on the detection of small objects is not good. The addition of the P4 layer increases the accuracy of object detection without introducing too many parameters and computational complexity. To consider the number of parameters and the detection accuracy comprehensively, the P2, P3, and P4 detection layers are used in our proposed model.

5.3. Ablation Experiments for the Proposed Modules

In order to evaluate the impact of the proposed three modules on the performance of object detection algorithms for UAV capture scenes under identical experimental conditions, ablation experiments are conducted on the VisDrone2019 dataset, and YOLOv8s is selected as the baseline model in these experiments. The performance comparison between different modules is shown in Table 6.

Table 6. Ablation experiments on VisDrone2019.

Baseline	PMSE	SCFPN	WIOU	mAP_{50}	mAP	Parameters (M)	FLOPs (G)
✓				36.4	21.6	11.1	28.5
✓	✓			37.6	22.4	11.9	31.7
✓		✓		45.6	27.8	7.4	59.1
✓			✓	38.2	22.4	11.1	28.5
✓	✓	✓	✓	39.1	23.2	11.9	31.7
✓	✓	✓		46.4	28.3	10.2	64.9
✓		✓	✓	46.6	28.4	7.4	59.1
✓	✓	✓	✓	47.1	28.7	10.2	64.9

According to Table 6, it can be seen that the PMSE module leads to an improvement in the mAP_{50} by 1.2% in comparison to the baseline model. The introduction of the SCFPN module results in a further improvement in mAP_{50} , with an increase of 9.2%. This is

achieved while simultaneously reducing the parameters to 9/10 of the baseline model. The use of the WIOU loss function results in an improvement in mAP_{50} in comparison to the baseline model of 1.8%, demonstrating the effectiveness of this approach. The combination of the three improvement methods yields an improvement of 10.7% in the mAP_{50} in comparison to the baseline model. The final experimental results demonstrate that all three module improvements are effective in enhancing the detection of small target objects.

5.4. Qualitative Result Analysis

To demonstrate the performance of the algorithm in practical scenarios, images from the VisDrone2019 and DOTA datasets are tested separately. Here, only the results of the proposed method and YOLOv8s (the baseline model of our method) are given for comparison; see Figure 7 and Figure 8, respectively.



Figure 7. Comparison of detection results on the Visdrone2019 dataset. (a) Based on YOLOv8s. (b) Based on the proposed model.

As shown in Figure 7, in the first column of the image, the YOLOv8s algorithm misses detection when there are obstacles such as trees and vehicles, while our improved algorithm can detect pedestrians and vehicles that are obstructed. In the detection images in the second column, the YOLOv8s algorithm performs poorly in detecting dense pedestrians in the distance, while our improved algorithm can effectively distinguish pedestrians in the distance.

As shown in Figure 8, in the first column of images, the YOLOv8s algorithm has obvious limitations in detecting small vehicles, while our improved algorithm can effectively detect small vehicles. In the detection images in the second column, the YOLOv8s algorithm performs poorly on storage tanks captured from different angles, while our improved algorithm can effectively detect different storage tanks.



Figure 8. Comparison of detection results on the DOTA dataset. (a) Based on YOLOv8s. (b) Based on the proposed model.

6. Conclusions

In this study, we present an improved YOLOv8s algorithm. Its objective is to improve the accuracy of the detection of small target objects under different conditions. Firstly, a parallel multi-scale feature extraction module (PMSE) is proposed, which inputs features into two parallel dilated convolutions and deformable convolutions. Adaptive weights for different receptive fields are generated by compressing the channel and expanding the self owned module. The generated receptive field weights are integrated into the initial feature map through non-linear mapping to improve the representation ability of small target object information. Then, a scale compensation feature pyramid network (SCFPN)

is proposed, which inputs the shallow feature maps of the network backbone layer by layer into the neck to further extract small target feature information from the high-level dimensional features. The original large-target detection head in the prediction head is replaced with an ultra-small-object detection head to improve the perception ability of small target position and size information. Finally, the WIOU loss function is applied to optimize the dynamic focusing of small objects in drone images, enhancing the detection accuracy and stability. The final experiments show that the improved YOLOv8s UAV object detection algorithm has an improvement in the actual detection accuracy. However, the floating-point operation is increased. The real-time detection performance is reduced, considering the limited resources for UAV detection; therefore, the model with high real-time performance can be further investigated.

Author Contributions: Conceptualization, S.Z. and J.N.; methodology, S.Z.; validation, J.N., C.K. and T.W.; writing—original draft preparation, S.Z.; writing—review and editing, G.T., C.K. and T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61873086) and Jiangsu Province Key R&D Program (BE2023340).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: <https://github.com/VisDrone/VisDrone-Dataset> (accessed on 27 July 2023), and <https://captain-whu.github.io/DOTA/index.html> (accessed on 17 December 2023).

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

References

1. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A Survey of Object Detection for UAVs Based on Deep Learning. *Remote Sens.* **2024**, *16*, 149. [\[CrossRef\]](#)
2. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Vehicle Detection From UAV Imagery With Deep Learning: A Review. *IEEE Trans. Neural Networks Learn. Syst.* **2022**, *33*, 6047–6067. [\[CrossRef\]](#)
3. Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.C.; Shi, P. Deep learning-based scene understanding for autonomous robots: A survey. *Intell. Robot.* **2023**, *3*, 374–401. [\[CrossRef\]](#)
4. Bo, W.; Liu, J.; Fan, X.; Tjahjadi, T.; Ye, Q.; Fu, L. BASNet: Burned Area Segmentation Network for Real-Time Detection of Damage Maps in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5627913. [\[CrossRef\]](#)
5. Zhu, J.; Yang, G.; Feng, X.; Li, X.; Fang, H.; Zhang, J.; Bai, X.; Tao, M.; He, Y. Detecting Wheat Heads from UAV Low-Altitude Remote Sensing Images Using Deep Learning Based on Transformer. *Remote Sens.* **2022**, *14*, 5141. [\[CrossRef\]](#)
6. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient Detection of UAV Image Based on Cross-Layer Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608911. [\[CrossRef\]](#)
7. Sun, Y.; Shao, Z.; Cheng, G.; Huang, X.; Wang, Z. Road and Car Extraction Using UAV Images via Efficient Dual Contextual Parsing Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5632113. [\[CrossRef\]](#)
8. Zhao, M.; Li, W.; Li, L.; Wang, A.; Hu, J.; Tao, R. Infrared Small UAV Target Detection via Isolation Forest. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5004316. [\[CrossRef\]](#)
9. Yang, Y.; Yang, F.; Sun, L.; Xiang, T.; Lv, P. Echoformer: Transformer Architecture Based on Radar Echo Characteristics for UAV Detection. *IEEE Sens. J.* **2023**, *23*, 8639–8653. [\[CrossRef\]](#)
10. Deng, A.; Han, G.; Chen, D.; Ma, T.; Liu, Z. Slight Aware Enhancement Transformer and Multiple Matching Network for Real-Time UAV Tracking. *Remote Sens.* **2023**, *15*, 2857. [\[CrossRef\]](#)
11. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards balanced learning for object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; Volume 2019, pp. 821–830.
12. Wang, X.; Wang, A.; Yi, J.; Song, Y.; Chehri, A. Small Object Detection Based on Deep Learning for Remote Sensing: A Comprehensive Review. *Remote Sens.* **2023**, *15*, 3265. [\[CrossRef\]](#)
13. Chen, Y.; Ni, J.; Tang, G.; Cao, W.; Yang, S.X. An improved dense-to-sparse cross-modal fusion network for 3D object detection in RGB-D images. *Multimed. Tools Appl.* **2023**, *in press*. [\[CrossRef\]](#)
14. Liu, H.; Sun, F.; Gu, J.; Deng, L. SF-YOLOv5: A Lightweight Small Object Detection Algorithm Based on Improved Feature Fusion Mode. *Sensors* **2022**, *22*, 5817. [\[CrossRef\]](#) [\[PubMed\]](#)

15. Qiu, Z.; Ma, Y.; Fan, F.; Huang, J.; Wu, L.; Du, Y. Improved DBSCAN for Infrared Cluster Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 5511905. [[CrossRef](#)]
16. Zhou, Y.; Chen, S.; Zhao, J.; Yao, R.; Xue, Y.; Saddik, A.E. CLT-Det: Correlation Learning Based on Transformer for Detecting Dense Objects in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4708915. [[CrossRef](#)]
17. Fang, H.; Xia, M.; Zhou, G.; Chang, Y.; Yan, L. Infrared Small UAV Target Detection Based on Residual Image Prediction via Global and Local Dilated Residual Networks. *IEEE Geosci. Remote Sens. Lett.* **2023**, *19*, 733–738. [[CrossRef](#)]
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; Volume 2015, pp. 1440–1448.
19. 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](#)]
20. Ni, J.; Chen, Y.; Chen, Y.; Zhu, J.; Ali, D.; Cao, W. A Survey on Theories and Applications for Self-Driving Cars Based on Deep Learning Methods. *Appl. Sci.* **2020**, *10*, 2749. [[CrossRef](#)]
21. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 936–944.
22. Ni, J.; Shen, K.; Chen, Y.; Yang, S.X. An Improved SSD-Like Deep Network-Based Object Detection Method for Indoor Scenes. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5006915. [[CrossRef](#)]
23. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Volume 2017, pp. 2980–2988.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 2016, pp. 779–788.
25. Ni, J.; Shen, K.; Chen, Y.; Cao, W.; Yang, S.X. An Improved Deep Network-Based Scene Classification Method for Self-Driving Cars. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5001614. [[CrossRef](#)]
26. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
27. Chen, L.; Shi, W.; Deng, D. Improved yolov3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images. *Remote Sens.* **2021**, *13*, 660. [[CrossRef](#)]
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16X16 words: Transformers for image recognition at scale. In Proceedings of the ICLR 2021—9th International Conference on Learning Representations, Virtual Online, 3–7 May 2021.
29. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 9992–10002.
30. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 548–558.
31. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Volume 12346 LNCS, pp. 213–229.
32. He, L.; Liao, K.; Li, Y.; Li, B.; Zhang, J.; Wang, Y.; Lu, L.; Jian, S.; Qin, R.; Fu, X. Extraction of Tobacco Planting Information Based on UAV High-Resolution Remote Sensing Images. *Remote Sens.* **2024**, *16*, 359. [[CrossRef](#)]
33. 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](#)]
34. Ye, T.; Qin, W.; Zhao, Z.; Gao, X.; Deng, X.; Ouyang, Y. Real-Time Object Detection Network in UAV-Vision Based on CNN and Transformer. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2505713. [[CrossRef](#)]
35. Xu, S.; Ji, Y.; Wang, G.; Jin, L.; Wang, H. GFSP-YOLO: A Light YOLO Model Based on Group Fast Spatial Pyramid Pooling. In Proceedings of the ICICN 2023–2023 IEEE 11th International Conference on Information, Communication and Networks, Xi'an, China, 11–17 August 2023; pp. 733–738.
36. Yang, Y.; Gao, X.; Wang, Y.; Song, S. VAMYOLOX: An Accurate and Efficient Object Detection Algorithm Based on Visual Attention Mechanism for UAV Optical Sensors. *IEEE Sens. J.* **2023**, *23*, 11139–11155. [[CrossRef](#)]
37. Liu, W.; Quijano, K.; Crawford, M.M. YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery With Improved YOLOv5 Based on Transfer Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8085–8094. [[CrossRef](#)]
38. Ye, T.; Zhang, J.; Li, Y.; Zhang, X.; Zhao, Z.; Li, Z. CT-Net: An Efficient Network for Low-Altitude Object Detection Based on Convolution and Transformer. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2507412. [[CrossRef](#)]
39. Shen, L.; Lang, B.; Song, Z. CA-YOLO: Model Optimization for Remote Sensing Image Object Detection. *IEEE Access* **2023**, *11*, 64769–64781. [[CrossRef](#)]

40. Xue, H.; Ma, J.; Cai, Z.; Fu, J.; Guo, F.; Weng, W.; Dong, Y.; Zhang, Z. NLFA: A Non Local Fusion Alignment Module for Multi-Scale Feature in Object Detection. In Proceedings of the 3rd International Symposium on Automation, Mechanical and Design Engineering, Beijing, China, 16–18 December 2023; Volume 138, pp. 155–173.
41. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 18–23 June 2018; pp. 8759–8768.
42. Sun, S.; Mo, B.; Xu, J.; Li, D.; Zhao, J.; Han, S. Multi-YOLOv8: An infrared moving small object detection model based on YOLOv8 for air vehicle. *Neurocomputing* **2024**, *588*, 127685. [[CrossRef](#)]
43. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Virtual, 13–19 June 2020; pp. 10778–10787.
44. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IOU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
45. Wang, Y.; Zou, H.; Yin, M.; Zhang, X. SMFF-YOLO: A Scale-Adaptive YOLO Algorithm with Multi-Level Feature Fusion for Object Detection in UAV Scenes. *Remote Sens.* **2023**, *15*, 4580. [[CrossRef](#)]
46. Zhao, D.; Shao, F.; Liu, Q.; Yang, L.; Zhang, H.; Zhang, Z. A Small Object Detection Method for Drone-Captured Images Based on Improved YOLOv7. *Remote Sens.* **2024**, *16*, 1002. [[CrossRef](#)]
47. Wang, K.; Wei, Z. YOLO V4 with hybrid dilated convolution attention module for object detection in the aerial dataset. *Int. J. Remote Sens.* **2022**, *43*, 1323–1344. [[CrossRef](#)]
48. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 11–17 October 2021; pp. 2778–2788.
49. Wu, Q.; Li, Y.; Huang, W.; Chen, Q.; Wu, Y. C3TB-YOLOv5: Integrated YOLOv5 with transformer for object detection in high-resolution remote sensing images. *Int. J. Remote Sens.* **2024**, *45*, 2622–2650. [[CrossRef](#)]
50. Ni, J.; Zhang, Z.; Shen, K.; Tang, G.; Yang, S.X. An improved deep network-based RGB-D semantic segmentation method for indoor scenes. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 589–604. [[CrossRef](#)]
51. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
52. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
53. 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](#)]
54. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
55. Xiong, G.; Qi, J.; Wang, M.; Wu, C.; Sun, H. GCGE-YOLO: Improved YOLOv5s Algorithm for Object Detection in UAV Images. In Proceedings of the Chinese Control Conference, CCC, Tianjin, China, 24–26 July 2023; pp. 7723–7728.
56. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9756–9765.
57. Wang, J.; Liu, W.; Zhang, W.; Liu, B. LV-YOLOv5: A light-weight object detector of Vit on Drone-captured Scenarios. In Proceedings of the International Conference on Signal Processing Proceedings, ICSP, Beijing, China, 21–24 October 2022; Volume 1, pp. 178–183.
58. Chen, L.; Liu, C.; Li, W.; Xu, Q.; Deng, H. DTSSNet: Dynamic Training Sample Selection Network for UAV Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5902516. [[CrossRef](#)]
59. Liu, J.; Zheng, K.; Liu, X.; Xu, P.; Zhou, Y. SDSDet: A real-time object detector for small, dense, multi-scale remote sensing objects. *Image Vis. Comput.* **2024**, *142*, 104898. [[CrossRef](#)]
60. Yang, Q.; Cao, L.; Huang, C.; Song, Q.; Yuan, C. A2Net: An Anchor-free Alignment Network for Oriented Object Detection in Remote Sensing Images. *IEEE Access* **2024**, *12*, 42017–42027. [[CrossRef](#)]
61. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-Scale Object Detection from Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602918. [[CrossRef](#)]
62. 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 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; Volume 2019, pp. 2844–2853.
63. 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](#)]

64. Zeng, S.; Yang, W.; Jiao, Y.; Geng, L.; Chen, X. SCA-YOLO: A new small object detection model for UAV images. *Vis. Comput.* **2024**, *40*, 1787–1803. [[CrossRef](#)]
65. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic Feature Pyramid Network for Object Detection. In Proceedings of the Conference Proceedings—IEEE International Conference on Systems, Man and Cybernetics, Banff, AB, Canada, 1–4 October 2023; pp. 2184–2189.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.