



## Article

# Remote Sensing Small Object Detection Network Based on Attention Mechanism and Multi-Scale Feature Fusion

Junsuo Qu <sup>1,\*</sup> , Zongbing Tang <sup>1,2</sup> , Le Zhang <sup>1</sup>, Yanghai Zhang <sup>1</sup> and Zhenguo Zhang <sup>1</sup>

<sup>1</sup> Xi'an Key Laboratory of Advanced Control and Intelligent Process, School of Automation, Xi'an Robotic Intelligent Systems International Science and Technology Cooperation Base, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

<sup>2</sup> School of Communication and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an 710121, China

\* Correspondence: qujunsuo@xupt.edu.cn

**Abstract:** In remote sensing images, small objects have too few discriminative features, are easily confused with background information, and are difficult to locate, leading to a degradation in detection accuracy when using general object detection networks for aerial images. To solve the above problems, we propose a remote sensing small object detection network based on the attention mechanism and multi-scale feature fusion, and name it AMMFN. Firstly, a detection head enhancement module (DHEM) was designed to strengthen the characterization of small object features through a combination of multi-scale feature fusion and attention mechanisms. Secondly, an attention mechanism based channel cascade (AMCC) module was designed to reduce the redundant information in the feature layer and protect small objects from information loss during feature fusion. Then, the Normalized Wasserstein Distance (NWD) was introduced and combined with Generalized Intersection over Union (GIoU) as the location regression loss function to improve the optimization weight of the model for small objects and the accuracy of the regression boxes. Finally, an object detection layer was added to improve the object feature extraction ability at different scales. Experimental results from the Unmanned Aerial Vehicles (UAV) dataset VisDrone2021 and the homemade dataset show that the AMMFN improves the AP<sub>s</sub> values by 2.4% and 3.2%, respectively, compared with YOLOv5s, which represents an effective improvement in the detection accuracy of small objects.

**Keywords:** small object detection; attention mechanism; loss function; remote sensing



**Citation:** Qu, J.; Tang, Z.; Zhang, L.; Zhang, Y.; Zhang, Z. Remote Sensing Small Object Detection Network Based on Attention Mechanism and Multi-Scale Feature Fusion. *Remote Sens.* **2023**, *15*, 2728. <https://doi.org/10.3390/rs15112728>

Academic Editor: Dusan Gleich

Received: 20 April 2023

Revised: 8 May 2023

Accepted: 18 May 2023

Published: 24 May 2023



**Copyright:** © 2023 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

With the continuous advancement of technology, images captured by drone are now widely used in remote sensing imagery, agriculture, wildlife conservation [1,2], and disaster surveillance. Although existing object detectors have made significant advancements in the object detection for natural scenes, the following problems remain when applying such general-purpose object detectors directly to remote sensing images: (i) The inconsistent flight altitude of the UAVs leads to different scale sizes of the same class of objects in the captured images. This is the case, for example, for the images in the dataset VisDrone2021. (ii) Small objects have problems, such as few effective pixels, limited feature expression, and a susceptibility to background effects. We also refer to them as spectral mixtures. (iii) The loss function based on the Intersection over Union (IoU) variant is more sensitive to the offset of small objects than that of larger objects. Therefore, small objects have the problem of being difficult to locate.

In practical applications, the large differences in scale between the various objects in remote sensing imagery present a greater challenge for object detectors. It is therefore vital to obtain a detection network that can detect objects at different scales. A prevalent approach to solve the varying scales of objects is to construct a multi-layer feature fusion, such as the Feature Pyramid Networks (FPN) [3] and the feature fusion modules Path

Aggregation Network (PANet) [4], Bi-directional Feature Pyramid Network (Bi-FPN) [5], Adaptively Spatial Feature Fusion (ASFF) [6], and Neural Architecture Search Feature Pyramid Networks (NAS-FPN) [7], which are all improved on the basis of FPN. However, small objects have fewer effective pixels, and more feature information is lost after passing through the backbone network, resulting in the model failing to correctly learn important spatial and semantic feature information about small objects. Therefore, it is necessary to increase the shallow branches as well as to improve the feature map resolution of the detection head in order to mitigate the loss of small object information.

In object detection, the regression loss function characterizes the extent of agreement between the model output box size and position and the true box size and position. The regression loss function has gone through L1/L2 loss, and smooth L1 loss [8] to the loss function, based on IoU [9–12] variants commonly used today. YOLOv5 [13] uses GIoU as its position regression loss. The GIoU is an improved version of IoU. Unlike IoU which focuses only on overlapping regions, GIoU focuses not only on overlapping regions but also on other non-overlapping regions, which can better reflect the overlap of both. However, this loss function is very sensitive to the positional bias of small objects, and a slight positional bias of small objects will cause a significant increase or decrease in the IoU value, which is thus unfriendly to small objects. Although other scholars have solved the regression problem for small objects to some extent using a variant of IoU, there is still the problem that this type of loss function is not friendly to small objects. Wang et al. [14] designed an NWD based on a two-dimensional Gaussian distribution to effectively alleviate the problem of low detection accuracy of commonly used object detection networks for small objects, but they failed to consider the advantage of IoU-based loss function for the detection of large and medium objects.

Regarding the problems above, this paper proposes a remote sensing small-object network detection based on the attention mechanism and multi-scale feature fusion, which can effectively improve the detection accuracy of the model for small objects in remote sensing images with the addition of fewer model parameters. First of all, the detection head contains information for the classification and regression of the final object. In order to make effective use of the feature information in the individual detection head, we propose a detection head enhancement module. Secondly, after multiple convolutions of the input image, feature redundancy occurs in the feature layer. To prevent this type of redundant information from interfering with small objects, we use an attention mechanism to design a channel cascade module. Then, to address the difficulty of detecting small objects with the three detection heads of the universal detector, we add a detection head with a higher-feature map resolution. Finally, we introduce a NWD loss function to calculate the similarity between two objects using a Gaussian distribution.

The main contributions of this paper can be summarized as follows:

1. We propose a detection head enhancement module DHEM to further achieve more accurate small-object detection by combining a multi-scale feature fusion module and an attention mechanism module to enhance feature characterization, at the cost of slightly increasing model parameters.
2. We design a channel cascade module based on an attention mechanism, AMCC, to help the model remove redundant information in the feature layer, highlight small-object feature information, and help the model learn more efficiently for small-object features.
3. We introduce the NWD loss function and combine it with GIoU as the location regression loss function to improve the optimization weight of the model for small objects and the accuracy of the regression boxes. Additionally, an object detection layer is added to improve the object feature extraction ability at different scales.
4. AMMFN is compared with YOLOv5s and other advanced models on the homemade remote sensing dataset and publicly available dataset VisDrone2021, with significant improvements in the  $AP_s$  values and mAP values.

The remaining sections are organized as follows: In Section 2, we summarize the literature on small-object detection; in Section 3, we describe the improved modules and the reasons for the improvements in detail, including a detection head enhancement module, a channel cascade module based on attentional features, a regression loss function, and the addition of a detection head; in Section 4, we describe the relevant steps involved in this experiment and analyze the results; in Section 5, we discuss the advantages and disadvantages of the proposed model; and in Section 6, we conclude this work and put forward directions for optimizing the model.

## 2. Related Works

Object detection has tremendous practical value and application promise, and it is the cornerstone of many vision algorithm tasks, such as face recognition and target tracking [15]. The existing networks for the detection of objects can be broadly divided into two categories. One class is the two-stage object detection networks based on the Region-CNN (RCNN) [16], Fast Region-based Convolutional Network (Fast-RCNN), Faster-RCNN [17], and the Region-based fully convolutional network (R-FCN) [18], which first perform feature extraction and create a good number of candidate boxes for images through a backbone network, and then perform classification tasks and regression tasks for objects. The detection accuracy of this type of network is high, but its real-time performance is very low. One class is the one-stage object detection networks represented by the Single Shot Multi-Box Detector (SSD) [19], You Only Look Once (YOLO) series [20–23], Fully Convolutional One-Stage Object Detector (FCOS) [24], and the RetinaNet [25], which directly perform semantic and spatial feature extraction on objects and then complete the classification and regression of objects. Although the overall performance of such networks is poor, their real-time performance is high and they have been broadly used in various scenarios. Academically, there are two ways to define a small object: relative size and absolute size. The relative size approach is to consider an object to be small if its aspect is 0.1 of the original image size [26], and the absolute size approach is to consider an object smaller than  $32 \times 32$  pixels to be small. This paper uses the definition of absolute size. Remote sensing images suffer from complex backgrounds, few effective pixels of objects, varying scales, and different morphologies, which make it difficult for existing general-purpose object detectors to extract accurate and effective feature information to classify and localize small objects. To address the challenge of the inaccurate detection of small objects in the field of remote sensing, this paper focuses on the work of other scholars from two aspects: multi-scale integration and attention mechanism.

Since the deep feature layer contains rich semantic object information and the large perceptual domain, while the shallow features contain more fine-grained information, the deep feature information and the shallow feature information can be reasonably used to increase the accuracy of the model for small-object detection by multi-scale fusion. Qu et al. [27] proposed a small-object detection model that is called the Dilated Convolution and Feature Fusion Single Shot Multi-box Detector (DFSSD) which improved the detection of remote sensing small objects to some extent by expanding the perceptual domain of features, obtaining contextual information of features at different scales, and enhancing the semantic information of shallow features. Deng et al. [28] designed an Extended Feature Pyramid Network (EFPN) specifically for small-object detection, which contained a Feature Texture Transfer (FTT) module that acted on the super-resolution feature map by extracting semantic information and texture features from the feature map of the FPN network, thus effectively improving the representation of small-object feature information and being efficient in both computation and storage. Deng et al. [29] proposed a multi-scale dynamic weighted feature fusion network, which adaptively assigns different weights to feature layers at different scales through network training to increase the contribution of shallow feature information in the whole network, which directs the model for small-object detection tasks.

Small objects have the problem of a low number of effective pixels and an easy background confusion, so highlighting the feature information of small objects is very necessary. The attention module helps the network to pay close attention to task-relevant foreground object feature information in a large amount of background information. Thus, the use of attention mechanisms can effectively improve the representation of small-object features. Zhu et al. designed a small-object detection model called the Transformer Prediction Heads-YOLOv5 (TPH-YOLOv5) [30] that combines YOLOv5 with the Transformer [31], which integrates the Convolutional Block Attention Model (CBAM) [32] module and self-attention mechanism [33] into the YOLOv5 model to help the network in extracting small-object feature information, which effectively improves its ability to detect small objects and means it can detect small objects under the perspective of UAV low-altitude flight. Shi [34] et al. proposed a feature enhancement module using the location attention mechanism, which improves the efficiency of the model at detecting small objects by highlighting the contribution of important features and suppressing the influence of irrelevant features on the overall feature information, by using the channel attention mechanism, after extracting the feature information from different sensory fields. Zhao et al. [35] proposed a feature fusion strategy based on the Efficient Channel Attention (ECA) module to enhance the expression of semantic information of shallow features by fusing object information at different scales, thus improving the performance of the model for small-object detection. Zhang et al. [36] designed a multi-resolution attention detector that captured useful location and contextual information through adaptive learning, which was used to obtain attention weights by calculating the cosine similarity between other output layers and the template layer, which were then weighted to fuse the first three layers of feature information of the backbone network to generate an attention graph, to highlight the feature information of small objects.

### 3. Our Work

In the paper, we improve on the YOLOv5s model by proposing the addition of AMMFN. Firstly, the prediction layer is optimized by the detection head enhancement module. Secondly, a channel cascade module based on an attention mechanism is designed to replace the generic cascade operation in the neck. Then, the NWD and the GIoU are merged to improve the weight losses of the small objects and to increase the accuracy of the regression boxes. Finally, the detection head of the shallow network is increased for detecting small objects. These four improvements are effective in improving the ability of the model to detect small objects. Figure 1 shows the overall structure of the proposed network model.

#### 3.1. Detection Head Enhancement Module

The detection head part of YOLOv5 contains the final object classification information as well as the regression information of the object box; therefore, for small-object detection, the detection head has a huge impact. During the model training process, the detection head detects too few small objects due to their weak feature representation and fewer pixels, resulting in interference with the optimized weights of small objects. For this reason, it is very necessary to significantly enhance the feature-expression capability of the foreground.

Using multi-scale feature fusion, features of different sizes can be obtained and the perceptual domain can be expanded to strengthen the description of small-object features, thus improving the detection performance of the model for small objects. To this end, this paper proposes a detection head enhancement module through this idea, as shown in Figure 2.

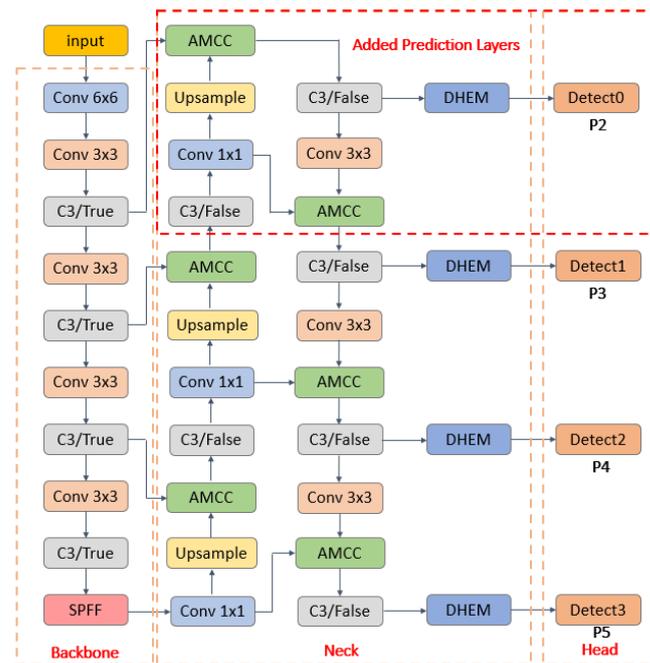


Figure 1. Network structure.

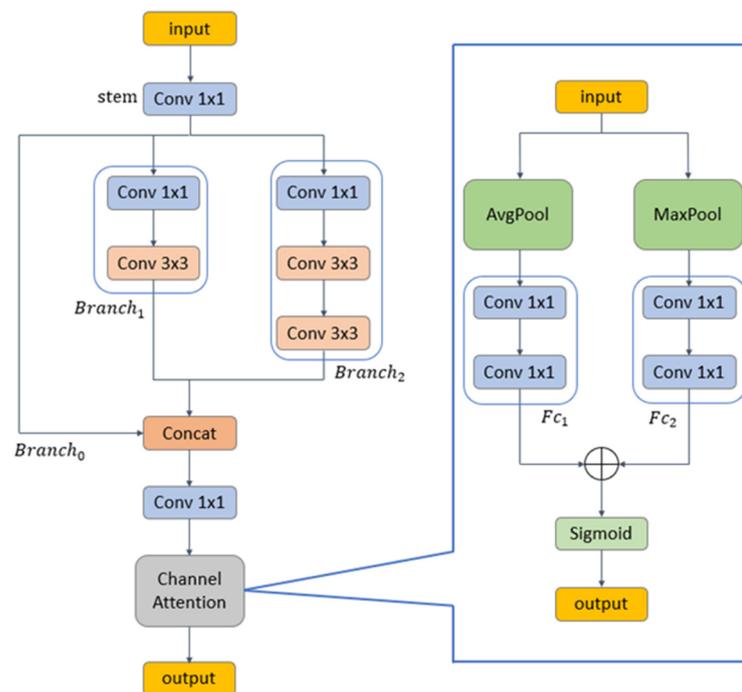


Figure 2. Detection head enhancement module.

The DHEM structure uses a multi-branch structure, wherein each branch uses convolution kernels of variable numbers and sizes of to obtain different scales of perceptual fields, and also uses the idea of residual connectivity. This approach improves the range of perceptual fields without adding too much computation and enables the model to obtain features with high discriminative power while being lightweight. However, there are semantic differences in the feature maps at different scales, and the fused feature layers may have a confounding effect which can cause the network to confuse localization and recognition tasks. To mitigate this negative impact, a lightweight channel attention mechanism module

is used in this paper. This module not only reduces the confusion between features but also significantly enhances the feature information of small objects.

Specifically, firstly,  $1 \times 1$  convolution is used to bring down the number of feature channels and thus reduce the computational effort; secondly, the information of different scales is extracted by three branches, respectively, and cascaded to obtain the feature map with multi-scale information, then  $1 \times 1$  convolution is used to organize the information in the feature map and decrease the number of channels; and finally, the model will obtain accurate and non-redundant feature information for final object detection. The formulae are shown in (1) and (2).

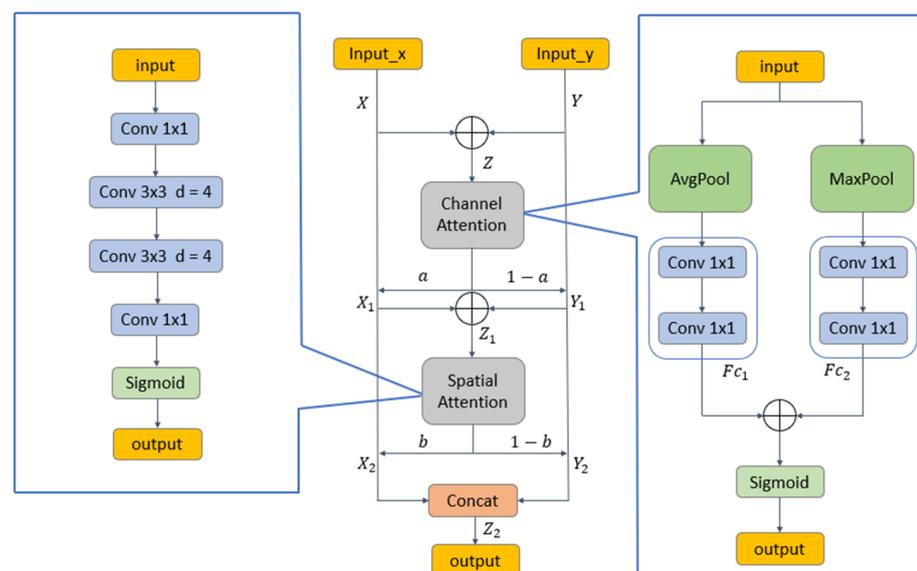
$$C = \text{Conv}([\text{Branch}_0(\text{stem}(x)); \text{Branch}_1(\text{stem}(x))]; \text{Branch}_2(\text{stem}(x))) \quad (1)$$

$$\text{Output} = \sigma(\text{Fc}_1(\text{Avg}(C)) + \text{Fc}_2(\text{Max}(C))) \times C \quad (2)$$

where  $[\cdot]$  represents the splicing operation,  $\text{Fc}_1$  and  $\text{Fc}_2$  represent two convolution operations with  $1 \times 1$  convolution kernels,  $\text{Branch}_0$ ,  $\text{Branch}_1$  and  $\text{Branch}_2$  are the convolution operations on the graph, and  $\sigma$  represents the Sigmoid activation function.

### 3.2. Channel Cascade Module Based on Attention Mechanism

Feature fusion is the combination of information from different scales or branches and is an essential part of the object detection network structure. A common method of feature fusion is to merge features by connecting channels of the feature map or by adding them element by element. Element-by-element addition can make the feature map more informative with the same dimensionality but less computationally intensive than the cascade approach. However, for the problem of semantic inconsistency or perceptual field inconsistency among input feature maps, this method may not be the best one. To prevent the imbalance problem caused by object scale variation and small-object feature information to the detection problem model in remote sensing images, in this paper, a channel cascade module AMCC based on the attention mechanism is designed according to the literature [37], as seen in Figure 3.



**Figure 3.** Channel cascade module.

The module consists mainly of a mechanism for paying attention to channels and a mechanism for paying attention to space. The channel attention mechanism adaptively learns and focuses on the channel weights that are more important for the task, thus enabling adaptive object selection and directing the network to focus more on important

objects. The spatial location attention mechanism can guide the model to learn and highlight task-relevant foreground objects on the feature map according to the spatial information of the feature map. Thus, the advantages of both are used to direct the network's attention to more regions about small objects, as shown in (3)–(7).

$$Z = X + Y \quad (3)$$

$$a = \sigma(\text{Fc}_1(\text{Avg}(Z)) + \text{Fc}_2(\text{Max}(Z))) \quad (4)$$

$$Z_1 = (a \times X_1) + ((1 - a) \times Y_1) \quad (5)$$

$$b = \sigma(\text{Conv\_all}(Z_1)) \quad (6)$$

$$Z_2 = (b \times X_2) + ((1 - b) \times Y_2) \quad (7)$$

where Avg and Max represent the average and maximum pooling operations, respectively, and Conv\_all indicates all convolution operations in the spatial attention mechanism.

Specifically, given two feature maps  $X, Y \in \mathbb{R}^{W \times H \times C}$ , firstly,  $X$  and  $Y$  are selected for initial feature fusion by element summation to obtain feature map  $Z$ . Secondly, feature map  $Z$  is input into the channel attention mechanism module to obtain weights  $a$  that help the network to focus on small-object information through pooling and convolution operations, and then weights  $a$  and weights  $1 - a$  are applied to feature maps  $X$  and  $Y$  to get  $X_1$  and  $Y_1$ , respectively. Then,  $X_1$  and  $Y_1$  are summed by elements for the second time to acquire the feature map  $Z_1$ , and the feature map  $Z_1$  is input into the spatial attention mechanism module to obtain the spatial weights associated with the object task  $b$ . The attention mechanism uses expanded convolution to expand the perceptual field and aggregate the contextual information. The weights  $b$  and  $1 - b$  are then applied to the feature maps  $X_1$  and  $Y_1$  to obtain  $X_2$  and  $Y_2$ , respectively. finally, the channels of  $X_2$  and  $Y_2$  are stitched to obtain effective feature information at different scales.

### 3.3. Optimization of the Loss Function

The position regression loss function of the YOLOv5s is GIoULoss, as seen in Equation (8). For small objects, a small positional offset will cause a sharp decrease in the IoU value, but for large objects, only a small change in the IoU value occurs for the same positional offset. Therefore, the IoU-based loss function is very sensitive to the positional shift of small objects, which reduces the overall detection accuracy of the object detector, as seen in Figure 4. To solve the problem, we introduce the position regression loss function based on normalized Wasserstein distance (NWD), which has become a new method for small-object detection and optimization in recent years. NWD uses a two-dimensional Gaussian distribution to model the bounding box and calculate the similarity between the predicted object box and the labelled object box by its corresponding Gaussian distribution, i.e., the normalized Wasserstein distance between them based on Equation (10). The method consistently reflects the distance between distributions for objects detected by the model, regardless of whether they overlap or not. NWD is insensitive to the scale of objects; it is therefore more appropriate to use it to measure the similarity between the predicted object boxes and the labelled boxes in the remote sensing images. However, this paper does not simply replace this GIoU with the NWD loss function, because the GIoU is better at detecting large- and medium-sized objects. Therefore, in the paper, NWD is fused with GIoU by scaling so that the model can improve the optimization weights and the accuracy of the regression boxes, according to Equation (11), as a loss function for the location

regression of AMMFN, where the coefficients  $a$  and  $b$  of GIoU and NWD are chosen as shown in the ablation experiment section.

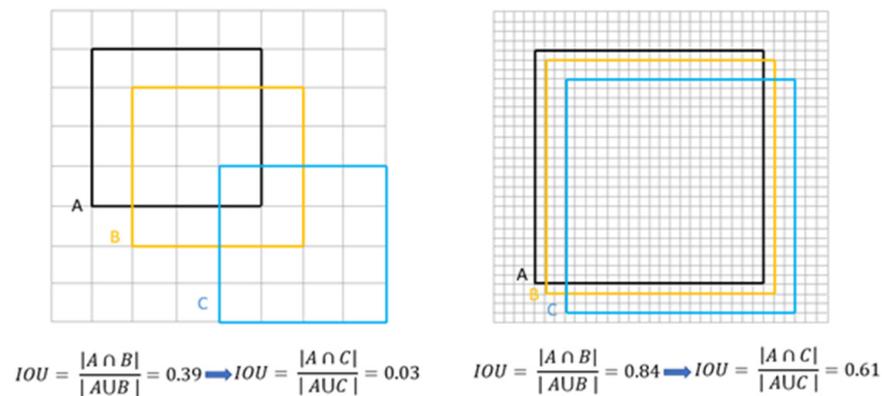
$$\text{GIoULoss} = 1 - \left( \text{IoU} - \frac{A^c - (A^p \cup A^g)}{A^c} \right) \quad (8)$$

$$\text{NWD}(N_a, N_b) = \exp \left( - \frac{\sqrt{W_2^2(N_a, N_b)}}{C} \right) \quad (9)$$

$$W_2^2(N_a, N_b) = \left\| \left( \left[ cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[ cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2 \quad (10)$$

$$\text{Loss} = 2 - (a * \text{GIoU} + b * \text{NWD}) \quad (11)$$

where IoU represents the ratio of the intersecting areas of two rectangular boxes to the sum of their areas,  $A^p$  indicates the area enclosed by the prediction box,  $A^g$  indicates the area surrounded by the label box,  $A^c$  represents the area of the smallest outer rectangle of the prediction box and label box, and  $C$  denotes the constant associated with the dataset. In this paper, the value of  $C$  is the number of categories in the dataset.  $W_2^2(N_a, N_b)$  denotes the distance measure, and  $N_a$  and  $N_b$  denote the Gaussian distribution modeled by  $A = (cx_a, cy_a, w_a, h_a)$  and  $B = (cx_b, cy_b, w_b, h_b)$ .



**Figure 4.** Schematic diagram of the sensitivity analysis of IoU for small and large objects, where each grid represents a pixel, the left diagram shows the small-object schematic, the right diagram shows the large-object schematic, A indicates the label box, and B and C indicate the prediction box with different degrees of program offset, respectively.

### 3.4. Optimization of the Prediction Feature Layer

In remote sensing images, small objects have the drawbacks of small size and insufficient effective information. In the YOLOv5s object detector, the effective pixels of small objects gradually decrease in the process of mapping the input image into feature maps of different scales after repeated down-sampling operations, which makes the network unable to learn the important spatial feature information of small objects well, thus influencing the detection accuracy of the detectors for small objects.

As seen in Figure 1, compared with the original three detection layers of YOLOv5s, detection layer P2 contains richer texture and more detailed information due to fewer down-sampling operations, which can help the model to detect small objects in remote sensing images more effectively, so we propose the addition of detection layer P2 to detect small objects on feature maps.

## 4. Experiments

### 4.1. Dataset

In the experimental work, we used two datasets. One was the open-source dataset VisDrone2021 [38], which contains 10 categories and is shown schematically in Figure 5. The number of each type of object in this dataset varied considerably, so we selected two of these categories, car and pedestrian, to train, validate, and test. A total of 8629 images were allocated according to the original proportion of the dataset, with 6471, 548, and 1610 images in the training, validation, and testing sets, respectively. One was a homemade dataset with 1156 images containing two categories: airplane and car, of which the training set, validation set, and test set had 739, 185, and 323 images, respectively, as shown in Figure 6. We counted the object sizes in both datasets according to the criteria of the MS COCO dataset [39]. Additionally, we also counted the number of tiny objects with less than  $16 \times 16$  pixels, as shown in Table 1.

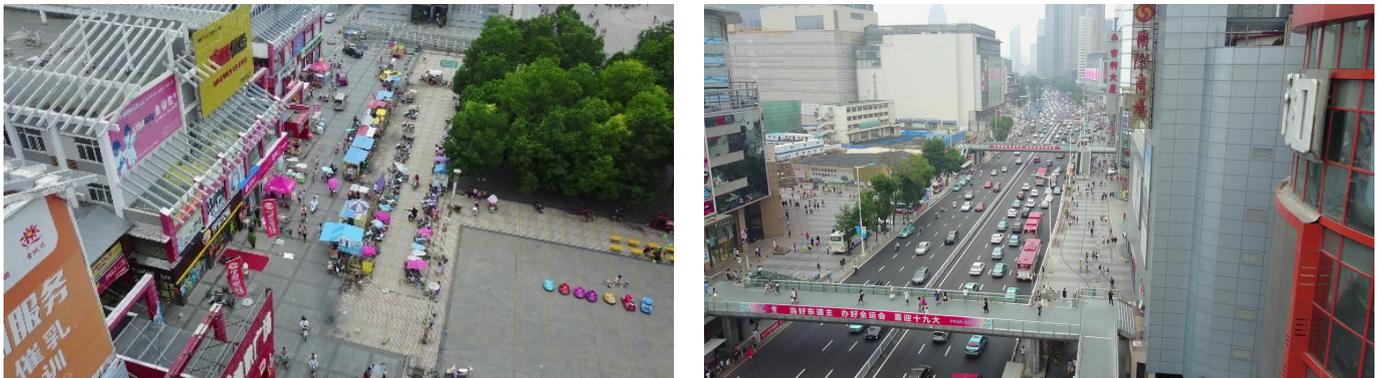


Figure 5. Sample VisDrone2021 dataset image.



Figure 6. Sample picture of the homemade dataset.

**Table 1.** The number of objects of each different size in the two datasets.

Type	VisDrone2021	Homemade Dataset
Tiny object (0, 16 <sup>2</sup> ]	94,424	294
Small object (16 <sup>2</sup> , 32 <sup>2</sup> ]	93,655	4901
Medium object (32 <sup>2</sup> , 96 <sup>2</sup> ]	94,273	6507
Large objects (96 <sup>2</sup> , +∞)	13,840	442

#### 4.2. Experimental Environment Configuration and Parameter Setting

This research work used the PyTorch framework to complete a series of tasks using GPUs for accelerated training, and the specific relevant environment configurations are seen in Table 2.

**Table 2.** Experimental environment configuration.

Types	Environment
Operating System	Ubuntu18.04
GPU (Video memory size, memory size)	NVIDIA GeForce GTX 4090Ti (24 G, 128 G)
PyTorch Versions	1.8.0
CUDA	12.1

AMMFN was improved from YOLOv5s, with the learning rate set to 0.001 in the training phase and the weight decay value set to 0.0005. To optimize the parameters of the model, we used the stochastic gradient descent algorithm and the momentum optimization algorithm, and the input image had a length and width of 640 and a batch size of 16; the epoch of iteration was 300, and the momentum factor was 0.937. The hyperparameters of the models in the comparison experiments were slightly different, but the overall hyperparameters were similar.

#### 4.3. Experimental Evaluation Metrics

In this experiment, mAP was used as an evaluation metric, where mAP refers to the  $P - R$  curve based on the accuracy and recall of each class in a multi-class object detection task, and the formulae for accuracy and recall are shown in (12) and (13). We also listed the specific size of parameters, the size of floating-point operations for each model, and frames per second.

$$P = \frac{TP}{TP + FP} \quad (12)$$

$$R = \frac{TP}{TP + FN} \quad (13)$$

where TP denotes the number of positive cases where the prediction is a correct prediction, FP denotes the number of predicted results that are misclassified as positive cases, and FN denotes the number of predicted results that are incorrectly classified as negative cases.

The area obtained by intersecting the  $P - R$  curve with the coordinate axis is the average accuracy. mAP is calculated as shown in (15):

$$AP = \int_0^1 P(R)d(R) \quad (14)$$

$$mAP = \sum_{i=1}^N \frac{AP(i)}{N} \quad (15)$$

#### 4.4. Results of Ablation Experiments

##### 4.4.1. Proposed Modules

To demonstrate that the module designed in this paper is valid for the model, we conducted ablation experiments on each improved module under the same conditions and verified the effect of each module on the model as a whole. Table 3 shows that the proposed module made a significant improvement compared to YOLOv5s. The best results regarding the loss function and the addition of detection layers for boosting small objects show that our optimization of the loss function was effective and that the shallow feature information can indeed help the model detect small objects. Although the enhancement of the detection head enhancement module and channel cascade module was less, the improvement in the small-object detection accuracy had a considerable influence.

**Table 3.** Effect of each module on YOLOv5s.

DHEM	AMCC	NWD	Head	mAP	AP <sub>0.5</sub>	AP <sub>0.75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>	Params
				0.221	0.440	0.204	0.146	0.395	0.556	7.025 M
✓				0.224	0.442	0.204	0.153	0.408	0.547	7.405 M
	✓			0.225	0.442	0.206	0.151	0.397	0.543	7.088 M
		✓		0.232	0.453	0.213	0.156	0.412	0.554	7.025 M
			✓	0.232	0.458	0.213	0.158	0.408	0.568	7.169 M
✓			✓	0.236	0.466	0.215	0.162	0.422	0.579	7.651 M
	✓		✓	0.235	0.464	0.216	0.160	0.415	0.572	7.236 M
		✓	✓	0.239	0.469	0.218	0.163	0.423	0.576	7.169 M
	✓	✓	✓	0.241	0.474	0.224	0.165	0.427	0.591	7.236 M
✓	✓	✓	✓	0.243	0.478	0.223	0.167	0.433	0.594	7.651 M
✓	✓		✓	0.239	0.475	0.217	0.163	0.429	0.588	7.717 M
✓	✓	✓	✓	0.247	0.481	0.229	0.170	0.436	0.601	7.717 M

\* ✓ indicates the selected module.

##### 4.4.2. Finding the Appropriate Scale Factor in the Loss Function

We performed the above by simply referencing NWD as the regression loss function of the model, but it did not greatly improve the overall detection capability of YOLOv5s. Although the ability of the model to detect small objects was improved, the detection accuracy for medium and large objects had the opposite effect, which was not the original intention behind the design of AMMFN. Therefore, we chose to retain the GIoU-based regression loss function and conduct experiments by continuously adjusting the scaling relationship between GIoU and NWD under the condition of four detection heads. The experimental results show that the use of GIoU combined with NWD can indeed bring great positive effects to the model, effectively improving the overall performance of the model as well as the effectiveness of small-object detection. Considering the object-detection capability of the detector for large, medium, and small objects, we finally choose the coefficients of GIoU and NWD as 1.2 and 0.8, respectively, and the results are listed in Table 4.

**Table 4.** Effect of different ratios of GIoU and NWD on YOLOv5s with four detection heads.

GIoU	NWD	mAP	AP <sub>0.5</sub>	AP <sub>0.75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
0	1	0.234	0.468	0.210	0.161	0.401	0.556
1	0	0.232	0.458	0.213	0.157	0.408	0.558
1	1	0.237	0.468	0.214	0.161	0.414	0.568
0.8	1.2	0.236	0.469	0.212	0.160	0.415	0.569
1.2	0.8	0.239	0.469	0.218	0.163	0.423	0.576
1.6	0.4	0.237	0.466	0.216	0.150	0.420	0.581
0.4	1.6	0.237	0.470	0.214	0.164	0.413	0.575

#### 4.5. Comparison Experiments

To validate the detection capability of AMMFN, we compared it with some benchmark networks. To ensure the fairness and reasonableness of the experiments, we retrained, validated, and tested all the comparison models. The results are seen in Table 5, and it can be seen that the overall evaluation index values of the AMMFN are higher than those of the YOLOv5s network. Compared with other networks, although AMMFN had slightly lower  $AP_1$  values and slightly lower  $AP_m$  values than the three models YOLOv8s, YOLOXs, and FOCs, the AMMFN had a smaller number of parameters and smaller GFLOPS than the three models YOLOv8s, YOLOXs, and FOCs. Additionally, the  $AP_s$  value, mAP value,  $AP_{0.5}$  value, and  $AP_{0.75}$  values of the AMMFN were higher than those of other compared model networks. This shows that the AMMFN can obviously improve the detection performance of small objects with lower parameters and lower arithmetic power, and can detect more small objects while the prediction boxes are more accurate.

**Table 5.** Experimental results of various detection models on VisDrone2021 dataset.

Model	mAP	$AP_{0.5}$	$AP_{0.75}$	$AP_s$	$AP_m$	$AP_l$	Params	GFLOPS	FPS
YOLOv5s	0.221	0.440	0.204	0.146	0.395	0.556	7.025 M	15.954 G	137.91
RetinaNet-50	0.102	0.168	0.114	0.008	0.258	0.505	36.351 M	145.652 G	82.54
Efficientnet-d2	0.126	0.207	0.140	0.017	0.325	0.526	8.007 M	14.281 G	39.44
Efficientnet-YOLOv3	0.120	0.297	0.076	0.072	0.189	0.347	6.999 M	9.039 G	102.47
YOLOv4-tiny	0.129	0.306	0.087	0.075	0.234	0.415	5.876 M	6.836 G	286.61
Mobilenetv2-YOLOv4	0.108	0.268	0.069	0.059	0.184	0.315	10.381 M	18.270 G	99.78
YOLOv7-tiny	0.220	0.432	0.212	0.143	0.417	0.613	6.017 M	13.190 G	277.13
YOLOXs	0.246	0.478	0.227	0.157	0.445	0.646	8.938 M	26.759 G	85.12
YOLOv8s	0.243	0.473	0.224	0.155	0.452	0.642	11.136 M	28.649 G	228.83
FCOS	0.231	0.434	0.221	0.138	0.444	0.610	32.113 M	161.174 G	81.95
AMMFN	0.247	0.481	0.229	0.170	0.436	0.601	7.717 M	25.782 G	84.28

To demonstrate the influence of AMMFN in a real scene in a more intuitive way, according to the comparative results in Table 5, we chose YOLOv5s, FCOS, YOLOv7-Tiny, YOLOv8s, YOLOXs, and AMMFN to predict the same remote sensing image with the same parameter settings, where the image contains a mass of small objects, and the results are shown in Figure 7 and Table 6. A total of 38 objects were detected by YOLOv8s, 39 objects were detected by YOLOXs, and 48 objects were detected by AMMFN, indicating that AMMFN can detect more small objects, and that the accuracy of the regression boxes was improved. Additionally, AMMFN is able to detect more distant objects compared to several other models. Therefore, it can be demonstrated that AMMFN can improve the effectiveness of small-object detection.

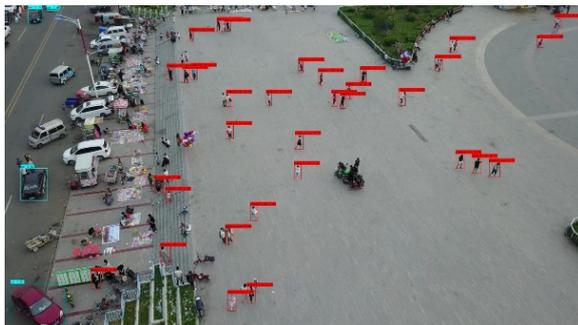
The modules proposed in this paper introduce attention mechanisms. To demonstrate whether the attention mechanism has a considerable impact on the models, we visualized the detection results of four models, YOLOv5s, YOLOv7-Tiny, YOLOv8s, and AMMFN, using a heat map. The results are shown in Figure 8. The darker the color, the more important the region is to the model. Compared to other models, it can be seen that AMMFN is able to focus precisely on task-relevant objects. Other models treat the railing as a car category and as important information; in contrast, AMMFN is able to competently mitigate the interference that this background information brings to the model, as shown in the yellow boxes in the figure. The reason for this is that the attention mechanism used in the AMMFN enhances the feature information of the objects and improves the ability of the model to perceive them.



Original



Label box



YOLOv5s



YOLOv7-Tiny



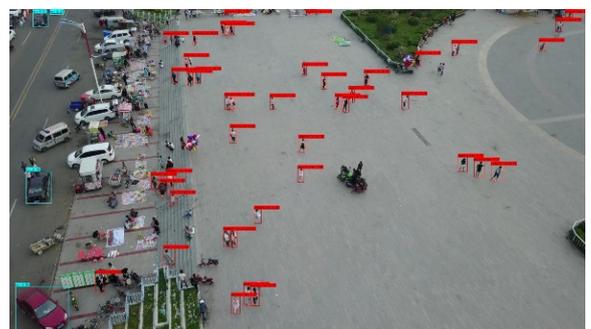
FCOS



YOLOv8s



YOLOXs

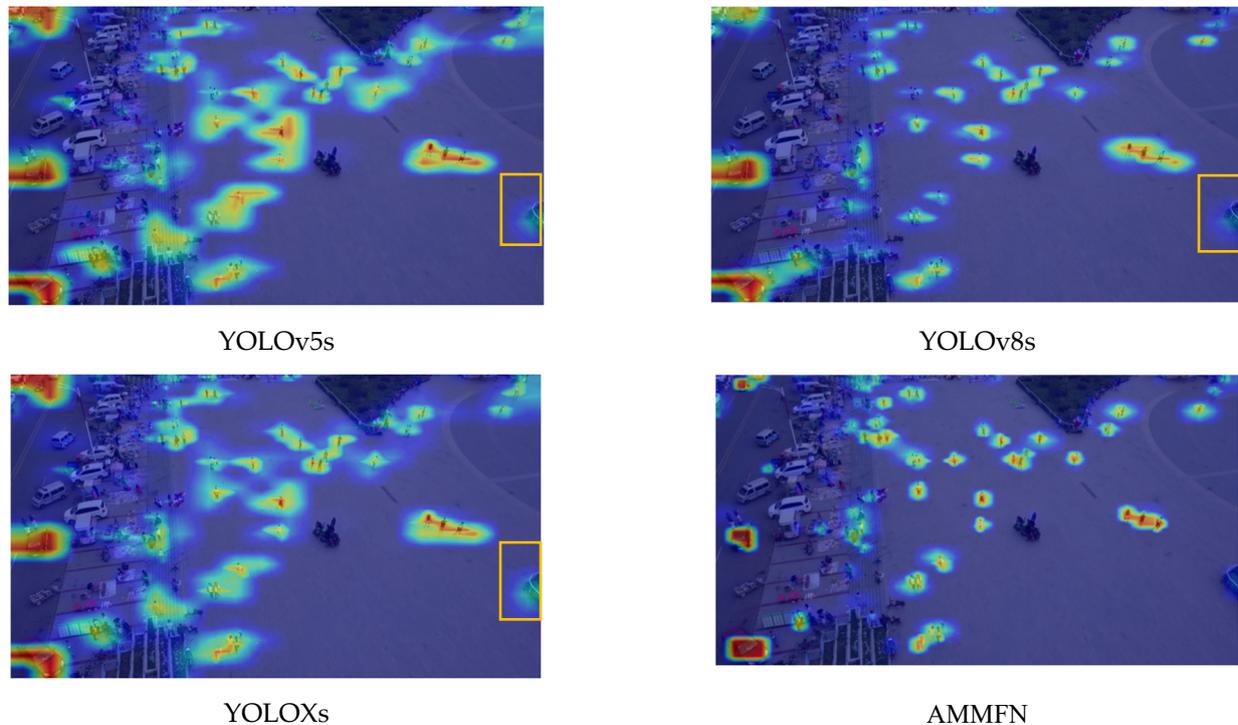


AMMFN

**Figure 7.** Pictures of the detection effect of each comparison model. The green and blue lines indicate the box and category of the label, respectively. The red and light blue lines indicate the prediction boxes for the pedestrian category and the car category, respectively.

**Table 6.** Total number of detected objects per comparison model.

Categories	YOLOv5s	YOLOv7-Tiny	FCOS	YOLOv8s	YOLOXs	AMMFN	Label
Car	4	4	4	4	4	4	12
Pedestrian	33	27	30	34	35	44	91

**Figure 8.** Visualization of the heat map for each comparison model.

In addition, we also used the log-average miss rate of four networks, YOLOv5s, YOLOv8s, YOLOXs, and AMMFN, to illustrate the leakage detection performance of our proposed model. The log-average miss rate indicates the missed detection rate of the model; as shown in Table 7, compared to the other three advanced object detection models, AMMFN has a lower log-average miss rate for both types of objects, but unfortunately, its log-average miss rate is still on the high side.

**Table 7.** Log-average miss rate for each comparison model.

Log-Average Miss Rate	YOLOv5s	YOLOXs	YOLOv8s	AMMFN
Pedestrian	0.91	0.88	0.89	0.86
Car	0.78	0.75	0.76	0.75

Finally, we use the example image in Figure 7 to investigate the effect that differently sized images have on the test results when fed into the model, and the results are shown in Table 8. The pedestrian category in this example image is almost exclusively small or extremely small objects. When the image is scaled down, the pedestrians in the image occupy fewer pixels, making the model less effective in detecting them, which was consistent with our theory. When the image is scaled up twice by interpolation, the small objects in the image become medium-sized objects and the extremely small objects become small objects. We believe that the reason for the lack of significant improvement in the model's detection results is that the interpolation method resulted in a change in the object's characteristics, which affected the detection results of the test model.

**Table 8.** The effect of different input sizes of images on the results during the test.

Input Size	Car	Pedestrian	Sum
224 × 224	2	6	8
320 × 320	3	18	21
416 × 416	4	25	29
512 × 512	4	32	36
640 × 640	4	44	48
1280 × 1280	3	46	49

#### 4.6. Comparison of the Homemade Dataset

To demonstrate whether the proposed modules perform well in other remote sensing datasets, we also conducted ablation experiments on a homemade dataset. This is shown in Table 1. The percentage of the quantity of tiny objects in the homemade dataset was not as large as the percentage of tiny objects in the VisDrone2021 dataset, so we only integrated the three improvements in NWD, DHEM, and AMCC into the YOLOv5s model to train, validate, and test. The results are shown in Table 9, where our proposed module still manages to perform well compared to the benchmark model, thus proving the usefulness of our proposed modules.

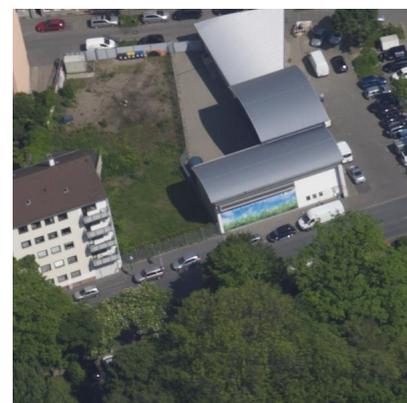
**Table 9.** Ablation results of each innovation module on the homemade dataset.

NWD	DHEM	AMCC	mAP	AP <sub>0.5</sub>	AP <sub>0.75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>	Params	GFLOPS
			0.441	0.865	0.389	0.302	0.491	0.649	7.025 M	15.954 G
✓			0.478	0.877	0.465	0.314	0.541	0.661	7.025 M	15.954 G
	✓		0.466	0.874	0.440	0.311	0.527	0.679	7.405 M	17.803 G
		✓	0.450	0.871	0.425	0.313	0.518	0.651	7.088 M	16.030 G
✓	✓		0.485	0.877	0.479	0.330	0.555	0.710	7.405 M	17.803 G
	✓	✓	0.472	0.879	0.451	0.320	0.537	0.674	7.468 M	17.878 G
✓		✓	0.487	0.883	0.485	0.329	0.554	0.669	7.088 M	16.030 G
✓	✓	✓	0.492	0.896	0.489	0.334	0.557	0.703	7.468 M	17.878 G

In addition, we randomly selected two typical images in the test set for testing the detection effect of YOLOv5s and AMMFN, and the original images and detection results are shown in Figure 9. AMMFN is able to detect small objects that both resemble the background or are obscured, although there is a case of missed detection, as shown in the yellow boxes. In addition, the number of objects detected by AMMFN and the accuracy of the regression boxes were better than those of YOLOv5s. Therefore, it can be emphasized that AMMFN can improve the model's ability to detect small objects.



Original (a1)

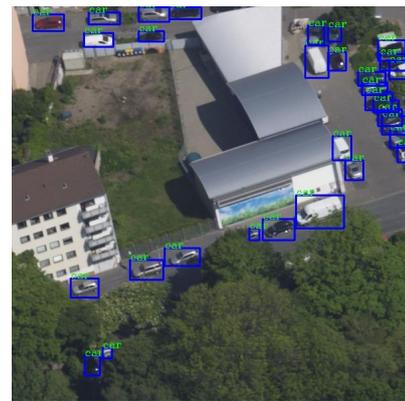


Original (a2)

**Figure 9.** Cont.



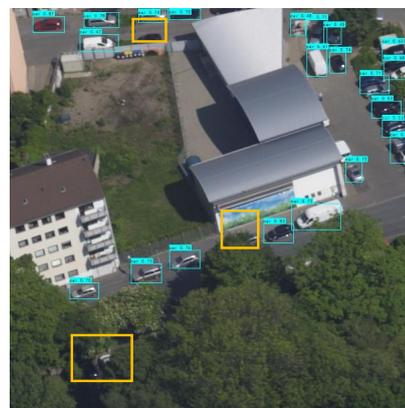
Label box (b1)



Label box (b2)



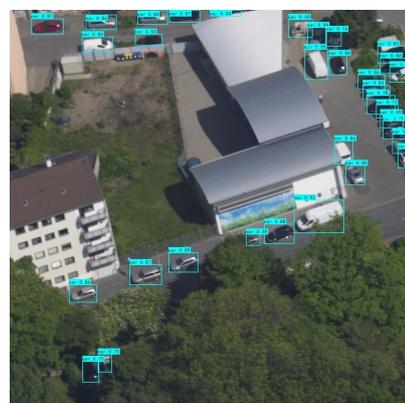
YOLOv5 (c1)



YOLOv5 (c2)



AMMFN (d1)



AMMFN (d2)

**Figure 9.** Pictures of detection results of YOLOv5s and our model. The green and blue lines indicate the box and category of the label, respectively. The red and light blue boxes indicate the forecast boxes for the aircraft category and the car category, respectively. The yellow box indicates the objects missed by the model.

## 5. Discussion

Detecting small objects is a highly challenging problem in object detection, and is also essential in some fields. The aim of this paper was the improvement in the accuracy of the detection of small objects in remote sensing images, which is of certain practical importance. Additionally, through the improvement of four aspects, the accuracy of the detection of small objects will be improved to a certain extent.

### 5.1. Discussion of Comparison with Other Advanced Models

Firstly, as seen in Table 5, AMMFN is slightly better than the two advanced object-detection networks, YOLOv8s and YOLOXs, in terms of overall performance, but its detection of large- and medium-sized objects is poor and not as good as the advanced models, such as YOLOv8s and YOLOXs. Such results are also acceptable in terms of the object distribution in the dataset, because most objects in remote sensing images belong to small objects; so, AMMFN has some advantages. Secondly, as seen in Table 7, in terms of the leakage rate, AMMFN is better than other advanced models, but the overall leakage rate is still relatively high. This is because the backbone network of YOLOv5s is small and the features of the objects are not extracted sufficiently, resulting in insufficient learning of object features by the model. Therefore, our next task will be to design a lightweight but extraction-capable backbone network.

### 5.2. Discussion of the Proposed Innovation Module

As can be seen in Table 4, simply referencing NWD does not improve the detection of large objects, but the advantages of both NWD and GIoU can be exploited through our methods to improve the model performance. However, the shortcoming is that there is variability in the object classification of different datasets, thus causing a change in the scaling relationship between NWD and GIoU. As can be seen from Table 3, the addition of the detection head and detection head enhancement module increases the number of parameters of the model, but brings a greater improvement in the small-object detection performance of the model, which makes it worthwhile.

### 5.3. Speed of Inference

As can be seen from Table 5, since our main task in this study was not to go in the direction of lightweighting, our model has much room for improvement regarding speed. It should be acknowledged, furthermore, that the module we designed wastes a lot of time on reading and writing data, which leads to slower inference of the model. We were surprised to find that YOLOv8s was still able to maintain a very fast speed with large parameters. We believe this may have been the reason for the gradient shunt, which will be one of the next focuses of our research.

## 6. Conclusions

In this paper, we proposed a remote sensing small-object detection network based on the attention mechanism and multi-scale feature fusion, to address the problem that existing object detectors are poor at detecting small objects due to the object's size, unsatisfactory feature extraction, and large-scale variation in UAV images.

In terms of the network structure, we firstly designed a detection-head-enhancement-module (DHEM) to enhance the weight information of foreground objects. Secondly, we proposed a feature cascade module with a multi-scale attention mechanism (AMCC) to reduce the redundant information in the feature layer and enhance the feature representation of small objects. Finally, a new detection head was added to predict small objects using shallow fine-grained information. In terms of loss functions, we introduced the NWD loss function to address the problems of small-object optimization weights and inaccurate small-object prediction boxes.

Although the detection performance of small objects in remote sensing images can be improved with the network described in this paper, there is still more room for improvement, such as the existence of large-model arithmetic power, missed objects, and its poor detection performance of large objects. The next step is to investigate ways to improve the overall performance of the object detector in a more efficient and lightweight manner.

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

**Funding:** This research was supported by Xi'an Key Laboratory of Advanced Control and Intelligent Process under grant No. 2019220714SYS022CG04 and the Key R&D plan of Shaanxi Province under grant No. 2021ZDLGY04-04.

**Data Availability Statement:** A vast amount of the research in this paper was based on the publicly available dataset VisDrone2021, and the role of the homemade dataset is for validation purposes only, so the homemade data in this study are available from the corresponding authors upon request.

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

## References

1. Kellenberger, B.; Marcos, D.; Tuia, D. Detecting Mammals in UAV Images: Best Practices to address a substantially Imbalanced Dataset with Deep Learning. *Remote Sens. Environ.* **2018**, *216*, 139–153. [[CrossRef](#)]
2. Kellenberger, B.; Volpi, M.; Tuia, D. Fast animal detection in UAV images using convolutional neural networks. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017.
3. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society, Honolulu, HI, USA, 21–26 July 2017.
4. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
5. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
6. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2018**, arXiv:1911.09516.
7. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.
8. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
9. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016.
10. Rezaatofghi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
11. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-iou loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019.
12. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient iou loss for accurate bounding box regression. *Neurocomputing* **2021**, *506*, 146–157. [[CrossRef](#)]
13. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Changyu, L.; Hogan, A.; Hajek, J.; Diaconu, L.; Kwon, Y.; Defretin, Y.; et al. Ultralytics/Yolov5: v5.0–YOLOv5-P6 1280 Models, AWS, Supervise.ly and YouTube integrations. *Zenodo*, 2021.
14. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized gaussian wasserstein distance for tiny object detection. *arXiv* **2021**, arXiv:2110.13389.
15. Yan, J.; Jiao, H.; Pu, W.; Shi, C.; Dai, J.; Liu, H. Radar Sensor Network Resource Allocation for Fused Target Tracking: A Brief Review. *Inf. Fusion* **2022**, *86–87*, 104–115. [[CrossRef](#)]
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
18. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
19. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Midtown Manhattan, NY, USA, 2016.

20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Redmon, J.; Farhadi, A. Yolo9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
22. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint* **2018**, arXiv:1804.02767.
23. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
24. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
25. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [[CrossRef](#)]
26. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016.
27. Qu, J.; Su, C.; Zhang, Z.; Razi, A. Dilated convolution and feature fusion SSD network for small object detection in remote sensing images. *IEEE Access* **2020**, *8*, 82832–82843. [[CrossRef](#)]
28. Deng, C.; Wang, M.; Liu, L.; Liu, Y.; Jiang, Y. Extended Feature Pyramid Network for Small Object Detection. *IEEE Trans. Multimed.* **2021**, *24*, 1968–1979. [[CrossRef](#)]
29. Deng, T.; Liu, X.; Mao, G. Improved YOLOv5 Based on Hybrid Domain Attention for Small Object Detection in Optical Remote Sensing Images. *Electronics* **2022**, *11*, 2657. [[CrossRef](#)]
30. 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/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 2778–2788.
31. 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. *arXiv* **2020**, arXiv:2010.11929.
32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 17–24 May 2018; pp. 3–19.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
34. Shi, T.; Gong, J.; Hu, J.; Zhi, X.; Zhang, W.; Zhang, Y.; Zhang, P.; Bao, G. Feature-Enhanced CenterNet for Small Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5488. [[CrossRef](#)]
35. Zhao, P.; Xie, L.; Peng, L. Deep-level Small Target Detection Algorithm Based on Attention Mechanism. *J. Comput. Sci. Explor.* **2022**, *16*, 927–937.
36. Zhang, F.; Jiao, L.; Li, L.; Liu, F.; Liu, X. MultiResolution Attention Extractor for Small Object Detection. *arXiv* **2020**, arXiv:2006.05941.
37. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional Feature Fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Village, CO, USA, 1–5 March 2020.
38. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; IEEE: Montreal, QC, Canada, 2021; pp. 2847–2854.
39. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Zitnick, C.L.; Dollar, P. Microsoft COCO: Common Objects in Context. *arXiv* **2015**, arXiv:1405.0312.

**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.