applied sciences

MDPI

Article

KCFS-YOLOv5: A High-Precision Detection Method for Object Detection in Aerial Remote Sensing Images

Ziwei Tian [1,*], Jie Huang [2], Yang Yang [2] and Weiying Nie [3]

[1] School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China
[2] College of Data Target Engineering, PLA Information Engineering University, Zhengzhou 450001, China
[3] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
* Correspondence: tianziwei2014@gs.zzu.edu.cn

Abstract: Aerial remote sensing image object detection, based on deep learning, is of great significance in geological resource exploration, urban traffic management, and military strategic information. To improve intractable problems in aerial remote sensing image, we propose a high-precision object detection method based on YOLOv5 for aerial remote sensing image. The object detection method is called KCFS-YOLOv5. To obtain the appropriate anchor box, we used the K-means++ algorithm to optimize the initial clustering points. To further enhance the feature extraction and fusion ability of the backbone network, we embedded the Coordinate Attention (CA) in the backbone network of YOLOv5 and introduced the Bidirectional Feature Pyramid Network (BiFPN) in the neck network of conventional YOLOv5. To improve the detection precision of tiny objects, we added a new tiny object detection head based on the conventional YOLOv5. To reduce the deviation between the predicted box and the ground truth box, we used the SIoU Loss function. Finally, we fused and adjusted the above improvement points and obtained high-precision detection method: KCFS-YOLOv5. This detection method was evaluated on three datasets (NWPU VHR-10, RSOD, and UCAS-AOD-CAR). The comparative experiment results demonstrate that our KCFS-YOLOv5 has the highest accuracy for the object detection in aerial remote sensing image.

Keywords: aerial remote sensing image; object detection; coordinate attention mechanisms; feature fusion; SIoU loss; tiny object detection

check for updates

Citation: Tian, Z.; Huang, J.; Yang, Y.; Nie, W. KCFS-YOLOv5: A High-Precision Detection Method for Object Detection in Aerial Remote Sensing Images. *Appl. Sci.* **2023**, *13*, 649. https://doi.org/10.3390/app13010649

Academic Editor: Seokwon Yeom

Received: 3 November 2022
Revised: 27 December 2022
Accepted: 29 December 2022
Published: 3 January 2023

1. Introduction

In recent years, with the continuous development of wireless communication, materials science, artificial intelligence, track analysis, and other technologies, the amount of valuable information in aerial remote sensing images has increased significantly. With the improvement of imaging quality, the multi-scale object which exists in the image contains more detailed and strategic information; some multi-scale targets in aerial remote sensing images such as vehicles, civil facilities, and aircraft can also be found easily by the search equipment, for which the aerial remote sensing images have become a hot spot for scholars in geology [1], agricultural [2], military [3], and forestry [4]. In the era of artificial intelligence, the economic and strategic values of aerial remote sensing images are gradually reflected. For example, Yao et al. [5] proposed an improved algorithm based on Mask RCNN to improve the accuracy of identifying special geological structures, and the unique geographical structure can be protected to achieve the role of disaster reduction. Meng et al. [6] used an object detection algorithm based on visual text information to search for the point of the gas leakage in aerial remote sensing images, their research can alleviate environmental pollution and prevent economic losses. Therefore, the high-precision detection method of multi-scale objects in aerial remote sensing images has high research value.

On the research of the object detection, the high-precision detection of multi-scale objects in aerial remote sensing images remains a formidable challenge. First, the multi-scale object is readily disturbed by environmental factors such as light intensity, topographic influences, and prevailing climatic conditions in the background of the aerial remote sensing images, objects are easily ignored by detection methods. Second, the ability of object feature extraction and fusion of existing detection methods still have room to be improved, because the loss of feature information will lead to a decrease in detection accuracy. Third, due to the large size of aerial remote sensing images, the feature of the tiny object is easily covered by the feature of the large object, which will result in the reduction of detection accuracy, if the algorithm pays too much attention to the object with large size, the detection accuracy of objects that have tiny sizes will be reduced. Fourth, the existing loss function usually ignores the vector angle between the ground truth box and the prediction box, which affects the detection accuracy in the detection task.

The detection ability of the traditional object detection algorithm based on the template matching method [7] is not efficient, because this method needs to manually extract and classify object features, its speed and accuracy are difficult to guarantee when dealing with a large number of tiny objects. Machine learning algorithms [8] often need to describe a large number of features by complex mathematical statistical methods, and their detection performance and generalization ability are obviously insufficient.

With the development of deep learning methods in the research of object detection, a large number of object detection methods based on deep neural networks have been proposed. Among them, the representative two-stage algorithm are R-CNN serial algorithms [9–11], while the representative single-stage algorithms are YOLO serial algorithms [12–17] and SSD serial algorithms [18,19]. The emergence of these algorithms based on deep neural networks has greatly improved the performance of object detection. The aerial remote sensing image as a kind of image is also used in the research of object detection. Yan et al. [20] proposed an improved Faster-RCNN algorithm to improve the recognition accuracy of mineral resources. Luo et al. [21] improved the detection precision of aircraft objects in remote sensing images by reducing the feature fusion output ports of neck network in YOLO algorithm. Although the above algorithm has excellent detection effect in a single category of object detection, the applicability is poor because remote sensing images often contain different categories of objects.

Aiming at the problem of object detection in aerial remote sensing images, an improved YOLOv5 object detection method called KCFS-YOLOv5 was proposed. The improved algorithm realizes the high-precision detection of multi-scale objects in aerial remote sensing images such as aircraft, vehicles, baseball fields and other common remote sensing objects. We used KCFS-YOLOv5 detection method in three aerial remote sensing image datasets [22–24], and achieved better detection average accuracy than the existing algorithms.

The innovation of our KCFS-YOLOv5 is reflected in the following aspects:

- We adopted the K-means++ clustering method to optimize the initial clustering points of anchor box.
- We embed the CA attention module into the backbone network of the YOLOv5 to improve the feature extraction ability of the algorithm for objects in aerial remote sensing images.
- We embedded BiFPN in the neck network of YOLOv5 to optimize the network framework to strengthen the fusion ability of the neck network for different dimensional features.
- Based on main characteristics of aerial remote sensing images which contain many tiny objects, we added a new tiny object detection head based on the conventional YOLOv5 to preserve the texture features of tiny objects and improve the detection accuracy for tiny objects.
- We selected the SIoU loss function to replace the conventional GIoU loss function to reduce the deviation between the predicted box and the ground truth box which can improve the detection performance.

The rest of our article is arranged as follows: Section 2 recommends the universal advanced object detection modules and their practical value of object detection in aerial remote sensing images, and the conventional YOLOv5 object detection algorithm model is also introduced in this section. The proposed KCFS-YOLOv5 object detection method is covered in Section 3. Section 4 presents the simulation environment and the preparation for experiments. The simulation results and the discussion are demonstrated in Section 5 and Section 6, respectively. The conclusion is presented in Section 7.

2. Related Works

2.1. Current Mainstream Object Detection Algorithm

Many studies have been investigated to address these tricky challenges. For example, before 2010, although traditional object detection methods such as template matching [7] and machine learning [8] are convenient to implement, the detection performance was poor. In 2012, Krzyzewski et al. [25] proposed a Deep Convolutional Neural Network (D-CNN) called AlexNet to bring object detection into the era of deep learning. Girshick [9] proposed a two-stage object detection algorithm called the region convolutional neural network (R-CNN), it surpasses other algorithms in detection accuracy, but the detection speed is slower. On this basis, Sun et al. [10] and Ren et al. [11] proposed the Fast Region-Based Convolutional Neural Network (Fast R-CNN) and the Faster Region-Based Convolutional Neural Network (Faster R-CNN) which improve the speed and accuracy of the R-CNN algorithm. Although the improved two-stage algorithm has improved the detection efficiency, it is still difficult to meet the needs of practical detection. Therefore, to intensify the practicability of the algorithm, Anguelov et al. [18] proposed a single-stage detection (SSD) algorithm; SSD serial algorithms have high detection accuracy and speed, but a disadvantage is that they require many manual parameters, which causes difficulties in practical applications. Since 2016, Redmon et al. [12] introduced the YOLO module (You Only Look Once), and YOLOv2 [13]. They balance the detection accuracy and detection speed to realize real-time detection of ordinary-sized objects. Meanwhile, because of its unique grid cell framework, the detection performance of tiny objects is relatively poor. In 2018, Redmon et al. [14] proposed YOLOv3, by adding Mosaic data augmentation to strengthen the network's ability to surpass the two-stage model in accuracy and YOLOv4 [15] was born in 2020. YOLOv5 is the most recognized sequel to the YOLO serial algorithms. Although the average detection accuracy and the detection efficiency have been improved compared to the previous works, the detection effect of multi-scale objects with complex backgrounds still needs to be improved.

2.2. Object Detection in Aerial Remote Sensing Images

In recent years, with the development of deep convolutional network applications in images, many scholars performed much valuable research in the field of multi-scale object detection in aerial remote sensing images. Da et al. [26] embedded the transition module, the residual network and the spatial pyramid pooling structure into the original YOLOv3 network, which could improve the detection precision of tiny ships in remote sensing images. Luo et al. [21] improved the detection precision of aircraft objects in remote sensing images by reducing the feature fusion output ports of neck network in YOLO algorithm. Cao et al. [27] optimized the CSP framework in the YOLO backbone network, enhanced the non-linear classification ability of the network with Mish activation function, and embedded the pyramid pooling framework in the neck network. The above methods strengthened the ability of feature fusion, and made the detection accuracy of objects increase. Li et al. [28] proposed an effective object detection method in low altitude environment for remote sensing images. Although it has high detection accuracy for low-altitude UAV objects, the detection result of aerial remote sensing images is poor. Wang et al. [29] embedded CBAM [30] attention module into the conventional YOLOv5 network to strengthen the ability of the algorithm to detect tiny objects in 2021. Although the algorithm makes the detection accuracy improve, It requires a lot of computation and reduces the operation efficiency. On this basis, Yang et al. [31] introduced the ECA [32] module and the SAHI [33] framework in YOLOv5 and tested them on three aerial remote sensing image datasets, achieving excellent detection results. However, they ignored the initial clustering points of anchor box and position features of multi-size objects, therefore, the detection accuracy could still be improved.

2.3. Attention Mechanism in Object Detection Task

The principle of the conventional attention mechanism is based on the visual selectivity of humans. In daily life, human beings always focus on the more important parts of a picture using their own previously-obtained information. The attention mechanism module adaptively updates the weights and screens the information to focus the attention on the more noteworthy information, ignores the useless information, and further converges the algorithm vision. Currently the main attention mechanisms are channel domain attention mechanisms, spatial domain attention mechanisms and fusion attention mechanisms.

The spatial attention mechanism obtains the attention weight mainly through the relative position of all objects in the image. STNs (Spatial Transformer Networks) [34] converts the object in reverse space to reduce the degree of image information loss and improve the accuracy in the detection task.

The representative algorithm for channel attention module is the SEnet (Squeeze-and-Excitation Networks) [35]. The method of the global average pool is used to extract and classify the information of different channels, and generates the probability statistics of each channel, and then relies on the fully connected layer to generate the weights between different channels, and finally outputs the weighted channel weights.

The mixed domain attention mechanism is a fusion algorithm which fuses the spatial attention mechanism and the channel attention mechanism. The representative algorithm is the CBAM [30] module. This module proves that input sequence of attention mechanisms has an impact on the experimental results, The authors input the feature map into the channel attention mechanism and then the spatial attention mechanism, and achieved better feature extraction results.

In recent years, various advanced attention mechanisms have been widely used in object detection. For example, Sun et al. [36] embedded the channel attention mechanism into YOLOv5 framework to achieve the high-precision detection of the leaf disease. Wang et al. [29] embedded a mixed attention mechanism into the Two-Stream Convolution Network to strengthen detection ability of the algorithm for human action recognition tasks. The above studies prove that adding an appropriate attention mechanism in the object detection task is helpful in improving the performance of the algorithm.

Based on valuable previous studies, a novel network model, we called KCFS-YOLOv5, is developed to improve the detection precision of multi-scale objects in aerial remote sensing images.

2.4. Principles of the Conventional YOLOv5 Detection Framework

The YOLOv5 object detection framework proposed in 2020 by Ultralytics LLC is an improved framework based on the YOLO series. Structurally, it is a one-stage detection framework composed of four units: the Input, the Backbone network, the Neck network, and the Output. After drawing on the advantages of earlier versions of the YOLO series and other detection algorithms, YOLOv5 embeds the Focus layer into the input for data augmentation. Meanwhile, it used the DarkNet53 in the backbone to extract the main features from the image. A feature fusion framework which incorporates the feature pyramid structure [37] (FPN) and the bottom-up Path Aggregation Network [38] is also embedded in the neck network to strengthen the short-circuit linking and cross-layer fusion in multi-scale features. The complete YOLOv5 framework is shown in Figure 1. The four constituent units in the YOLOv5 network are demonstrated as follows:

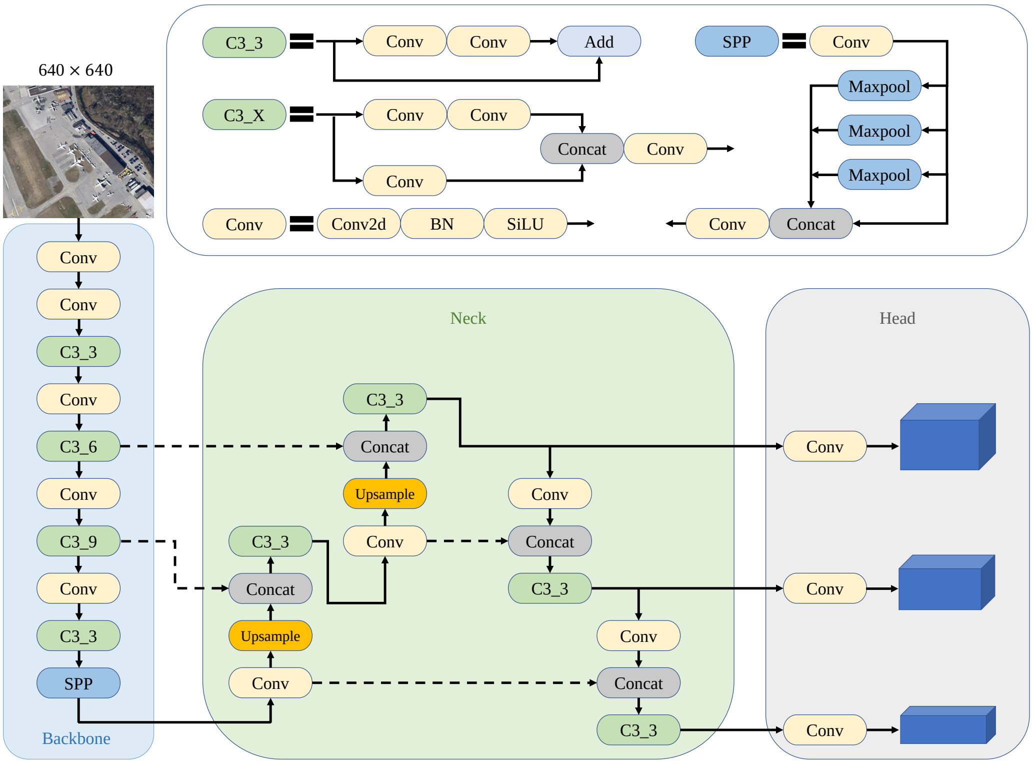

Figure 1. The conventional YOLOv5 algorithm framework.

Input: similar to YOLOv4, YOLOv5 uses the Mosaic module to augment the data. It uses four photographs, thus significantly increasing the amount of data between different pictures. Fusion enhances the detection generalization ability of background information and multi-scale objects and reduces the computational burden. The improved framework adjusts the size of input images to a unified 640 × 640 pixels through the adaptive image scaling module to improve data complexity. Meanwhile, in the process of network parameter training, the input framework generates the preset anchor box of the object through a K-means adaptive anchor box algorithm, calculates the deviation between the preset anchor and the ground truth box of the object, and update the network weight through the reverse transmission of the fusion framework.

The backbone network: YOLOv5 adopts CSPDarknet53 [14] as the backbone network; the backbone network consists of Focus layer, CSPNet framework and Spatial Pyramid Pooling (SPP) module [39]. The Focus layer is a method for data enhancement on the data side. When dealing with larger feature maps, YOLOv5 can first split and splice a feature map, then pass concatenating layers to stack images so that feature representations at different levels can then be extracted through convolutional layers. The CSPNet framework forms the backbone network and enhances the feature fusion ability of feature maps with different dimensions through residual connections, which is the basis of backpropagation in the network. The SPP module performs maximum pooling in four different dimensions: 1×1, 5×5, 9×9, and 13×13, to strengthen the network's receptive ability of pictures and differentiate feature information.

The Neck network: this fuses the texture information and position information in the feature map to strengthen the ability of information fusion on multi-scale objects, the neck network of YOLOv5 adopts the fusion structure of PAFPN [40]. The FPN structure further enhances the features of different dimensions in the network through Up-sampling, Graph fusion ability and multi-size object detection ability. The PAN framework can bring the information of the shallow layer to the bottom layer through short-circuit links, which improves the detection results of disturbing objects.

The Output: The output consists of the NMS module and the loss function, the original YOLOv5 used CIoU loss function in the output. It overcomes the problem where IoU is not steerable in special cases and improves the detection effect when the prediction frames overlap. Weighted NMS is used to consolidate the detection performance in multi-objective environment, thus obtaining the optimal detection framework.

Although YOLOv5 has shown excellent performance in different tasks of object detection, it still has some problems to be optimized in the object detection task for remote sensing images. The problems are as follows:

- The conventional YOLOv5 algorithm uses K-means algorithm to cluster anchor boxes at the output, but the K-means clustering algorithm randomly selects the center points of the K pre-training marked boxes in the data set as the center of each cluster. If the location of the initial points is not selected properly, the final clustering result should be very negative.
- Aerial remote sensing images contain rich information and multi-size objects. The traditional deep convolution structure makes it difficult to effectively extract the texture features of objects in the aerial detection task, and some important feature information will be ignored by the conventional feature extraction network.
- The conventional YOLOv5 uses the PAN+FPN framework as the neck network. This feature fusion network focuses on strict hierarchical information fusion, and ignores the information fusion between the shallow network and the bottom network, which will lead to the loss of feature information in the deep network.
- Aerial remote sensing images are rich in multi-size objects. For tiny objects, the output ports of the feature fusion framework of the original YOLOv5 algorithm are in three sizes, which makes some tiny objects output from the medium size port, and reduces the detection accuracy.
- The conventional YOLOv5s uses the GIoU loss function as the function of the predicted box, it ignores the vector angle between the predicted box and the ground truth box, which affects the detection accuracy in the detection task.

3. Improvements of the KCFS-YOLOv5

Aiming at the above problems, we put forward several measures to improve YOLOv5.

- We use the K-means++ algorithm [41] to optimize the initial clustering points of the anchor box, which makes the initial clustering points of anchor box distributed on the whole feature map as much as possible. In this method, the object on the feature map can be detected as much as possible, and the probability of missing detection can be reduced.
- We embed the Coordinate Attention (CA) module [42] into the backbone network of YOLOv5 to improve the feature extraction ability of objects in aerial remote sensing images, especially tiny objects.
- We embedded the Bidirectional Feature Pyramid Network (BiFPN) [43] in the neck network of YOLOv5 to optimize the network framework to improve the fusion ability of the neck network for different dimensional features.
- Based on the characteristics of aerial remote sensing images which contain many tiny objects, we added a new tiny object detection head based on the conventional YOLOv5 to preserve the texture features of tiny objects and improve the detection accuracy of tiny objects.
- We replaced the GIoU function with the SIoU [44] function, which can reduce the deviation between the predicted box and the ground truth box and improve the detection performance.

3.1. The Clustering Anchor Box Optimization

In the conventional YOLOv5 algorithm, the detection anchor box will have a certain effect on detection results. The initial anchor boxes in the conventional algorithm are nine anchor boxes of random sizes designed and generated using the K-means clustering algorithm, it covers objects of various sizes. However, the K-means algorithm randomly selects the center points of the K pre-training marked boxes in the data set as the center of each cluster. If the location of the initial points is not selected properly, the final clustering result should be very negative. It significantly affects the mean average precision (Map) of training. We therefore propose a K-means++ algorithm-based [41] anchor framework generation mechanism for optimization.

First, we set the number of center points K that need to be clustered, and randomly select a center point C_{first} of a pre-training marker framework C from the graph as the initial clustering center.

Second, we reselect all points except the initial cluster point we picked R_{rest} in the dataset and compute the IOU distance $D(C_{first}, R_{rest})$ of R_{rest} and C_{first} can be denoted by:

$$D(C_{first}, R_{rest}) = 1 - \text{IOU}(C_{first}, R_{rest}) \tag{1}$$

Then, the probability $P_{next\ 1}$ that each reselect point R_{rest} to become the next cluster center R_{next} in the first iteration can be expressed as:

$$P_{next\ 1} = \frac{D^2\left(C_{first}, R_{next}\right)}{\sum_{R_{rest} \in data} D^2\left(C_{first}, R_{rest}\right)} \tag{2}$$

Furthermore, using the Roulette method, the probability P_{next} of selecting the next cluster point C_{second} will be evaluated, and the process is shown in Figure 2.

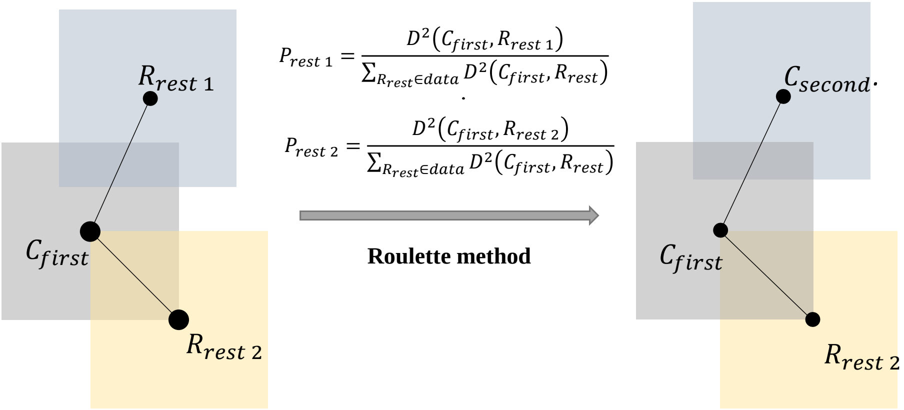
$R_{rest\ 1}$
C_{first}
$R_{rest\ 2}$
$P_{rest\ 1} = \frac{D^2(C_{first}, R_{rest\ 1})}{\sum_{R_{rest} \in data} D^2(C_{first}, R_{rest})}$
$P_{rest\ 2} = \frac{D^2(C_{first}, R_{rest\ 2})}{\sum_{R_{rest} \in data} D^2(C_{first}, R_{rest})}$
Roulette method
C_{second}.
C_{first}
$R_{rest\ 2}$

Figure 2. The process of generating the second cluster center.

After applying a K-mean++ algorithm, two cluster centers C_{first} and C_{second} in the graph are generated. We change the probability $P_{next\ 2}$ that each reselected point R_{rest} to become the next cluster center R_{next} in subsequent iterations which can be expressed as:

$$P_{next\ 2} = \frac{D^2(C_{close}, R_{next})}{\sum_{R_{rest} \in data} D^2(C_{close}, R_{rest})} \tag{3}$$

where C_{close} represents the cluster center closest to each reselect point R_{rest}, and the process is shown in Figure 3.

Like the first loop, we will evaluate the probability $P_{next\ 1}$ to select the next cluster point C_{third} by the Roulette method. After iterating the K-means++ algorithm K times, we can generate K cluster centers such as: $C_{first} C_{second} \cdots C_k$. Based on the K optimised cluster centers we have obtained, we use the traditional K-means clustering method that comes with YOLOv5 to generate anchors.

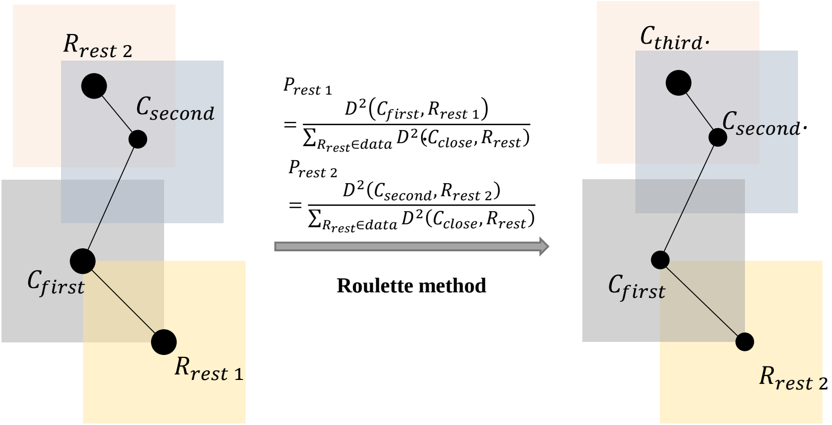
$R_{rest\ 2}$
C_{second}
C_{first}
$R_{rest\ 1}$
$P_{rest\ 1} = \frac{D^2(C_{first}, R_{rest\ 1})}{\sum_{R_{rest} \in data} D^2(C_{close}, R_{rest})}$
$P_{rest\ 2} = \frac{D^2(C_{second}, R_{rest\ 2})}{\sum_{R_{rest} \in data} D^2(C_{close}, R_{rest})}$
Roulette method
C_{third}
C_{second}
C_{first}
$R_{rest\ 2}$

Figure 3. The process of generating subsequent cluster centers.

This method ensures that there is a certain distance between the initial K cluster centers, which can effectively alleviate the problem that the cluster center points tend to local optimal solutions due to the chance of random selection, and at the same time it can improve the detection accuracy.

3.2. Coordinate Attention Module

The attention mechanism module converges the algorithm vision through the network adaptive update weight and information screening [30]. The algorithm used in the present research, The Coordinate Attention (CA) [42] modules is embedded in the bottom of the backbone network of YOLOv5. The CA module is built as depicted in Figure 4. The feature representation of objects in various environments can be enhanced with the help of these modules.

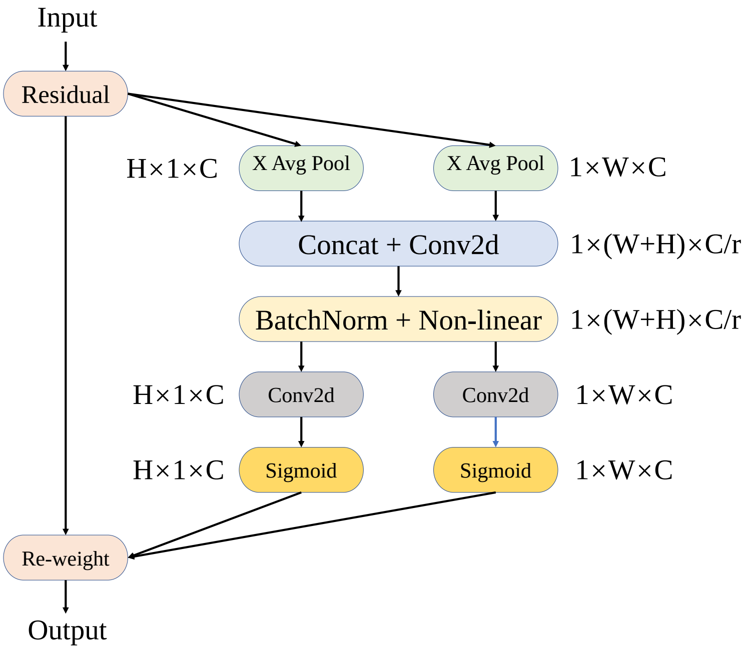

Figure 4. The conventional Coordinated Attention module.

CA is a convolutional attention mechanism that combines the channel attention and position information, which realizes attention extraction in different dimensions of channel and space of feature maps. In contrast to channel attention, which uses two-dimensional global pooling to convert the input into a single feature vector, CA disassembles the channel attention into one-dimensional features in two distinct directions. The input feature map is a tensor of arbitrary size X = $[x_1, x_2, \cdots, x_C]$ where C is the number of channels in the feature map, H is the height of the feature map, and W stands for the width of the feature map.

First, an average pooling kernel measuring (1, W) and (H, 1) is used to encode each channel in horizontal and vertical coordinates, respectively. At height h and width w, the outputs of the *c-th* channel are formulated as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} x_c(h, i) \tag{4}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le i \le H} x_c(j, w) \tag{5}$$

These two transformations result in feature fusion in the height and width directions of the space. The convolution kernel which has size one is used to transform the two feature maps, and an intermediate feature map $f \in \mathbb{R}^{C/r \times (H+W)}$ containing both horizontal and vertical spatial information is generated. The formula is shown below:

$$f = \delta(F_1([z^h, z^w])) \tag{6}$$

where r is the ratio of Down-sampling.

Second, the spatial dimension of the middle feature map f will be divided into two distinct tensors, f^h and f^w. Furthermore, a convolution kernel which has size one will be used to convert the two processed feature maps f^w and f^h to the same number of channels as X, as given by:

$$g^h = \sigma(F_h(f^h)) \tag{7}$$

$$g^w = \sigma(F_w(f^w)) \tag{8}$$

where σ is the activation function called sigmoid. The processed feature map f^h and f^w will be generalized as the attention weight of the system, and the final output is given by:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{9}$$

Coordinate attention integrates more location information and semantic information on the previous attention mechanism such as CBAM, so that the network can learn the feature map and achieve higher detection accuracy in the multi-size object detection task.

Here, the CA module is inserted into the bottom of the backbone network to strengthen the capabilities of the backbone network to extract and learn semantic features of different dimensions. This structure can make the backbone network better obtain the pre-training weight, and Figure 5 displays the model comparison results. As Figure 5 shows, the colors of different regions indicate the degree of the attention module in the network to different positions of the image. The more the color tends to blue, the higher the attention of the network to the region. From the results, adding the CA module enhances the capability of feature extraction of the object in deep network. Aiming at the distribution characteristics of tiny objects and important location information in the aerial remote sensing images. Based on the conventional feature extraction network, more location information and semantic information are retained in the specific multi-size object feature extraction, which lays the foundation for the subsequent detection process.

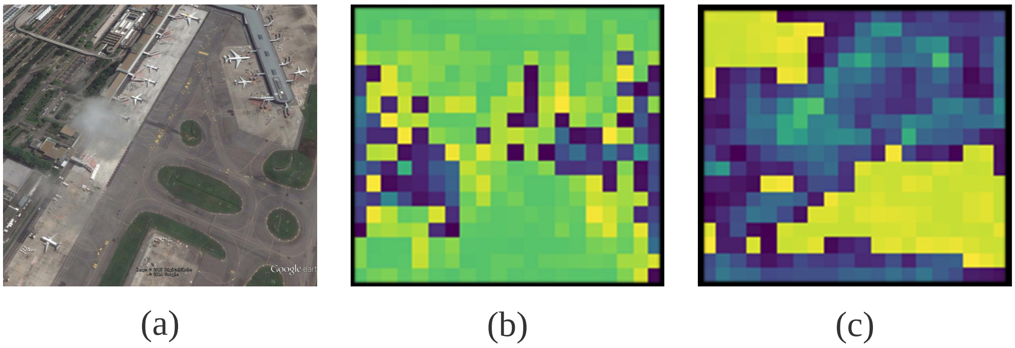

Figure 5. The feature map comparison of YOLOv5 and YOLOv5 with Coordinate Attention module. Where, (**a**) is the original remote sensing image; (**b**) is the feature map without the Coordinate Attention module; (**c**) is the feature map with the Coordinate Attention module.

3.3. Optimize the Feature Fusion Framework

In the conventional YOLOv5 framework, PAN+FPN framework is used in the feature fusion framework of neck network, this framework can make the information fusion more effective and realize information communication between the shallow layer and deep layer of the network. Compared with the ordinary optical image, the aerial remote sensing image has more tiny objects scattered randomly and the complex background. Therefore, we hope to fuse more multi-size object features in different dimensions. However, the framework of PAN+FPN emphasizes strict Up-sampling and Down-sampling in the process of feature fusion, and does not focus on the fusion of multi-size object features in aerial remote sensing images. The Bidirectional Feature Pyramid Network (BiFPN) [43] framework can effectively solve the above problems. In this paper, we embed the BiFPN framework into the neck network of the YOLOv5. The framework of the BiFPN is shown in Figure 6. Compared with the framework of PAN+FPN, instead of emphasizing strict stepwise sampling, BiFPN uses a short-circuit link method to better integrate information from different layers. Meanwhile, the conventional YOLOv5 uses Concat fusion method, which cannot extract features effectively. BiFPN introduces a weighted feature fusion method, which can adjust the importance of different features by training weights, emphasizing the object features and ignoring the background information. BiFPN normalizes the weight of each feature to a value between 0 and 1, which indicates the importance of the feature. Furthermore, the process is as follows:

$$O = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} I_i \tag{10}$$

where, the feature map is represented as the O. the ω_i and ω_j are the weights of the feature map. The value of ε is 0.0001, which ensures the stability of weight update. In the improved neck network, short-circuit links are added to P3 and P4 layers for weighted feature fusion of different dimensions. Furthermore, the fusion process of layer P4 is shown as follows:

$$P_4^{td} = Conv\left(\frac{\omega_1 P_4^{in} + \omega_2 Sample\left(P_5^{in}\right)}{\omega_1 + \omega_2 + \varepsilon}\right) \tag{11}$$

$$P_4^{out} = Conv\left(\frac{\omega_1^n P_4^{in} + \omega_2^n P_4^{in} + \omega_3 Sample\left(P_3^{out}\right)}{\omega_1^n + \omega_2^n + \omega_3^n + \varepsilon}\right) \tag{12}$$

where $Conv$ is the process of convolution. Furthermore, $Sample$ represents the process of Up-sampling and Down-sampling.

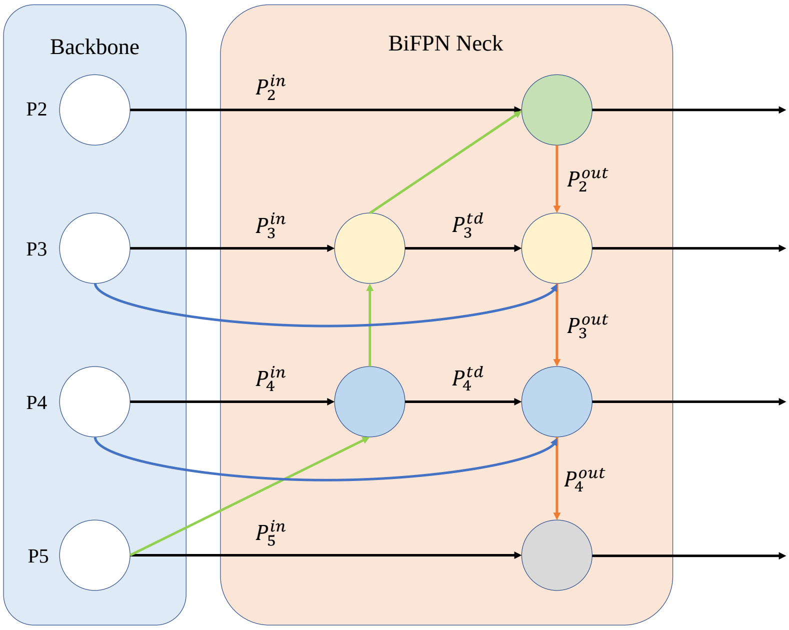

Figure 6. The conventional framework of the BiFPN.

3.4. The New Tiny Object Detection Head

In the original YOLOv5 framework, the image is subjected to 8, 16, and 32-fold downsampling to generate different levels of feature maps which be sent to the detection head, that is, the detection head obtains 20 × 20 (small), 40 × 40 (medium), and 80 × 80 (large) feature maps if the size of the input image is 640 × 640. Compared with the ordinary optical image, the remote sensing image has more tiny objects. Excessive downsampling will destroy the features of tiny objects in aerial remote sensing images such as aircraft and vehicles. Feature extraction for large feature maps can reduce the receptive field of input feature maps, reduce the loss of feature information in networks, and strengthen the tiny object detection ability. We added a new tiny object detection head based on the original YOLOv5 to preserve the texture features of tiny objects. The improved framework is shown in Figure 7, and we will obtain the new detection head with the size of 160 × 160 by 4 times downsampling. The framework of the improved object detection module is illustrated in Figure 8. Figure 9 presents the comparison of the detection results using the improved object detection module and the conventional object detection module in YOLOv5.

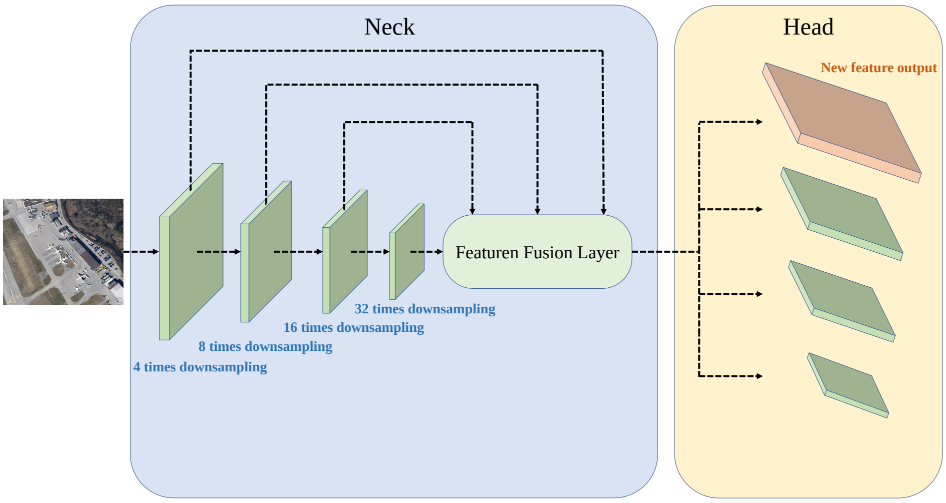

Figure 7. The new feature fusion module.

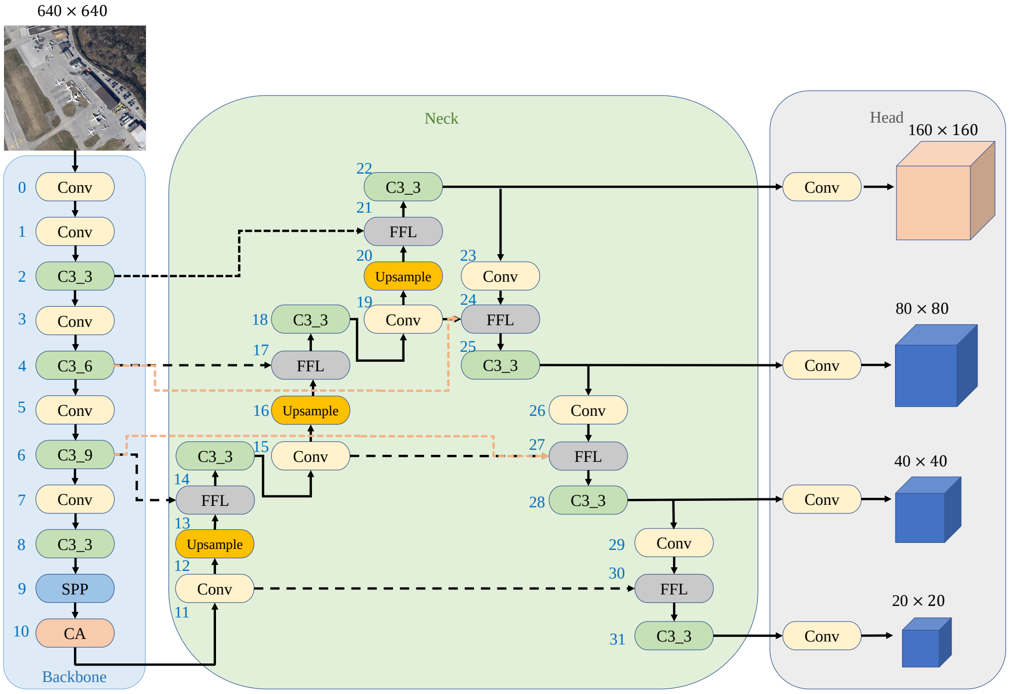

Figure 8. The framework of the improved object detection module.

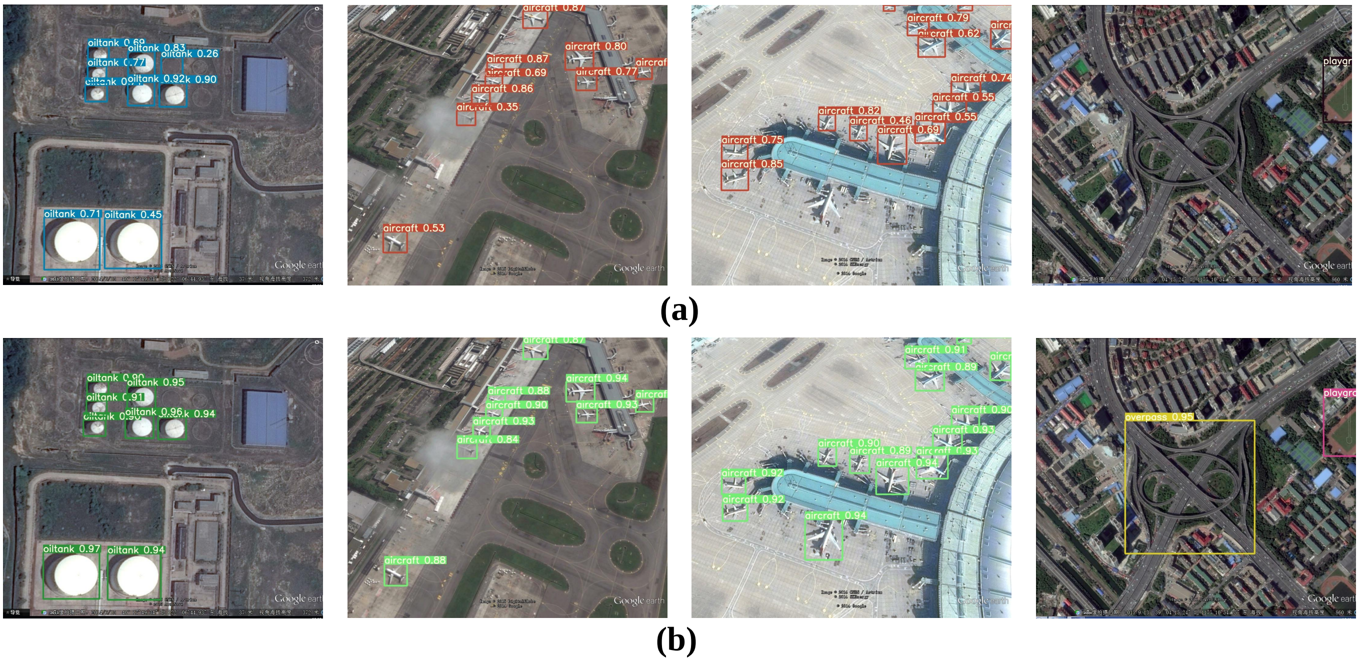

Figure 9. This is the comparison results using the improved object detection module and the original YOLOv5. (**a**) is the result of the conventional YOLOv5, (**b**) is the result of the improved object detection module.

3.5. Optimization of the Loss Function

Tiny objects in aerial remote sensing images place extremely onerous requirements on the precision of the loss function in the network because of their tiny feature objects and important texture information. The loss function is used at the output of the original YOLOv5s framework, relying on it to calculate the deviation value between the predicted box and the ground truth box of the detection object. The conventional YOLOv5s uses the GIoU loss function [45] as the function of the predicted box; the complete function is represented as follows: first, the minimum circumscribed rectangle C of the two AB detection boxes can be calculated, and then the ratio of the size of the AB area difference set to the area of the circumscribed rectangle C is calculated, and then the GIoU is attained by subtracting this ratio from the IoU value of AB. Compared with the IoU function of the early YOLO series, although GIoU solves the problem that IoU is not steerable in some cases and cannot be judged when the IoU is the same, but when the predicted box overlaps with the ground truth box, GIoU will return to IoU, repeating the above problems; the subsequently proposed CIoU loss function [46] takes into account information such as the height-width ratio of the frame and the position of the center. Although it solves the problem of box overlap, it ignores the deviation in the orientation between the ground truth box and the predicted box, which leads to "wandering" of the predicted frame in the training process, and finally affects the updating of the weight, so as to reduce the detection performance.

$$GIOU_{Loss} = (\mathrm{IOU} - \frac{C - (A \cup B)}{C}) \tag{13}$$

The SIoU [44] is an improved loss function based on the CIoU and GIoU function. It considers the vector angle between ground truth box and the predicted box, providing regression direction guidance for predicted box, and redefining the penalty function. In the tiny object detection-based remote sensing images, the performance of object detection results can be improved. Its working principle is to calculate the angle loss value between the ground truth box and the predicted box, as defined by:

$$\Lambda = 1 - 2 \times sin^2(arcsin(\frac{C_h}{\sigma}) - \frac{\pi}{4}) = cos(2 \times (arcsin(\frac{C_h}{\sigma}) - \frac{\pi}{4})) \tag{14}$$

where C_h is the height difference between the center point between the predicted box and the ground truth box, and σ is the deviation in distance between the center point of the predicted box and the ground truth box. The process is shown in Figure 10.

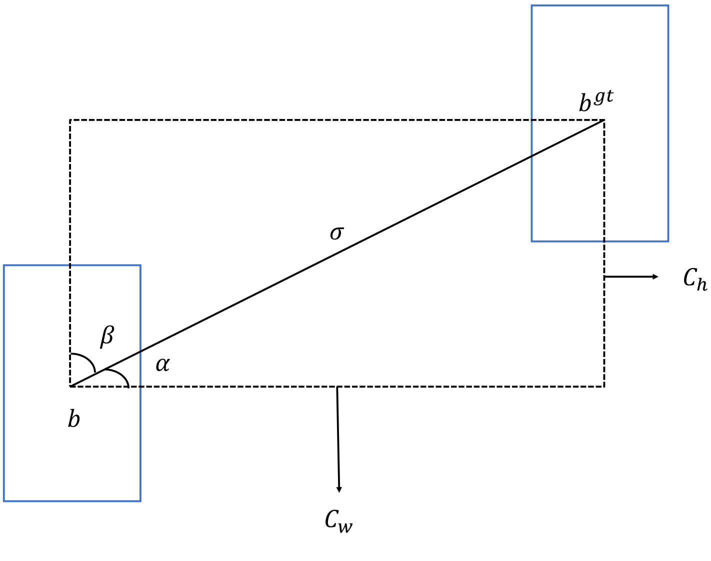

Figure 10. Calculation of the angle loss value.

The deviation in distance between the prediction box and the detection box can be defined as follows:

$$\triangle = \sum_{t=x,y} (1 - e^{-\gamma\rho t}) = 2 - e^{-\gamma\rho x} - e^{-\gamma\rho y} \tag{15}$$

$$\rho_x = (\frac{b^{gt}_{C_x} - b_{C_x}}{C_w})^2, rho_y = (\frac{b^{gt}_{C_y} - b_{C_y}}{C_h})^2, gamma = 2 - \Lambda \tag{16}$$

where h_w and h_h are the minimum enclosing rectangle of the ground truth box and the predicted box, respectively. The process is shown in Figure 11.

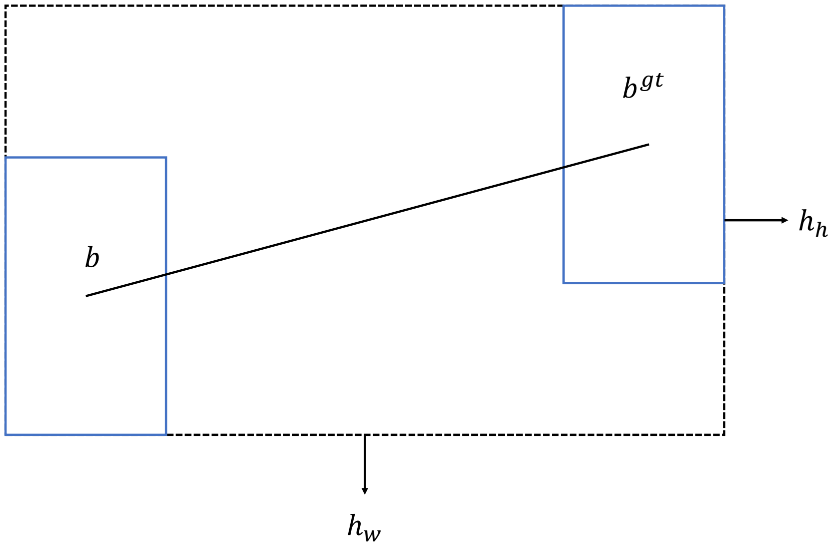

Figure 11. The process of redefining the distance deviation.

Evaluation of loss shape is defined as:

$$\Omega = \left(1 - e^{-w_w}\right)^{\theta} + \left(1 - e^{-w_h}\right)^{\theta} \tag{17}$$

$$w_w = \frac{\left|w - w^{gt}\right|}{max\left(w, w^{gt}\right)}, w_h = \frac{\left|h - h^{gt}\right|}{max\left(h, h^{gt}\right)} \tag{18}$$

where (w,h) and (w^{gt},h^{gt}) denote the width and height of the predicted box and the ground truth box, respectively, and θ represents the degree of attention that the loss function pays to the shape loss.

Finally, based on the IoU values of the predicted box and the ground truth box and the above parameters, the SIoU loss function is determined as follows:

$$Loss_{\text{SIoU}} = 1 - \text{IoU} + \frac{\triangle + \Omega}{2} \tag{19}$$

Based on the diversity of shapes and sizes of target objects in remote sensing images, especially to improve the algorithm performance of tiny objects in aerial remote sensing images and reduce the deviation between the predicted box and the ground truth box at the output of the object detection model. This paper strengthens the detection ability of the overall framework, by replacing the original GIoU loss function with a more accurate SIoU loss function that considers direction factors in the improved object detection module.

Therefore, based on the above improvement points, we proposed the KCFS-YOLOv5: a High-Precision detection method for object detection in aerial remote sensing images, and the detection process of KCFS-YOLOv5 is shown in Figure 12.

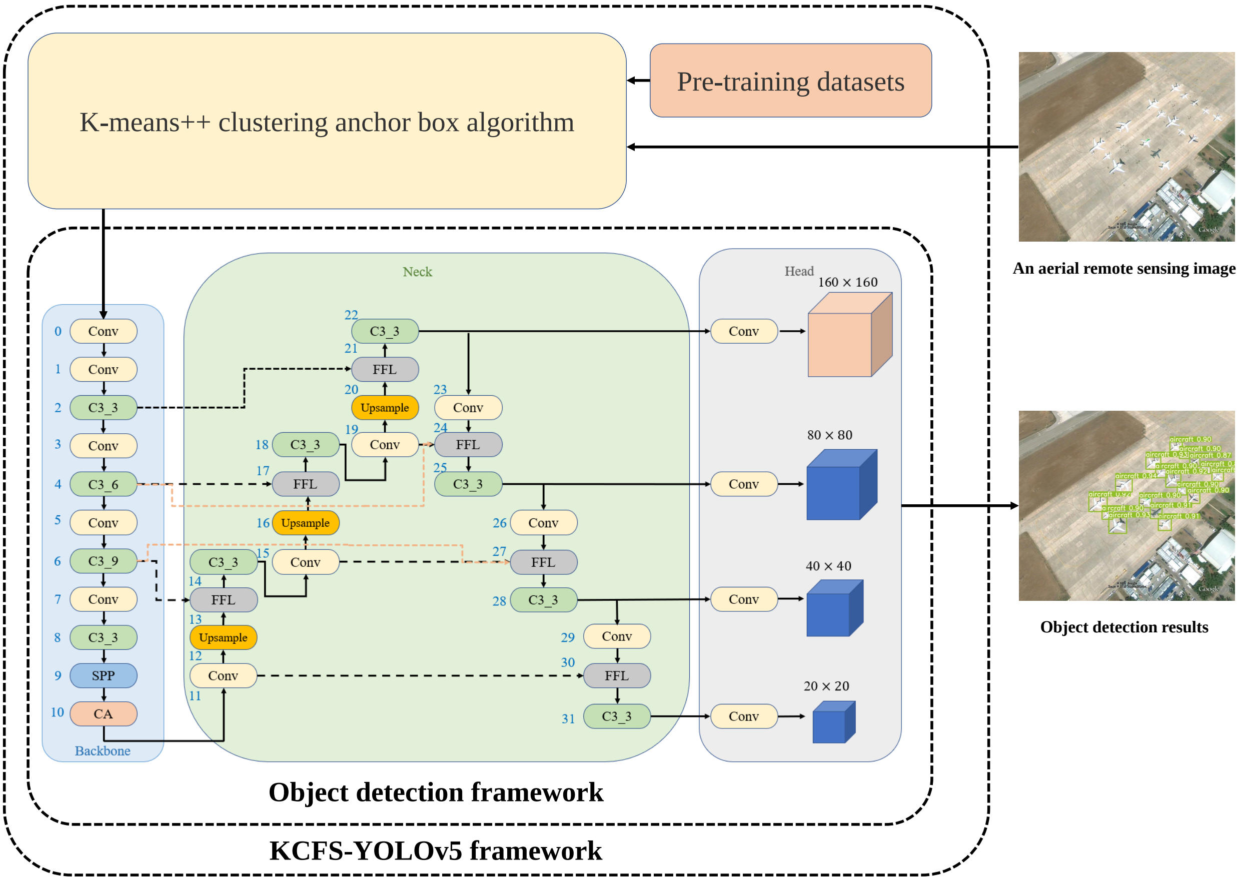

Figure 12. The detection process of KCFS-YOLOv5.

4. Simulation

4.1. Simulation Environment

These experiments build a deep network framework based on Pytorch platform. Meanwhile, the hardware configuration of the experiment is as follows: the graphics card is NVIDIA GTX2060 with 8 GB of memory, the core processor is AMD Ryzen 7 5800H with Radeon Graphics. To alleviate the computational burden in terms of the GPU in the experiment, CUDA is introduced to accelerate the experiment (Python is used as a compilation language).

4.2. Evaluation Indicator

For precision, we recall that the *Map* is often used as an experimental indicator of algorithm performance in the task of object detection. Among them, the accuracy rate (P_r) is the ratio of the number of samples classified as correct by the system to the total number of samples, which is expressed as follows:

$$P_r = \frac{P_T}{P_T + P_F} \times 100 \tag{20}$$

where P_T denotes the correct samples which are assigned correctly; and P_F represents the positive samples that are incorrectly assigned. The recall rate (R_e) is the ratio of positive samples in a sample which are predicted to be right, and its expression is:

$$R_e = \frac{P_T}{P_T + N_F} \times 100 \tag{21}$$

where P_T refers to the correct samples which are categorized correctly, N_F represents the negative samples which are assigned to the negative answer. However, in general, recall and precision are difficult to maintain at high levels at the same time. Therefore, the function of the parameter map is to integrate these two parameters to achieve a comprehensive evaluation of the system's performance. Meanwhile, the expression can be expressed as:

$$Map = \frac{\sum_{k=1}^{N} P(k)\Delta R(k)}{C} \tag{22}$$

where N is the total number of samples in the test set, $P(k)$ denotes the accuracy rate when K samples are identified simultaneously, $\Delta\ R(k)$ represents the change in recall rate after differentiation, and C is the number of object categories in the multi-class object detection task.

On the other hand, in order to further explore the detection speed of the algorithm, Latency and FPS of each algorithm are counted and compared. Where, latency represents the average inference time to detect an image, and FPS means the number of image frames detected per second. Their relationship is defined as:

$$FPS = \frac{1000}{Lantency} \tag{23}$$

In addition, we will also count the computation of each module, which is used to measure the size of the network and is represented as the NP.

4.3. Dataset

We evaluated the proposed improved YOLO5 model on three aerial remote sensing image datasets: NWPU VHR-10 [23], UCAS-AOD-CAR [24], and RSOD [22].

The NWPU VHR-10 dataset is a geographic remote sensing dataset for spatial object detection. It contains 650 images containing objects and 150 background images, a total of 10 categories of objects, which were taken from Google Earth. In the process of data processing, we uniformly adjust each image to a size of 640 × 640 pixels and divide the image into training set, validation set and test set with a ratio of 7:2:1. Among them, there are 560 images in the training set, 160 images in the verification set and 80 images in the test set.

The RSOD dataset contains 976 images, four categories, and 6950 labeled objects. It has a uniform size of 1000 × 1000 pixel. Although the RSOD dataset is widely used in object detection tasks, the number of objects in different categories is not balanced. The further to enrich the content of the dataset, data augmentation is performed based on the original dataset; the original image is cut, flipped, and pixel-filled under the premise of keeping the image aspect ratio unchanged. In the end, the size of the processed dataset is increased from 976 images to 1500 images.

The UCAS-AOD dataset is an aerial remote sensing image dataset for vehicle and aircraft detection. It contains 300 vehicle images and 600 aircraft images, which were taken from Google Earth in 2014. All images in the dataset are 1280 × 659 pixels in size and evenly distributed in orientation. To verify the detection effect of the improved YOLOv5 framework on dense tiny objects, we selected vehicle datasets with denser distribution in the dataset for simulation experiments. We uniformly adjust each image to a size of 640 × 640 pixels and randomly assigned the image into training set, validation set and test set with a ratio of 7:2:1. After image processing, the number of training sets, test sets and validation sets is 210 images, 30 images and 60 images, respectively.

5. Simulation Results

5.1. Effects of Kmeans++ Clustering Anchor Box

In order to prove the effect of K-means++ clustering anchor box algorithm, we add K-means++ clustering anchor box algorithm to the conventional YOLOv5 algorithm, and evaluate the effect of the improved algorithm on three different datasets. The results are shown in Table 1.

Table 1. Comparison of the detection performance of the improved algorithm embedded with K-means++ algorithm and the conventional algorithm.

Algorithm	**NWPU VHR-10**				**RSOD**				**UCAS-AOD-CAR**			
	Map **(%)**	$\boldsymbol{P_r}$ **(%)**	$\boldsymbol{R_e}$ **(%)**	***FPS***	***Map*** **(%)**	$\boldsymbol{P_r}$ **(%)**	$\boldsymbol{R_e}$ **(%)**	***FPS***	***Map*** **(%)**	$\boldsymbol{P_r}$ **(%)**	$\boldsymbol{R_e}$ **(%)**	***FPS***
YOLOv5	91.08	91.08	86.26	54.3	90.67	90.59	89.56	56.5	86.48	85.30	83.36	60.6
YOLOv5+K-means++	91.71	95.86	87.82	52.1	93.97	95.27	91.41	54.3	88.17	91.04	83.68	55.9

The experimental result is the average value calculated by several experiments. It can be seen that the improved algorithm has different improvement effects on the three test data sets. The improved result is approached to the global optimal value of the algorithm. Therefore, we believe that adding the Kmeans++ clustering anchor box algorithm in YOLOv5 is conducive to improving detection performance.

Meanwhile, the method of the K-means++ clustering anchor box will increase the computation and detection speed of the improved algorithm. Since it runs independently of the improved algorithm, its number of parameters is not counted in the total number of parameters of the improved object detection.

5.2. Effects of Coordinate Attention Module

To compare the results of different attention modules, we added several advanced attention modules in the YOLOv5 algorithm for contrast experiments. Among them, SE [35] is a classical channel attention module, and it is the basis for other attention models. The CBAM is a fusion attention model that integrates channel attention and spatial attention, which has excellent performance on different datasets. NAM [47] is a lightweight fusion attention module based on the CBAM module. The ECA is an improved attention mechanism based on SE module. We added these attention modules and the Coordinate attention module to the back of the backbone network and sent the processed feature map to the same feature fusion layer. We conducted simulation experiments on the RSOD dataset which has abundant object categories, and the results are shown in the following Table 2:

Table 2. The effects of different attention mechanism module on RSOD dataset.

Algorithm	***Map*** **(%)**	$\boldsymbol{P_r}$ **(%)**	$\boldsymbol{R_e}$ **(%)**	**P (M)**	***Lantency*** **(ms)**	***FPS***
YOLOv5	90.67	90.59	89.56	7.072	17.7	56.5
YOLOv5+SE	91.56	91.64	87.65	7.200	18.3	54.6
YOLOv5+CBAM	93.76	95.25	91.80	7.302	19.5	51.3
YOLOv5+NAM	92.14	92.98	89.54	7.237	20.3	49.3
YOLOv5+ECA	93.07	94.01	90.33	7.222	18.7	53.5
YOLOv5+CA	94.17	95.27	91.41	7.253	19.6	51.0

From Table 2, we can see that after embedding coordinate attention module into YOLOv5 algorithm, the improved algorithm achieves the best detection effect on the RSOD dataset. Where, *Map* P_r and R_e reach 94.17%, 95.27% and 91.41%, respectively. The precision of the algorithm framework can be improved obviously without dramatically increasing the computation and the *Lantency*. Therefore, we finally decided to embed the coordinate attention module in the YOLOv5.

To further explore the effect of coordinate attention module on task of object detection, we used the YOLOv5 algorithm embedded with the coordinate attention module for object detection on three different datasets, and evaluated the detection results. The results are shown in Table 3.

Table 3. Effects of Coordinate Attention module on three datasets.

Algorithm	NWPU VHR-10				RSOD				UCAS-AOD-CAR			
	Map (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*
YOLOv5	91.08	91.08	86.26	54.3	90.67	90.59	89.56	56.5	86.48	85.30	83.36	60.6
YOLOv5+CA	93.00	94.18	88.27	50.0	94.17	95.27	91.87	51.0	87.80	85.54	84.24	52.1

It can be seen from Table 3, the YOLOv5 algorithm embedded with Coordinate Attention module has achieved improved detection accuracy on three different datasets. Meanwhile, the detection speed of the improved algorithm is not significantly reduced. Therefore, we finally decided to use the Coordinate Attention module in the improved YOLOv5 algorithm.

5.3. Effects of the Bidirectional Feature Pyramid Network Framework

Compared with the ordinary optical image, the aerial remote sensing image has more multi-size objects which scattered randomly in the complex background. To further fuse more object features in different dimensions and ignore background interference. We embed the BiFPN framework into the neck network of the YOLOv5, and tested the improved algorithm on three different datasets. The results are shown in Table 4.

Table 4. Effects of the bidirectional feature pyramid network framework on three different datasets.

Algorithm	NWPU VHR-10				RSOD				UCAS-AOD-CAR			
	Map (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*
YOLOv5	91.08	91.08	86.26	54.3	90.67	90.59	89.56	56.5	86.48	85.30	83.36	60.6
YOLOv5+BIFPN	91.92	92.77	88.67	52.4	92.29	91.19	90.85	54.9	88.66	89.75	84.38	55.6

The experimental result is the average value calculated by several experiments. It can be seen from the results that the improved algorithm has excellent improvement effects on the three test datasets. Meanwhile, the improved YOLOv5 algorithm with BiFPN framework still has high detection speed. Therefore, we finally decided to use the Bidirectional Feature Pyramid Network(BiFPN) framework in the improved YOLOv5 object detection method.

5.4. Effects of the New Tiny Object Detection Head

Compared with the conventional YOLOv5 framework, the new tiny object detection head is inclined to extract and detect features of tiny objects. Therefore, we embed the new tiny object detection head into the neck network of the YOLOv5 to further achieve the feature of tiny objects, and tested the improved algorithm on three different aerial remote sensing datasets. We introduced the results in Table 5.

Table 5. Effects of the new tiny object detection head on three different datasets.

Algorithm	NWPU VHR-10				RSOD				UCAS-AOD-CAR			
	Map (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*
YOLOv5	91.08	91.08	86.26	54.3	90.67	90.59	89.56	56.5	86.48	85.30	83.36	60.6
YOLOv5 + New tiny object detection head	93.47	93.58	89.04	45.2	91.79	93.30	88.83	46.7	87.92	87.52	83.51	48.8

The new tiny object detection head improves the extraction and detection ability of tiny object features by UP-sampling. Therefore, the improvement effect is obvious for the dataset rich in tiny objects. The three datasets used in the experiment are all aerial remote sensing datasets, which contain a plenty of tiny objects such as aircarft and ships. It can be seen that the addition of the new tiny object detection improves the detection accuracy of the three datasets. Therefore, we decided to embed the new tiny object detection in the improved YOLOv5 object detection method.

To further explore the improvement effect of the improved neck network on the accuracy, we fused the Bidirectional Feature Pyramid Network(BiFPN) and the new tiny object detection head into the conventional YOLOv5 algorithm, and simulated the improved fusion neck network in the above dataset. Table 6 is the results.

Table 6. Effects of the improved fusion neck network on three different datasets.

Algorithm	NWPU VHR-10				RSOD				UCAS-AOD-CAR			
	Map (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e(%)	*FPS*
YOLOv5	91.08	91.08	86.26	54.3	90.67	90.59	89.56	56.5	86.48	85.30	83.36	60.6
YOLOv5+Improved neck network	93.77	93.69	89.71	42.6	93.77	93.60	92.87	42.9	89.86	90.02	86.85	45.5

The improved fusion neck network embedded with the conventional YOLOv5 network achieves the best detection performance which is better than adding the Bidirectional Feature Pyramid Network(BiFPN) framework and the new tiny object detection head alone. Meanwhile, the *FPS* of the improved algorithm is not significantly reduced. Therefore, the YOLOv5 algorithm with the improved fusion neck network has excellent performance. To further measure the practical value of the improved neck network, we count the parameters of the above improved fusion neck network. The results are shown in Table 7.

Table 7. The parameters of the above improved neck network.

Algorithm	P (M)
YOLOv5	7.072
YOLOv5+BIFPN	7.137
YOLOv5+New tiny object detection head	7.224
YOLOv5+Improved fusion neck network	7.298

Compared with the original algorithm, the number of network parameters of the improved fusion neck network has been improved, which will increase the network's demand for hardware. Since aerial remote sensing images are rich in tiny objects, the increased number of parameters and detection latency are tolerable compared to the improved detection accuracy. Therefore, we finally decided to embed the improved fusion neck network into the improved YOLOv5 object detection method.

5.5. Effects of the SIoU Loss Function

The confidence loss function which used in the YOLOv5 model is the GIOU loss. The GIoU loss and its improved version DIoU loss [48] and CIoU loss ignore the vector angle between the ground truth box and the prediction box, which affects the detection accuracy in the detection task. We used the SIoU function to improve the detection accuracy on above three datasets and we will present the results in Table 8.

It can be seen from Table 8 that when we applied the SIoU loss function in YOLOv5 algorithm, the best detection results were obtained. Compared with the conventional YOLOv5 framework, the value of the P_r and the *FPS* have been greatly improved. Therefore, we decided to choose SIoU as the loss function of the improved YOLOv5 object detection method.

Table 8. The detection results after adding several advanced loss functions in YOLOv5.

Algorithm	NWPU VHR-10				RSOD				UCAS-AOD-CAR			
	Map (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*	*Map* (%)	P_r (%)	R_e (%)	*FPS*
YOLOv5+GIoU	91.08	91.08	86.26	53.1	90.67	90.59	89.56	54.9	86.48	85.30	83.36	59.1
YOLOv5+DIoU	91.75	92.78	87.27	54.0	90.99	91.73	89.10	57.3	86.72	86.71	83.64	58.8
YOLOv5+CIoU	92.58	93.22	88.89	54.3	91.46	93.13	89.45	56.5	88.16	89.36	83.74	59.5
YOLOv5+SIoU	92.99	94.76	89.73	56.2	91.77	94.60	90.87	58.1	88.58	89.55	84.22	63.3

5.6. Ablation Experiment

To further measure the availability of the improved algorithm which introduced above, ablation experiments are conducted on the NWPU VHR-10 dataset. The experimental results are summarized in Table 9.

Table 9. The simulation results of adding each module to YOLOv5 in turn on NWPU VHR-10 dataset.

Algorithm	*Map* (%)	P_r (%)	R_e (%)	P (M)	*FPS*
YOLOv5	91.08	91.08	86.26	7.072	54.3
+K-means++	91.71 (+0.63)	95.86	87.82	7.156	52.1
+CA	93.11 (+1.40)	93.44	88.22	7.278	49.8
+Improved fusion neck network	93.97 (+0.86)	93.51	89.33	7.341	41.3
+SIoU(Improved Algorithm)	94.21 (+0.24)	94.20	89.42	7.341	43.9

Although the improved YOLOv5 network increases the number of parameters in the network and slightly reduces the detection speed, it achieved a 3.13% increase in *Map* value on the NWPU VHR-10 dataset. Meanwhile, the accuracy rate and the recall rate are also increased by 3.12% and 3.16%, respectively. Based on the above improved methods, we propose our new detection method in aerial remote sensing images: KCFS-YOLOv5.

5.7. Evaluation Model

Furthermore, in order to measure the excellent detection accuracy of the KCFS-YOLOv5 object detection method, we will present the results of a series of comparative experiments in the following article. The results from Table 10 show that compared with the conventional object detection method, the KCFS-YOLOv5 method significantly improves the detection performance of tiny objects in aerial remote sensing images.

Table 10. Comparison of the KCFS-YOLOv5 with the common algorithms on the NWPU VHR-10 dataset.

Category	SSD/%	Faster-RCNN/%	YOLOv3/%	YOLOv5/%	CAD [49]/%	ATTS [50]/%	CANet [51]/%	Ours/%
Airplane	85.1	91.6	89.8	98.6	97.0	99.8	99.9	99.5
Ship	72.7	81.9	77.9	83.3	77.9	92.5	86.0	85.1
Storage tank	80.6	83.0	83.3	97.6	95.6	97.1	99.3	98.1
Baseball diamond	85.1	85.8	89.4	97.5	93.6	92.7	97.3	98.8
Tennis court	80.2	80.1	84.5	92.4	87.6	93.7	97.8	96.4
Basketball court	70.6	82.9	74.6	82.9	87.1	96.7	84.8	92.4
Ground track field	83.4	94.0	88.1	98.8	99.6	98.4	98.4	99.5
Harbor	67.6	78.1	71.6	80.3	99.9	99.7	90.4	84.0
Bridge	83.4	61.3	87.4	93.0	86.2	71.5	89.2	94.3
Vehicles	72.2	76.1	76.3	86.4	89.9	91.6	90.3	93.8
Mean AP	78.1	81.5	82.3	91.1	91.5	93.4	93.3	94.2

The KCFS-YOLOv5 achieves the best Map detection result compared with other advanced algorithms on the NWPU VHR-10 dataset. Meanwhile, the detection results of each object category in the dataset are greatly improved compared with the conventional algorithm. The simulation results show the excellent performance of the improved YOLOv5 algorithm, and Table 10 shows the detection effect comparison of the KCFS-YOLOv5 with the conventional algorithms on the NWPU VHR-10 dataset. Our KCFS-YOLOv5 has achieved the highest detection accuracy, and some detection results are shown in Figure 13.

In order to further measure the performance of KCFS-YOLOv5 object detection method on other datasets, we used the KCFS-YOLOv5 algorithm to detect the object on RSOD dataset and the UCAS-AOD-CAR dataset, respectively, and the results of comparative experiments on the RSOD dataset are shown in Table 11.

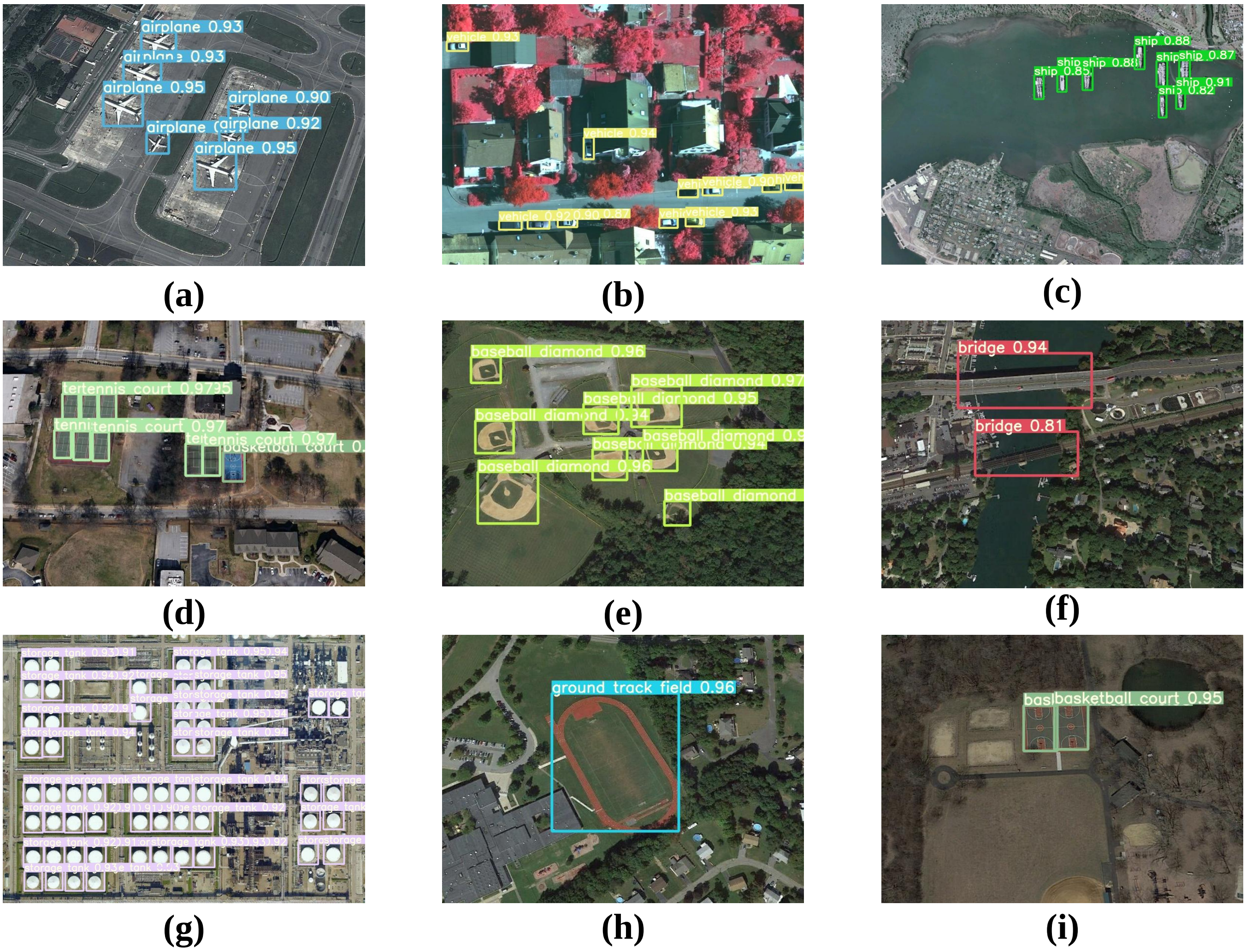

Figure 13. There are some detection results of the KCFS-YOLOv5 on the NWPU VHR-10 dataset. (**a**) is the airplane, (**b**) is the vehicle, (**c**) is the ship, (**d**) is the tennis court, (**e**) is the baseball diamond, (**f**) is the bridge, (**g**) is the storage tank, (**h**) is the ground track field, and (**i**) is the basketball court.

Table 11. Comparison of the KCFS-YOLOv5 with other advanced algorithms on the RSOD dataset.

Category	Faster-RCNN/%	YOLOv4/%	YOLOv5/%	RSyoloX [31]/%	FCOS [52]/%	DA^2FNet [50]/%	Ours/%
Aircraft	90.2	88.8	92.6	-	91.0	95.7	97.4
Oil tank	88.3	89.6	94.0	-	97.6	97.7	98.6
Playground	72.1	76.4	81.2	-	86.1	90.6	85.6
Overpass	80.9	91.5	94.9	-	99.7	95.1	99.5
Mean AP	82.9	86.6	90.7	93.1	93.6	94.8	95.3

- These data are not presented in citation [31].

Based on the above data, the improved object detection method improves the detection accuracy of each object in the RSOD dataset. Compared with the current advanced object detection method, it also has the highest detection accuracy, and can be competent for multi-size object detection tasks with excellent practicability. The detection results are illustrated in Figure 14.

The KCFS-YOLOv5 also has excellent detection performance on the UCAS-AOD-CAR dataset. Figure 15 shows its detection results on the UCAS-AOD-CAR dataset.

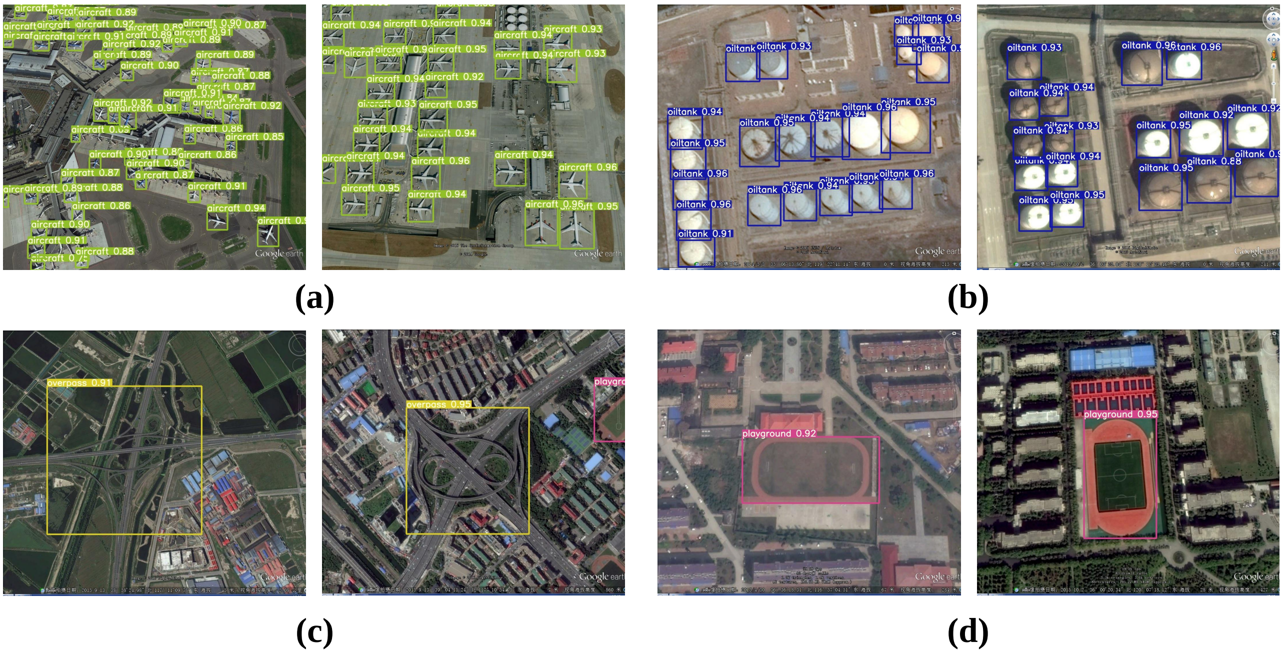

Figure 14. This is the detection result of the KCFS-YOLOv5 on the RSOD dataset. Where, (**a**) is the aircraft, (**b**) is the oiltank, (**c**) is the overpass, (**d**) is the playground.

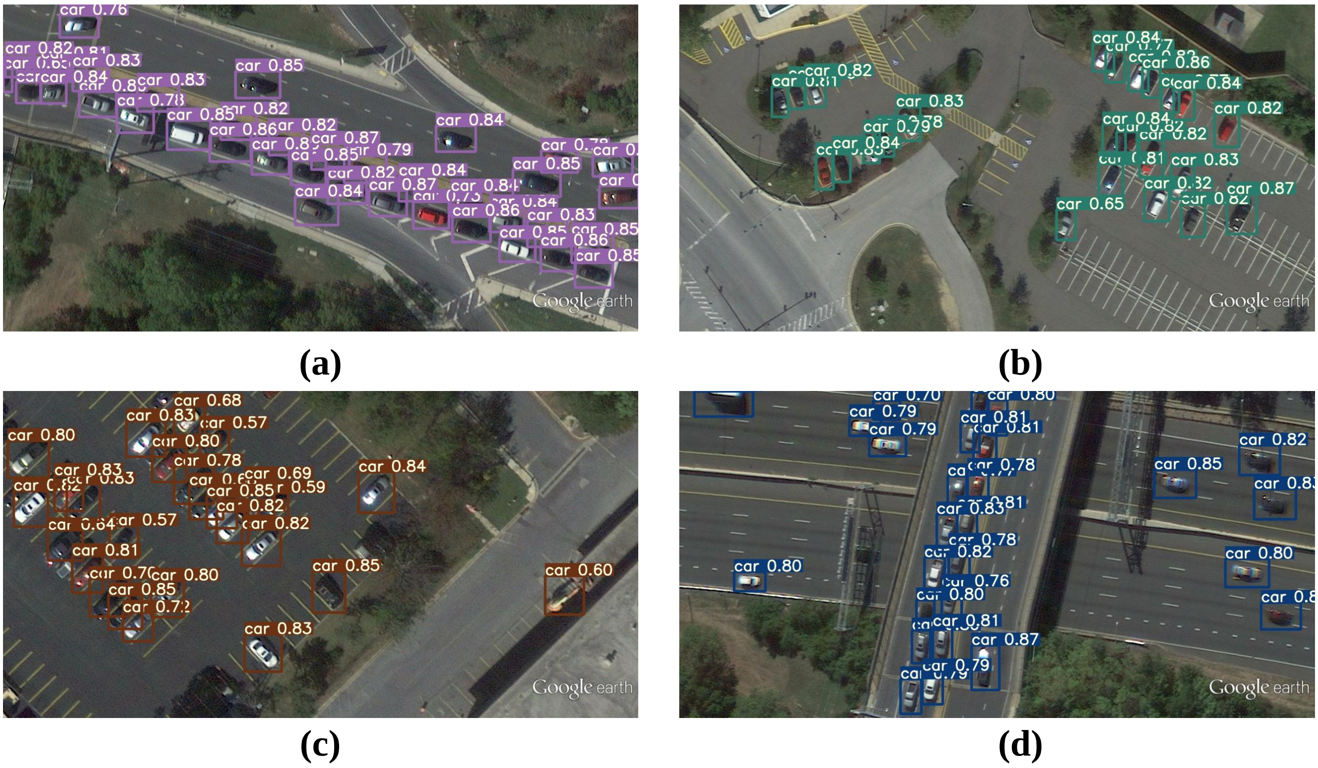

Figure 15. This is the detection result of the KCFS-YOLOv5 on the UCAS-AOD-CAR dataset. (**a**–**d**) are the vehicle.

In the end, we used some advanced remote sensing image detection algorithms to test these three datasets, and we put the simulation datasets in Table 12.

Table 12. Comparison of the KCFS-YOLOv5 with the common algorithms on all three datasets.

Dataset	Evaluation Metrics	Faster-RCNN	YOLOv3	YOLOv4	YOLOv5	Ours
NWPU VHR-10	*Map* (%)	81.39	82.29	86.31	91.08	94.21
	Precision (%)	90.23	88.12	89.97	91.08	94.20
	Recall (%)	85.36	84.43	86.68	86.26	89.42
	parameters (*M*)	60.684	61.949	62.725	7.072	7.341
	Latency (*ms*)	139.3	24.8	23.3	18.4	22.8
	FPS	7.2	40.3	42.9	54.3	43.9
RSOD	*Map* (%)	82.88	84.73	86.62	90.67	95.27
	Precision (%)	86.46	87.14	87.68	90.59	96.11
	Recall (%)	85.41	86.18	86.59	89.56	93.05
	parameters (*M*)	60.684	61.949	62.725	7.072	7.341
	Latency (*ms*)	126.4	23.7	22.9	17.7	22.6
	FPS	7.9	42.2	43.7	56.5	44.2
UCAS-AOD-CAR	*Map* (%)	84.54	81.66	85.03	86.48	90.13
	Precision (%)	84.75	83.47	84.15	85.30	89.80
	Recall (%)	82.81	83.38	83.11	83.36	86.51
	parameters (*M*)	60.684	61.949	62.725	7.072	7.341
	Latency (*ms*)	114.8	23.1	22.3	16.5	21.4
	FPS	8.7	43.3	44.8	60.6	46.7

Compared with other advanced object detection algorithms, our KCFS-YOLOv5 object detection algorithm achieves the best detection results on all three datasets which demonstrates excellent detection accuracy.

6. Discussion

Based on the above research, in Section 6, we will discuss the connection between previous research and our KCFS-YOLOv5 algorithm, and demonstrate the contribution of each improvement point and the future plan of the algorithm based on the experimental data obtained in Section 5.

6.1. About the K-Means++ Algorithm

The detection anchor framework will have a favorable effect on detection results. In the conventional YOLOv5 algorithm, the initial anchor boxes are designed and generated using the K-means clustering algorithm which randomly selects the clustering center points. We use a K-means++ algorithm for optimization. The optimized results are shown in Table 1. Compared with the conventional algorithm, KCFS-YOLOv5 has improved in each evaluation indicator on the three datasets. Although this method can improve the detection ability, it will reduce the detection speed. In the future, we will try to simplify this model.

6.2. About the Backbone Network

The backbone network extracts the texture information and position information in the feature map. To further enhance the feature extraction ability of the backbone network, many studies have been investigated to address these tricky challenges. Li et al. [28] integrated the SE attention mechanism module into YOLOv5 network to achieve the accuracy improvement of the high detection accuracy for low-altitude UAV objects. Yang et al. [31] embedded ECA attention mechanism into YOLOX [17] network, proposed RS-YOLOX, and realized the high-precision detection of satellite remote sensing images. Based on the above research, we decided to embed CA attention mechanism module into the YOLOv5 backbone network to strengthen the feature extraction capacity of the network.

We compared the detection performance with some advanced attention mechanisms (the results are shown in Tables 2 and 3). The CA attention mechanism module achieves the excellent detection results on the test dataset. The detection accuracy of YOLOv5 embedded with CA attention mechanism module is greatly improved compared with the conventional algorithm on three datasets. Therefore, we decided to embed the CA attention mechanism in the improved YOLOv5 backbone network.

Although this method can improve the detection accuracy, it will reduce the detection speed. In the future, we will try to simplify this model. Many scholars have performed valuable research in this field. For example, Yi et al. [53] added Mobilenet to the YOLO network to lightweight the detection network. However, this improved method will lead to the reduction in detection accuracy. Therefore, we do not adopt this method.

6.3. About the Neck Network

The function of the neck network is the information fusion of feature maps with different dimensions. Many studies have been investigated to improve the information integration capabilities of the neck network. Li et al. [20] embedded feature pyramid network (FPN) into the Faster-RCNN to improve the neck network's ability of different dimensions. Li et al. [54] embed BiFPN module into the proposed AB-DLM detection method to implement the driver distraction behavior detection and obtain excellent detection accuracy.

We embed BiFPN in the improved YOLOv5 based on its excellent performance. The detection results on the three datasets are shown in Table 4. Furthermore, the detection accuracy of the YOLOv5 detection method embedded in BIFPN has been significantly improved on three datasets.

On the other hand, the abundant tiny objects in aerial remote sensing images make us have to improve the framework for the detection results of tiny objects. In order to enhance the detection capacity of tiny objects, we added the new tiny object detection head to the original YOLOv5 network, and tested the improved YOLOv5 on three datasets. The results in Table 5 show that the improved YOLOv5 network has higher detection accuracy.

The improved YOLOv5 proposed by Luo et al. [21] integrated PAN and FPN networks to enhance feature fusion ability and achieve better detection accuracy. Therefore, we also fused the BiFPN and the new tiny object detection head to obtain the new improved neck network. We tested the new improved neck network on three datasets. It can be seen from Table 6, the new improved neck network achieves the best detection accuracy, which is better than using the BiFPN or the new tiny object detection head alone. Therefore, we decided to use the new improved neck network.

In the future, we will continue to explore the framework of the neck network to improve the feature fusion capability.

6.4. About the Loss Function

The conventional YOLOv5 uses the GIoU loss as the positioning loss function. The GIoU loss and its improved version ignore the vector angle between the ground truth box and the prediction box, which affects the detection accuracy in the detection task. The SIoU [44] is an improved loss function based on the CIoU function. It considers the vector angle between ground truth box and the predicted box, providing regression direction guidance for predicted box, and redefining the penalty function. Therefore, we used the SIOU loss function on the three datasets and obtained the best detection results (the results are shown in Table 8).

In the future, we will continue to embed other advanced loss functions in YOLOv5 to improve the detection performance of the algorithm.

7. Conclusions

We aimed at the characteristics of object detection methods in aerial remote sensing images, such as intensive distribution, complex background, and poor generality. We carried out a series of improvements based on YOLOv5, tested the improved model on three datasets, and achieved excellent detection results. Finally, we proposed an advanced high-precision aerial object detection method: KCFS-YOLOv5.

To achieve excellent detection performance, we first used the K-means++ clustering algorithm to optimize the center of the anchor, which improves the detection accuracy of YOLOv5. Second, the CA attention mechanism module is embedded to strengthen feature extraction ability, retain more semantic information and reduce the interference of complex backgrounds. Third, the neck network is optimized by adding Bidirectional Feature Pyramid Network (BiFPN) and new tiny object detection head, which can enhance the feature fusion ability in different dimensions of the neck feature fusion network to improve the detection accuracy. Finally, the loss function is replaced with SIoU Loss to improve the fitting degree of the predicted box and the ground truth box, and reduce the influence of the deviation thereon.

Although our KCFS-YOLOv5 object detection method realizes high-precision detection in aerial remote sensing images, compared with the conventional YOLOv5 method, it has slightly more calculation parameters. In future research, we will try to embed more lightweight modules and residual frameworks into the KCFS-YOLOv5 object detection method, so as to reduce the network scale and improve its detection accuracy.

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

Funding: This research received no external funding.

Data Availability Statement: The three datasets that support this study are openly available online. RSOD: [22], NWPU VHR-10: [23], UCAS-AOD-CAR: [24].

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive suggestions, which comprehensively improved the quality of the paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

D-CNN	Deep Convolutional Neural Network
R-CNN	Region Convolutional Neural Network
Fast R-CNN	Fast Region-Based Convolutional Neural Network
Faster R-CNN	Faster Region-Based Convolutional Neural Network
SSD	Single Shot MultiBox Detector
YOLO	You Only Look Once
CSP	Cross Stage Partial
FPN	Feature Pyramid Network
PAN	Path Aggregation Network
NMS	Non Maximum Suppression
CA	Coordinate Attention
BiFPN	Bidirectional Feature Pyramid Network
SE	Squeeze and Excitation
NAM	Normalization-based Attention Module
ECA	Efficient Channel Attention
DIoU	Distance-IoU
CIoU	Complete IoU
GIoU	Generalized Intersection over Union
SIoU	Convolutional Block Attention Module
UAV	Unmanned Aerial Vehicle
SAHI	Slicing Aided Hyper Inference
STN	Spatial Transformer Networks

References

1. Zhao, D.; Xie, D.; Yin, F.; Liu, L.; Feng, J.; Ashraf, T.T.M. Estimation of Pb Content Using Reflectance Spectroscopy in Farmland Soil near Metal Mines, Central China. *Remote Sens.* **2022**, *14*, 2420. [CrossRef]
2. Chen, Z.; Su, R.; Wang, Y.; Chen, G.; Wang, Z.; Yin, P.; Wang, J. Automatic Estimation of Apple Orchard Blooming Levels Using the Improved YOLOv5. *Agronomy* **2022**, *12*, 2438. [CrossRef]
3. Wahyudi Sumari, A.D.; Pranata, A.S.; Mashudi, I.A.; Syamsiana, I.N.; Sereati, C.O. Automatic Target Recognition and Identification for Military Ground-to-Air Observation Tasks using Support Vector Machine and Information Fusion. In Proceedings of the 2022 International Conference on ICT for Smart Society (ICISS), Virtual, 10–11 August 2022; pp. 1–8. [CrossRef]
4. wei Wang, Z.; Li, X.; Mao, Y.; Li, L.; Wang, X.; Lin, Q. Dynamic simulation of land use change and assessment of carbon storage based on climate change scenarios at the city level: A case study of Bortala, China. *Ecol. Indic.* **2022**, *134*, 108499. [CrossRef]
5. Liu, Y.; Yao, X.; Gu, Z.; Zhou, Z.; Liu, X.S.; Chen, X.; Wei, S. Study of the Automatic Recognition of Landslides by Using InSAR Images and the Improved Mask R-CNN Model in the Eastern Tibet Plateau. *Remote Sens.* **2022**, *14*, 3362. [CrossRef]
6. Meng, J.; Yan, J.; Zhao, J. Bubble Plume Target Detection Method of Multibeam Water Column Images Based on Bags of Visual Word Features. *Remote Sens.* **2022**, *14*, 3296. [CrossRef]
7. Jin, S.; Li, X.; Yang, X.; Zhang, J.A.; Shen, D. Identification of Tropical Cyclone Centers in SAR Imagery Based on Template Matching and Particle Swarm Optimization Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 598–608. [CrossRef]
8. Jian, S.; Jiang, J.; Lu, K.; Zhang, Y. SEU-tolerant Restricted Boltzmann Machine learning on DSP-based fault detection. In Proceedings of the 2014 12th International Conference on Signal Processing (ICSP), Hangzhou, China, 19–23 October 2014; pp. 1503–1506. [CrossRef]
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
10. 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. [CrossRef]
11. 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]
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
13. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
14. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
16. Kasper-Eulaers, M.; Hahn, N.; Berger, S.; Sebulonsen, T.; Myrland, Ø.; Kummervold, P.E. Short Communication: Detecting Heavy Goods Vehicles in Rest Areas in Winter Conditions Using YOLOv5. *Algorithms* **2021**, *14*, 114. [CrossRef]
17. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
18. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the ECCV, Amsterdam, The Netherlands, 11–14 October 2016.
19. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD : Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
20. Yan, D.; Li, G.; Li, X.; Zhang, H.; Lei, H.; Lu, K.; Cheng, M.; Zhu, F. An Improved Faster R-CNN Method to Detect Tailings Ponds from High-Resolution Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2052. [CrossRef]
21. Luo, S.; Yu, J.; Xi, Y.; Liao, X. Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv5. *IEEE Access* **2022**, *10*, 5184–5192. [CrossRef]
22. Long, Y.G. RSOD Dataset. [EB/OL]. Available online: https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset- (accessed on 1 November 2022).
23. Cheng, G.; Han, J. NWPU VHR-10 Dataset. [EB/OL]. Available online: https://github.com/chaozhong2010/VHR-10_dataset_coco (accessed on 1 November 2022).
24. UCAS-AOD Dataset. [EB/OL]. Available online: https://hyper.ai/datasets/5419 (accessed on 1 November 2022).
25. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
26. Da, Y.; Gao, X.; Li, M. Remote Sensing Image Ship Detection Based on Improved YOLOv3. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022; pp. 1776–1781. [CrossRef]
27. Cao, C.; Wu, J.; Zeng, X.; Feng, Z.; Wang, T.; Yan, X.; Wu, Z.; Wu, Q.; Huang, Z. Research on Airplane and Ship Detection of Aerial Remote Sensing Images Based on Convolutional Neural Network. *Sensors* **2020**, *20*, 4696. [CrossRef]
28. Li, Z.; Namiki, A.; Suzuki, S.; Wang, Q.; Zhang, T.; Wang, W. Application of Low-Altitude UAV Remote Sensing Image Object Detection Based on Improved YOLOv5. *Appl. Sci.* **2022**, *12*, 5784. [CrossRef]

29. Wang, Z.; Lu, H.; Jin, J.; Hu, K. Human Action Recognition Based on Improved Two-Stream Convolution Network. *Appl. Sci.* **2022**, *37*, 5784. [CrossRef]
30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
31. Yang, L.; Yuan, G.; Zhou, H.; Liu, H.; Chen, J.; Wu, H. RS-YOLOX: A High-Precision Detector for Object Detection in Satellite Remote Sensing Images. *Appl. Sci.* **2022**, *12*, 8707. [CrossRef]
32. Wang, Q.; Wu, B.; Zhu, P.F.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
33. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. *arXiv* **2022**, arXiv:2202.06934.
34. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the NIPS, Montreal, QC, Canada, 12 December 2015.
35. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
36. Li, Y.; Sun, S.; Zhang, C.; Yang, G.; Ye, Q. One-Stage Disease Detection Method for Maize Leaf Based on Multi-Scale Feature Fusion. *Appl. Sci.* **2022**, *12*, 7960. [CrossRef]
37. Lin, T.Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
38. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J.W. PanNet: A Deep Network Architecture for Pan-Sharpening. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1753–1761.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
41. Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding. In Proceedings of the SODA '07, New Orleans, LA, USA, 7–9 January 2007.
42. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
43. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
44. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
45. Rezatofighi, S.H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.D.; 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; pp. 658–666.
46. Chen, D.; Miao, D. Control Distance IoU and Control Distance IoU Loss Function for Better Bounding Box Regression. *arXiv* **2021**, arXiv:2103.11696.
47. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based Attention Module. *arXiv* **2021**, arXiv:2111.12419.
48. 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, New York, NY, USA, 7–12 February 2020.
49. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
50. Guo, Y.; Tong, X.; Xu, X.; Liu, S.; Feng, Y.; Xie, H. An Anchor-Free Network With Density Map and Attention Mechanism for Multiscale Object Detection in Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
51. Shi, L.; Kuang, L.; Xu, X.; Pan, B.; Shi, Z. CANet: Centerness-Aware Network for Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
52. Lu, X.; Wang, W.; Shen, J.; Crandall, D.J.; Van Gool, L. Segmenting Objects From Relational Visual Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7885–7897. [CrossRef]
53. Li, X.; Qin, Y.; Wang, F.; Guo, F.; Yeow, J.T.W. Pitaya detection in orchards using the MobileNet-YOLO model. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 6274–6278.
54. Li, T.; Zhang, Y.; Li, Q.; Zhang, T. AB-DLM: An Improved Deep Learning Model Based on Attention Mechanism and BiFPN for Driver Distraction Behavior Detection. *IEEE Access* **2022**, *10*, 83138–83151. [CrossRef]

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.