Improved YOLOv5 Based on Hybrid Domain Attention for Small Object Detection in Optical Remote Sensing Images

Tianmin Deng *, Xuhui Liu and Guotao Mao *

School of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China
* Correspondence: dtianmin@cqjtu.edu.cn (T.D.); mguotao@sina.com (G.M.)

Abstract: The object detection technology of optical remote sensing images has been widely applied in military investigation, traffic planning, and environmental monitoring, among others. In this paper, a method is proposed for solving the problem of small object detection in optical remote sensing images. In the proposed method, the hybrid domain attention units (HDAUs) of channel and spatial attention mechanisms are combined and employed to improve the feature extraction capability and suppress background noise. In addition, we designed a multiscale dynamic weighted feature fusion network (MDW-Net) to improve adaptive optimization and deep fusion of shallow and deep feature layers. The model is trained and tested on the DIOR dataset, and some ablation and comparative experiments are carried out. The experimental results show that the mAP of the proposed model surpasses that of YOLOv5 by a large margin of +2.3 and has obvious advantages regarding the detection performance for small object categories, such as airplane, ship, and vehicle, which support its application for small target detection in optical remote sensing images.

Keywords: hybrid domain attention unit; small object detection; remote sensing images; weighted feature fusion

1. Introduction

In recent years, with the development of remote sensing technology, high-quality remote sensing images are increasing in abundance, which lays a foundation for the application of remote sensing. Remote sensing images are widely used in military investigation, traffic planning, environmental monitoring, and other fields, which is of great significance for social and economic development. In target detection, as one of the applications of remote sensing image processing, the location of targets according to specific categories is obtained and mapped, and common targets include aircraft, airports, ships, bridges, and automobiles, so it has very important applications in military and civil fields. In the military field, the detection and acquisition of these types of information is conducive to quickly and accurately locking the location of attack targets, analyzing the war situation, and formulating military operations. In the civil field, the positioning of ships is conducive to maritime rescue operations, and the positioning of vehicles is conducive to vehicle counting and analysis of road congestion. Therefore, the accurate detection of targets in remote sensing images is very important.

Traditional remote sensing image object detection methods mostly rely on substantial prior knowledge to design manual feature extractors. It is difficult for this kind of method to meet the requirements of remote sensing image object detection because of cumbersome manual feature design, poor robustness, and computational redundancy. With the rapid development of convolutional neural networks (CNNs) in the field of computer vision [1], object detection methods based on convolutional neural networks have made a new breakthrough in the field of remote sensing image object detection by virtue of their strong adaptive learning ability and feature extraction ability. Therefore, an increasing number of excellent works are being reported in remote sensing image object detection [2–6]. An
attention mechanism has been introduced for small object detection in remote sensing images in numerous reports, and satisfactory performance has been achieved. Attention mechanisms can be generally divided according to different attention domains as channel domain attention mechanisms, spatial domain attention mechanisms, and hybrid domain attention mechanisms. The channel attention mechanism focuses on the long-range dependencies among the channels of the feature map. The spatial attention mechanism focuses on the pixel regions that play a decisive role in classification in the feature map. The hybrid domain attention mechanism utilizes both spatial domain and channel domain information, and each element in the feature map of each channel has a corresponding attention weight. These plug-and-play attention models can be seamlessly integrated into various deep learning networks to guide object detection tasks.

The motivation, novelty and contributions of this paper are as follows:

- There are still some problems in remote sensing image target detection. Because the remote sensing images are taken from a top view, there are a large number of small-scale objects, as shown in Figure 1. Within a large range of spatial resolutions, irrelevant background targets are inevitable. Because the objects in remote sensing images have small-scale and dense distribution, achieving ideal detection is difficult when the object detection method is applied to general scenes in remote sensing images. In addition, due to the complex and chaotic background noise interference in remote sensing images, traditional methods have difficulty effectively extracting the features of small-scale objects in these images. To cope with the above challenges, we designed a simple and efficient small object detection network to accurately capture small-scale objects in remote sensing images.

- We develop a hybrid domain attention unit (HDAU) combining channel and spatial attention, which is innovatively proposed for small object detection when complex background noise is present, and the method adaptively captures the important feature information of the feature space from the two dimensions of channel and spatial to suppress interference feature information. By embedding it into the backbone network, the network can extract the key information of the task object area more directionally. Furthermore, combined with the multiscale feature fusion idea [7], a multiscale dynamic weighted feature fusion network (MDW-Net) is proposed, which realizes the efficient fusion of deep semantic information and shallow detail information by dynamically adjusting the importance distribution weight of each feature layer.

- In this paper, a method is proposed for solving the problem of small object detection in optical remote sensing images. In the proposed method, the hybrid domain attention units (HDAUs) of channel and spatial attention mechanisms are combined and employed to improve the feature extraction capability and suppress background noise. In addition, we designed a multiscale dynamic weighted feature fusion network (MDW-Net) to improve adaptive optimization and deep fusion of shallow and deep feature layers. The proposed method has obvious advantages in the detection performance of small target categories.

The remainder of this paper is organized as follows. Section 3 presents the small object detection method based on hybrid domain attention and a multiscale dynamic feature fusion network. Section 4 details the experiments and results for applying our method on the DIOR dataset, and the results are discussed and analyzed. The conclusion is presented in Section 5.
2. Related Works

2.1. Region-Based Object Detection of Remote Sensing Images

At present, remote sensing image object detection methods based on deep learning can be divided into two categories according to whether region proposals are needed. Regression-based object detection of remote sensing images classifies and regresses the preset candidate object regions by extracting the feature information of several candidate regions and then obtains the location and classification of objects, among which the more typical ones are Faster R-CNN [8], Mask R-CNN [9], and Cascade R-CNN [10]. The methods proposed by many scholars in recent years for small object detection methods in remote sensing images are based on such methods. For example, Dong et al. [11] proposed replacing the traditional non-maximum suppression (NMS) in the region proposal network (RPN) with Sig-NMS, which effectively improved the problem of missed detections of small objects. Zheng et al. [12] proposed a multiscale attention feature pyramid network (MAFPN), which uses the feature map obtained by the feature pyramid introduced into the attention module as input for the RPN to obtain the final candidate region, which effectively improves the object detection ability of small-scale remote sensing images. However, the computational overhead of the region proposal network needs to be improved. Yao et al. [13] designed a multiresolution feature extraction network (MFE) based on a Faster R-CNN network to reduce the semantic difference among scales by refining the features with different resolutions of each scale, but the method is not ideal for small object detection when there is obvious background noise interference. Li et al. [14] proposed a new small sample target detection method combining local features and a convolutional neural network (LF-CNN) with the aim of detecting small numbers of unevenly distributed ground object targets in remote sensing images. The LF-CNN method can more effectively integrate the local features of ground object samples and improve the accuracy of target identification and detection in small samples of remote sensing images than traditional target detection methods. However, there are shortcomings in terms of time and complexity.
2.2. Regression-Based Object Detection of Remote Sensing Images

Regression-based object detection of remote sensing images completes end-to-end object detection without region proposal and directly locates the object and predicts the category through the initial anchor box, among which the more typical examples are You Only Look Once (YOLO) series methods [15], Single-Shot MultiBox Detector (SSD) [16], and RetinaNet [17]. To improve real-time object detection, a regression-based object detection method is employed in remote sensing images. For example, Courtrai et al. [18] proposed a super-resolution method based on adaptive learning and used the residual module integrated in the Wasserstein generative adversarial network (WGAN) for small object detection in remote sensing images. In the training process, they focused the method on the object to be detected and verified the effectiveness of small object detection based on YOLOv3. Yin et al. [19] proposed a novel encoding–decoding module to accurately detect small-scale objects. The module effectively combines shallow and deep features through upsampled and summation operations, which greatly enriches the semantic information of small objects in remote sensing images. Aiming at the problem of insufficient small object detection performance of regression-based object detection methods, Yu et al. [20] proposed a two-way convolution network (TWC-Net), which effectively avoids the loss of small object feature information in the feature extraction process. Li et al. [21] designed a lightweight multiscale aggregated model based on YOLOv3 for remote sensing image object detection, which combined the channel attention module and the spatial pyramid pooling module to extract features of different scales for fusion, which effectively improved the detection of small objects, but was poorly effective in detecting dim and small objects with unclear features. Based on the YOLOv3-tiny network, Nong et al. [22] proposed a lightweight, embedded, remote sensing image object real-time detection method by optimizing the multiscale prediction layers and introducing the spatial attention mechanism, in which the number of parameters was effectively reduced. However, due to the lack of shallow semantic features, the detection of small objects in remote sensing images is poor. To solve the problem of detecting small objects and high-resolution objects in remote sensing images, Liu et al. [23] proposed an improved YOLOv5 method based on a needed residual transformer (NRT-YOLO). NRT-YOLO adds an extra prediction head and adopts a novel nested residual transformer module and a nested residual attention module. It demonstrates adequate ability in detecting small sample objects. Fang et al. [24] proposed an improved S$^2$ANET-SR model based on S$^2$A-NET network to solve the problem of detection difficulty in cases where there are many small-scale objects in remote sensing images. Firstly, the original image and the restored image are sent to the detection network at the same time, and a super-resolution enhancement module is then designed for the restored image to enhance the feature extraction of small objects.

In this paper, we present our proposed method for small target detection method to solve the problem of small target detection in optical remote sensing images and compare this with the state-of-the-art approaches above. In our method, the hybrid domain attention unit and spatial attention mechanism of the channel are combined to improve the ability of feature extraction and suppress background noise. In addition, a multiscale dynamic weighted feature fusion network (MDW network) is designed to improve the adaptive optimization and deep fusion of shallow and deep feature layers. We demonstrate that this method is effective in the detection of small target categories in remote sensing images. The comparisons of related works are shown in Table 1.
Table 1. The related work comparisons.

| Paper            | Technique                  | Work Goals                                                                 |
|------------------|----------------------------|-----------------------------------------------------------------------------|
| 2019 Dong et al. | Faster R-CNN and sig NMS  | Improve the problem of missing detection of small targets.                   |
| 2021 Zheng et al.| FAGNet                    | Improve object detection capabilities for small-scale remote sensing images. |
| 2021 Yao et al.  | MFE-Net                    | Improve the performance of remote sensing image target detection.            |
| 2022 Li et al.   | LF-CNN                     | Improve the accuracy of object detection in small sample remote sensing images. |
| 2020 Courtrai et al. | Wasserstein GAN       | Improve the accuracy of small target detection in remote sensing images.     |
| 2020 Yin et al.  | AF-SSD                     | Improve the accuracy of small-scale target detection in remote sensing images. |
| 2021 Yu et al.   | TWC-Net                    | Solve the problem of sparsity of image objects and complexity of interference.|
| 2021 Li et al.   | DSYOLOv3                   | Improve the detection effect of small objects in aerial images.             |
| 2021 Nong et al. | Embedded-based Method      | Reduce the number of parameters.                                            |
| 2022 Liu et al.  | NRT-YOLO                   | Solve the problem of detecting small and high resolution objects in remote sensing images. |
| 2022 Fang et al. | S^2ANET-SR                 | Solve the problem that many small-scale objects in remote sensing images are difficult to detect. |

This paper HDAU-YOLO Improve the accuracy of small target detection in remote sensing images.

3. Proposed Improved YOLOv5 Network for Small Object Detection

3.1. Proposed Network Architecture

The core idea of a HDAU-YOLO is to fully capture key feature information beneficial to small object detection on the premise of ensuring the real-time detection performance of the object detector to the greatest possible extent. The performance of the baseline network (YOLOv5) in the small object detection task of remote sensing images is improved by a HDAU and a MDW-Net. The network structure of a HDAU-YOLO is shown in Figure 2. In the figure, C1–C5 represent the convolutional features of each stage of the feature extraction network; P3–P5 represent the convolutional features obtained from up–bottom feature fusion layer; and N3–N5 represent the convolutional features obtained from the bottom–up feature fusion layer.

By assigning different attention weights to different feature channels and spatial regions of input features, a HDAU screens out the more important information for the remote sensing image object detection task from a large amount of feature information. The MDW-Net can further capture the semantic information of the deep features and the fine-grained information of the shallow features in the feature pyramid and then combine the rich context information to strengthen the expression ability of the object detector.

3.2. Hybrid Domain Attention Unit for Feature Extraction

The HDAU is composed of a coordinate attention module (CAM) [25] and a spatial attention module (SAM) [26]. On the one hand, the long-range dependencies among channels are obtained along the channel dimension. On the other hand, the contextual information of feature maps is further mined by emphasizing the region of interest of the spatial dimension. The algorithm flow of a HDAU is shown in Figure 3. The input feature
space \( X \) is weighted by the attention in the channel domain through the CAM as the input feature \( X' \) of the SAM, and \( Y \) is then further weighted by the attention weight map of the spatial domain generated by the SAM to obtain the final output feature space \( Y \).

Figure 2. Architecture of the proposed HDAU-YOLO.

The network structure of the CAM is shown in Figure 4. Unlike the squeeze-and-excitation block (SEB) [27], which only focuses on the long-range dependencies among channels, the CAM also considers the long-range dependencies among spatial locations. The CAM aims to capture the long-range dependencies among input and output feature channels by generating attention weight maps with horizontal and vertical directions that maintain accurate spatial location information.

Figure 3. Algorithm flow chart of a HDAU.
Suppose that the input feature space of the CAM is $X = [x_1, x_2, \cdots, x_c] \in \mathbb{R}^{C \times H \times W}$, where $C$ is the number of channels of the input feature, $H \times W$ is the scale size of the input feature, and $x_c \in \mathbb{R}^{H \times W}$ is the c-th channel of the input feature space. The CAM includes two parts: coordinate information embedding and coordinate attention generation. First, the 1D horizontal global pooling layer is used to carry out an average pooling operation on the input space feature $X$ to obtain the c-th horizontal direction feature $z^h_c$ [25] with height $h$ in the global feature space $Z \in \mathbb{R}^{C \times H \times W}$:

$$z^h_c = F_{\text{ie}}(h) = \frac{1}{W} \sum_{i=1}^{W} x_c(h, i) \quad (1)$$

where $F_{\text{ie}}(\cdot)$ is the mapping function of coordinate information embedding. Similarly, the 1D vertical global pooling layer is used to carry out an average pooling operation on the input space feature $X$ to obtain the c-th vertical direction feature $z^w_c$ [25] with width $w$ in the global feature space $Z \in \mathbb{R}^{C \times H \times W}$:

$$z^w_c = F_{\text{ie}}(w) = \frac{1}{H} \sum_{j=1}^{H} x_c(j, w) \quad (2)$$

Furthermore, coordinate attention generation is used to reduce computational overhead and obtain an efficient adaptive learning attention map. First, the feature maps of the horizontal and vertical directions are concatenated in the spatial dimension. Subsequently, a $1 \times 1$ convolution kernel is used for the convolution operation $F_{\text{conv}}$ to achieve the effect of dimensionality reduction, and a batch normalization (BN) layer and ReLU activation function are used to encode $\delta$ features in the horizontal and vertical directions to generate the intermediate feature maps $F \in \mathbb{R}^{C/r \times (H+W)}$ [25] in the two directions, where $r$ is the proportion of downsampled features:

$$F = \delta(F_{\text{conv}}((Z^h, Z^w))) \quad (3)$$

After that, $F$ is split into two feature maps $F^h \in \mathbb{R}^{C/r \times H}$ and $F^w \in \mathbb{R}^{C/r \times W}$ along the spatial direction and then restored to the original input dimension through two $1 \times 1$ convolution operations $F_{\text{convh}}$ and $F_{\text{convw}}$, respectively. Finally, the horizontal direction attention map $M^h$ [25] and vertical direction attention map $M^w$ [25] can be obtained by the normalization operation of the sigmoid activation function $\sigma$:

$$M^h = \sigma(F_{\text{convh}}(F^h)) \quad (4)$$

$$M^w = \sigma(F_{\text{convw}}(F^w)) \quad (5)$$
Finally, the horizontal and vertical attention weight maps were multiplied by the original input feature space $X$ to obtain the output feature space $X' = [x'_1, x'_2, \ldots, x'_c] \in \mathbb{R}^{C \times H \times W}$:

$$x'_c(i, j) = x_c(i, j) \times g^h_c(i) \times g^v_c(j)$$  \hspace{1cm} (6)

where $x'_c(i, j)$ [25] is the $c$-th convolution kernel of the output feature space $X'$, and $g^h_c(i)$ and $g^v_c(j)$ denote the weight of the $c$-th channel in the horizontal direction and vertical direction, respectively.

The network structure of the SAM is shown in Figure 5. The SAM aims to use the feature information of the input feature space to generate a spatial attention weight map and weight the input features with spatial attention to enhance the feature expression of the region of interest.

![Figure 5. Architecture of the SAM.](image)

The input feature space of the SAM is the output feature space of the CAM $X' = [x'_1, x'_2, \ldots, x'_c] \in \mathbb{R}^{C \times H \times W}$, and two feature maps $X'_{\text{avg}} \in \mathbb{R}^{1 \times H \times W}$ and $X'_{\text{max}} \in \mathbb{R}^{1 \times H \times W}$ are obtained after compression by global max pooling and global average pooling along channel dimensions, respectively. After concatenating the two feature maps, the convolution operation $F^7_{\text{conv}}$ was carried out by a $7 \times 7$ convolution kernel with a large receptive field and then normalized by the sigmoid activation function $\sigma$ to obtain the spatial attention weight map $S \in \mathbb{R}^{1 \times H \times W}$ [26]:

$$S = \sigma(F^7_{\text{conv}}([X'_{\text{avg}}, X'_{\text{max}}]))$$  \hspace{1cm} (7)

Finally, the spatial attention weight map $S$ is weighted with the input feature space $X'$ to obtain the final output feature space of a HDAU $Y = [y_1, y_2, \ldots, y_c] \in \mathbb{R}^{C \times H \times W}$:

$$Y = [y_1, y_2, \ldots, y_c] \in \mathbb{R}^{C \times H \times W}$$  \hspace{1cm} (8)

where $F_{\text{scale}}$ is the mapping function for weighting spatial attention and $\otimes$ is the operation symbol of weighted multiplication of the feature map.

The SAM gives the same spatial attention weighting to each channel in the input feature space, so it ignores the information interaction among channels. The CAM weights the information in the channel globally, ignoring the information interaction in spatial regions. Therefore, in this paper, the CAM and the SAM are connected in a cascade way, aiming to capture key information inside the input feature space along the two dimensions of channel and spatial from the perspective of global feature information and then screen out the important information related to the remote sensing image object detection task to weaken the interference from irrelevant information.
3.3. Multiscale Dynamic Weighted Network for Feature Fusion

In the feature extraction stage, the network scales the feature image through the lower sampling layer to enlarge the receptive field. As shown in Figure 6, each downsampled operation compresses the resolution of the feature image, resulting in the loss of fine-grained feature information of the object. A shallow network can extract the texture and edge features of objects with a more detailed content description. The deep network extracts the rich semantic features of the object but, at the same time, weakens the perception of the location information and detail information of the small object such that the feature information of the small object in the feature map is lost.

As shown in Figure 7, the three methods generate a new feature map \( N_{i+1} \) by laterally connecting a higher resolution feature map \( N_i \) and a coarser feature map \( P_{i+1} \). Each feature map \( N_i \) of PANet first goes through the convolutional layer to reduce the spatial size, and each element of feature map \( P_{i+1} \) and the downsampled map are then added through lateral connection. YOLOv5 concatenates the feature map \( P_{i+1} \) and the downsampled map of \( N_i \) through lateral connection. These two methods integrate feature information of different depths with equal relationships, but the contribution of feature layers of different depths to the small object detection task is actually different, and shallow feature information occupies a more important position in the process of small object detection. To solve the above problems, a multiscale dynamic weighted feature fusion method, MDW-Net, is proposed in this section. By assigning different proportion weights \( w_i \) to different scale feature layers, the contribution of shallow feature information in the feature pyramid is strengthened, and the enhanced shallow fine-grained information is then fused to guide the task of small object detection in remote sensing images.

To enable the network to dynamically learn the importance of different input features, the additional weight value of input features at layer \( i \) in the feature pyramid is defined as \( w_i \):\n
\[
    w_i = \frac{\delta(\lambda_i)}{\mu + \sum_{j=1}^{n} \delta(\lambda_j)}
\]

where \( n \) is the number of input feature maps and \( \lambda \) is the weight that can be learned. In the initial stage, each \( \lambda \) is given an initial value of 1, and the ReLU activation function \( \delta \) ensures that all \( \lambda \geq 0 \). Then, let \( \mu = 0.0001 \) to avoid numerical instability.

After the weight that can be learned dynamically is obtained, each input feature layer \( x_i \) is endowed with the corresponding weight \( w_i \) for weighted fusion, and the ReLU activation function \( \delta \) is used to ensure the scale consistency of the feature data. After the convolution operation \( F_{\text{conv}} \), the feature fusion result \( F_{\text{weight}} \) is obtained:
The MDW-Net fully captures the rich feature information of feature layers at different depths. In the training process, the weight values of feature layers at each scale are dynamically adjusted according to the contribution of feature layers at each scale to small object detection tasks in remote sensing images so that the fused features take into account both rich semantic information and fine-grained information.

\begin{equation}
F_{\text{weight}} = F_{\text{conv}}(\delta(\sum_{i=1}^{n} w_i x_i))
\end{equation}

Figure 7. Comparison of feature fusion methods: (a) PANet; (b) YOLOv5; (c) ours.

3.4. Algorithm

According to the proposed network and improved blocks in each module, the pseudocode of the proposed method is designed, and the process of the one step-by-step training method is illustrated in Algorithm 1.

\begin{algorithm}[H]
\begin{footnotesize}
\textbf{Algorithm 1}: Algorithm pseudocode of applying the proposed HDAU-YOLO model in small object detection.
\begin{enumerate}
\item \textbf{Input}: Dataset samples \( S = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\} \). The \( S \) is classified into a training set (Train \( X \), Train \( Y \)), a validation set (Val \( X \), Val \( Y \)), a testing set (Test \( X \), Test \( Y \)) in a certain proportion (\( X \)—remote sensing image, \( Y \)—the remote sensing image label). The number of learning epochs is \( T \).
\item \textbf{Output}: Well-trained network.
\item \textbf{1.} Load the training set and validation set;
\item \textbf{2.} Augmentation the training set;
\item \textbf{3.} \textbf{Begin:}
\item \textbf{4.} Initialize all weights and biases.
\item \textbf{5.} \textbf{For} \( m = 1, 2, \ldots, T \):
\item \textbf{6.} Extract features through Conv (CSPDarknet);
\item \textbf{7.} Input the feature map to the HDAU;
\item \textbf{8.} Generate the attention weight map through the HDAU;
\item \textbf{9.} Concatenate the input feature map and output attention weight of the HDAU;
\item \textbf{10.} Input the weighted feature map into the MDW-Net;
\item \textbf{11.} Concatenate the feature maps of different scales;
\item \textbf{12.} Generate weight coefficients corresponding to feature maps of different scales;
\item \textbf{13.} Weight the multiscale feature maps, and calculate the output of the MDW-Net;
\item \textbf{14.} Model fit (SGD, (train \( X \), train \( Y \)) \( \rightarrow \) \( M(m) \))
\item \textbf{15.} Model evaluate (\( M(m), (\text{Val } X, \text{Val } Y) \) \( \rightarrow \) \( m\text{AP}(m) \)).
\item \textbf{16.} \textbf{End For}
\item \textbf{17.} Save the optimal model which has max \( m\text{AP} \) in \( T \) epochs.
\item \textbf{18.} \textbf{End}
\item \textbf{19.} Load the testing set;
\item \textbf{20.} Load the optimal model in terms of object detection performances.
\end{enumerate}
\end{footnotesize}
\end{algorithm}
4. Experiment and Discussion

4.1. Experimental Dataset

In this study, the publicly available optical remote sensing image dataset DIOR was adopted for evaluating the presented approach [28]. The DIOR dataset includes airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill objects corresponding to a total of 20 object categories, expressed as C1–C20, respectively. The DIOR dataset contained 23,463 remote sensing images with a size of 800 × 800, and the spatial resolutions ranged from 0.5 to 30 m, including 5862 training set images, 5863 validation set images, and 11,738 test set images and a total of 192,472 object instances. The numbers of object instances in the training set, validation set, and test set are 32,592, 35,437, and 124,443, respectively. Figure 8 shows the distribution of the number of object instances per class in the DIOR training set, validation set, and test set. The number of small-scale objects, such as ships, vehicles, and storage tanks, represents the majority in the DIOR dataset, and the number of object instances of airplanes, vehicles, and storage tanks is much larger in the test set than in the training set. Such a small number of training samples brings certain difficulties to the training process of our model.

![Figure 8. Number of object instances per class in the DIOR dataset.](image-url)

4.2. Implementation Details of the Experiment

In this study, YOLOv5 was used as the baseline network of the experiment, and the specific experimental details are as follows:

- **Experiment platform**
  
  The hardware configuration used in this experiment is an Nvidia RTX3060 GPU and Intel i5-10400 2.90 GHz CPU, and the software environment is the PyTorch deep learning framework in the Windows 10 system.

- **Implementation details**
  
  The stochastic gradient descent (SGD) method was used to optimize the network for training. The initial learning rate was set to 0.01, momentum factor to 0.937, weight decay factor to 0.0005, batch size to 12, and maximum number of epochs to 300.

- **Evaluation metrics**
  
  To evaluate the effectiveness of the proposed method, the model size and number of parameters are selected to evaluate the complexity of the detector, and the average detection time of each image is used to evaluate the speed of the detector. At the same time, mean average precision (mAP) was used as the evaluation index of the model’s comprehensive detection performance for multiple object categories, and average precision (AP) was used to evaluate the model’s detection performance for a single object category. The confidence threshold is set to 0.001. If the confidence score is greater than this threshold, it will be regarded as a positive sample; otherwise, it will be regarded as a negative sample. The intersection over union (IoU) threshold between the ground-truth box and the prediction box.
• Implementation details
The stochastic gradient descent (SGD) method was used to optimize the network for training. The initial learning rate was set to 0.01, momentum factor to 0.937, weight decay factor to 0.0005, batch size to 12, and maximum number of epochs to 300.

• Evaluation metrics
To evaluate the effectiveness of the proposed method, the model size and number of parameters are selected to evaluate the complexity of the detector, and the average detection time of each image is used to evaluate the speed of the detector. At the same time, mean average precision (mAP) was used as the evaluation index of the model’s comprehensive detection performance for multiple object categories, and average precision (AP) was used to evaluate the model’s detection performance for a single object category. The confidence threshold is set to 0.001. If the confidence score is greater than this threshold, it will be regarded as a positive sample; otherwise, it will be regarded as a negative sample. The intersection over union (IoU) threshold between the ground-truth box and the predicted box is set to 0.5. When the IoU value is higher than this threshold, the detection is considered correct.

4.3. Ablation Experiment
To verify the effectiveness of the proposed HDAU and the MDW-Net in remote sensing image object detection tasks, this study conducted a series of ablation studies on the DIOR test set, with the mAP, model size, and number of parameters as evaluation indexes, and the final results are shown in Table 2.

| Method | mAP (%) | Model Size | Parameters | Time [ms] |
|--------|---------|------------|------------|-----------|
| YOLOv5 (baseline) | 73.9 | 357 MB | 46.73 M | 26.3 |
| YOLOv5 + CAM | 74.8 | 373 MB | 48.80 M | 27.8 |
| YOLOv5 + SAM | 74.3 | 364 MB | 47.68 M | 27 |
| YOLOv5 + SAM&CAM * | 75.2 | 380 MB | 49.75 M | 28.6 |
| YOLOv5 + CAM&SAM * | 75.3 | 380 MB | 49.75 M | 28.6 |
| YOLOv5 + CAM//SAM ** | 75.1 | 380 MB | 49.75 M | 28.6 |
| YOLOv5 + MDW-Net | 75.1 | 372 MB | 48.64 M | 27.8 |
| YOLOv5 + CAM&SAM + MDW-Net (ours) | 76.2 | 392 MB | 51.37 M | 29.4 |

Note: * the SAM and the CAM are cascade; ** the SAM and the CAM are parallel.

Compared with the baseline network, the number of parameters for a HDAU-YOLO increased by 4.64 M, the model size increased slightly to 392 MB due to the increase in parameter number, the average detection time of each image increased, and the mAP increased by 2.3% to 76.2%. Further ablation studies showed that the baseline network after CAM embedding obtained 74.8% mAP, an increase of 0.9% mAP compared with the baseline network, while the model size and the number of parameters were 373 MB and 48.80 M, respectively. It can be seen that the CAM is very effective for remote sensing image object detection by capturing the long-range dependencies from the spatial dimension. Similarly, the baseline network after SAM embedding obtained 74.3% mAP, which increased by 0.4% mAP, and the number of parameters increased by 0.95. The model size also increased to 364 MB. Compared with the CAM, the mAP obtained by the baseline network after SAM embedding was lower, but a smaller model size and number of parameters are also guaranteed. Therefore, the SAM can be effectively used to capture the feature information of the region of interest from the channel dimension.

Considering the functional differences between the SAM and the CAM, different combination methods may affect the overall performance of the hybrid domain attention mechanism. Therefore, this experiment compared the performance of different hybrid domain attention mechanisms by starting with two combination methods of cascade (&&)
and parallel (//). The experimental results show that the sequential CAM&SAM (HDAU) performs better than the sequential SAM&CAM when the number of parameters is the same, and 75.3% mAP is obtained. It can be seen that with the same number of parameters, the hybrid domain attention generated by the parallel method is not as effective as the cascade method. This experiment analyzed the detection performance of the above hybrid domain attention and selected the sequential CAM&SAM (HDAU) as the hybrid domain attention of this paper. After adding a HDAU to the baseline network, the mAP increased by 1.4%, while the number of parameters increased by only 3.02. The proposed HDAU generates more efficient attentional maps than other combinations.

After embedding the proposed HDAU into the backbone network of the baseline network, the model size and the number of parameters increased by 23 MB and 3.02 M, respectively, and 75.3% mAP was obtained. It can be seen that the HDAU effectively combines the CAM and the SAM through the cascade method, which can effectively improve the detection performance of remote sensing image objects while ensuring a smaller model complexity. Furthermore, the MDW-Net proposed by us improves the baseline network by 1.2% mAP and only increases the model size by 15 MB and the number of parameters by 1.91. The MDW-Net can fully capture the feature information of remote sensing image objects after integrating rich semantic information and fine-grained information from deep and shallow layers of the network. In conclusion, our proposed HDAU-YOLO achieves better detection performance under the premise of maintaining a small computational overhead and can effectively guide the remote sensing image object detection task.

### 4.4. Comparison of Different Models

To demonstrate the effectiveness of a HDAU-YOLO in detecting various objects in remote sensing images, this study compared the performance of our method with that using several other state-of-the-art object detection methods, such as Faster R-CNN [28], SSD [29], YOLOv3 [29], YOLOv4 [30], DFPN-YOLO [30], RetinaNet [29], T-TRD-DA [31], CANet [32], FRPNet [33], and FSoD-Net [29], on the DIOR dataset. Table 3 shows the performance comparison of different methods on the DIOR test set.

| C1: AIRPLANE | SSD [29] | YOLOv3 [29] | YOLOv4 [30] | DFPN-YOLO [30] | RetinaNet [29] | T-TRD-DA [31] | CANet [32] | FRPNet [33] | FSoD-Net [29] | Ours |
|---------------|---------|-------------|-------------|----------------|----------------|---------------|-------------|-------------|---------------|------|
|               |         |             |             |                |                |               |             |             |                |      |
| C2: AIRPORT   | 53.6    | 92.9        | 67.5        | 79.1           | 80.2           | 63.4          | 77.9        | 70.3        | 64.5          | 88.9 |
| C3: BASEBALL FIELD | 66.2 | 77.2        | 86.8        | 88.4           | 89.1           | 78.4          | 86.3        | 87.8        | 81.7          | 90.2 |
| C4: BASKETBALL COURT | 52.0 | 53.7        | 54.2        | 54.2           | 54.2           | 54.2          | 54.2        | 54.2        | 54.2          | 54.2 |
| C5: BRIDGE    | 47.8    | 48.7        | 49.8        | 50.8           | 51.8           | 52.8          | 53.8        | 54.8        | 55.8          | 56.8 |
| C6: CHIMNEY   | 67.9    | 67.9        | 67.9        | 67.9           | 67.9           | 67.9          | 67.9        | 67.9        | 67.9          | 67.9 |
| C7: DAM       | 72.3    | 94.4        | 94.4        | 94.4           | 94.4           | 94.4          | 94.4        | 94.4        | 94.4          | 94.4 |
| C8: EXPRESSWAY SERVICE AREA | 69.0 | 47.9        | 55.7        | 58.8           | 59.8           | 53.2          | 59.0        | 83.5        | 80.0          | 86.9 |
| C9: EXPRESSWAY TOLL STATION | 55.2 | 68.8        | 68.8        | 68.8           | 68.8           | 68.8          | 68.8        | 68.8        | 68.8          | 68.8 |
| C10: GOLF FIELD | 56.9 | 33.7        | 68.9        | 72.4           | 71.6           | 59.1          | 73.9        | 83.6        | 77.4          | 77.3 |
| C11: GROUND TRACK FIELD | 50.2 | 47.9        | 54.3        | 57.5           | 63.1           | 47.6          | 49.2        | 56.0        | 78.7          | 53.6 |
| C12: HARBOR   | 50.1    | 39.9        | 51.7        | 57.2           | 58.7           | 45.7          | 57.8        | 63.6        | 82.4          | 59.7 |
| C13: OVERPASS | 27.7    | 59.6        | 86.8        | 78.8           | 81.5           | 72.0          | 74.2        | 81.0        | 62.9          | 78.3 |
| C14: SHIP     | 73.0    | 77.5        | 40.3        | 38.8           | 40.1           | 82.4          | 61.1        | 79.8        | 72.6          | 69.9 |
| C15: STADIUM  | 39.8    | 80.8        | 67.8        | 70.7           | 74.2           | 80.7          | 69.8        | 70.8        | 67.6          | 75.0 |
| C16: STORAGE TANK | 75.2 | 88.3        | 83.9        | 85.4           | 85.8           | 90.1          | 84.0        | 88.2        | 81.2          | 91.4 |
| C17: TENNIS COURT | 38.6 | 55.6        | 32.3        | 64.4           | 73.6           | 55.0          | 58.8        | 67.6        | 65.2          | 52.3 |
| C18: TRAIN STATION | 23.6 | 38.8        | 49.1        | 46.6           | 49.7           | 47.7          | 50.5        | 51.2        | 52.7          | 52.0 |
| C19: VEHICLE  | 45.4    | 86.7        | 73.6        | 83.5           | 86.5           | 92.9          | 77.2        | 89.6        | 89.1          | 90.6 |
| C20: WINDMILL | 54.1    | 60.9        | 59.9        | 66.7           | 69.3           | 65.2          | 66.8        | 74.3        | 71.8          | 71.8 |

mAP | 54.1 | 60.9 | 59.9 | 66.7 | 69.3 | 65.2 | 66.8 | 74.3 | 71.8 | 71.8 | 76.2 |
It can be found that the HDAU-YOLO achieved optimal detection results compared to several state-of-the-art object detectors, 1.9% mAP higher than the suboptimal CANet. For airports, chimneys, and golf fields with large-scale and fewer object instances, our method achieves optimal APs of 83.9%, 80.4%, and 82.6%, respectively. It can be seen that feature information can be fully learned even with fewer object instances. Our method also achieves the best detection performance particularly for the detection of small object categories such as ships and vehicles, reaching a 90.5% and 55.5% AP, respectively, a 3.7% and 2.8% AP higher than the suboptimal YOLOv3 and FRPNet, respectively. In general, the HDAU-YOLO proposed by us has great advantages in processing small object detection tasks in remote sensing images, and its detection effectiveness is quite considerable.

Then, to further determine the full-scale object detection performance of the proposed HDAU-YOLO, the F1 score, precision and recall curves were assessed and are shown in Figure 9. Evaluating an object detection model using precision and recall can provide valuable insight into how the model performs at various confidence values. Similarly, the F1 score is especially helpful in determining the optimum confidence that balances the precision and recall values for that given model. In general, as the confidence threshold increases, the precision increases and the recall decreases. The F1 score can be regarded as a weighted average of model precision and recall. It takes into account both precision and recall of the model to achieve the best balance between them. The F1 score has a maximum of 1 and a minimum of 0, and a higher value indicates better model performance. By comparing Figure 9a,b, it can be found that the F1 score of the HDAU-YOLO increases rapidly, reaching approximately 0.7 when the confidence threshold is 0.2, while the F1 score of the baseline network is only approximately 0.6. In addition, the F1 score of the HDAU-YOLO starts to decline when the confidence threshold is approximately 0.7, and the F1 score of the baseline network starts to decline when the confidence threshold is approximately 0.6. The F1 score of the HDAU-YOLO peaks at 0.76 when the confidence threshold is 0.582. Compared with the baseline network, after reaching the peak, the F1 score curve of the HDAU-YOLO becomes more stable. By comparing Figure 9c,d, it can be found that the precision of the baseline network reaches the peak value of 1.0 when the confidence threshold is 0.962, which is slightly faster than the HDAU-YOLO. However, a closer look shows that the increase in precision of the HDAU-YOLO is initially faster, and its precision performance is better than that of the baseline network at the confidence threshold range of 0.1 to 0.8. The comparison between Figure 9e,f shows that the recall curve of the HDAU-YOLO tends to gradually stabilize after an initial rapid decline, becoming smoother than the baseline network. In general, the curves of our methods are more stable and show better performance than the baseline network.

Figure 10 shows the detection results of 20 remote sensing image object categories on the DIOR test set for the HDAU-YOLO method. (a–t) are the airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill, respectively. From the figure, the advantages of the HDAU-YOLO in overall detection performance can be seen more intuitively. Our method can not only better detect large-scale categories such as airports, golf fields, ground track fields, and stadiums but also achieve ideal detection results for situations with large-scale differences among different categories, such as Figure 10h,k,l. In addition, the HDAU-YOLO also showed excellent detection performance for densely packed and small object categories such as airplane, ships, storage tank, and vehicle. In conclusion, the HDAU-YOLO can effectively suppress the interference of remote sensing image background noise information and selectively capture important feature information beneficial to remote sensing image object detection tasks.
Figure 9. F1 score curves, precision curves, and recall curve analysis on the DIOR test set.

To further evaluate the performance between the baseline network and the proposed HDAU-YOLO in tasks of small object detection in remote sensing images, four small object samples (including ships, cars, aircraft, and windmills) in the DIOR test set are randomly selected for testing. The results of detection by the HDAU-YOLO and the baseline network are shown in Figure 11.

As shown in Figure 11, it can be seen that the detection of small-scale objects is effectively improved using the HDAU-YOLO compared with the baseline network. Through the comparison of Figure 11a,c, it is found that the baseline network performs unnecessary repeated detection on the same ship object, resulting in the overlap of two prediction boxes, and this object can be effectively located using the HDAU-YOLO.
Comparing Figure 11b,d, it can be seen that the baseline network is disturbed by background noise information, and the traffic marking (left red circle) and the ground (right red circle) are incorrectly recognized as vehicles. The HDAU-YOLO shows strong anti-interference ability in the face of the complex background information of remote sensing images and demonstrates more effective performance with fewer false alarms by weakening the noise interference. In addition, comparing Figure 11e with Figure 11g, the baseline network could detect airplane objects though mistakenly detected the airplane tail as a vehicle, while the HDAU-YOLO can accurately detect the object as an airplane. Further comparing Figure 11f,h, it is found that the HDAU-YOLO accurately detects all windmill objects in the figure, while repeated detection (left red circle) and missed detection (right red circle) resulted from the baseline network. In general, when dealing with the task of small object detection in remote sensing images, the HDAU-YOLO proposed in this paper has more obvious advantages than the baseline network. It has a higher ability to correctly recognize small objects in remote sensing images with dense arrangements and complex backgrounds and effectively avoids the phenomena of false alarms, missed detections, and repeated detection.
In this study, a method for the detection of small objects in optical remote sensing images is proposed, HDAU-YOLO, which is based on a hybrid domain attention mechanism. According to the characteristics of the complex background of remote sensing images, a hybrid domain attention unit (HDAU) combined with cascade mode is proposed. Through the efficient combination of spatial and channel attention, the key feature information of remote sensing image objects is extracted along the channel and spatial directions, and irrelevant background noise information is weakened, which is beneficial to improving the performance with respect to small object detection in remote sensing images. In view of the lack of effective feature information of small-scale objects in remote sensing images, a multiscale dynamic weighted feature fusion network (MDW-Net) is proposed, which dynamically adjusts the weight of each input feature layer through adaptive learning, emphasizing the importance of shallow fine-grained feature information in the process of feature fusion and effectively improving the ability of object detectors in perceiving the details and location information of small objects. The experimental results on the DIOR test set are randomly selected for testing. The results of detection by the HDAU-YOLO show that, compared with several state-of-the-art object detection methods, the HDAU-YOLO has achieved the best detection results in the two small target detection tasks, with an AP of 79.2% and 74.1%, respectively. The HDAU-YOLO surpasses the baseline network by a large margin of +2.3 mAP, demonstrating its effectiveness in handling the remote sensing image objects.

\textbf{Figure 11.} Comparison of small object detection on the DIOR test set.

\textbf{5. Conclusions}

In this study, a method for the detection of small objects in optical remote sensing images is proposed, HDAU-YOLO, which is based on a hybrid domain attention mechanism. According to the characteristics of the complex background of remote sensing images, a hybrid domain attention unit (HDAU) combined with cascade mode is proposed. Through the efficient combination of spatial and channel attention, the key feature information of remote sensing image objects is extracted along the channel and spatial directions, and irrelevant background noise information is weakened, which is beneficial to improving the performance with respect to small object detection in remote sensing images. In view of the lack of effective feature information of small-scale objects in remote sensing images, a multiscale dynamic weighted feature fusion network (MDW-Net) is proposed, which dynamically adjusts the weight of each input feature layer through adaptive learning, emphasizing the importance of shallow fine-grained feature information in the process of feature fusion and effectively improving the ability of object detectors in perceiving the details and location information of small objects. The experimental results on the DIOR test set show that, compared with several state-of-the-art object detection methods, the HDAU-YOLO has achieved the best detection results in the two small target detection tasks, with an AP of 79.2% and 74.1%, respectively. The HDAU-YOLO surpasses the baseline network by a large margin of +2.3 mAP, demonstrating its effectiveness in handling the remote sensing image objects.
dataset show that, compared with several state-of-the-art object detection methods, the HDAU-YOLO has achieved the best detection results in the two small target categories, vehicle and ship, with an AP reaching 55.5% and 90.5%, respectively, while it ranked fourth and fifth for detection results in the two small target categories aircraft and storage tank, with an AP of 79.2% and 74.1%, respectively. The HDAU-YOLO surpasses the YOLOv5 by a large margin of +2.3 mAP, demonstrating it can effectively handle the task of small object detection in remote sensing images.

However, although our method has achieved good performance on the DIOR dataset, its detection performance on some high-noise remote sensing images is poor, and achieving accurate detection in high-noise remote sensing images is still the main challenge of remote sensing object detection. We plan to conduct further research in the future to address this issue.

Author Contributions: Conceptualization, T.D. and G.M.; methodology, T.D.; validation, X.L. and G.M.; formal analysis, G.M.; data curation, X.L.; writing—original draft preparation, G.M.; writing—review and editing, T.D. and X.L.; visualization, G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments: This work was supported by the National Key Research and Development Program of China (grant number: SQ2020YFF0418521), Chongqing Science and Technology Development Foundation (grant number: cstc2020jcx-dxwtBx019 and cstc2021jcx-gkshX0058), and the Joint Key Research & Development Program of Sichuan and Chongqing (CN) (grant number: cstc2020jcx-cyhlX0007 and cstc2020jcx-cyhlX0005). Tianmin Deng and Guotao Mao gave the idea, Guotao Mao and Xuhui Liu performed the experiments, and Tianmin Deng, Guotao Mao and Xuhui Liu wrote the paper and interpreted the results.

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

References

1. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 2013, 35, 1798–1828. [CrossRef] [PubMed]
2. Xiao, Z.; Qian, L.; Shao, W.; Tan, X.; Wang, K. Axis Learning for Oriented Objects Detection in Aerial Images. *Remote Sens.* 2020, 12, 908. [CrossRef]
3. Zhang, X.; Zhu, K.; Chen, G.; Tan, X.; Zhang, L.; Dai, F.; Liao, P.; Gong, Y. Geospatial Object Detection on High Resolution Remote Sensing Imagery Based on Double Multi-Scale Feature Pyramid Network. *Remote Sens.* 2019, 11, 755. [CrossRef]
4. Du, B.; Cai, S.; Wu, C. Object Tracking in Satellite Videos Based on a Multiframe Optical Flow Tracker. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2019, 12, 3043–3055. [CrossRef]
5. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 27–28 October 2019; pp. 8231–8240.
6. Peng, C.; Li, Y.; Jiao, L.; Chen, Y.; Shang, R. Densely Based Multi-scale and Multi-Modal Fully Convolutional Networks for High-Resolution Remote-Sensing Image Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2019, 12, 2612–2626. [CrossRef]
7. 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 Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
8. 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.* 2015, 28, 91–99. [CrossRef]
9. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
10. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
11. Dong, R.; Xu, D.; Zhao, J.; Jiao, L.; An, J. Sig-NMS-Based Faster R-CNN Combining Transfer Learning for Small Target Detection in VHR Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* 2019, 57, 8534–8545. [CrossRef]
12. Zheng, Z.; Lei, L.; Sun, H.; Kuang, G. FAGNet: Multi-Scale Object Detection Method in Remote Sensing Images by Combining MAFPN and GVR. *J. Comput.-Aided Des. Comput. Graph.* 2021, 33, 883–894. [CrossRef]
13. Yao, Y.; Cheng, G.; Xie, X.; Han, J. Optical remote sensing image object detection based on multi-resolution feature fusion. *Natl. Remote Sens. Bull.* 2021, 25, 1124–1137. [CrossRef]

14. Li, C.; Liu, L.; Zhao, J.; Liu, X. LF-CNN: Deep Learning-Guided Small Sample Target Detection for Remote Sensing Classification. *Cmes-Comput. Model. Eng. Sci.* 2022, 131, 429–444. [CrossRef]

15. 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.

16. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.

17. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 2018, 42, 318–327. [CrossRef]

18. Courtrai, L.; Pham, M.T.; Friguet, C.; Lefevre, S. Small Object Detection from Remote Sensing Images with the Help of Object-Focused Super-Resolution Using Wasserstein GANs. In Proceedings of the IGARSS-IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 19–24 July 2020; pp. 260–263.

19. Yin, R.; Zhao, W.; Fan, X.; Yin, Y. AF-SSD: An Accurate and Fast Single Shot Detector for High Spatial Remote Sensing Imagery. *Sensors* 2020, 20, 6530. [CrossRef] [PubMed]

20. Yu, L.; Wu, H.; Zhong, Z.; Zheng, L.; Deng, Q.; Hu, H. TWC-Net: A SAR Ship Detection Using Two-Way Convolution and Multiscale Feature Mapping. *Remote Sens.* 2021, 13, 2558. [CrossRef]

21. Li, Z.; Liu, X.; Zhao, Y.; Liu, B.; Huang, Z.; Hong, R. A Lightweight Multi-scale Aggregated Model for Detecting Aerial Images Captured by UAVs. *J. Vis. Commun. Image R* 2021, 77, 103058. [CrossRef]

22. Nong, Y.; Wang, J. Real-Time Object Detection in Remote Sensing Images Based on Embedded System. *Acta Opt. Sin.* 2021, 41, 179–186. [CrossRef]

23. Liu, Y.; He, G.; Wang, Z.; Li, W.; Huang, H. NRT-YOLO: Improved YOLOv5 Based on Nested Residual Transformer for Tiny Remote Sensing Object Detection. *Sensors* 2022, 22, 4953. [CrossRef]

24. Fang, X.; Hu, F.; Yang, M.; Zhu, T.; Bi, R.; Zhang, Z.; Gao, Z. Small object detection in remote sensing images based on super-resolution. *Pattern Recognit. Lett.* 2022, 153, 107–112. [CrossRef]

25. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. *arXiv* 2021, arXiv:2103.02907.

26. Wu, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Chab: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.

27. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.

28. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* 2020, 159, 296–307. [CrossRef]

29. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-Scale Object Detection from Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* 2021, 60, 1–18. [CrossRef]

30. Sun, Y.; Liu, W.; Gao, Y.; Hou, X.; Bi, F. A Dense Feature Pyramid Network for Remote Sensing Object Detection. *Appl. Sci.* 2022, 12, 4997. [CrossRef]

31. Li, Q.; Chen, Y.; Zeng, Y. Transformer with Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* 2022, 14, 984. [CrossRef]

32. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2021, 14, 2148–2161. [CrossRef]

33. Wang, J.Y.; Wang, Y.Z.; Wu, Y.L.; Zhang, K.; Wang, Q. FRPNet: A Feature-Reflowing Pyramid Network for Object Detection of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* 2022, 19, 1–5. [CrossRef]