*remote sensing*

MDPI

*Article*

# Multi-Feature Information Complementary Detector: A High-Precision Object Detection Model for Remote Sensing Images

**Jiaqi Wang, Zhihui Gong, Xiangyun Liu, Haitao Guo *, Jun Lu, Donghang Yu and Yuzhun Lin**

Institute of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
* Correspondence: ghtgjp2002@163.com

**Abstract:** Remote sensing for image object detection has numerous important applications. However, complex backgrounds and large object-scale differences pose considerable challenges in the detection task. To overcome these issues, we proposed a one-stage remote sensing image object detection model: a multi-feature information complementary detector (MFICDet). This detector contains a positive and negative feature guidance module (PNFG) and a global feature information complementary module (GFIC). Specifically, the PNFG is used to refine features that are beneficial for object detection and explore the noisy features in a complex background of abstract features. The proportion of beneficial features in the feature information stream is increased by suppressing noisy features. The GFIC uses pooling to compress the deep abstract features and improve the model's ability to resist feature displacement and rotation. The pooling operation has the disadvantage of losing detailed feature information; thus, dilated convolution is introduced for feature complementation. Dilated convolution increases the receptive field of the model while maintaining an unchanged spatial resolution. This can improve the ability of the model to recognize long-distance dependent information and establish spatial location relationships between features. The detector proposed also improves the detection performance of objects at different scales in the same image using a dual multi-scale feature fusion strategy. Finally, classification and regression tasks are decoupled in space using a decoupled head. We experimented on the DIOR and NWPU VHR-10 datasets to demonstrate that the newly proposed MFICDet achieves competitive performance compared to current state-of-the-art detectors.

**Keywords:** remote sensing imagery; object detection; multi-scale; complementary information



**Citation:** Wang, J.; Gong, Z.; Liu, X.; Guo, H.; Lu, J.; Yu, D.; Lin, Y. Multi-Feature Information Complementary Detector: A High-Precision Object Detection Model for Remote Sensing Images. *Remote Sens.* **2022**, *14*, 4519. https://doi.org/10.3390/rs14184519

Academic Editor: Józef Lisowski

Received: 15 August 2022
Accepted: 6 September 2022
Published: 9 September 2022



## 1. Introduction

Remote sensing technology is one of the most important ways to observe the Earth [1] as it provides a broader perspective. Remote sensing image object detection has been applied in several fields. Recently, with the development of deep learning technology, substantial progress has been made in object detection methods, which has led to a new wave of object detection processing and applications based on remote sensing images.

Object detection via deep learning models can be classified into two categories: two-stage series models based on region proposal and one-stage series models based on regression. The main difference between the two is that the former pre-generates candidate boxes that may contain objects using heuristics and then performs fine-grained classification and regression on the candidate boxes, whereas the latter is densely sampled directly on the feature map, and the sampled regions contain different scales and aspect ratios, which in turn leads to the direct classification and regression of the extracted features. A representative algorithm of the two-stage approach is R-CNN [2]. An improved version of R-CNN, Faster R-CNN [3], provided the idea for subsequent two-stage methods. The two-stage algorithm generates candidate boxes in advance, followed by an exact regression;

thus, its accuracy is higher. However, this group of models is computationally complex and slow to detect. The one-stage model regresses directly to the object boundary, thus saving a lot of computational resources while obtaining sufficient accuracy. At present, one-stage models are commonly used in industry, and classical algorithms include the YOLO [4] series and the SSD [5] series. General-purpose object detection algorithms have achieved excellent performance; however, there are still outstanding problems in applying these models directly to remote sensing images.

Compared with images in natural scenes, remote sensing images are characterized by strong background information interference, large variations in object scale, and rich appearance and shapes. These characteristics pose considerable challenges during classification and regression. Many scholars have proposed solutions to the difficulties in remote sensing image object detection. However, there is still room for improvement in the accuracy of these models applied to complex remote sensing image datasets. The wide field of view of remote sensing images leads to strong interference from the background in the object detection task. Several methods have recently been proposed to solve complex backgrounds. Liu et al. [6] proposed a relationally connected attention module that obtains global information by stacking features and relative features between features to enhance the foreground information and weaken the background information, making the features of interest more distinguishable. Bai et al. [7] proposed an object detection method based on time-frequency analysis, which enables the detector to focus on the object region rather than the background region through a discrete wavelet multi-scale attention mechanism. Cheng et al. [8] applied a priori scene information and Bayesian criteria to infer the relationship between scenes and objects, using various scene semantics as a specific prior to improve the performance of object detection in remote sensing images.

From the above information, the current methods used to solve the complex background of remote sensing images are as follows.

(1) Using feature enhancement methods, such as the attention mechanism, to improve the feature representation of the object, thus indirectly weakening the background information. This method is the current mainstream approach and offers a substantial improvement in accuracy. However, this approach requires a targeted design for the corresponding modules and is relatively computationally complex.
(2) The relationship between the object and background selectively eliminates background features or enhances features of the object. This approach considers the features of the object and attends to background features. However, a better strategy is needed to distinguish the beneficial background from the interfering background; otherwise, this will lead to confusion between the object and the background.
(3) Using prior information, the impact of complex background information on detector performance is reduced manually. This method is simple and easy to use but results in a limited improvement in accuracy and requires a considerable labor force to select data with a single background for pretraining, which increases the cost of training.

The large variation in object scales in remote sensing images poses a great challenge to the task of object detection and regression. To address this difficulty, extensive research has been conducted from various perspectives. Lin et al. [9] proposed an adaptive feature pyramid network to effectively solve multi-scale and dense object detection; this contains a selective refinement module to selectively refine different feature maps. Wu et al. [10] proposed a feature refinement module that combines different branches to convolve multiple receptive fields for object instances with drastic changes in scale and shape, thus further refining the features and improving feature discrimination at different scales. Ma et al. [11] proposed a feature split-and-merge module to distribute large and small objects in a scene across multiple levels of feature maps for subsequent detection, alleviating feature confusion between multi-scale objects, and proposed an offset-error rectification module to correct inconsistencies in the spatial layout of objects among multiple levels of feature maps. The above models perform well but suffer from the following shortcomings: (1) a lack of more effective feature-aware strategies for small-scale objects; (2) failure to

deeply mine contextual information for object detection; (3) most of the proposed modules target different scale objects on different images but lack the consideration of the same class of objects with different scales in the same image.

To address these issues, we propose a high-precision object detection model, which we term the multi-feature information complementary detector (MFICDet). The detector is designed based on an efficient one-stage detection architecture that combines multi-feature supervision and information complementation to achieve the highly accurate detection of objects with multiple scales and complex backgrounds in remote sensing images. First, MFICDet uses CSPDarkNet53 [12] as the backbone of the model to extract features. Second, we propose a global feature information complementation module (GFIC) for the deep semantic enhancement of abstract features. Then, considering the effect of the complex background on the object detection performance, we propose a positive and negative feature guidance module (PNFG) to remove the interference of noisy features in the background. Finally, the spatially decoupled head (DHead) is used to decouple the object localization task from the classification task in spatial terms, improving the object localization accuracy and reducing the training time of the network.

The main contributions of this work are as follows.

(1) A global feature information complementary (GFIC) module which combines the advantages of pooling and dilated convolution to deeply fuse the primary features and enhance the semantic representation of the model. Aimed at the characteristics of remote sensing images with large-scale changes in objects, a dual multi-scale feature fusion strategy is used to solve the challenges posed by different scale objects in the same image.
(2) A positive and negative feature guidance (PNFG) module. We define noise information in a complex background that is useless for object detection as negative features. In contrast, the features that provide valuable information for object detection are defined as positive features. Because positive and negative features are coupled with the features extracted by the backbone network, a PNFG strategy is designed to eliminate negative features while enhancing and refining positive features.
(3) A highly accurate object detection model for remote sensing images that achieves state-of-the-art performance on publicly available remote sensing image object detection datasets.

The remainder of this paper is organized as follows. Section 2 presents the latest research results on solving the complex background of remote sensing images and multi-scale objects. Section 3 provides a detailed description and analysis of the proposed model. An experimental analysis and discussion are presented in Section 4. Section 5 discusses future research work and the limitations of the proposed detector. Finally, Section 6 gives the conclusions of this study.

## 2. Related Work

### *2.1. Object Detection for Complex Backgrounds*

In recent years, many scholars have contributed to the solution to the complex background problem in remote sensing images. Li et al. [13] proposed a saliency pyramid module that combines a saliency algorithm with a feature pyramid network to suppress background noise information and reduce the influence of complex backgrounds. Shi et al. [14] proposed a position attention module for error detection caused by complex backgrounds in remote sensing images. The complex internal structure feature representation was extracted by calculating the similarity of the features between any two pixels on the target feature map to improve the ability to distinguish between the background and foreground. Cheng et al. [15] proposed a diverse contextual information fusion framework based on convolutional neural networks to improve object detection and recognition in complex backgrounds using structured object-level relationships. Song et al. [16] designed an enhancement network to overcome the diversity and complexity of the background and object through adaptive multi-scale anchors and improved the loss function. Other scholars have

used the attention mechanism to process fused multi-scale features to enhance the feature information of the object region [17–19]. Yu et al. [20] used deformable convolutional layers to extract high-level features of an object from the background by the irregular sampling of locally pooled features. Zhang et al. [21] designed a feature relationship module to improve the distinction between the foreground regions of the feature map using a contextual representation of the foreground. The foreground regions were weighted to alleviate the background–foreground imbalance problem. Wang et al. [22] used a multi-scale, feature-focused attention module to enhance the ability of the network to represent features in different regions, weakening the information interference from the background and negative sample objects. Zhu et al. [23] reduced the interference of complex backgrounds by pretraining images with a single and uniformly distributed background. In addition, the decoupling of the background and the object was achieved.

### *2.2. Object Detection of Multi-Scale Objects*

The feature pyramid network (FPN) is a common method for solving multi-scale objects in object detection tasks; many scholars have applied it to remote sensing object detection and have produced targeted improvements. Wang et al. [24] proposed a feature reflow pyramid structure that improves the detection performance of multi-scale objects by fusing fine-grained features from adjacent lower levels to generate a high-quality feature representation for each scale. Liu et al. [25] proposed a gated trapezoidal FPN to construct a more representative feature pyramid to detect objects of different sizes in optical remote sensing images. Cheng et al. [26] found that feature pyramids have features at different levels that interact when performing top-down operations. Therefore, a perceptual FPN was proposed to improve the detection performance of multi-scale objects.

In addition, some scholars have separately enhanced the features of different scale objects in feature maps. Zhou et al. [27] introduced three parallel convolutional branches with the same structure and a cascaded feature-fusion module to generate effective multi-scale features. A code—decode architecture was used to treat depth-feature fusion as a decoding process and integrate multi-scale depth features in a progressive manner. Cong et al. [28] proposed a parallel multiscale attention module to efficiently recover detailed information and resolve scale variations of salient objects using low-level features refined by parallel multi-scale attention.

Scholars have also addressed the problem of multi-scale object detection in remote sensing images from a receptive field viewpoint. Han et al. [29] proposed a multiscale residual block to capture multi-scale contextual information and designed a multiscale receptive field enhancement module to enhance the multi-scale feature representation of remote sensing objects. Liu et al. [6] constructed a new multi-receptive field feature extraction module that enables the network to aggregate multi-receptive field information for multi-scale object features, providing a powerful representation of feature multi-scale objects. Zhang et al. [30] proposed a receptive field enhancement module that focuses on different receptive fields by multi-branching different convolutions; this is committed to more stable multi-scale feature extraction.

Finally, this study summarizes the work related to solving the complex background and diverse object scales of remote sensing images, as shown in Table 1.

**Table 1.** A summary of recently proposed object detectors.

| Targeted Questions | Methods | Literatures | Advantages | Unresolved Issues |
| --- | --- | --- | --- | --- |
| Complex backgrounds | Diminish background features and highlight object features | [7,11,13,14,16,20,22,23] | 1. Using attention mechanism or priori information to highlight the features of the object for interest effectively improves the performance of the detector.<br>2. The designed feature enhancement strategies are targeted to discriminate different feature regions and effectively suppress the influence of noise information on the detector. | 1. Focusing too much on the features of the objects, the role of background information may be ignored. The detection accuracy of objects strongly coupled with the background is limited.<br>2. Detection of objects with extremely complex backgrounds remains unsatisfactory. |
| | Explore the relationship between background and object | [6,8,15,21] | 1. Contextual information about the object is considered and explores the dependency relationships between features.<br>2. The scene information is better utilized, and the relationship between the object and the backgrounds is analyzed from the semantic level. | 1. The effectiveness of the feature relationship mining strategies is required highly.<br>2. The suppression of useless noise information is relatively poor. |
| Scale diversity | Feature pyramid network | [9,11,24–27] | 1. It can effectively enhance the semantic information of the multi-scale feature map, and significantly improve the detection performance of the detector.<br>2. The structure is simple and suitable for various detectors. | 1. Due to the increase in parameters, it is more expensive to occupy memory during training. |
| | Increase the receptive field of multi-scale features | [6,29,30] | 1. The actual receptive field of the network is enlarged, and the stability of multi-scale features extraction is improved. | 1. Most of them are built based on a feature pyramid network.<br>2. Compared to FPN, the detector performance is not significantly improved. |
| | Refine multi-scale features | [10,24,28] | 1. The feature awareness of the network for the object is improved and obtains high-quality feature representation. | |

### *2.3. One-Stage Remote Sensing Image Object Detection*

Although the two-stage detectors achieved high accuracy, the algorithms operated with low efficiency, limiting their practical application. YOLO was proposed in 2015 as the first one-stage detector, which achieved real-time object detection and substantially improved the operational efficiency of the detector, although its detection accuracy was inferior to that of the two-stage detectors. SSD was proposed as another classical one-stage detector, which has been widely used in engineering because of its excellent performance in terms of detection speed and accuracy by the regression and classification of default boxes on feature maps of different scales. In 2018 RetinaNet [31] was proposed, and this detector proposed focal loss as a classification loss function, which greatly improves the samples' imbalance problem that exists in one-stage detectors.

To achieve a balance between the operational efficiency and accuracy of remote sensing image object detectors, many scholars have conducted an in-depth exploration of one-stage detectors. Yang et al. [32] proposed a unified framework of aggregation and detection for the two problems of dense small objects and sparse with non-uniform object distribution in aerial images, which effectively improves the detection accuracy of small objects. Wang et al. [24] proposed an end-to-end multiscale object detection method based on a dual-attention mechanism, which enhances feature reuse by a multiscale, feature-focused attention module and improves the correlation between feature sets using a two-level depth feature fusion module to achieve the accurate detection of multiscale and multi-pose

objects. There exist more large aspect ratio objects in remote sensing images, Li et al. [33] proposed a double-aligned, one-stage rotational detector, which achieves feature alignment at the image level by adjusting the receptive fields of neurons, and then uses a rotational feature alignment module to achieve feature alignment at the instance level. Hou et al. [34] designed asymmetric convolutional blocks embedded in an asymmetric feature pyramid network for processing objects with extreme aspect ratios. Xu et al. [35] proposed an efficient feature-aligned, one-stage detector that solves the spatial misalignment between the anchor and its corresponding object, and then uses the contextual feature alignment module to adaptively adjust the sampling points of the convolution kernel to collect contextual information. Huang et al. [36] designed a channel separation aggregation structure to simplify the convolutional complexity and developed a dynamic receptive field mechanism to dynamically customize the convolutional kernel and its perceptual range to maintain high accuracy while reducing the network complexity. Wei et al. [37] proposed to represent the object as a pair of midlines, based on which an oriented detection network was proposed to encode and detect the paired middle lines. Liu et al. [38] proposed a semantic supervised branch used only during training, which extracts additional key point features from boundary points and interior points to help the network localize the objects. Huang et al. [39] proposed a refined U-shaped module for the multi-level fusion of features to form a feature pyramid for object detection.

In recent years, anchor-free detectors have been widely studied. Shi et al. [40] proposed a center-aware network based on the observation that remote sensing image objects typically maintain symmetry, and the proposed multiscale center descriptor and feature selection module can select the best semantic information around the center region, allowing the network to gradually fit the symmetric shape of remote sensing objects. Law et al. [41] detected the objects by treating them as pairs of key points and introduced corner point pooling to help the network better localize the corner point locations.

## 3. Methodology

### *3.1. Network Architecture*

Figure 1 illustrates the overall architecture of the MFICDet proposed in this study. As shown in Figure 1, MFICDet comprises four parts: a feature extraction backbone network, a PNFG, a GFIC, and a spatial decoupled head (DHead). Specifically, CSPDarkNet53 was used as the feature extraction backbone network to acquire the primary features of the image. The successful application of the YOLOv4 object detection model demonstrated that CSPDarkNet53 can effectively extract semantic information for feature enhancement. The PNFG module proposed in this work can adaptively perceive the spatial location of beneficial features and the contribution of semantic information. The background noise information is removed from the feature stream of the model to suppress the interference of noisy features in the complex background and highlight the refined object features. The PNFG module effectively reduces the impact of complex backgrounds in remote sensing images on model performance while improving the sensitivity of the model to difficult classification objects. Subsequently, the GFIC module is designed to extract the abstract semantic information of the objects and noise information in the background. The module achieves a deep fusion of contextual information and the multi-level perception of background information. Therefore, it solves problems such as large feature differences and sparse detailed information on remote sensing image objects. Finally, the network fuses different scale feature maps using a top-down strategy. This strategy forms a dual multi-scale feature fusion with multi-scale feature stitching in the GFIC module, which improves the detection performance of different scale objects in the same image. Furthermore, a spatially decoupled head is introduced to resolve feature conflicts between classification and regression tasks in object detection.

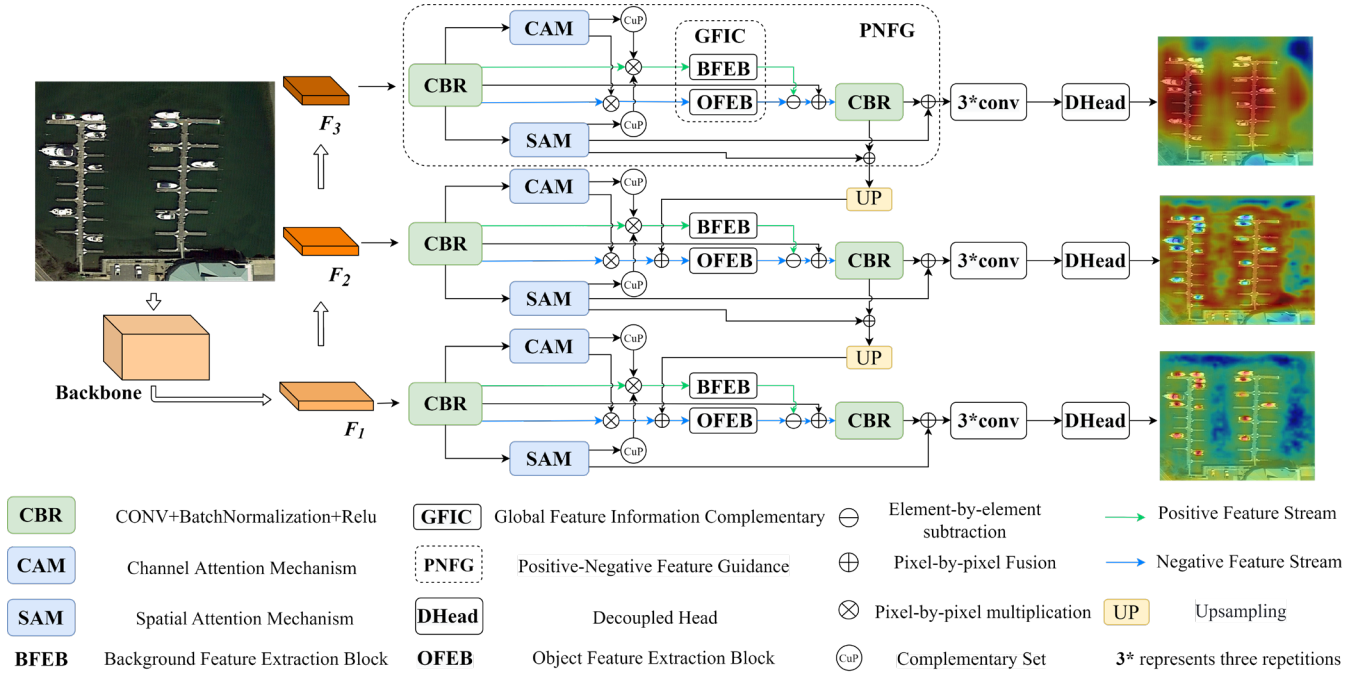


**Figure 1.** MFICDet overall architecture.

### *3.2. Global Feature Information Complementary Module*

To improve the ability of the model to perceive a global scene and enhance the expression of features, we proposed a GFIC module. Inspired by the pyramid pooling module in PSPNet [42], this effectively aggregates non-neighborhood multi-scale contextual information to improve the global feature extraction capability of the model. The pyramid pooling module is implemented using the global average pooling method. Global average pooling leads to a slower convergence of the model, and the operation of global averaging of the feature map leads to the loss of many textural features; therefore, it is not suitable for object detection tasks. The GFIC module proposed in this study enhances the detailed feature information of the object while considering global semantic information. Furthermore, background noise features are extracted and removed during downstream processing. The structure of the GFIC module is shown in Figure 2; it comprises an object feature extraction block (OFEB) and a background feature extraction block (BFEB) with an identical structure. The difference between the two methods is that the OFEB in the pooling branch uses maximum pooling to extract abstract features, whereas the BFEB uses average pooling for information compression. The OFEB is used to extract beneficial object features from the positive feature information stream and perform a multi-scale enhancement. The BFEB is used to explore the background noise information from the negative feature information stream, which interferes heavily with the detection task.

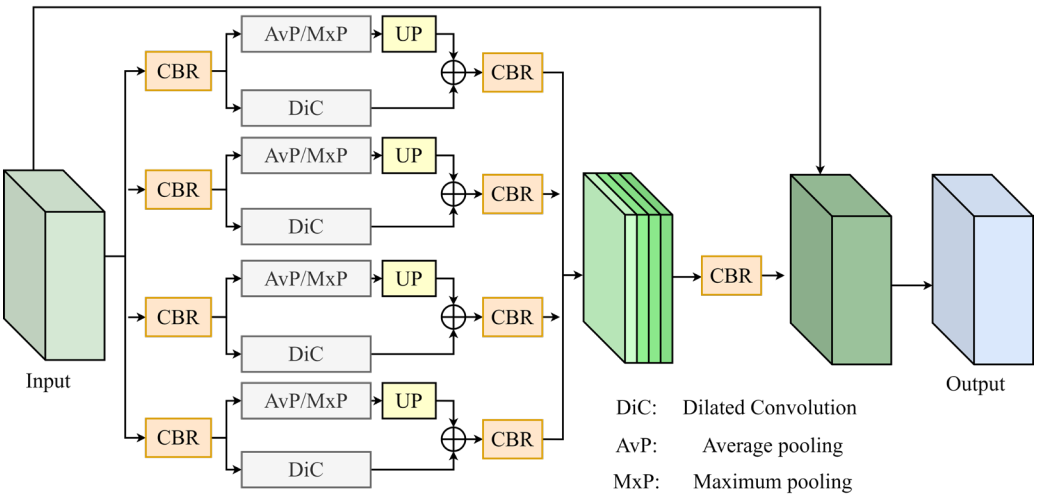


**Figure 2.** Global feature information complementary module structure.

(1) Object Feature Extraction Block

The OFEB is applied to extract the feature information of the object with interest in the positive feature information stream and perform a multi-scale semantic enhancement. This section first uses the maximum pooling method for sibling multi-scale feature information compression to increase the receptive field of the network. Maximum pooling can resist the translation and rotation of features, so it can overcome the effects of spatial position transformations, such as the displacement and rotation of objects, on detection. The critical features of the object are efficiently extracted using maximum pooling to filter redundant feature information. Maximum pooling loses numerous detailed features during downsampling to alleviate the redundancy of network information. Therefore, to address feature overcompression, we introduce a dilated convolution. Dilated convolution contributes to a larger receptive field per pixel than the output of the pooling operation without changing the image resolution. The beneficial features lost in the pooling process are supplemented by dilated convolution, and the spatial location relationships between the global context features are established using dilated convolution. In addition, dilated convolution with different dilation factors can capture multi-scale contextual information, thus further enhancing the receptive field of multi-scale features and improving the ability of the model to recognize multi-scale information.

(2) Background Feature Extraction Block

A BFEB is used in the negative feature information stream. The purpose is to improve the sensitivity of the negative feature information stream to background noise information. Specifically, the multi-scale global semantic information of the feature maps is extracted using adaptive average pooling. Average pooling can maintain the local smoothing of feature information and improve the problem of increasing the variance of estimates caused by the restricted domain size, retaining more information of the image, and is more suitable for the extraction of background information. However, average pooling has the disadvantage of obfuscating detailed features, resulting in pooled deep features that obscure semantic information during upsampling. Therefore, in line with the OFEB, dilated convolution is also introduced in the BFEB for information addition. However, the dilation factor of the dilated convolution is different for the BFEB compared to the OFEB.

We introduce the structure of the GFIC module, with the BFEB as an example, as shown in Figure 2. The channels of the input features are reduced in dimension using a $1 \times 1$ convolution. The results are then normalized using batch normalization and further activated using the ReLU function. In this study, the above process is defined as CBR. The output features are passed into the adaptive average pooling branch and upsampled (UP) to obtain global semantic information. Then, the feature map outputs from the adaptive

pooling branch are fused with the feature maps extracted from the dilated convolutional branch in the pixel space to obtain deep-feature maps with complementary information. Adaptive average pooling and dilated convolution use different pooling kernel sizes and dilation factors, respectively, thus generating multi-scale depth-feature maps. The feature maps on each scale are concatenated in the channel dimension after passing through CBR. Finally, the concatenated feature maps are passed through CBR again, and then fused with the initial feature maps in the spatial dimension to obtain the output of the module. The OFEB operates through the same process as described, except that adaptive average pooling is replaced by maximum pooling in the pooling branch.

### *3.3. Positive and Negative Feature Guidance Module*

The complex background of remote sensing images has a serious impact on object classification and localization, and it is difficult to identify objects with geometric and radiometric features extremely similar to the background. The spatial resolution of different sensors affects the size of object imaging; therefore, remote sensing images with poor imaging quality pose a challenge for object detection tasks. Furthermore, there are large differences in the radiometric resolutions of different sensors, resulting in the same objects showing a diversity of detailed features in different images. To mitigate the impact of these problems on remote sensing image object detection, in this study, a PNFG is designed. The design of the PNFG is inspired by the focus module for refinement prediction in PFNet proposed by Mei et al. [43]. The PNFG comprises two major steps: (1) Generate positive and negative features using the spatial attention mechanism (SAM) and the channel attention mechanism (CAM); (2) Use parallel information streams to achieve positive and negative feature guidance.

(1) Generation of Positive and Negative Features.

Positive and negative features in this study are generated by spatial attention and CAMs. Specifically, spatial attention is sensitive to the spatial location information of an object, whereas channel attention focuses more on the semantic information of the input image. Therefore, combining spatial and channel attention can effectively explore the critical features of an object and improve the flow of beneficial information in the network.

Although each channel of the feature map provides feature information for the prediction task, the effective features contributed by different channels are limited [44]. Therefore, channel attention is used to calculate the weight of the contribution of each channel to the final predicted result. The convolutional block attention module (CBAM) [45] is a lightweight attention module widely used, and in this study, only the channel attention mechanism is used in CBAM. The structure is shown in Figure 3. For an input $X$ with channel number $C$, the CAM first aggregates spatial information to generate spatial context descriptors using the maximum pooling and average pooling methods. Then, a channel attention map is generated using a multilayer perceptron (MLP). Finally, the attention map is fused using element-by-element summation and then passed through a sigmoid function. This formula is expressed as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}{}^c)) + W_1(W_0(F_{\max}{}^c))) \end{aligned} \quad (1)$$

where $M_c(F)$ represents the attention map of the channel attention output, $\sigma$ represents the sigmoid function, $W_0 \in \mathbb{R}^{C/r \times C}$, $W_1 \in \mathbb{R}^{C \times C/r}$ The weights, $W_0$ and $W_1$, are shared between maximum pooling and average pooling by the MLP.

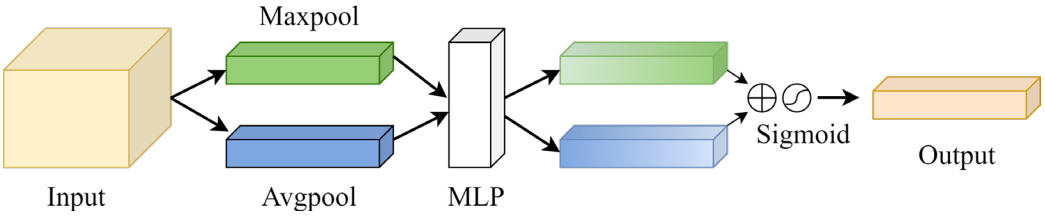


**Figure 3.** Structure of the channel attention mechanism.

The SAM can divide different regions of an image based on feature contribution. In this study, we use spatial attention in PFNet to achieve an effective discovery of the spatial location of an object. The structure of the SAM is shown in Figure 4, where the input features are convolved through three $1 \times 1$ convolution layers to produce three matrices, $Q$, $K$, and $V$, respectively. Then, matrix multiplication is performed between the transpose matrices of $Q$ and $K$. To accelerate convergence, the above results are normalized using the softmax function. Subsequently, the spatial attention map, $X_a$, is obtained. This process can be expressed as follows:

$$X_a = x_{ij} = \frac{exp(Q_{:i} \cdot K_{:j})}{\sum_{j=1}^{N} exp(Q_{:i} \cdot K_{:j})} \tag{2}$$

where $x_{ij}$ represents the effect of the $j$th position on the $i$th position, and $Q_{:i}$ represents the $i$th column of the matrix. Finally, matrix multiplication is performed between $V$ and the transpose of the attention weight matrix to obtain the output of spatial attention. To balance the learning process, a learnable parameter, $\gamma$, is introduced into the above results. Furthermore, the input feature maps are fused with the output in a shortcut manner to obtain the final output of the SAM.

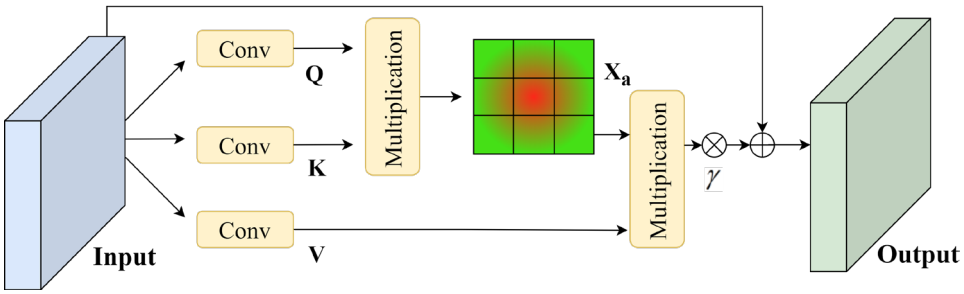


**Figure 4.** Structure of the spatial attention mechanism.

Complementary spatial attention and channel attention improve the ability of the model to explore deep semantic information and establish a correlation of features in the channel dimension. Information beneficial to the object detection task is markedly enhanced in the feature output by the two attention mechanisms, which are defined as positive features in this study. The absolute complementary set of positive features in the generated information space set is then defined as a negative feature used to accurately discriminate the background noise features.

(2) Positive and Negative Features Guidance

Spatial attention and channel attention aggregate the spatial and semantic information used for feature mapping and extract the positive and negative features of the object of interest. This information enables the model to efficiently explore the spatial location and deep semantic features of an object and remove irrelevant background noise. To achieve the above process, we here propose a positive and negative feature guidance strategy. As shown in Figure 5, this strategy comprises parallel information streams with positive and negative features. The positive feature stream is shown in blue, and the negative feature stream is shown in green.

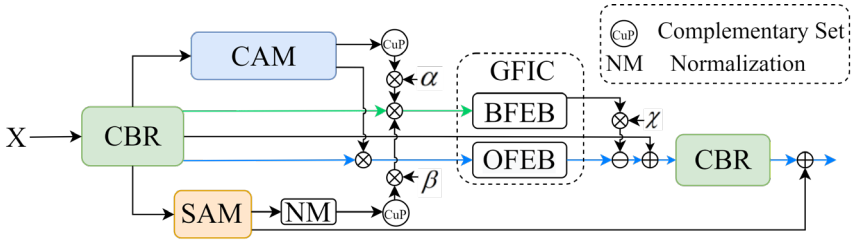


**Figure 5.** Positive and negative features guidance strategy.

Specifically, the output of CAM is applied to the input feature map $X$, and then normalized to generate a weight matrix $M_{trc}$. $M_{trc}$ is multiplied with $X$ to produce a positive feature information stream. The attention map output from the SAM is normalized (NM) to generate the weight matrix $M_{trs}$. Then, $M_{trc}$ and $M_{trs}$ are subtracted element-by-element from $M_n$, where $M_n$ is a matrix with all elements of 1. The procedure is to obtain the absolute complement set (CuP) of the positive feature information. The result obtained is also multiplied by the input feature map, $X$, to generate the negative feature stream. To balance the learning process, the learnable parameters, $\alpha$ and $\beta$, are introduced into the above positive and negative feature information stream generation processes, respectively. The information streams are then fed into the GFIC module. The positive feature information is used to guide the model to learn useful information strongly related to the object, whereas the negative feature information is used to discover noise in the background and remove useless features that interfere with the model. Finally, the negative feature information stream is excluded from the positive feature information stream, whereas a learnable parameter, $\chi$, is introduced. To prevent the disappearance of features, the input features are fused with the above results, before being passed through a CBR. Finally, the position information generated by the SAM is added to the information flow to assist the model in localizing the objects.

Using the PNFG strategy, the model adaptively determines the features that are beneficial for the task of subsequent feature enhancement. Moreover, the PNFG module efficiently discovers useless noise features with low contributions and eliminates them in the subsequent feature stream. Therefore, the weight of valid information in the feature stream increases. In addition, the PNFG strategy can effectively enhance the features of the object while eliminating the negative impact of complex background information on the object detection task. This improves the regression accuracy and feature awareness of the model for the object.

### *3.4. Decoupled Head*

The spatial feature mismatch problem for classification and boundary regression tasks in object detection is first proposed in IoU-Net [46], whereby features that improve classification confidence are not conducive to the prediction of bounding boxes; the IoU-Net model solves this problem by calculating the localization accuracy of the detected boxes. Song et al. [47] explored the nature of spatial feature misalignment for classification and localization and proposed task-aware spatial disentanglement to decompose the gradient flow of classification and localization in space. Through task-aware proposal estimation and the detection head, task-specific feature representation can be generated to eliminate compromises between classification and localization. Several experiments have demonstrated the effectiveness of decoupling features in space to improve the performance and convergence speed of object detection [31,48–50].

In this contribution, we introduce the decoupled head proposed in YOLOX [51], which is a prediction head that improves the detection accuracy of the model and accelerates model convergence. The structure of the decoupled head is shown in Figure 6. Specifically, the channels are first reduced to 256 dimensions for each input feature layer. Then, the classification and localization subtasks of the model are performed in two parallel branches

using 3 × 3 convolutional layers, whereas the IoU computation branch is implemented in parallel in the regression branch.

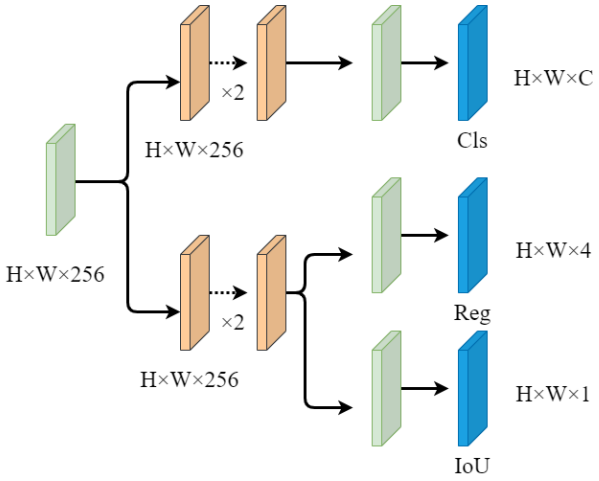


**Figure 6.** Decoupled head structure.

### *3.5. Loss Function*

The MFICDet loss function contains localization loss and classification loss. Complete-IoU (CIOU) [52] is used as the localization loss function. CIOU further optimizes distance-IoU (DIOU) [52] considering the overlap area, centroid distance, and aspect ratio between the prediction and truth boxes to improve the accuracy and stability of the regression. Specifically, the penalty term of CIOU introduces an impact factor, $\alpha v$, which is based on DIOU. The penalty term is expressed as follows:

$$R_{CIOU} = \frac{\rho^2(b, b^{gt})}{C^2} + \alpha v \tag{3}$$

where $\rho^2(b, b^{gt})$ represents the Euclidean distance between the center point of the prediction box and that of the truth box. $C$ denotes the diagonal distance of the smallest closed area that encompasses both the prediction and the truth boxes. In $\alpha v$, $\alpha$ is the parameter used for balancing and $v$ is used to measure the consistency of the prediction box aspect ratio. The formulas are as follows:

$$\alpha = \frac{v}{(1 - Iou) + v} \tag{4}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{5}$$

The overall formula for the CIOU loss function is:

$$L_{CIOU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{C^2} + \alpha v \tag{6}$$

In [31], it was pointed out that one-stage object detection methods suffer from severe category imbalances during the training process, leading to inefficient training because easy-to-classify useless information dominates the gradient. In addition, negative samples can guide the training process and lead to model degradation. Therefore, in this study, focal loss is used to calculate classification loss. The formula to calculate the focal loss is as follows:

$$L_{conf} = FL(P_t) = -\alpha(1 - P_t)^{\gamma} \log(p_t) \tag{7}$$

where $P_t$ represents the probability that the sample belongs to positive samples, $(1 - P_t)^{\gamma}$ is an adjustment factor introduced based on a balanced cross-entropy loss function, and $\alpha$ controls the weight of the positive and negative sample contributions to the loss.

The focal loss function effectively improves the problem of training instability caused by the imbalance between positive and negative samples and reduces the influence of easy-to-classify samples that dominate the gradient descent during model training.

## 4. Experiments and Analysis

### *4.1. Data Introduction*

To evaluate the effectiveness of MFICDet, experiments are conducted on two widely used benchmark datasets for remote sensing image object detection: DIOR [53] and NWPU VHR-10 [54]. The NWPU VHR-10 dataset contains 10 categories with 650 annotated images; herein, this dataset is divided into training, validation, and testing sets in a ratio of 6:2:2.

DIOR is one of the largest and most diverse open-source datasets in remote sensing image object detection. The dataset contains 23,463 images covering 20 common categories: aircraft, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, harbor, ship, golf course, ground track field, overpass, stadium, storage tank, tennis court, train station, vehicle, and windmill. The training set contained 5862 images, the validation set contained 5863 images, and the remaining 11,738 images are used as the testing set. A schematic of each category is shown in Figure 7.


**Figure 7.** Examples of DIOR dataset categories.

### *4.2. Evaluation Metrics*

In this study, four commonly used metrics are employed to evaluate the performance of the method: precision, recall, harmonic mean ($F_1$), and mean average precision ($mAP$). Precision refers to the ratio of the number of correctly detected positive samples to the number of all predicted positive samples among all objects predicted by the model on the test dataset. Recall reflects the probability that positive samples are correctly identified among all detection results and measures the false detection of true objects by the detector. The formulas are as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{8}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{9}$$

where $TP$ represents the number of samples correctly classified as positive, $FN$ represents the number of samples incorrectly classified as negative, and $FP$ represents the number of samples incorrectly classified as positive.

The above precision and recall metrics are contradictory in practice, as reflected in recall being usually lower when precision is high and vice versa. Therefore, $F_1$ is proposed, which combines the two metrics as follows:

$$F_1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

The $mAP$ is a comprehensive metric for evaluating model performance in object detection and represents the average of all categories of average precision ($AP$). The $AP$ values for each category are calculated using the area under the precision–recall ($PR$) curve, composed of precision and recall using the following formulas:

$$AP_i = \int_0^1 P_i(R_i) dR_i = \sum_{k=0}^{n} P_i(k) \Delta R_i(k) \tag{11}$$

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_i \tag{12}$$

### *4.3. Training Details*

All experiments in this study are performed using PyTorch architecture, and the hardware environment for training and testing is NVIDIA GeForce RTX 3090. The mosaic data enhancement strategy is used to improve the data diversity, and the regularization method of label smoothing is effectively applied to suppress model overfit during the training phase. The parameters of the backbone network of this model are initialized using the pre-weights of YOLOv4. The Adam optimizer combines the advantages of the adaptive learning rate gradient descent algorithm with those of the momentum gradient descent algorithm. Therefore, applicability to sparse gradients is achieved, and the gradient oscillation problem can be effectively improved. Moreover, in this study, an Adam optimizer with a weight decay of $1 \times 10^{-5}$ is adopted. The initial learning rate is set to 0.001, and the learning rate decay strategy is implemented using a combination of cosine annealing and equal-interval methods.

### *4.4. Ablation Experiments*

Detailed ablation experiments are performed on the DIOR and NWPU VHR-10 datasets to evaluate the effectiveness of different modules. Our ablation experiments use CSPDarknet53 as the backbone network, and the same pretrained model is used for the initialization of the backbone network. The other parameters are set as initial values by fitting a normal distribution. We introduce the feature pyramid strategy as a baseline based

on the CSPDarknet53 backbone network. The effectiveness of each module is compared separately using the control-variable method.

(1) Ablation Experiments on the DIOR Dataset

Impact of the PNFG module: As shown in Table 2, the overall *mAP* of the model is improved by 1.21% using the PNFG strategy. The results show PNFG can effectively promote the network to learn the features of the object of interest and weaken the negative influence of background noise information on the model. The spatial and channel attention mechanisms can explore the location of the object and the semantic content of the feature information stream. Figure 8 shows the AP of each category for the different models, highlighting a substantial improvement in accuracy for objects with high feature similarity or severe background interference. Examples include bridges and dams, bridges, and overpasses, expressway toll stations and service areas, baseball fields, and aircraft. We have demonstrated that the PNFG strategy enables the model to effectively distinguish object features from background noise in the feature response region, suppress background noise interference, and adaptively locate the spatial position where the object is located.

**Table 2.** Results of ablation experiments on the DIOR dataset.

| Model | *Recall* | *Precision* | $mF_1$ | *mAP* |
|---|---|---|---|---|
| Baseline | 56.51 | 87.61 | 67.50 | 66.17 |
| Baseline + DHead | 57.33 | 89.98 | 68.45 | 66.77 |
| Baseline + GFIC + DHead | 63.18 | 87.58 | 72.70 | 70.87 |
| MFICDet | 62.77 | 88.55 | 72.65 | 72.08 |

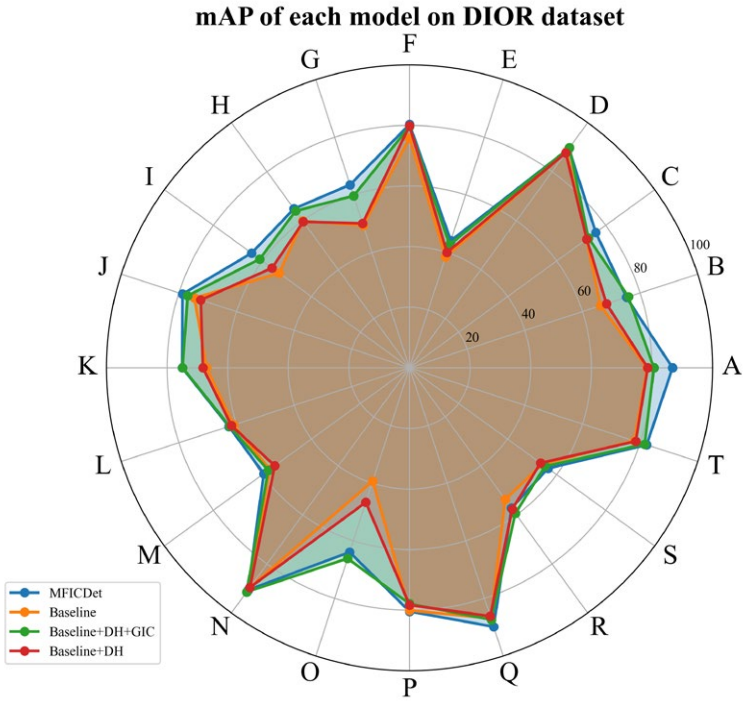


**Figure 8.** *mAP* of each model for the DIOR dataset. Note: A–T represent the following categories: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill, respectively.

Impact of the GFIC module: The GFIC module deeply fuses the global multi-scale feature information and improves the *mAP* from 66.77% to 70.87%. For the classification task, the model focuses on whether a feature appears without focusing on the specific

location of its appearance. The pooling operation desensitizes the network to the location of the object, obtaining a powerful prior. Dilated convolution complements the internal data structure during upsampling to avoid the loss of spatially hierarchical information. This obtains a larger receptive field while maintaining a lower number of parameters. The multi-scale feature maps obtained using different dilation factors contain more detailed features. As shown in Figure 8, this module offers a relatively large improvement for objects with complex geometric and radiometric features. For example, the internal features of the stadium and the ground track field are almost the same, and only the edge features are different. Moreover, there are large variations in size between the two images. Therefore, the advantages of a larger receptive field of the GFIC module and the ability to cope with different scales of the same object are exploited. The improvement in the accuracy of objects such as expressway toll stations and dams proves that the GFIC module has a powerful detail-capturing capability. Furthermore, with the addition of detailed information, the model has better resistance to the interclass similarity and intraclass diversity of objects.

Impact of the DHead: To verify the effect of decoupling the two tasks of classification and localization, we conduct comparison experiments of decoupled and shared heads separately. As shown in Table 2, this module improves *mAP* by only 0.6%. However, the model obtains the highest precision after introducing the decoupled head at the baseline. As shown in Figure 8, the model with the introduction of the decoupled head exhibits a substantial performance improvement for objects with diverse and fuzzy boundaries, such as airports, train stations, and stadiums. Therefore, decoupling the classification and localization tasks allows the model to perform better feature alignment; thus, the boundary localization accuracy of the object is improved.

In this study, the log-average miss rate is used to evaluate the impact of the different modules of the detector on the object discovery ability; this is calculated using the false positive per image (FPPI) as the horizontal axis and the logarithm of the miss rate as the vertical axis of the curve. Specifically, nine FPPI ratios between 0.01 and 1 are uniformly selected in the logarithmic space, the logarithm of the corresponding miss detection rate is obtained and averaged, and the percentage is reduced by the exponential operation as the evaluation index.

As shown in Figure 9d, the leakage detection of all object categories except the ship, train station, storage tank, and stadium, is improved after adding the PNFG modules. The ship is closely associated with the harbor, and the appearance of the docked ship is small compared to the wharf. The introduction of the PNFG strategy resulted in the harbor being modeled with excessively weak background information, thus weakening the features of the ship. The appearance of the train station is like other buildings, and there is a great similarity between the features of the object and the background; this resulted in a contradiction during model learning, causing the missed detection of the train station. As shown in Figure 9c, the GFIC module enhances the discovery capability of the network for all objects because of its powerful deep-feature mining capability. Therefore, the effectiveness of the information complementation strategy is demonstrated.

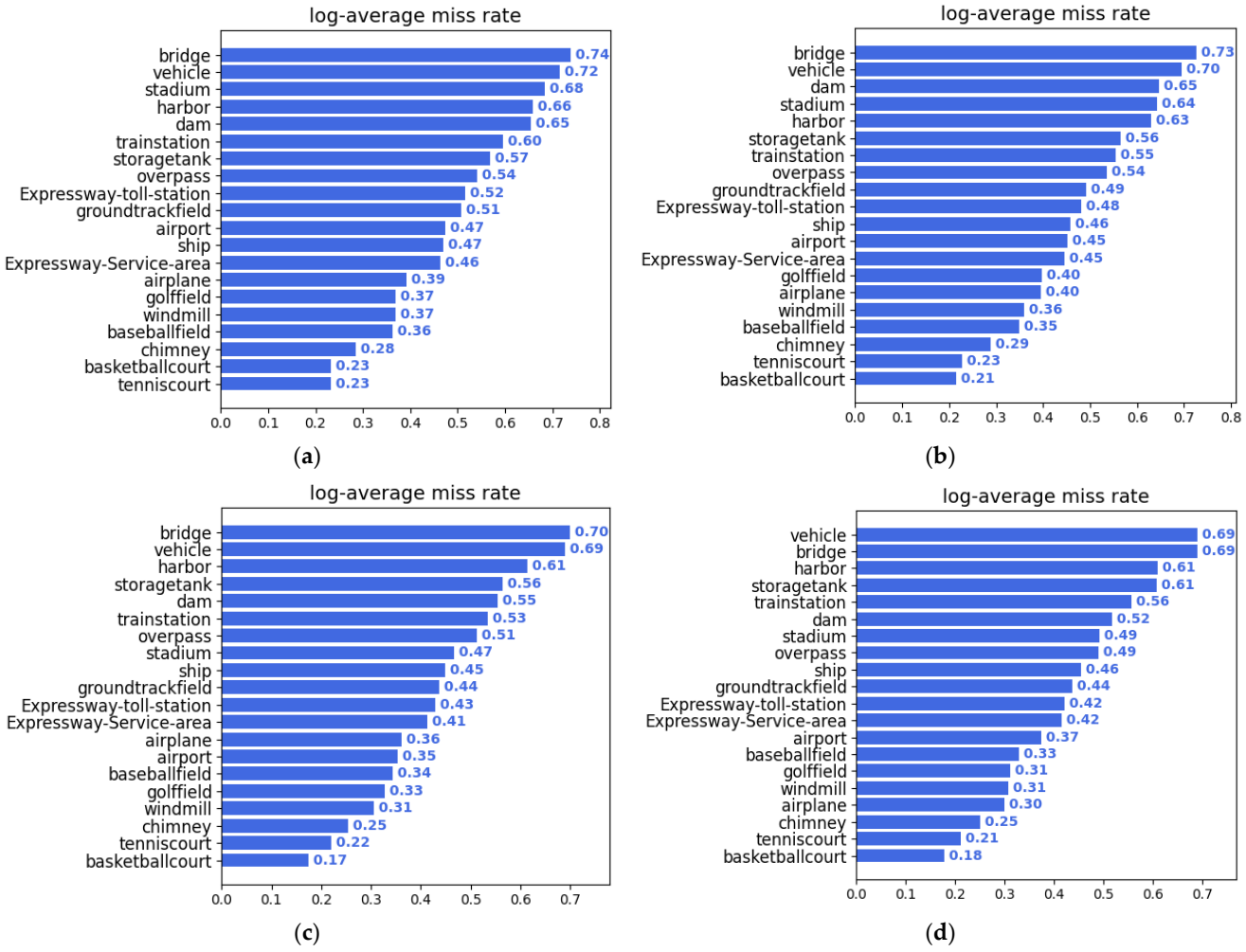


**Figure 9.** Log-average miss rate of each model. (**a**) Baseline; (**b**) Baseline + DHead; (**c**) Baseline + GFIC + DHead; (**d**) MFICDet.

(2) Ablation Experiments on the NWPU VHR-10 Dataset

To demonstrate the generality of MFICDet in the remote sensing image object detection task, the ablation experiment is repeated on the NWPU VHR-10 dataset. The results are presented in Table 3, where the evidence shows that *mAP* improves by 1.03% after decoupling the classification and regression tasks in space. The GFIC and PNFG proposed in this study improve the *mAP* to 95.35% and 96.41%, respectively. As the designed modules are increased sequentially, both the recall and the mF1 of the model improve steadily, indicating that each module remains valid for different datasets. However, when the baseline detection head is replaced by DHead, the precision decreases. We speculate that this may result from the overfitting of the classification caused by decoupling the classification from the regression, which increases the false detection rate. The results of the ablation experiments on NWPU VHR-10 are like those on the DIOR dataset.

**Table 3.** Results of ablation experiments in the NWPU VHR-10 dataset.

| Model | *Recall* | *Precision* | $mF_1$ | *mAP* |
|---|---|---|---|---|
| Baseline | 87.80 | 90.56 | 88.70 | 92.57 |
| Baseline + DHead | 90.26 | 89.42 | 89.60 | 93.60 |
| Baseline + GFIC + DHead | 93.66 | 91.03 | 92.30 | 95.35 |
| MFICDet | 95.47 | 90.59 | 92.60 | 96.41 |

To visually compare the gains brought to the model by different modules, the PR curves for each category of different ablation experiments on the NWPU VHR-10 dataset are visualized; the results are shown in Figure 10, where Model_1–Model_4 correspond to Baseline, +DHead, +GFIC, and MFICDet, respectively.

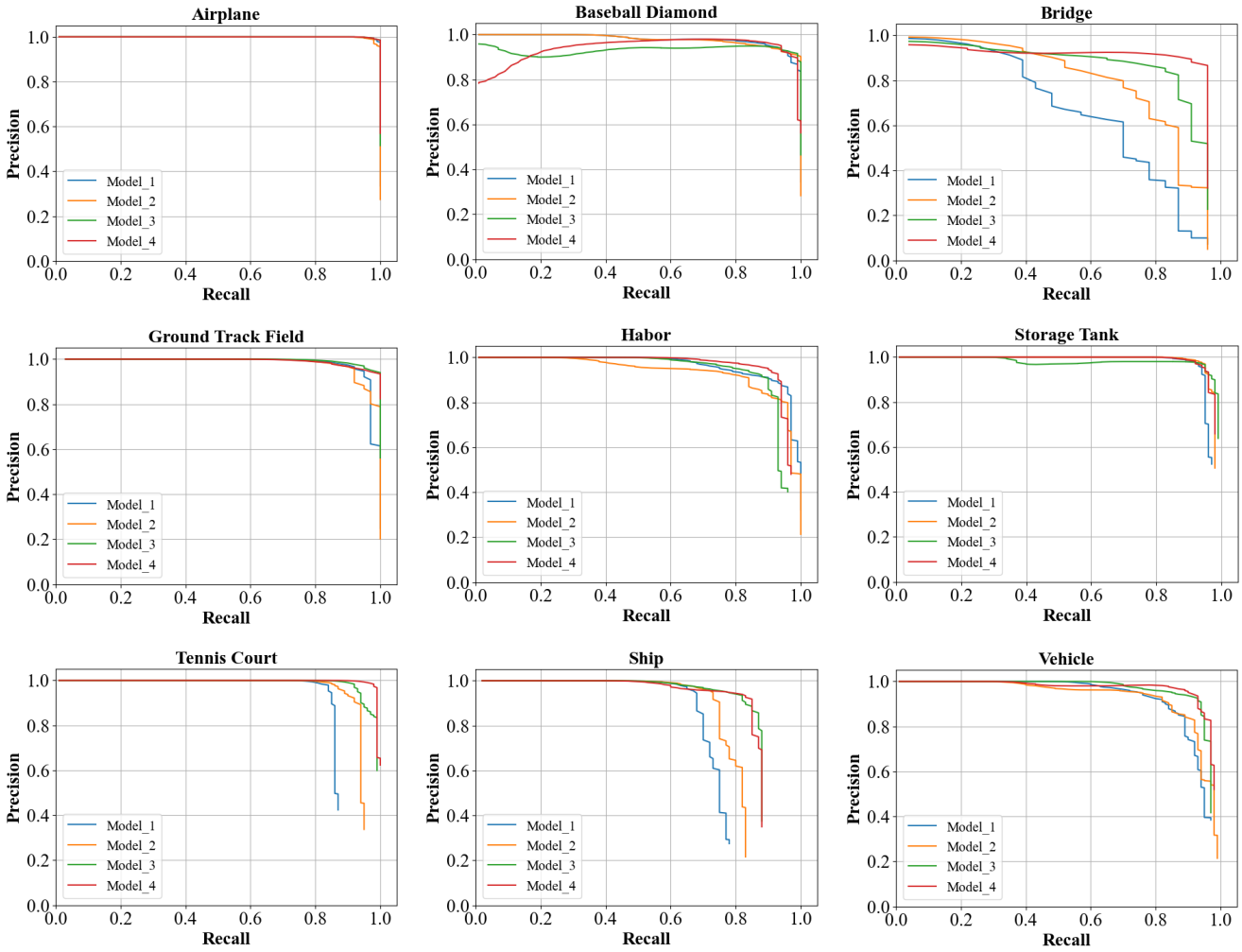


**Figure 10.** Visualization of *PR* curves for each model.

### *4.5. Quantitative Comparison and Analysis*

Comparison experiments are conducted on two commonly used remote sensing image datasets to verify the progress of the proposed model. Widely used and recently proposed models that include not only general object detection methods but also models designed for remote sensing image object detection are selected as comparison groups. These methods contain three common object detection architectures: one-stage and two-stage, as well as anchor-free.

(1) Comparison and Analysis Using the DIOR Dataset

The most challenging DIOR dataset for remote sensing image object detection is used to test the proposed model in detail. Comparison experiments are conducted with the classical general object detection model and the latest remote sensing image object detection model; the results are shown in Table 4. First, widely used, general-purpose object detectors without solutions for remote sensing image characteristics cannot obtain satisfactory results on the complex and variable DIOR dataset. For example, Faster-RCNN is a two-stage model that lacks an effective feature enhancement strategy, such as multi-scale feature fusion. The features of multi-scale objects cannot be modeled effectively,

and thus the detection accuracy is lower. The YOLOv4 one-stage detector model has multi-scale feature fusion capability; however, its feature enhancement method, by stacking multiple convolutional layers, only focuses on the features of the object, so it cannot achieve advanced performance when facing remote sensing images with complex backgrounds. This proves that there are still shortcomings in general object detection models directly applied to remote sensing images.

Furthermore, our proposed approach is compared with advanced remote sensing image object detection models (i.e., CF2PN [39], FENet [55], ASSD [35], and CSFF [56]). As shown in Table 4, the model proposed in this study obtains the highest detection accuracy because the PNFG module and the GFIC module effectively solve the problems of complex background information and the multi-scale variation of the object. CF2PN uses only the loss function to solve the problem of complex backgrounds within remote sensing images, leading to poor results. Our model achieved a 4.83% increase in *mAP* compared to the CF2PN model. FENet uses an attention mechanism to enhance object features and introduces contextual feature enhancement methods to achieve advanced performance in two-stage detectors. The accuracy of the one-stage detector proposed in this study is considerably better than a two-stage detector.

As seen in the last four rows in Table 4, anchor-free detectors suffer from semantic ambiguity because of the lack of more effective coupled semantic relations. Therefore, such detectors perform poorly with complex datasets such as DIOR. The recently proposed MSFC-Net [57] incorporates a composite semantic feature fusion method to handle complex scenes in remote sensing images and achieves excellent performance in anchor-free detectors. However, the *mAP* of MFICDet is still higher than MSFC-Net, although MSFC-Net expands the DIOR dataset using multiple data enhancement methods.

**Table 4.** Accuracy comparison of our proposed model with other advanced models available, using the DIOR dataset.

| Model | *mAP* | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Faster R-CNN [3] | 63.10 | 54.10 | 71.40 | 63.30 | 81.00 | 42.60 | 72.50 | 57.50 | 68.70 | 62.10 | 73.10 | 76.50 | 42.80 | 56.00 | 71.80 | 57.00 | 63.50 | 81.20 | 53.00 | 43.10 | 80.90 |
| YOLOv4 [12] | 66.71 | 75.27 | 69.95 | 70.95 | 88.78 | 39.99 | 76.61 | 54.02 | 59.94 | 60.65 | 67.68 | 70.15 | 58.76 | 57.34 | 87.71 | 50.21 | 75.66 | 86.58 | 52.62 | 52.74 | 78.62 |
| SSD [5] | 58.60 | 59.50 | 72.70 | 72.40 | 75.70 | 29.70 | 65.80 | 56.60 | 63.50 | 53.10 | 65.30 | 68.60 | 49.40 | 48.10 | 59.20 | 61.00 | 46.60 | 76.30 | 55.10 | 27.40 | 65.70 |
| CF2PN [39] | 67.25 | 78.32 | 78.29 | **76.48** | 88.4 | 37.00 | 70.95 | 59.9 | 71.23 | 51.15 | 75.55 | 77.14 | 56.75 | 58.65 | 76.06 | 70.61 | 55.52 | 88.84 | 50.83 | 36.89 | 86.36 |
| FENet [55] | 68.30 | 54.10 | 78.20 | 71.60 | 81.00 | **46.50** | 79.00 | **65.20** | 76.50 | 69.60 | 79.10 | **82.20** | 52.00 | 57.60 | 71.90 | 71.80 | 62.30 | 81.20 | 61.20 | 43.30 | 81.20 |
| ASSD [35] | 71.10 | 85.60 | 82.40 | 75.80 | 89.50 | 40.70 | 77.60 | 64.70 | 67.10 | 61.70 | **80.80** | 78.60 | 62.00 | 58.00 | 84.90 | 65.30 | 65.30 | 87.90 | **62.40** | 44.50 | 76.30 |
| CSFF [56] | 68.00 | 57.20 | 79.60 | 70.10 | 87.40 | 46.10 | 76.60 | 62.70 | **82.60** | 73.20 | 78.20 | 81.60 | 50.70 | 59.50 | 73.30 | 63.40 | 58.50 | 85.90 | 61.90 | 42.90 | **86.90** |
| CornerNet [41] | 64.90 | 58.80 | **84.20** | 72.00 | 80.80 | 46.40 | 75.30 | 64.30 | 81.60 | 76.30 | 79.50 | 79.50 | 26.10 | **60.60** | 37.60 | 70.70 | 45.20 | 84.00 | 57.10 | 43.00 | 75.90 |
| AOPG [58] | 64.41 | 62.39 | 37.79 | 71.62 | 87.63 | 40.90 | 72.47 | 31.08 | 65.42 | **77.99** | 73.20 | 81.94 | 42.32 | 54.45 | 81.17 | **72.69** | 71.31 | 81.49 | 60.04 | 52.38 | 69.99 |
| $O^2$-DNet [37] | 68.40 | 61.20 | 80.10 | 73.70 | 81.40 | 45.20 | 75.80 | 64.80 | 81.20 | 76.50 | 79.50 | 79.70 | 47.20 | 59.30 | 72.60 | 70.50 | 53.70 | 82.60 | 55.90 | 49.10 | 77.80 |
| MSFC [57] | 70.08 | 85.84 | 76.24 | 74.38 | **90.10** | 44.15 | 78.12 | 55.51 | 60.92 | 59.53 | 76.92 | 73.68 | 49.55 | 57.24 | 89.62 | 69.21 | 76.52 | 86.74 | 51.82 | 55.23 | 84.31 |
| Our | **72.08** | **86.78** | 75.28 | 75.96 | 89.46 | 44.13 | **80.33** | 63.53 | 64.88 | 64.40 | 78.76 | 75.01 | **62.67** | 59.45 | **90.65** | 63.97 | **80.41** | **89.86** | 57.22 | **56.49** | 82.30 |

Note: The bold font in the table represents the optimal value. A–T represent the following categories: 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.

(2) Comparison and Analysis Using the NWPU VHR-10 Dataset

Generalization experiments are conducted on the NWPU VHR-10 dataset to verify the generality of the proposed model for the remote sensing image object detection task. Comparison experiments are also conducted with the latest proposed remote sensing image object detectors, as shown in Table 5. Compared with the latest models, the model in this study achieves higher detection accuracy. ABNet is an excellent detector based on Faster R-CNN improvements; however, our model achieves a 2.2% higher *mAP* compared with ABNet. SMENet and MSGNet [59] are the latest high-precision one-stage detectors proposed to solve the problems of complex backgrounds and diverse object scales in remote sensing images. As shown in Table 4, the model in this study obtains a higher accuracy than both SMENet and MSGNet. MPFPNet [60] is the latest weakly supervised detector proposed for multi-scale objects, and the model in this study achieves a 1.84% higher *mAP* compared with it. Furthermore, our model achieves the highest detection accuracy for all four categories. The above results are similar to those obtained using the DIOR dataset.

These results demonstrate the effectiveness and generality of the proposed method for multicategory remote sensing image object detection tasks.

**Table 5.** Accuracy comparison of our proposed model with other advanced models available, using the NWPU VHR-10 dataset.

| Model | *mAP* | Airplane | Basketball | Bridge | Ground | Harbor | Ship | Storage | Tennis | Vehicle | Baseball |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Yolov4 | 90.39 | 99.93 | 95.73 | 69.79 | 99.26 | 93.25 | 75.98 | 97.88 | 84.24 | 90.16 | 97.72 |
| ABNet [9] | 94.21 | 100 | 95.98 | 69.04 | 99.86 | 94.26 | 92.58 | 97.77 | 99.26 | 95.62 | 97.76 |
| SMENet [11] | 95.64 | 99.06 | 98.56 | 99.06 | 100 | 93.98 | 95.65 | 91.92 | 98.15 | 81.28 | 98.76 |
| MPFPNet [60] | 94.57 | 99.84 | 91.69 | 92.30 | 99.73 | 94.82 | 92.63 | 96.98 | 89.83 | 89.15 | 98.49 |
| MSGNet [59] | 95.53 | 98.93 | 92.02 | 91.07 | 99.98 | 99.09 | 93.68 | 97.90 | 91.82 | 92.22 | 98.60 |
| MRNet [28] | 92.50 | 99.50 | 95.40 | 82.20 | 99.20 | 98.60 | 88.40 | 90.20 | 89.20 | 92.90 | 98.70 |
| EVCP [61] | 94.10 | 98.80 | 91.60 | 87.80 | 99.70 | 91.80 | 92.50 | 99.80 | 91.10 | 88.60 | 99.80 |
| Our | 96.41 | 99.99 | 99.99 | 92.13 | 99.62 | 95.17 | 86.43 | 97.86 | 99.62 | 95.90 | 97.42 |

### *4.6. Visualization*

To verify the ability of the proposed model to perceive object features more visually, the predicted feature maps of the model for several typical objects are visualized; the results are shown in Figure 11. The transition from blue to red indicates an increase in the sensitivity of the model. The visualization results show that the method proposed in this study can better adapt to various optical remote sensing image objects in multi-scale, irregular aspect ratio, and complex background environments.

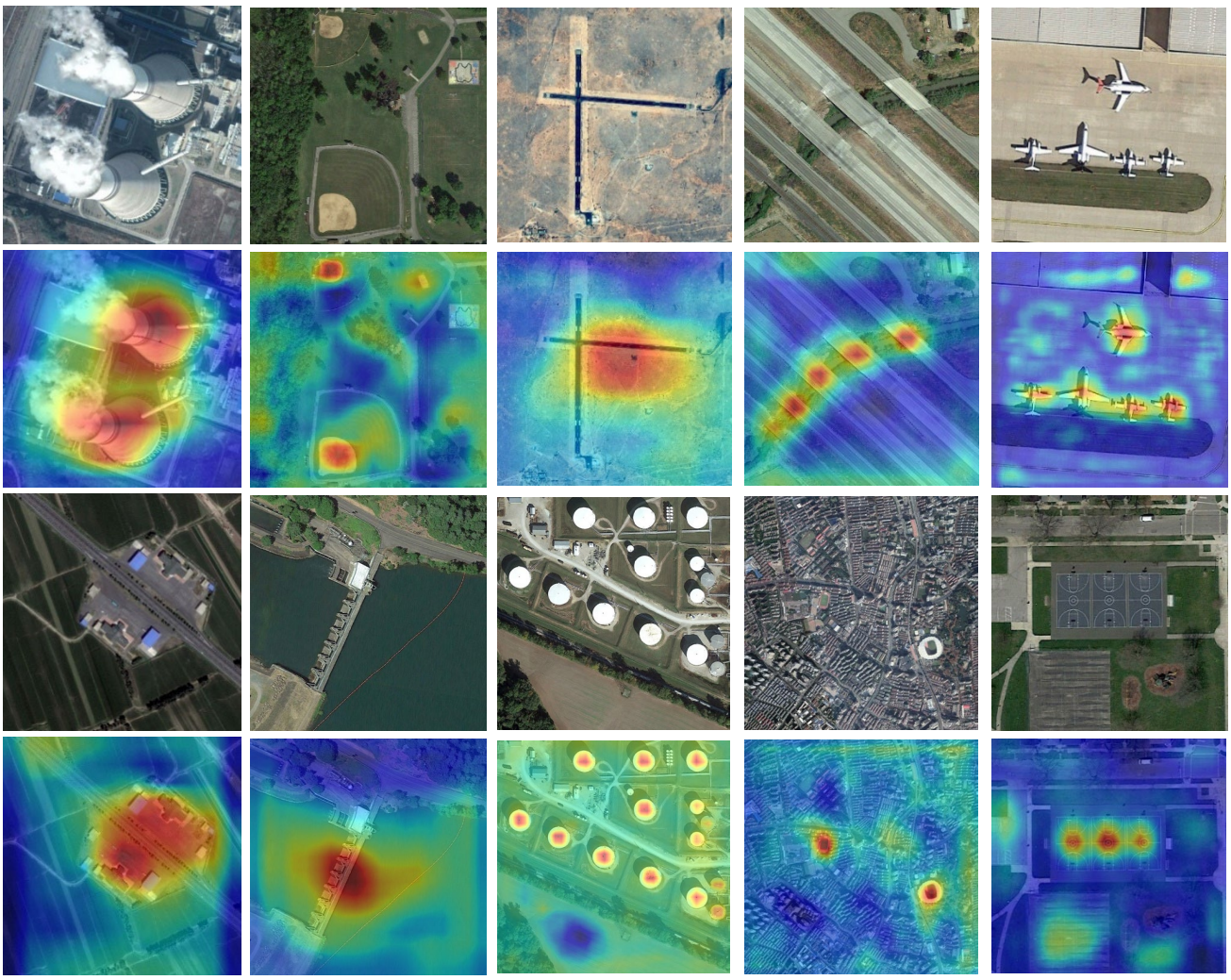

**Figure 11.** Visualization of typical object features.

For example, there is a substantial scale difference between the chimney and the oil tank, but the model can accurately locate both. This result indicates that the model has excellent multi-scale detection capability. Airports typically have irregular aspect ratios. The feature heat map indicates that the model in this work can accurately identify the center of the airport, which verifies the adaptability of the model to objects with irregular aspect ratios. The athletic field is a complex environment with strong background interference, but the model accurately recalls it, demonstrating the strength of the detector in processing complex background information interference.

Remote sensing images usually have a large field of view, resulting in objects with large differences in the same image. In this study, this image is visualized and analyzed. As shown in Figure 12, the proposed method can accurately respond to images with different object sizes and types. Simultaneously, a good coupling between the semantic information of the scene and the object is established; thus, it can be concluded that the model adequately considers the contextual information of the region of interest.

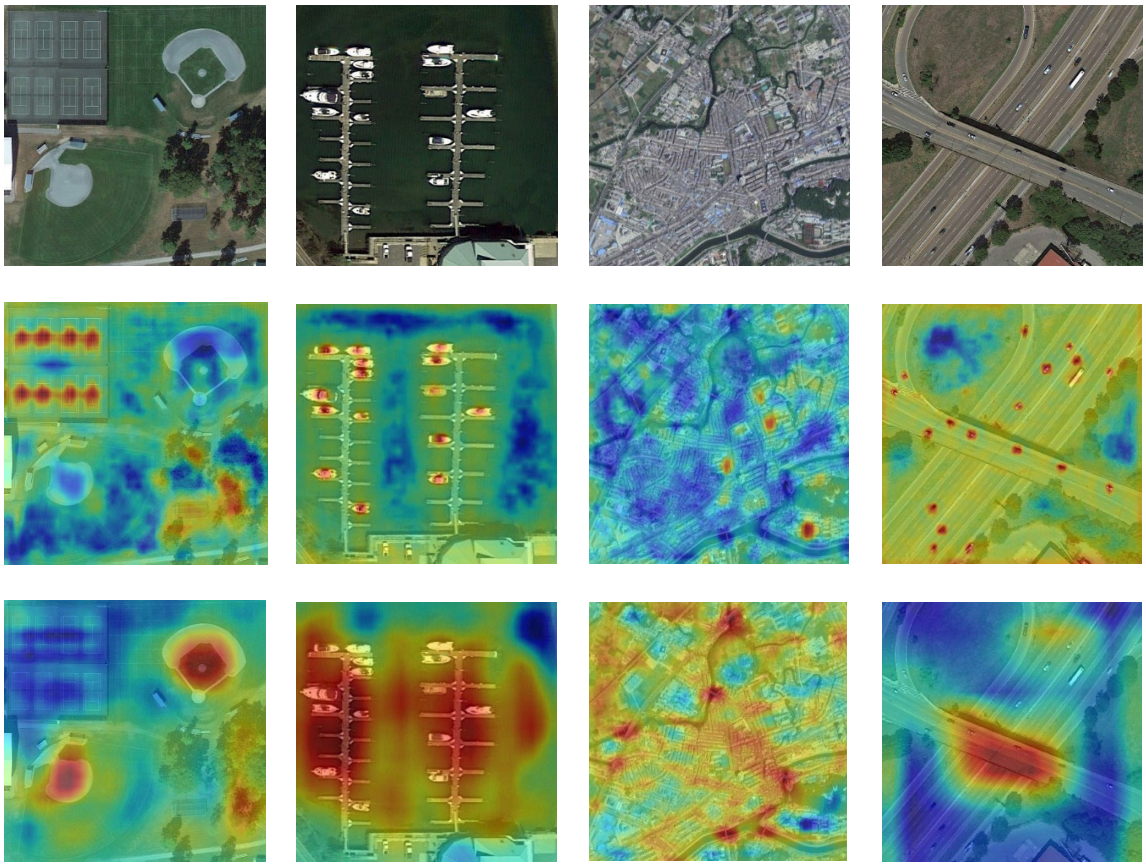

**Figure 12.** Visualization of objects with different scales in the same image.

Figure 13 shows a visualization of the representative test results. The large gap in the spatial resolution of different sensors in remote sensing images leads to large-scale differences and different background interference information for the same object, which poses a great challenge to the detector. As seen in Figure 13a, the method proposed in this study can achieve excellent detection results for objects with large-scale variations, which indicates that the proposed network has a strong generalization capability for object-scale variations. Another challenge of remote sensing images is strong interclass interference. Differences in the size and shape of objects of the same class are not negligible. As shown in Figure 13b, MFICDet can effectively distinguish bridges from dams, which indicates that the method proposed in this study can effectively extract the salient detailed features of an object and has a certain resistance to interclass interference. The detector also accurately

identified chimneys with large differences in appearance, which proves that MFICDet can accurately model the feature space of an object. Remote sensing imaging can be affected by numerous factors, such as sensor height, radiation resolution, atmospheric conditions, imaging season, and ground radiation. Therefore, the quality of images acquired under different imaging conditions will vary. Better robustness to image quality is also an important measure of network performance. Figure 13c,d shows that the method proposed in this study is highly robust to image quality and can resist the interference of different imaging conditions.

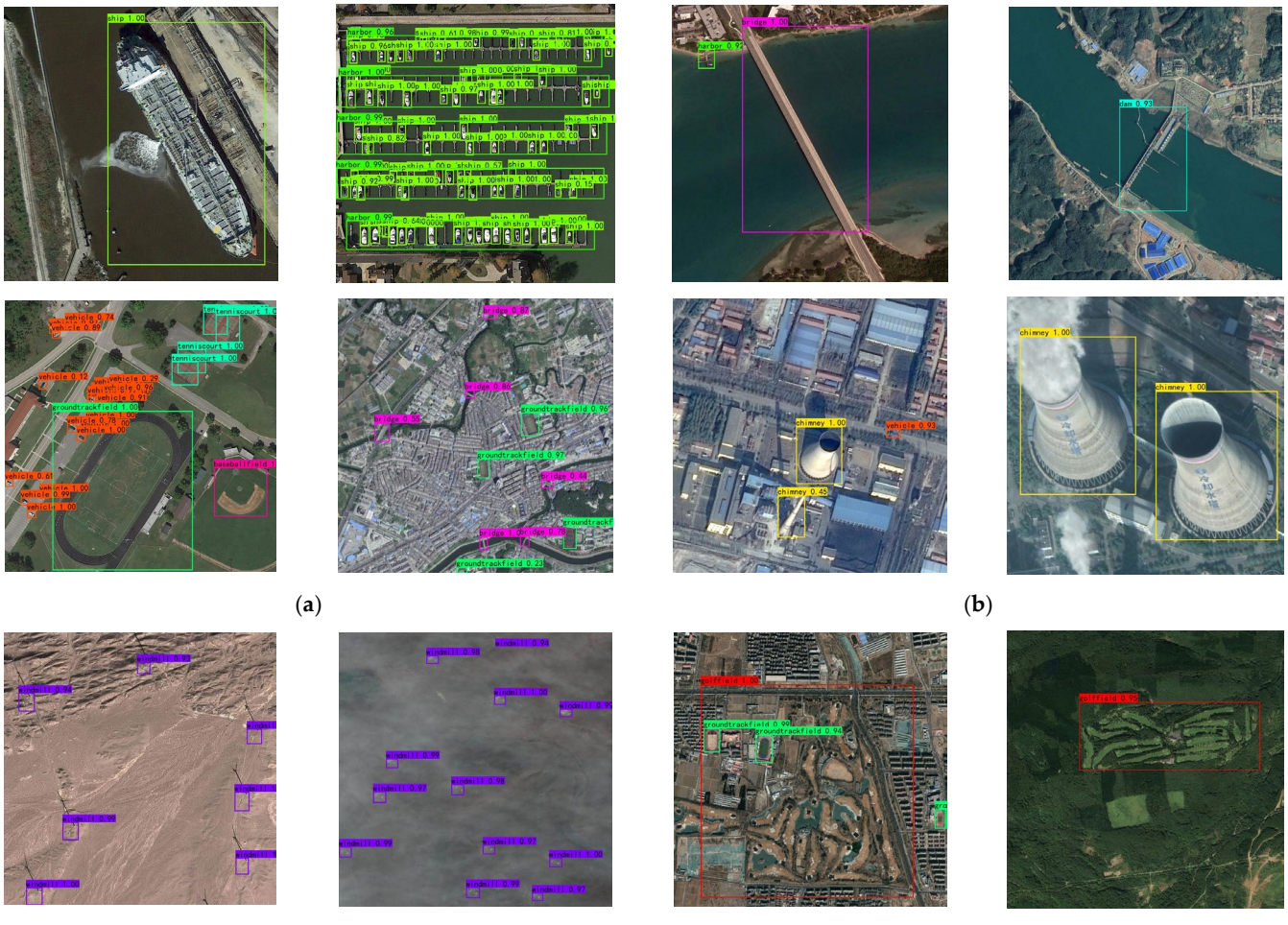

**Figure 13.** Detection results for diverse objects within the DIOR dataset. (**a**) Different scales for the same type of object; (**b**) Interclass similarity and intraclass variation; (**c**) Various imaging conditions; (**d**) Different seasonal images.

## 5. Discussion

### *5.1. Limitations*

We demonstrated the effectiveness and advancement of our method using extensive experiments in Section 4. During the experiments, we discovered the shortcomings of our detector. In scenes where the objects are extremely similar to the background, the detector cannot maintain advanced performance. We analyzed the performance degradation caused by the failure of the PNFG module in such scenarios. The reason for that is the lack of targeted samples in this scene during the training process, resulting in a number of images with significant differences dominating the model. Therefore, when the object is extremely similar to the background, the PNFG module may incorrectly confuse the object feature with the

background feature. In addition, our detector is more computationally intensive, although it obtains strong performance. This limits its deployment in lightweight mobile devices.

### *5.2. Future Works*

Our research will focus on maintaining the current performance of the detector while enabling it to be deployed on edge computers, which is critical to the practical application of the detector. For example, drones are widely used. However, limited by their size and workload, only lightweight computing devices can be deployed. Moreover, based on the limitations of the detector proposed in this study, camouflaged object detection is also one of the directions of our research.

## 6. Conclusions

In this study, we aimed to solve the challenges posed by the complex background and large variation in object scales in remote sensing image object detection. A one-stage object detection model based on the concept of divide and conquer was proposed: a multi-feature information complementary detector. First, we proposed a PNFG module for the complex background problem of remote sensing images. The module was used to extract features from coupled feature information beneficial for the detection and suppression of invalid noise information. Subsequently, a GFIC module was proposed to solve the problem of the large-scale variation of objects, which combines pooling operations with dilated convolution. The abstract features were compressed by a pooling operation to improve the resistance of the model to object translation and rotation. Detailed features lost during the pooling operation were complemented using dilated convolution while increasing the receptive field of the model. In addition, we proposed a dual multi-scale feature fusion strategy to solve the problem of low detection accuracy caused by multi-scale objects in the same image. Extensive experiments were conducted on two remote sensing image datasets, DIOR and NWPU VHR-10. The results showed that the detector proposed in this work overcomes the effect of the complex background on performance. Moreover, the challenges posed by multi-scale objects, especially multi-scale objects in the same image, are solved. Compared with other models, the model proposed here achieved state-of-the-art performance. However, this study did not consider the operational efficiency of the model, resulting in a relatively expensive computation. Future research should focus on improving the model's detection speed while maintaining its current performance.

**Author Contributions:** Conceptualization, Z.G.; methodology, J.W.; software, J.W. and D.Y.; formal analysis, X.L. and Y.L.; writing—original draft preparation, J.W.; writing—review and editing, J.L.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Science Foundation of China, Grant number 41876105; 41671410.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest.

## References

1. Zhang, X.; Zhou, Y.N.; Luo, J. Deep learning for processing and analysis of remote sensing big data: A technical review. *Big Earth Data.* **2021**, 1–34. [CrossRef]
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, USA, 7–12 December 2015; Volume 28.
4. 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.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.

6. Liu, J.; Yang, D.; Hu, F. Multiscale Object Detection in Remote Sensing Images Combined with Multi-Receptive-Field Features and Relation-Connected Attention. *Remote Sens.* **2022**, *14*, 427. [CrossRef]
7. Bai, J.; Ren, J.; Yang, Y.; Xiao, Z.; Yu, W.; Havyarimana, V.; Jiao, L. Object Detection in Large-Scale Remote-Sensing Images Based on Time-Frequency Analysis and Feature Optimization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
8. Cheng, B.; Li, Z.; Xu, B.; Dang, C.; Deng, J. Target Detection in Remote Sensing Image Based on Object-and-Scene Context Constrained CNN. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
9. Liu, Y.; Li, Q.; Yuan, Y.; Du, Q.; Wang, Q. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
10. Wu, Y.; Zhang, K.; Wang, J.; Wang, Y.; Wang, Q.; Li, X. GCWNet: A Global Context-Weaving Network for Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
11. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature Split–Merge–Enhancement Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
12. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
13. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
14. Shi, L.; Tang, Z.; Wang, T.; Xu, X.; Liu, J.; Zhang, J. Aircraft detection in remote sensing images based on deconvolution and position attention. *Int. J. Remote Sens.* **2021**, *42*, 4241–4260. [CrossRef]
15. Cheng, B.; Li, Z.; Xu, B.; Yao, X.; Ding, Z.; Qin, T. Structured Object-Level Relational Reasoning CNN-Based Target Detection Algorithm in a Remote Sensing Image. *Remote Sens.* **2021**, *13*, 281. [CrossRef]
16. Song, Z.; Sui, H.; Hua, L. A hierarchical object detection method in large-scale optical remote sensing satellite imagery using saliency detection and CNN. *Int. J. Remote Sens.* **2021**, *42*, 2827–2847. [CrossRef]
17. Hou, L.; Lu, K.; Xue, J.; Hao, L. Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
18. Yang, F.; Li, W.; Hu, H.; Li, W.; Wang, P. Multi-Scale Feature Integrated Attention-Based Rotation Network for Object Detection in VHR Aerial Images. *Sensors* **2020**, *20*, 1686. [CrossRef] [PubMed]
19. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: Piscataway, NJ, USA, 2022.
20. Yu, D.; Ji, S. A New Spatial-Oriented Object Detection Framework for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
21. Zhang, T.; Zhang, X.; Zhu, P.; Chen, P.; Tang, X.; Li, C.; Jiao, L. Foreground Refinement Network for Rotated Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
22. Wang, J.; He, X.; Faming, S.; Lu, G.; Jiang, Q.; Hu, R. Multi-Size Object Detection in Large Scene Remote Sensing Images Under Dual Attention Mechanism. *IEEE Access* **2022**, *10*, 8021–8035. [CrossRef]
23. Zhu, D.; Xia, S.; Zhao, J.; Zhou, Y.; Niu, Q.; Yao, R.; Chen, Y. Spatial hierarchy perception and hard samples metric learning for high-resolution remote sensing image object detection. *Appl. Intell.* **2021**, *52*, 3193–3208. [CrossRef]
24. Wang, J.; Wang, Y.; Wu, Y.; Zhang, K.; Wang, Q. FRPNet: A Feature-Reflowing Pyramid Network for Object Detection of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2020**. [CrossRef]
25. Liu, N.; Celik, T.; Li, H.-C. Gated Ladder-Shaped Feature Pyramid Network for Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
26. Cheng, G.; He, M.; Hong, H.; Yao, X.; Qian, X.; Guo, L. Guiding Clean Features for Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
27. Zhou, X.; Shen, K.; Liu, Z.; Gong, C.; Zhang, J.; Yan, C. Edge-Aware Multiscale Feature Integration Network for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
28. Cong, R.; Zhang, Y.; Fang, L.; Li, J.; Zhao, Y.; Kwong, S. RRNet: Relational Reasoning Network With Parallel Multiscale Attention for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
29. Han, W.; Kuerban, A.; Yang, Y.; Huang, Z.; Liu, B.; Gao, J. Multi-Vision Network for Accurate and Real-Time Small Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5. [CrossRef]
30. Zhang, K.; Wu, Y.; Wang, J.; Wang, Y.; Wang, Q. Semantic Context-Aware Network for Multiscale Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
31. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
32. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8311–8320.
33. Li, Y.; Kong, C.; Dai, L.; Chen, X. Single-Stage Detector with Dual Feature Alignment for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
34. Hou, L.; Lu, K.; Xue, J. Refined One-Stage Oriented Object Detection Method for Remote Sensing Images. *IEEE Trans. Image Process* **2022**, *31*, 1545–1558. [CrossRef]

35. Xu, T.; Sun, X.; Diao, W.; Zhao, L.; Fu, K.; Wang, H. ASSD: Feature Aligned Single-Shot Detection for Multiscale Objects in Aerial Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
36. Huang, Z.; Li, W.; Xia, X.-G.; Wang, H.; Jie, F.; Tao, R. LO-Det: Lightweight Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
37. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented objects as pairs of middle lines. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 268–279. [CrossRef]
38. Liu, N.; Celik, T.; Zhao, T.; Zhang, C.; Li, H.-C. AFDet: Toward More Accurate and Faster Object Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 12557–12568. [CrossRef]
39. Huang, W.; Li, G.; Chen, Q.; Ju, M.; Qu, J. CF2PN: A Cross-Scale Feature Fusion Pyramid Network Based Remote Sensing Target Detection. *Remote Sens.* **2021**, *13*, 847. [CrossRef]
40. 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.* **2021**, *60*, 1–13. [CrossRef]
41. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2019**, *128*, 642–656. [CrossRef]
42. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
43. Mei, H.; Ji, G.P.; Wei, Z.; Yang, X.; Wei, X.; Fan, D.-P. Camouflaged object segmentation with distraction mining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8772–8781.
44. Zeiler, M.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.
45. Woo, S.; Park, J.; Lee, J.; Kweon, I. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
46. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–799.
47. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11563–11572.
48. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9627–9636.
49. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
50. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
51. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
52. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
53. Su, H.; Wei, S.; Yan, M.; Wang, C.; Shi, J.; Zhang, X. Object Detection and Instance Segmentation in Remote Sensing Imagery Based on Precise Mask R-CNN[C]. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1454–1457.
54. 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]
55. Cheng, G.; Lang, C.; Wu, M.; Xie, X.; Yao, X.; Han, J. Feature enhancement network for object detection in optical remote sensing images. *J. Remote Sens.* **2021**, *2021*, 9805389. [CrossRef]
56. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 431–435. [CrossRef]
57. Zhang, T.; Zhuang, Y.; Wang, G.; Dong, S.; Chen, H.; Li, L. Multiscale Semantic Fusion-Guided Fractal Convolutional Object Detection Network for Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–20. [CrossRef]
58. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
59. Zhu, S.; Zhang, J.; Liang, X.; Guo, Q. Multiscale Semantic Guidance Network for Object Detection in VHR Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
60. Shamsolmoali, P.; Chanussot, J.; Zareapoor, M.; Zhou, H.; Yang, J. Multipatch Feature Pyramid Network for Weakly Supervised Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
61. Li, W.T.; Li, L.W.; Li, S.Y.; Mou, J.C.; Hei, Y.Q. Efficient Vertex Coordinate Prediction-Based CSP-Hourglass Net for Object OBB Detection in Remote Sensing. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]