remote sensing

MDPI

Article

AF-EMS Detector: Improve the Multi-Scale Detection Performance of the Anchor-Free Detector

Jiangqiao Yan [1,2,3,4], Liangjin Zhao [1,2], Wenhui Diao [1,2,*], Hongqi Wang [1,2,3,4] and Xian Sun [1,2,3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; yanjiangqiao16@mails.ucas.ac.cn (J.Y.); zhaolj004896@aircas.ac.cn (L.Z.); wanghq@aircas.ac.cn (H.W.); sunxian@mail.ie.ac.cn (X.S.)
[2] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: diaowh@aircas.ac.cn; Tel.: +86-13126730220

Abstract: As a precursor step for computer vision algorithms, object detection plays an important role in various practical application scenarios. With the objects to be detected becoming more complex, the problem of multi-scale object detection has attracted more and more attention, especially in the field of remote sensing detection. Early convolutional neural network detection algorithms are mostly based on artificially preset anchor-boxes to divide different regions in the image, and then obtain the prior position of the target. However, the anchor box is difficult to set reasonably and will cause a large amount of computational redundancy, which affects the generality of the detection model obtained under fixed parameters. In the past two years, anchor-free detection algorithm has achieved remarkable development in the field of detection on natural image. However, there is no sufficient research on how to deal with multi-scale detection more effectively in anchor-free framework and use these detectors on remote sensing images. In this paper, we propose a specific-attention Feature Pyramid Network (FPN) module, which is able to generate a feature pyramid, basing on the characteristics of objects with various sizes. In addition, this pyramid suits multi-scale object detection better. Besides, a scale-aware detection head is proposed which contains a multi-receptive feature fusion module and a size-based feature compensation module. The new anchor-free detector can obtain a more effective multi-scale feature expression. Experiments on challenging datasets show that our approach performs favorably against other methods in terms of the multi-scale object detection performance.

Keywords: remote sensing imagery; anchor free; multi-scale detection; scale-aware feature

Citation: Yan, J.; Zhao, L.; Diao, W.; Wang, H.; Sun, X. AF-EMS Detector: Improve the Multi-Scale Detection Performance of the Anchor-Free Detector. *Remote Sens.* **2021**, *13*, 160. https://doi.org/10.3390/rs13020160

Received: 23 November 2020
Accepted: 1 January 2021
Published: 6 January 2021

1. Introduction

In recent years, with the development of deep learning research, many convolutional neural networks (CNN)-based detection frameworks were proposed and they improve the detection performance in different aspects [1–8]. Although these methods have obtained impressive achievements in various application scenarios, complex scale variations in many real-world scenarios are still a fundamental challenge to achieve satisfied performance, such as ship or plane detection in remote sensing images [9,10], diseased organ detection in medical images, and traffic sign detection [11] of autonomous driving. As the increasing complexity of application scenarios and the continuous improvement of image quality, the scale variations in the image has become more and more obvious, especially in remote sensing images. Therefore, how to effectively deal with the problem of multi-scale object detection is of great significance for expanding the application range of detection models.

According to the approaches used to obtain candidate regions of objects in the image, the CNN-based detection methods can be mainly divided into two types: anchor-based

detection methods and anchor-free methods. The former, such as Faster-RCNN [12], YOLO [13] and SSD [14], predefines anchors or default boxes with specific size and aspect ratios to find the location which is most likely to have an object. While the latter, such as FCOS [15] and RepPoint [16], does not require pre-set anchors or default boxes designed manually. To relieve the multi-scale detection issue, a variety of solutions are proposed under the anchor-based detection framework. It is of critical importance to design good features for multi-scale stimuli for object detection. Thus, when the network solves multi-scale object detection problems, these methods extract better multi-scale features through well-designed network structures, or pre-process input data to simplify the learning of parameters.

Furthermore, the methods of getting better multi-scale detection performance on anchor-based detectors can be roughly divided into the following categories:

1. Methods using data preprocessing, which can be represented by multi-scale training and testing methods and algorithms such as SNIP [17] and SNIPER [18]. These methods leverage multi-scale image pyramids which have been proved useful in both hand-crafted feature-based methods and current CNN-based detection framework [19]. In the data preprocessing stage, these methods scale origin images to different sizes and send them to the network. Input images of different sizes ensure that objects of different sizes acquire sufficient training samples. SNIP and SNIPER improve the training strategy and simplify multi-scale object detection problems. They only detect objects in a specific scale range in input images with various sizes, and solve multi-scale problems without changing the network structure. However, these methods consume more computing and storage resources, and cannot be applied when resources are limited.
2. Methods based on feature layer design, including Feature Pyramid Network (FPN), which is widely used to solve the problem of multi-scale object detection. Also, the method based on feature layer design is widely used in the field of remote sensing object detection. In [20], Zhang et al. indicate that the commonly used ResNet [21] detection backbone structure reduces the spatial resolution of deep feature maps, which makes the information inevitably weaker and affects the detection of small objects. Therefore, they proposed a new DetNet-FPN to effectively enhance the feature representation ability. To obtain better multi-scale feature representation, Fu et al. propose a new feature fusion structure in [5] and use bottom-up or top-down paths to combine different information into each level of feature maps.
3. Methods based on the design and selection of the candidate region, including face detection algorithms such as S3FD [22] and DSFD [23], and some multi-class object detection method used in remote sensing imagery [24,25]. In order to effectively detect targets of different sizes in images, this type of methods promote targets of different sizes to obtain more accurate anchors in the process of network training, and this improves the algorithm's multi-scale detection performance.

In the past two years, anchor-based detection methods that need to set fixed anchors densely have poor migration in different application scenarios; it is difficult to effectively match all objects with larger differences among their sizes. As a result, anchor-free detection algorithms have been researched continuously. In an anchor-free detection framework, we can get the position and size of the target in the image through feature learning at the key points of the object or grouping the key points detected. As we can obtain the final detection result through adaptive learning of the location and size of the candidate region, we do not need to select appropriate hyperparameters for the data distribution in the specific application scenario, which is inevitable for anchor-based methods. It not only can solve the problem that the densely anchor boxes cannot effectively match all targets with large differences in sizes and aspect ratios in complex scenarios, but also can avoid a lot of redundant calculations. However, how to effectively deal with multi-scale object detection problems and how to achieve better detection performance in remote sensing images with anchor-free detection framework still need further study.

In this paper, we use the common anchor-free detector RepPoints Detector (RPDet) as our baseline, and study how to further improve the multi-scale detection performance of the anchor-free detection framework. On the one hand, we analyze and select a more suitable attention mechanism for feature map layers of different sizes, and we define it as Specific-Attention FPN (SA-FPN). On the other hand, we analyze some problems in the structure of original detection head: Under the original framework, the detection head has 3 stacked 3×3 convolutions to process each position on the feature map of different sizes, so as to process the distinguishing features of various-size objects. However, even for objects of different sizes on the same feature map layer, the optimal receptive field size is different. Furthermore, as shown in Figure 1, in anchor-free detection framework, the feature expression of the target in the entire candidate region is represented based on the features of key points, which will be insufficient when the detector is used to deal with large-scale objects. For multi-scale detection task, especially large-scale objects, when the key points are far away from the actual foreground, it is not a good choice to directly use the features on key points to obtain the feature expression of targets. In order to solve the above problems, we design two corresponding modules, a Multi-Receptive Feature Fusion module (MRFF) and a Size-based Feature Compensation module (SBFC), to obtain better multi-scale feature expression.

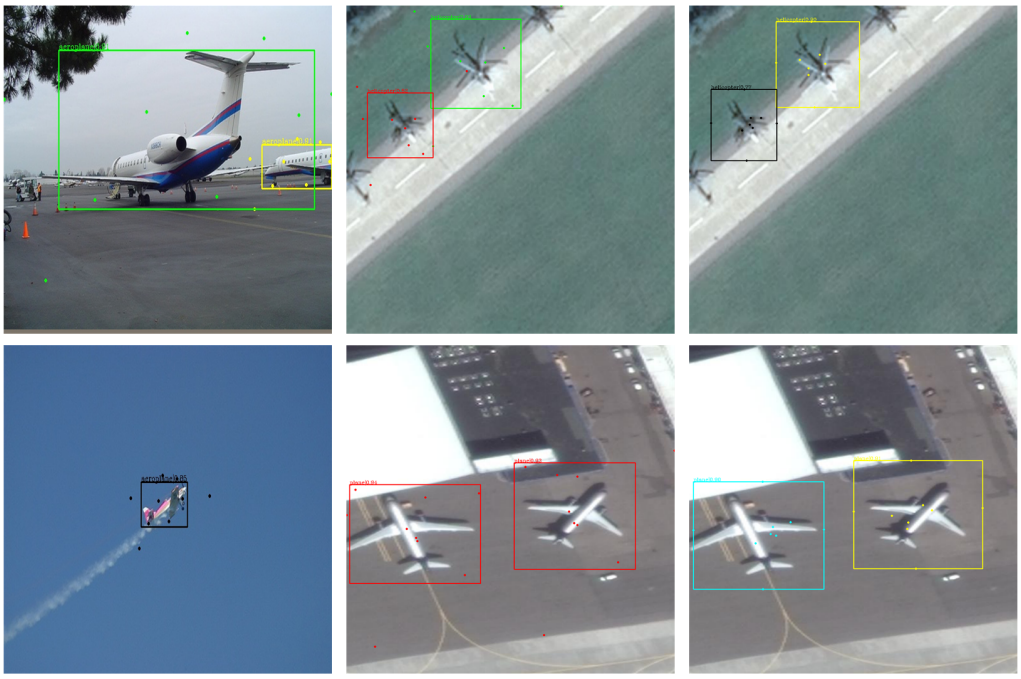

Figure 1. Visualization of reppoints and detection results, the first column is the result on the VOC dataset, the second column and the third column are the results on the DOTA dataset. The second column is the visualization result when using moment to obtain the pseudo box, and the third column is the visualization result when using minmax.

In summary, the main contributions of this paper can be summarized as follows:

1. We analyze the suitable attention mechanisms on feature map layers of different sizes, and combine actual experiment results to obtain a feature pyramid (SA-FPN) that is more suitable for multi-scale object detection.
2. To obtain better multi-scale feature expression, we design the MRFF module to replace the original stacked convolution structure in the detection head. With this modification, the model can perform more refined multi-scale features for objects corresponding to the same feature map layer. In addition, we propose a feature compensation module SBFC, which can be weighted based on object size, to solve the problem that the anchor-free algorithm is insufficient to express the objects in the

corresponding area based on the features on their key points. Combining with the differences in feature extraction of objects in different sizes, the model can express objects in the candidate regions more effectively.

3. Experiments on three challenging datasets show that our approach performs favorably against the baseline methods in terms of the popular evaluation metrics. In addition, our Anchor-Free Enhance Multi-Scale (AF-EMS) detector can get better multi-scale detection performance consistently when it is used on datasets of different distributions without any modification, which verifies the generalization and transferability of the proposed method.

The rest of this paper is arranged as follows. Section 2 briefly introduces some object detection algorithms related to this article. Section 3 introduces in detail the specific-attention FPN structure and the scale-aware detection head with two new modules we designed. We analyze the experiments and results in Section 4. In Section 5, we discuss some limitations of our model and some works that can be further studied in the future. Finally, in Section 6, we make a corresponding summary of this paper.

2. Related Work

Compared with the anchor-based detection method, the anchor-free detectors do not need to design the anchor box parameters, which typically need heuristic turning and many tricks. Therefore, these anchor-free detectors can effectively prevent the final detection performance from being affected by unreasonable hyperparameter settings and thus have better generality. In the past two years, anchor-free detection methods have been explored intensively. In the DetNet [26] network, the four corner points of the rectangular candidate area are respectively predicted based on the key point prediction method, and the initial proposals are obtained by enumerating all possible key point combinations. In the CornerNet detection framework [27], the upper left corner and lower right corner of the rectangular detection results are selected as key points for prediction, and the corner points are grouped based on the learning of the embedding vector to form the initial candidate area. Considering that the features at the upper left and lower right corners may not be able to effectively model the features inside the object, in [28], Duan et al. propose to introduce the features at the center point to construct better object features. In addition to the above-mentioned methods which obtain the position and size of the candidate area based on the corner detection, in the GA-RPN [29] algorithm, a method that can guide the generation of the candidate boxes based on local features is proposed. Considering that the size and aspect ratio of the object are related to the location of the object, a structure is proposed for learning the possible size and aspect ratio of the candidate area based on the predicted area where the object may exist. To reduce the inconsistency between the acquired feature and the candidate region, an adaptive module is proposed to adjust the corresponding feature map layer according to the predicted shape of the candidate region. In FCOS detection framework [15], Tian et al. propose an anchor-free detection method based on the fully convolutional networks (FCN). Based on this neat fully convolutional per-pixel prediction framework, FCOS achieves state-of-the-art detection result among all single-stage detectors. Yang et al. put forward the concept of representative points in RPDet [16]. According to the distribution of representative points, the size and aspect ratio of the object are obtained, and the features at the representative points are further used to express the candidate region.

Objects with variant scales are indeed common in many real-world scenarios, it is very important to obtain better multi-scale feature representations for objects of different sizes. In the field of object detection, how to obtain better multi-scale features has been studied for a long time. Earlier detection method is based on artificially designed local descriptors. For instance, the SIFT [30] local descriptor constructs a feature expression with scale-invariant characteristics designed for the detection of multi-scale objects. For detection method based on the convolutional neural network, the multi-layer stacking of convolution and pooling operations can naturally obtain feature map layers containing

features of different scales. However, the multi-scale feature obtained by simple module stacking is insufficient. In order to further improve the efficiency of the detection framework, a large number of algorithms have been proposed to optimize the backbone of the network or the multi-scale information acquisition to get better feature representations. In terms of the backbone of the framework, with the proposal of the residual connection, the problem of gradient disappearance during the training process has been solved. Based on the structure of AlexNet [31] and VGGNet [1], ResNet [21] and DenseNet [32] use skip connections to increase the depth of the network significantly, and to increase the size of the receptive field of the feature map. In addition, GoogLeNet [33] and Inception network structure [34,35] set multiple filters of different sizes on the same layer in parallel to process the input feature, which can increase the complexity of the receptive field and enhance the multi-scale representation capability in the width dimension of the model. The TridentNet network [36] proposes that objects in different size ranges should be processed with branches with different receptive field sizes. Thus, a parallel multi-branch architecture is constructed, and a scale-aware training strategy is used to specialize each branch by sampling object instances of proper scales for training. Inspired by the above jobs, Gao et al. propose the Res2Net [37] module, which uses a hierarchical residual connection structure on a single residual module to construct a multi-scale feature expression at a finer-grained level. In terms of multi-scale information acquisition, FPN network is the most commonly used method to solve the multi-scale object detection problem. Through the effective fusion of feature layers, a feature map pyramid containing certain semantic information is constructed. Based on the FPN structure, the M2Det [38] network further proposes a multi-level feature pyramid network to improve the size diversity of the features contained in the final output feature map layers. For two-stage detectors, Libra RCNN [39], PANet [40] and FAS-Net [41] all propose to obtain the feature representations of the candidate regions from all layers of the feature map pyramid, which can obtain better multi-scale detection performance.

By analyzing the above methods, in this paper, we propose an anchor-free detector called AF-EMS to deal with multi-scale object detection problem. Inspired by the application of the attention mechanism on the feature pyramid in FAS-Net, we propose a specific attention FPN module, and different attentions are constructed on feature map layers of different sizes. To obtain a better multi-scale representation, we propose a new scale-aware detection head. First, we propose a multi-receptive field feature fusion module, which can get more fine-grained multi-scale features for objects of different sizes on the same feature map layer. Furthermore, we propose a method to compensate the feature representation based on the size of the object, which can solve the feature deficiencies caused by getting the entire object representation from features of the limited key points of the object.

3. Method

More effective feature representations can be obtained based on features from more key points. In addition, as shown in Figure 1, compared with natural images, the targets in remote sensing images are all presented from downward perspectives, so the key points corresponding to targets in the remote sensing image are more fixed. Thus, we chose the RPDet as our baseline and made further improvements to it to get better multi-scale detection performance. We design our structure following the principles below: (1) To design corresponding attention mechanisms for feature maps of different sizes to promote the detectors to extract more effective multi-scale features. (2) To design a new module in detection head which can get more flexible features for objects with small size difference. (3) The feature extraction of objects should be scale-aware to make objects of different scales get more effective feature representations. As shown in Figure 2, AF-EMS mainly consists of three parts: First we extract image features based on the backbone network and use the specific-attention FPN (SA-FPN) structure to build a feature pyramid structure which is more suitable for object detection of different sizes. Then we obtain the initial position and size of the object based on RPDet structure. Finally, we use a new detection head

structure including MRFF and SBFC modules to get better multi-scale features to classify and fine-tune the locations of different candidate regions. More details about these core modules we designed and network configurations of AF-EMS detectors are introduced as follows.

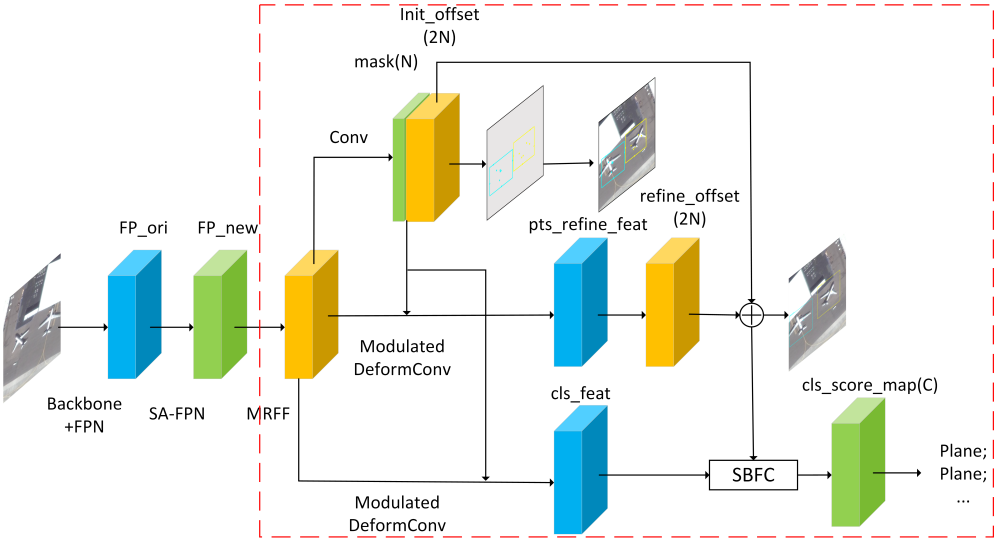

Figure 2. Overview of the AF-EMS detector. We only draw the detection head of one scale of FPN feature maps for clear illustration. In addition, all scales of FPN feature maps share the same afterwards network architecture (There are slight differences in SBFC module which we will explain in detail later) and the same model weights.

3.1. SA-FPN

Considering the important role that the attention mechanism plays in the human perception, how to construct and use attention in computer vision tasks has been studied extensively in previous research [41–44]. To be able to learn 'what' and 'where' in the channel dimension and the spatial dimension of the features to attend to, existing methods exploit spatial or channel-wise attention based on different efficient architecture to help decide which information to emphasize or suppress. Besides, in [41], Yan et al. propose that in addition to constructing and using attention in the backbone network, reconstructing the global context on the feature maps of different sizes in the feature pyramid to obtain specific attention is also helpful for dealing with multi-scale object detection problems. However, they all use the same module to obtain corresponding attention at different locations of the network, and to rescale or enhance the features.

In this paper, we propose the SA-FPN, in which we use different attention mechanisms for feature maps of different sizes to promote the detectors to extract more effective multi-scale features. We use existing context building blocks to model different types of attention. We choose the Simplified None-Local (SNL) module proposed in GCNet [43] to generate the channel-wise attention. The structure of SNL is shown in Figure 3a. Based on a large number of comparative experiments in GCNet, it is proved that the construction method of the global context and the integration method of the attention feature map and the input feature map have impact on the final performance. Therefore, we choose to use attention pooling to model the long-range dependency, and use the addition method instead of rescale method for feature aggregation. The SNL module can model global context as a weighted average of the features at all locations in feature map with a 1×1 convolution layer and a soft-max operator, and then recalibrate the importance of channels with another 1×1 convolution layer. This part is consistent with the global attention-based feature selection module (GAFSM) in FAS-Net. The output of the feature pyramid network at each level is represented as P2–P7. In our detection structure, attention module on the P2 layer is excluded and only used on the P3, P4, and P5 feature map layers. P6 and P7 are obtained

through down-sampling of P5 and P6 respectively, so there is no need to rebuild attention on them. However, we find that adding spatial-level attention to some of these feature map layers can achieve better detection performance which is ignored by FAS-Net detector. More importantly, considering that multi-scale object detection is performed in complex scenarios, it is difficult for small objects to be effectively focused when constructing spatial attention. When the spatial attention module is used improperly, the enhancement of the features of large objects in the image will bring difficulties to the detection of small objects which is verified in subsequent experiments. Therefore, we only construct spatial-level attention after constructing channel-level attention at the high-level feature map layer (P5). For the spatial-level attention, we use the spatial attention module of CBAM directly. The structure of this spatial attention module is be shown in Figure 3b.

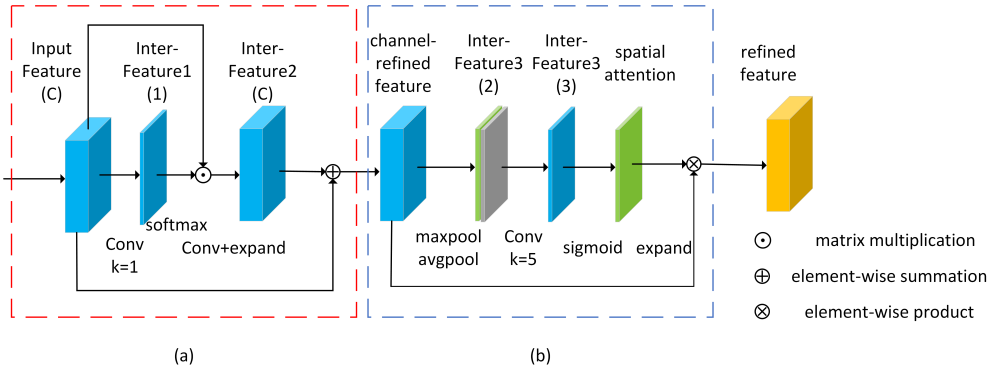

Figure 3. The structure of SA-FPN, (**a**) shows the structure of the SNL, which we use to build channel-level attention, and (**b**) shows the structure of the spatial-level attention building module. Note that on the low-level feature map, we only used part (**a**) of this module to obtain the refined features.

3.2. Scale-Aware Detection Head

The baseline anchor-free detector RPDet has a detection head similar to the RetinaNet [45] detection framework. In the original detection head, the classification branch and the regression branch are parallel that have similar small FCN structure, use separate parameters. On each branch, 3 stacked 3×3 vanilla convolution layers are used to further abstract the features on each feature map. On the regression branch, a regular 3×3 convolution is additionally used to obtain the initial bias value of reppoint at each position of the feature map. Then the classification branch and the regression branch use a deformable convolution to obtain the feature expression on the reppoints corresponding to the target at each position of the feature map and fuse them respectively. Finally, two branches use a 1×1 convolution respectively to classify the candidate regions and further fine-tune the position of key points. However, the fixed receptive field in the detection head structure is not conducive to extracting features which are more suitable for multi-scale object detection, especially for targets with smaller size difference on the same feature layer. In addition, expressing the candidate region using only the features on the limited key points may be insufficient in scenarios where there is occlusion or large target size. We propose a multi-receptive field feature fusion (MRFF) module and a size-based feature compensation (SBFC) module to solve the above-mentioned problems.

3.2.1. MRFF

The effect of receptive fields on the detection of different scale objects have been investigated by Li et al. in [36]. The baseline network uses feature maps with different receptive field sizes in the feature pyramid to detect objects within a specified size range. However, the range of the receptive field on the same feature map is fixed. When the receptive field of the output feature in the detection head structure is also fixed, multi-scale features extracted from the objects with small size difference are suboptimal, especially for objects on the same feature layer. On the other hand, the feature points on the high-level

feature map already have a large receptive fields. The stacked convolutional layers in the detection head structure to integrate features in a larger range may introduce unnecessary noise, which may bring difficulties to the extraction of effective features of the object.

Inspired by Res2Net proposed by Gao et al. in [37], we design a new feature extract module for detection head, which can get more fine-grained multi-scale features for objects with small size difference while getting the detection head with more suitable receptive field size for objects on different feature map layers. Res2Net constructs hierarchical residual-like connections on each residual block, which can get multi-scale features at a granular level and increase the range of receptive fields for each output features. As shown in Figure 4a, the Res2Net module sends the input to a 1×1 convolution, and evenly splits the output feature maps into s subsets along the channel dimension. Then each subset x_i is fed into a corresponding 3×3 convolution, denoted by K_i. The input x_i and output y_i corresponding to each 3×3 convolutional layer are shown in the following formula.

$$y_i = \begin{cases} x_i, & \text{i} = 0 \\ K_i(x_i), & \text{i} = 1 \\ K_i(x_i + y_{i-1}), & \text{others} \end{cases} \quad (1)$$

Finally, all outputs are concatenated and a 1×1 convolution is used to better integrate the effective information on different scale features.

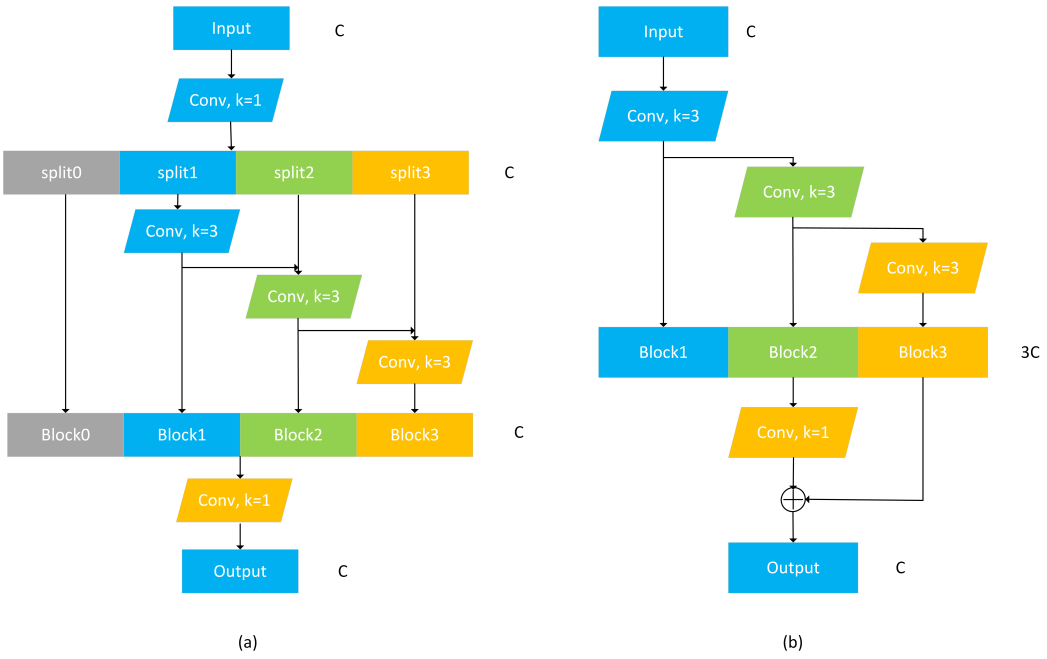

Figure 4. The structure diagram of the Res2Net module (**a**) and the MRFF module (**b**), the number on the right side of the feature map indicates the number of channels.

In contrast, considering that the output channel number of each feature map layer in the feature pyramid is limited, the split operator used in Res2Net may cause invalid aggregation of different types of features. Thus, we remove the split operation on the input feature map in our MRFF module, and use each feature map layer in the feature pyramid as input and send it to each feature extraction convolutional layer in the detection head structure sequentially. Different from the original structure, we retain the results obtained at each level and concatenate them. Finally, we use a 1×1 convolutional layer to select and integrate effective information from this feature map with different receptive field sizes. As shown in Figure 4b, we combine the selected and fused multi-scale receptive field features with the original feature output in the way of residual summation to ensure that the module will not bring adverse effects.

3.2.2. SBFC

In the baseline detection framework, the discriminant feature of the candidate region is integrated by the features of the key points predicted at each position. However, this approach has some problems when the targets in the application scene have large size differences or occlusions. For large objects, in order to predict the accurate candidate region size, the key points are more distributed at the edge of the objects, and some key points are even on the background area. For small objects, only a small number of pixels may be used for expression in the feature map, and a fixed number of key points may cause repeated overlap of some regional features, which in turn makes the extracted feature expressions have a certain offset. Finally, for partially occluded objects, the key points predicted at the same location may not be located on the same category of objects. Therefore, directly fusing the features on all key points indiscriminately may cause difficulties for network learning.

To solve the above problems, we first add an amplitude modulation mechanism to the deformable convolution of the original detection head structure. On the basis of the original structure, the feature amplitude at each sampling position is learned by DCNv2 [8]. With this modulation mechanism, we can remove the influence of the noisy information from key points irrelevant to the current foreground object. Then we propose a compensation module based on the predicted feature at the center point of the final candidate region. As shown in Figure 2, since the initial key point position is not accurate, we obtain the corresponding candidate region at each position according to the fine-tuned key points. Then the feature at the center of the prediction box is used to enhance the feature expression of the candidate region. Since the predicted center coordinates are not integer numbers, we adopt the bilinear interpolation method to obtain the features at the center point through the features at the four integer coordinate points around the center point. In addition, not all objects require feature compensation. For some objects, only a limited number of pixels are occupied on the feature map, it is sufficient to use features on repoints to learn feature expression. The additional feature compensation for the above objects will affect the learning of these key points. Therefore, we propose a method to weight the compensation features based on the size of the candidate region. The center point feature is weighted by calculating the ratio of the candidate region to the corresponding feature maps, so that the final feature can obtain different degrees of feature compensation according to the size difference. The feature compensation weight is formulated as follows:

$$w = 2 \times \sigma(\beta x) - 1 \tag{2}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

where x indicates the proportion of the candidate region at the current position to the feature map size. w represents the weight of features at the center point. $\sigma(.)$ is a sigmoid function and β is a hyper-parameter used to control the change speed of the weight with regard to the current size proportion. As shown in Figure 5, when we set β to different values, the w changes differently as x increases. It can be seen that the larger the β is, the more severe the weight value changes when x is in the range of $[0, 1]$, which also means the features of the center point have a greater influence on relatively small objects. In addition to the low-level detection feature map P3, on each feature map, the minimum number of pixels occupied by the target is 4×4 and the maximum is 8×8. Therefore, x can be calculated according to the corresponding stride of different feature maps. The calculation method is shown as follows:

$$x = (pred\_w/(s \times 8)) \times (pred\_h/(s \times 8)) - 0.25 \tag{4}$$

$$x = \begin{cases} 0, & x <= 0.25 \\ x, & 0.25 < x < 0.75 \\ 0.75, & \text{others} \end{cases} \tag{5}$$

$$x = x/0.75 \tag{6}$$

where *pred_w* and *pred_h* is the width and height of the predicted box on each position and s is the stride of the corresponding feature map. In addition, in order to solve the problem that the feature at the center point of the target is not the most expressive feature in some scenarios, we use 1×3 and 3×1 convolutional layer to process the feature maps before obtaining the compensated center point feature at the high-level layers (P6, P7).

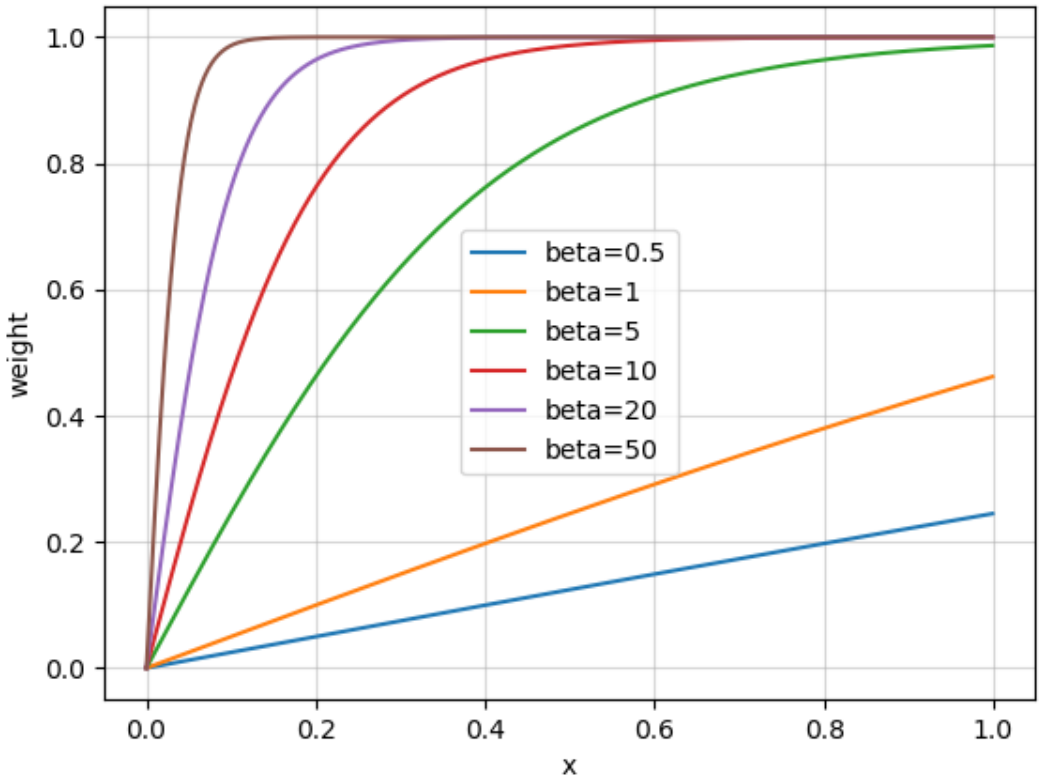

Figure 5. The influence of hyperparameter β on the change speed of the weight w with regard to the current size proportion x.

4. Experiment and Results

All experiments are implemented by using Pytorch and mmdetection on a server with GeForce RTX 2080 GPU and 11G memory. We conduct experiments on both aerial images and common objects images to verify the effectiveness and generalization of our proposed method. All hyper-parameters follow the default settings in mmdetection if not specifically noted. We conduct experiments with ResNet50-FPN backbones which is pre-trained on ImageNet. Furthermore, the batch size is 1 and we only use one GPU for all ablation experiments. Except for the simple horizontal flip, no additional data augmentation strategies are used in the data preprocessing stage.

4.1. Datasets and Evaluation

Most of our experiments are performed on DOTA [46], an aerial image object detection dataset. This dataset contains a total of 15 categories and 188,282 labeled instances in 2806 aerial images, whose size ranging from 800×800 to 4000×4000. The object categories include plane, baseball diamond (BD), bridge, ground field track (GTF), small vehicle (SV), large vehicle (LV), ship, tennis court (TC), basketball court (BC), storage tank (SC), soccer ball field (SBF), roundabout (RA), swimming pool (SP), helicopter (HC), and harbor. Each object is labeled with an arbitrary quadrilateral with a fixed starting point instead of the commonly used axis-aligned method. Therefore, the detection tasks on the dataset are also divided into horizontal bounding boxes (HBB) detection and oriented bounding boxes

(OBB) detection. Whether or not we get an OBB detection result is not related to the method described in our paper, so we only get HBB detection results in this paper. In order to facilitate the training and testing of the network, we use a 1024×1024 sliding window on the original image with an interval of 512 pixels to get image patches. Images with a size less than 1024 will be filled and scaled to a size of 1024×1024 before being sent to the network. Through the above method, we finally obtain all images for training. In addition, in order to further verify the generalization of the proposed method, we also conduct related experiments on the two most commonly used datasets in the object detection field, the PASCAL VOC dataset [47] and the COCO2017 dataset [48]. More details about these two common datasets can be seen in the review paper [4].

To quantitatively evaluate the performance of the proposed detector, we adopt the average-precision (AP), which is a standard metric for object detection. The AP can be obtained by calculating the approximate integral value of the corresponding precision when the recall value of the detection result changes from 0 to 1. Recall and precision can be derived from the number of true positive(TP), false positive(FP) and false negative(FN) results, which be formulated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

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

As AP is calculated in a category specific manner, the mean AP (mAP) averaged over all object categories is adopted as the final measure of performance. In order to compare the multi-scale object detection performance better, we also adopt the AP_S, AP_M and AP_L to evaluate the performance of detection results with different sizes separately. Consistent with the definition used in the MS COCO object detection challenges, AP_S is the mAP for small objects with an area smaller than 32^2, AP_M is the mAP for objects with area between 32^2 and 96^2, and AP_L is the mAP for large objects with an area bigger than 96^2. In addition, we also compare the amount of model parameters and detection speed of the detector, which have great significance to the practicality of the detector.

4.2. Ablation Experiments

To show the effectiveness of the modules proposed in this paper, we perform several ablation studies. The ablation experiments are performed on the DOTA train clip subset and evaluate on the DOTA val clip subset unless otherwise stated. The learning rate is initialized as 0.001 and changes on 8 and 11 epochs with a factor set as 0.1.

4.2.1. Specific Attention FPN Module

To use more suitable attention information in feature map layers of different sizes, we first try to construct and use channel-level attention and spatial-level attention on all FPN layers. As shown in rows 2, 3, and 4 of Table 1, we first apply the SNL module used in FAS-Net or the attention module in the CBAM algorithm to each layer of the feature pyramid directly. Rebuilding the global context on each feature map layer of the feature pyramid can improve the detection performance of the model, but the mAP improvement is extremely limited or even negligible. Considering that better context can be modeled by attention pooling and addition for feature aggregation, which is proved in GCNet, we try to replace the channel attention module in the CBAM with SNL module. The performance of the modified module CBAMv2 can be seen in row 5 of Table 1. In addition, through the comparative analysis of the detection performance on objects in different sizes, it can be found that combining the two types of attention information can achieve better detection performance on medium and large objects, while the detection performance on small objects will decrease. Considering that the spatial attention module on the low-level feature map is easier to enhance the local features of large objects instead of the feature of small objects, we further remove the use of spatial attention on the low-level feature

maps (P3 and P4). Finally, the SA-FPN module can achieve the best multi-scale detection performance compared to other attention mechanisms.

Table 1. The impact of different attention on FPN.

Neck Structure	AP_S	AP_M	AP_L	mAP
FPN	44.99	72.14	72.01	68.02
FPN+SNL	45.54	72.11	73.21	68.17
FPN+CBAM	44.57	72.15	73.36	68.23
FPN+CBAMv2	45.10	72.34	73.98	68.65
SA-FPN	47.36	72.48	73.63	68.77

4.2.2. Scale-Aware Detection Head

As shown in Table 2, the scale-aware detection head structure actually adds three parts: by using the MRFF module, compared to the baseline network, the detection performance of the model in most categories can be improved. Especially in BD, GTF, BC, and SBF, a large number of objects have similar sizes, and detection performances show a most obvious improvement. As shown in the row 4 of Table 2, the introduction of deformable convolution with modulation mechanism can improve the detection performance of categories where key points are easily located incorrectly, such as Bridge, GTF, and RA. Considering that the scale-aware feature compensation module mainly affects the features of large objects, we use SBFC module in combination with DCNv2 in the DOTA dataset, which mainly consists of medium-sized objects and small objects. The further analysis and structure design comparative experiments of this module are completed on the VOC dataset. Compared with the detection performance of the baseline network, although the difference in the final detection performance is small when each sub-module is added, the performance growth between these sub-modules is complementary. All sub-modules can be combined to form our scale-aware detection head, which can get a 1.8% improvement. Considering that the structure only introduces a small number of parameters (only 0.4 M additional parameters are introduced, which are about 1% of the original total parameters), the performance improvement is satisfactory and significant.

Table 2. The impact of different part in scale-aware detection head.

Method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC
baseline	92.29	71.00	48.70	64.16	48.22	74.71	91.29	91.56
baseline+MRFF	91.98	73.74	48.75	66.07	50.32	74.76	91.56	91.31
baseline+DCNv2	90.97	68.69	52.61	68.05	48.98	74.31	90.43	91.07
baseline+DCNv2+SBFC	92.46	73.65	49.33	63.39	50.22	74.59	90.54	92.12
baseline+scale-aware detection head	93.11	73.70	48.95	63.03	50.85	75.65	91.54	92.57
Method	**BC**	**ST**	**SBF**	**RA**	**harbor**	**SP**	**HC**	**mAP**
baseline	50.74	71.64	46.78	65.85	79.97	67.27	56.18	68.02
baseline+MRFF	53.86	72.44	49.29	66.24	80.97	68.48	51.38	68.74
baseline+DCNv2	54.60	70.14	47.32	69.46	79.81	68.64	51.16	68.42
baseline+DCNv2+SBFC	53.98	73.12	47.34	66.20	80.23	66.90	60.30	68.96
baseline+scale-aware detection head	59.29	73.17	49.87	69.00	81.79	66.99	58.33	69.86

4.2.3. Size-Based Feature Compensation

As shown in Table 3, we conduct comparative experiments of the SBFC module on the PASCAL VOC dataset, which focuses on large targets, to obtain a better feature compensation strategy. First, we use the center point feature of the initial candidate region in the Reppoint network to enhance the feature expression of all targets, and denote it as SBFCV1. It can be seen that the feature compensation at the center point has a positive effect on the detection performance of medium objects and large objects. Then, we try to weight the introduction of features at the center point to reduce the impact of features at the center point on small objects, and denote it as SBFCV2. It can be seen that the features

on small objects remain stable while the detection performance on other objects can be further improved. In addition, we also try to replace the center point of the initial candidate region with the coordinates of the center point of the refined detection results on the basis of SBFCV2, and denote it as SBFCV3. It can be seen that using the features at the center point after fine-tuning can obtain relatively better multi-scale target detection performance. Finally, we try to introduce 1×3 and 3×1 convolutional layers on the high-level feature map to solve the problem that the features at the center of some large objects are not the most expressive feature regions, and to further optimize the detection performance on objects with large sizes.

Table 3. The impact of different structure of SBFC.

SBFC	AP_S	AP_M	AP_L	mAP
None	45.55	64.77	85.92	81.16
V1	42.89	65.23	86.40	81.71
V2	45.37	66.12	87.07	82.17
V3	45.45	68.09	86.45	82.19
V4	49.53	66.27	86.76	82.36

β introduced in SBFC is a hyperparameter which allows us to control the change speed of the weight with regard to the current size proportion. We conduct experiments with the basic RPDet extended by our SBFCV4 for a range of different β values. As shown in Table 4, we find that setting $\beta = 1$ can achieve better detection performance on PASCAL VOC dataset. The center feature compensation weight changing too fast or too slow will cause a decrease in detection performance. Considering that the DOTA dataset contains more medium targets and small targets, we set $\beta = 35$ for all experiments reported in this work if not specifically noted.

Table 4. The impact of different β values in SBFC. We perform evaluation on the test dataset of PASCAL VOC.

β	AP_S	AP_M	AP_L	mAP
0.5	42.76	65.75	87.03	82.41
1	49.67	65.53	87.24	82.46
20	49.44	64.82	86.92	82.16
35	49.53	66.27	86.76	82.36
50	50.28	65.34	86.53	82.13
100	44.55	65.32	86.41	81.85

4.2.4. Parameters and the Detection Speed

We compare the number of parameters and the detection speed on DOTA dataset with the Res50-FPN backbone structure. The number of parameters is calculated when the input image is resized to 1024×1024. As shown in Table 5, it can be seen that compared with the use of a deeper backbone network, our method achieves better detection performance with negligible increase in the amount of parameters.

Table 5. Comparison of model size and detection speed.

Detector	Parameters (M)	Detection Speed (FPS)	mAP
Faster-RCNN	41.20	8.60	68.62
RPDet	36.61	14.80	68.02
RPDet+Res101backbone	55.60	9.30	68.57
AF-EMS	37.61	12.50	70.32

4.3. Comparison with State-of-the-Art Methods

We get a AF-EMS detector using ResNet101-FPN backbone and compare it with the state-of-the-art detector on the DOTA dataset. Furthermore, the multi-scale object detection performance of our method and the basic RPDet detector are compared on the VOC data set and COCO data set as shown in Tables 6 and 7. Both of them are trained with a ResNet50-FPN backbone for simplicity. Some detection results on the DOTA dataset are shown in Figure 6. It can be observed that our anchor-free detection model can achieve satisfactory detection performance in remote sensing images without any specific modification (In contrast, the anchor-based detection method needs to add more initial anchor boxes with more aspect ratio, such as [41]). The AF-EMS detector can obtain more accurate detection results especially on the bridge, ship and those objects with larger aspect ratio. In addition to the baseline detector, we compare with some algorithms on the HBB detection task of the DOTA dataset as shown in Table 8. The experimental results show the effectiveness and robustness of our method. It is worth noting that our detector has significantly faster detection speed than Yan et al.'s method [41]. Their method takes 0.95 s to process a single 800×800 image, while our method takes 0.08 s to process a single 1024×1024 image. In addition, we also compare our detector with the baseline network on the VOC dataset and the COCO dataset to prove its effectiveness and generalization in solving the multi-scale object detection problem.

Table 6. Detection result on VOC dataset.

	AP_S	AP_M	AP_L	AP0.5
baseline	45.55	64.77	85.92	81.16
AF-EMS	50.19	70.10	89.13	84.83

Table 7. Detection result on COCO dataset.

	AP_S	AP_M	AP_L	mAP	AP0.5	AP0.75
baseline	21.31	40.2	49.02	36.86	56.62	39.55
AF-EMS	22.29	41.3	49.46	37.62	58.35	40.49

Table 8. Quantitative comparison of the baseline and our method on the HBB task in test set of DOTA dataset.

Method	YOLOv2	SSD	Faster R-CNN	RFCN	Azimi et al. [49]	Yan et al. [25]	RPDet	AF-EMS
Plane	76.53	57.85	79.36	81.63	**90.00**	88.62	88.53	89.89
BD	34.26	32.78	79.32	78.69	77.71	**80.22**	73.94	77.45
Bridge	25.35	16.39	39.32	36.85	53.38	53.18	54.40	**54.61**
GTF	34.58	19.35	68.42	72.48	**73.26**	66.97	70.36	65.11
SV	36.58	7.36	62.14	59.60	73.46	**76.30**	54.35	66.05
LV	32.24	37.35	58.32	53.66	65.02	72.59	76.83	**80.12**
Ship	53.24	25.63	57.10	56.36	78.22	84.07	85.25	**87.00**
TC	61.65	81.14	89.36	**92.70**	90.79	90.66	90.87	90.88
BC	48.52	28.47	69.65	80.10	79.05	80.95	75.06	**82.87**
ST	35.54	48.35	59.12	67.58	84.81	86.50	80.74	**85.92**
SBF	28.94	15.36	59.23	**67.32**	57.20	57.12	56.87	60.89
RA	36.14	36.25	54.53	63.74	62.11	**66.65**	41.55	44.47
harbor	37.25	14.35	57.32	65.81	73.45	74.08	76.75	**77.62**
SP	38.21	9.36	56.23	63.29	70.22	66.36	65.63	**75.20**
HC	10.37	9.87	49.52	59.36	58.08	56.85	52.95	**71.53**
mAP	39.29	29.32	62.60	66.60	72.45	72.72	69.61	**73.97**

Figure 6. The detection result on DOTA test dataset.

5. Discussion

As shown in Section 4, our new detector, based on the characteristics of the anchor-free detection framework, can effectively improve the multi-scale object detection performance. On remote sensing detection task, the performance degradation on GTF is due to the fact that the centers of some GTF targets are also the centers of SBF targets. Our SBFC module can make the feature expression of GTF targets indistinguishable from the features of SBF targets, and we will try to solve this problem in our future research. We also find that using minmax instead of moment function to obtain pseudo-box from reppoints can obtain more reasonable key point positions and feature expressions of candidate regions in remote

sensing image, which is inconsistent with conclusions in natural image detection tasks. The key points obtained by the two methods can be seen in column 2 and 3 of Figure 1. As the anchor-free detectors do not need to design the corresponding anchors according to the actual application scenarios, which is inevitable using the anchor-based detector, the AF-EMS can be directly and effectively applied to different application scenarios without parameter adjustment. In order to obtain more effective feature expression in the high-level feature layer of the SFAC module, we add 1×3 and 3×1 convolutional layers before we get the center compensation feature. This structure can be replaced with the center pooling strategy which is proposed in CenterNet [28]. A similar function can be achieved while reducing the number of additional parameters. On the other hand, the size-aware weight calculation formula proposed in this paper is designed in a very simple way. We will conduct further research on how to design a more reasonable weight calculation method.

6. Conclusions

In this paper, a novel detection framework, which is defined as Anchor-Free Enhance Multi-Scale detector (AF-EMS) is proposed to solve the multi-scale object detection problem on remote sensing images. AF-EMS consists of the following improvements: First, we design the Specific-Attention FPN structure, which can get suitable attention for feature maps with different sizes to get better multi-scale feature layers. Then we modify the detection head to get a better local feature expression with the scale information of the candidate regions. In this scale-aware detection head, the Multi-Receptive Feature Fusion module can construct multi-scale features for objects with small size difference and the Size-based Feature Compensation module can enhance the feature representation and make features controlled by size information. Compared to the original model, our detector achieves consistent and significant boosts with negligible additional parameter overheads. Adequate experiments on three datasets demonstrate the effectiveness and the generality of the proposed detector. We hope that there will be more research conducted to improve the multi-scale detection performance of anchor-free detectors in remote sensing images.

Author Contributions: J.Y. conceived and designed the experiments; J.Y. performed the experiments and analyzed the data; J.Y. wrote the paper; W.D., X.S. and L.Z. contributed materials; H.W., X.S. and L.Z. supervised the study and reviewed this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 61725105 and 41701508.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://captain-whu.github.io/DOTA/dataset.html.

Acknowledgments: The authors would like to thank all their colleagues in the lab.

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

Abbreviations

The following abbreviations are used in this manuscript:

FPN	Feature Pyramid Network
IoU	Intersection-over-Union
SA-FPN	Specific-Attention FPN
MRFF	Multi-Receptive Feature Fusion
SBFC	Size-based Feature Compensation

References

1. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 2015 International Conference on Learning Representations (ICLR), Santiago, Chile, 7–9 May 2015.
2. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. *Comput. Vis. Pattern Recognit.* **2016**. [CrossRef]

3. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
4. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Pietikinen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
5. Kun, F.; Zhonghan, C.; Yue, Z.; Guangluan, X.; Keshu, Z.; Xian, S. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308.
6. Dai, J.; Qi, H.; Xiong, Y.; Yi, L.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
7. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
8. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
9. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position Detection and Direction Prediction for Arbitrary-Oriented Ships via Multitask Rotation Region Convolutional Neural Network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
10. Yang, X.; Liu, Q.; Yan, J.; Li, A. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), Virtual Conference, 2–9 February 2021. Available online: https://aaai.org/Conferences/AAAI-21/ (accessed on 23 November 2020).
11. Prasomphan, S.; Tathong, T.; Charoenprateepkit, P. Traffic Sign Detection for Panoramic Images Using Convolution Neural Network Technique. In Proceedings of the the 2019 3rd High Performance Computing and Cluster Technologies Conference, Guangzhou, China, 22–24 June 2019; pp. 128–133
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37
15. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636
16. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point Set Representation for Object Detection. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9657–9666
17. Singh, B.; Davis, L.S. An Analysis of Scale Invariance in Object Detection-SNIP. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3578–3587.
18. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient Multi-Scale Training. In Proceedings of the Thirty-second Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 2–8 February 2018; pp. 9333–9343.
19. Chen, Z.; Wu, K.; Li, Y.; Wang, M.; Li, W. SSD-MSN: An Improved Multi-Scale Object Detection Network Based on SSD. *IEEE Access* **2019**, *7*, 80622–80632. [CrossRef]
20. Zhang, M.; Chen, Y.; Liu, X.; Lv, B.; Wang, J. Adaptive Anchor Networks for Multi-Scale Object Detection in Remote Sensing Images. *IEEE Access* **2020**, *8*, 57552–57565. [CrossRef]
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
22. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. S3FD: Single Shot Scale-invariant Face Detector. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 192–201.
23. Li, J.; Wang, Y.; Wang, C.; Tai, Y.; Qian, J.; Yang, J.; Huang, F. DSFD: Dual Shot Face Detector. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5055–5064.
24. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Xian, S.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8232–8241
25. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. IoU-Adaptive Deformable R-CNN: Make Full Use of IoU for Multi-Class Object Detection in Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
26. Tychsen-Smith, L.; Petersson, L. DeNet: Scalable Real-Time Object Detection with Directed Sparse Sampling. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 428–436. [CrossRef]
27. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
28. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578
29. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region Proposal by Guided Anchoring. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2965–2974.

30. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157. [CrossRef]
31. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
32. Huang, G.; Liu, Z.; Laurens, V.D.M.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
33. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA,USA, 7–12 June 2015; pp. 1–9.
34. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the 2016 International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
35. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
36. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6054–6063.
37. Gao, S.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Torr, P.H.S. Res2Net: A New Multi-scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef] [PubMed]
38. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Cai, L.; Chen, Y.; Ling, H. M2Det: A Single-Shot Object detector based on Multi-Level Feature Pyramid Network. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019. [CrossRef]
39. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 821–830.
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
41. Yan, J.; Zhang, Y.; Chang, Z.; Zhang, T.; Sun, X. FAS-Net: Construct Effective Features Adaptively for Multi-Scale Object Detection. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020.
42. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
43. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
44. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
45. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
46. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.J.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
47. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
48. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
49. Azimi, S.M.; Vig, E.; Bahmanyar, R.; Korner, M.; Reinartz, P. Towards Multi-class Object Detection in Unconstrained Remote Sensing Imagery. In Proceedings of the 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 150–165.