remote sensing

MDPI

Article

MS-IAF: Multi-Scale Information Augmentation Framework for Aircraft Detection

Yuliang Zhao [1,2], Jian Li [1,2], Weishi Li [1,2], Peng Shan [1,2], Xiaoai Wang [1,2], Lianjiang Li [1,2] and Qiang Fu [3,*]

[1] Sensor and Big Data Laboratory, Northeastern University, Qinhuangdao 066000, China; zhaoyuliang@neuq.edu.cn (Y.Z.); 2172080@stu.neu.edu.cn (J.L.); 2172085@stu.neu.edu.cn (W.L.); peng.shan@neuq.edu.cn (P.S.); 2101929@stu.neu.edu.cn (X.W.); lilianjiang@qhd.neu.edu.cn (L.L.)
[2] Hebei Key Laboratory of Micro-Nano Precision Optical Sensing and Measurement Technology, Qinhuangdao 066000, China
[3] Shijiazhuang School, Army Engineering University of PLA, Shijiazhuang 050003, China
* Correspondence: 201801051714@sdust.edu.cn

Abstract: Aircrafts have been an important object of study in the field of multi-scale image object detection due to their important strategic role. However, the multi-scale detection of aircrafts and their key parts from remote sensing images can be a challenge, as images often present complex backgrounds and obscured conditions. Most of today's multi-scale datasets consist of independent objects and lack mixed annotations of aircrafts and their key parts. In this paper, we contribute a multi-scale aircraft dataset (AP-DATA) consisting of 7000 aircraft images that were taken in complex environments and obscured conditions. Our dataset includes mixed annotations of aircrafts and their key parts. We also present a multi-scale information augmentation framework (MS-IAF) to recognize multi-scale aircrafts and their key parts accurately. First, we propose a new deep convolutional module ResNeSt-D as the backbone, which stacks scattered attention in a multi-path manner and makes the receptive field more suitable for the object. Then, based on the combination of Faster R-CNN with ResNeSt-D, we propose a multi-scale feature fusion module called BFPCAR. BFPCAR overcomes the attention imbalance problem of the non-adjacent layers of the FPN module by reducing the loss of information between different layers and including more semantic features during information fusion. Based on AP-DATA, a dataset with three types of features, the average precision (AP) of MS-IAF reached 0.884, i.e., 2.67% higher than that of the original Faster R-CNN. The APs of these two modules were improved by 2.32% and 1.39%, respectively. The robustness of our proposed model was validated using the open sourced RSOD remote sensing image dataset, and the best accuracy was achieved.

Keywords: multi-scale; MS-IAF; ResNeSt-D; BFPCAR; object detection; remote sensing

Citation: Zhao, Y.; Li, J.; Li, W.; Shan, P.; Wang, X.; Li, L.; Fu, Q. MS-IAF: Multi-Scale Information Augmentation Framework for Aircraft Detection. *Remote Sens.* **2022**, *14*, 3696. https://doi.org/10.3390/rs14153696

Academic Editors: Weijia Li, Lichao Mou, Angelica I. Aviles-Rivero, Runmin Dong and Juepeng Zheng

Received: 26 June 2022
Accepted: 28 July 2022
Published: 2 August 2022

1. Introduction

Aircrafts play an important role in the fields of national defense and scientific research. Therefore, aircraft detection is of strategic importance. Currently, the different scales of aircrafts in images have become a major object of study in multi-scale object detection and remote sensing image detection [1,2]. Many researchers have been exploring high-precision methods for detecting multi-scale objects, especially those with complex backgrounds and under obscured conditions.

At present, multi-scale object detection technology is being widely used in human visual systems, aircrafts, ships [3], automotive [4], and remote sensing systems [5] to detect key parts of objects, which is important for unmanned driving [6], unmanned aircraft navigation, and defect detection [7]. Generative Adversarial Networks (GAN) [8] and multi-scale object detection [9] is now also widely used in the field of defect detection [9]. The main purpose of improving the multi-scale object detection technology is to develop

more accurate methods for locating the regions of interest (ROI) of objects of different scales in images.

However, there are two challenges to overcome when exploring and improving the multi-scale object detection technology for the detection of key parts. One is the lack of multi-scale datasets with mixed annotations of aircrafts and their key parts. The other is the lack of effective multi-scale object detection frameworks for detecting aircraft objects of different scales, especially those of small and medium scales interfered by complex backgrounds and obscured conditions [10].

Today's multi-scale aircraft datasets are still dominated by remote sensing images and whole aircrafts. Google Maps open-sourced the DOTA [11] remote sensing image dataset, which includes multi-scale objects such as aircrafts and ships. Google Earth and Vaihingen open sourced the NWPU VHR-10 [12] satellite image dataset, which includes multi-scale objects such as aircraft, ships and automobiles. Wuhan University open sourced the RSOD [13] remote sensing image dataset, which includes multi-scale objects such as aircrafts and oil tanks. Tsinghua University and Tencent jointly open sourced the TT100K [14] traffic sign dataset, which includes traffic signal board of different scales. However, most of today's multi-scale datasets contain independent objects, lacking objects with mixed annotations and severe environmental backgrounds and obscured conditions. Therefore, we open sourced a new multi-scale dataset AP-DATA, which contains aircrafts of different scales and their key parts.

Since the emergence of the multi-scale object detection technology, a great deal of attention has been focused on improving the detection performance of small and medium-scale objects. This is because large objects are easier to detect due to their rich feature information, and small and medium-scale ones are more difficult to detect as they take up a limited space in the image [10,15]. At present, detection accuracy improvement efforts for multi-scale aircraft objects are mainly about improving the detection accuracy of small and medium-scale objects, especially those in complex environments and remote sensing images [16]. Object detection networks currently available for researchers mainly include single-stage and two-stage networks. For two-stage networks, various improved Faster R-CNN models have achieved satisfactory performance in multi-scale object detection. At present, the performance of multi-scale aircraft detection based on a Faster R-CNN model can be improved by three main ways: by improving the backbone module, by improving the FPN module, and by improving the multitude of scales of the model.

Below are some examples of research efforts to improve the backbone module. Chen et al. [2] designed a novel MSCA module which improves the attention of original features in both the space and the channel, and improves the AP of multi-scale object detection by 2.59%. Gao et al. [17] proposed a Res2Net for multi-scale detection by following the idea of ResNet. This backbone module builds residual connections again in the residual block to extract multi-scale features at a tiny level, which is 1.84% lower than the top-1 error of ResNet-50. Wang and Sun et al. [18,19] proposed an HRNet module. This module adopts multiple parallel resolution branches while conducting information interaction between different branches to enhance the quality of semantic segmentation.

Efforts have also been made to improve the FPN module. Tan et al. [20] designed a new BIFPN which updates a scale that balances the introduced information by setting weights compared to the previous FPN processing features equally. This module can improve the AP by 1.65%, but it also requires more computation. Guo et al. [21] proposed an AugFPN module, which uses supervision to reduce the semantic gap of information at different scales before feature fusion and uses soft ROI selection to better learn features. Using AugFPN improves the AP by 2.3% when ResNet-50 is used as the backbone module. Cao et al. [22] introduced two modules, CEM and AM, into FPN and designed ACFPN. The CEM module uses the receptive field to enrich the feature information, and the AM module is used to deal with the cluttered task caused by the CEM module. This design improves the AP by 3.2% in object detection tasks, but the addition of two modules to the model leads to a significantly increased computation workload. Zhang et al. [23] proposed a method that

uses BFP as a pyramid feature fusion module for detecting ships in remote sensing images and enhances the multi-layer pyramid feature function by the same balanced semantic features. This method achieved a 7.15% higher AP for the detection of ships in remote sensing images.

Some of the research on improving the multitude of scales of the model are summarized below. A research team from the University of Maryland [24] proposed a SNIPER algorithm, which uses multi-scale training to detect objects of different scales in images. Although this method reduces the detection speed, its detection accuracy is greatly improved. Singh et al. [25] improved SNIPER and proposed a training scheme called SNIP. This scheme reduces the error of mapping migration caused by the large-scale difference among objects by reducing the loss of information within the specified area by several folds, and the performance of the COCO dataset is improved by 3%. Li et al. [26] proposed a TridentNet model that uses dilated convolution and parallel convolution to adapt the distribution of three objects of different scales through three different and parallel receptive fields. With the same training setting, TridentNet Fast achieves 41.0 AP for the same dataset, which show a 1.2 AP improvement over the baseline without computational cost.

Although the abovementioned methods proved effective at improving the performance of object detection, they still have limitations. As demonstrated by the working principles of human vision, due to the false detection problem caused by environments, the model extracts insufficient information for objects [27]. If the images are taken in complex backgrounds and obscured conditions and are polluted, there will be a large loss of information, making it difficult to detect objects in these images. In addition, a single scale is inadequate in addressing the low accuracy in detecting targets of different scales in natural and remote sensing images taken in complex backgrounds, and single scale attention is inefficient when it comes to detecting such multi-scale features [2]. Considering the information loss problem in multi-scale object detection, we designed a multi-scale information augmentation framework for aircraft detection. Starting from backbone and multi-scale feature fusion perspectives, we focused more effective attention on the ROIs in the image, reduced the loss of information in critical objects, and improved the detection accuracy for small- and medium-scale objects with complex backgrounds and obscured conditions. The main contributions of this paper are as follows.

(1) We provided a multi-scale object dataset AP-DATA (containing 7000 images of multi-scale aircrafts taken from different angles and in complex backgrounds, where a portion of the multi-scale objects are obscured and the whole aircrafts and their key parts come with mixed annotation).
(2) We proposed a ResNeSt-D backbone, which stacks scattered attention in a multi-path manner and makes the receptive field more suitable for the object.
(3) We proposed a BFPCAR feature fusion module to overcome the loss of information during information fusion in non-adjacent layers and introduced a larger receptive field to obtain semantic features.

2. AP-DATA Data Set

Although existing multi-scale object datasets contain enough numbers of images, they lack mixed annotations of different objects. Therefore, we acquired aircraft images from airliners and completed the mixed annotation independently.

When it comes to object detection, multi-scale objects are classified into three main categories: small objects with a pixel area less than 32^2, medium objects with a pixel area greater than 32^2 and less than 96^2, and large objects with a pixel area greater than 96^2 [28]. Usually, better detection results are achieved for large-scale objects due to their rich information. However, challenges remain to detect small- and medium-scale objects due to the limited representation of feature information inside the images. We established our own dataset Aircraft Dataset (AP-DATA), which contains five types of objects: aircrafts, engines, landing gears, air foils, and tail planes. The dataset contains images taken from

different angles and there is overlap between the objects. Some of the typical images in the dataset are shown in Figure 1.

Figure 1. AP-DATA (includes 7000 aircraft images of different attitudes).

In our experiments, we classified the objects into three scales based on the pixel area they take up. The numbers of the five key parts in the dataset are shown in Table 1.

Table 1. Statistics of the annotations (Number of five categories of objects).

Object	Quantity
Aircraft	7850
Airfoil	8160
Engine	10,127
Tail Plane	7703
Landing gear	17,759

The aircraft types contained in the dataset include passenger aircrafts, helicopters, carrier aircrafts and fighter aircrafts. Some statistics of objects contained in the dataset are shown in Figure 2. From Figure 2a,b, it can be seen that the aircraft and airfoil objects show a rectangular object distribution due to the longer marked length. As shown in Figure 2c–e, the tail planes, landing gears, and engines each have relatively uniform dimensions. According to the statistics in Figure 2f, large and medium objects take up most of the sample data, and small-scale objects take up a small fraction. The scale statistics of multi-scale objects provide supporting data for setting an appropriate model anchor size [16].

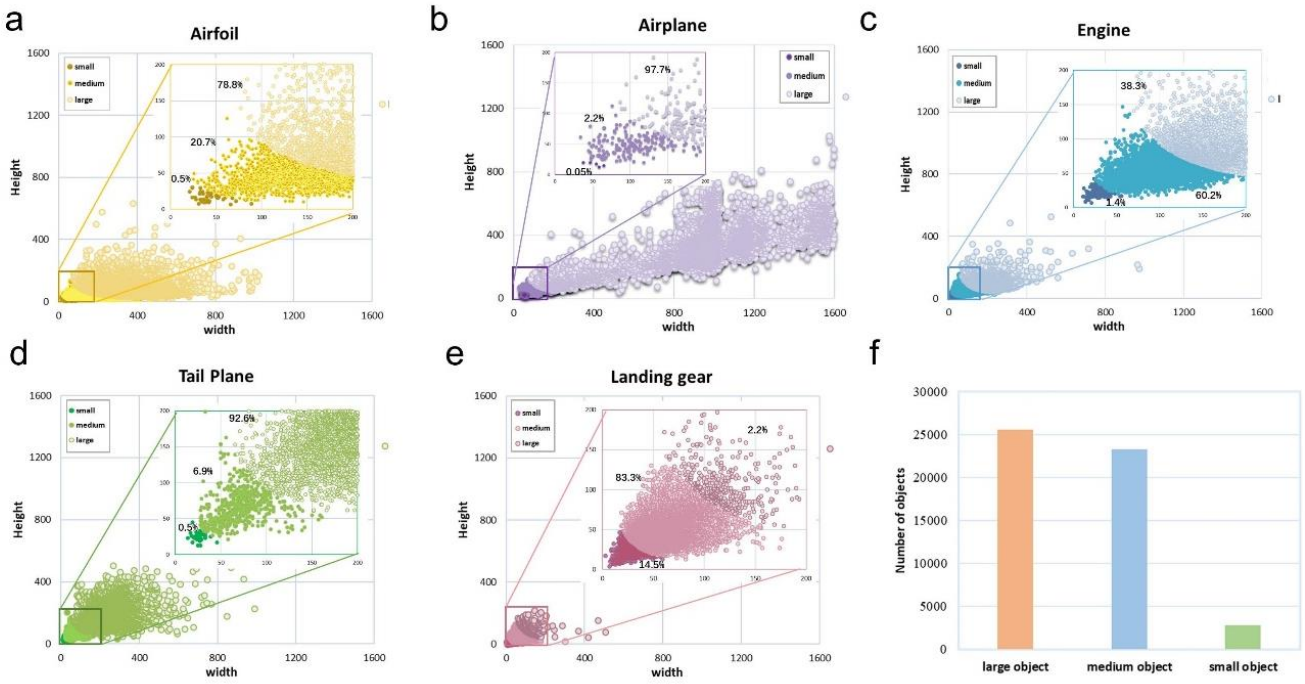

Figure 2. Statistics of multi-scale objects in AP-DATA. (**a**) Statistics of scales of airfoils by length and width. (**b**) Statistics of scales of airplanes by length and width. (**c**) Statistics of scales of engines by length and width. (**d**) Statistics of scales of tail plane by length and width. (**e**) Statistics of scales of landing gear by length and width. (**f**) Statistics of the numbers of objects of different scales.

3. Framework of the Model

Our proposed MS-IAF model follows the original Faster R-CNN [29]. The input images were resized to 1333 × 800 and then sent to our new backbone module. First, a new backbone (ResNetSt-D) based on ResNeSt and ResNet was proposed [30], which stacked scattered attention in a multi-path manner and made the receptive field more suitable for the object. Second, a Balanced Feature Pyramid with Content-Aware Reassembly of Features (BFPCAR) module was used to improve the connection between non-adjacent layers during feature fusion. The feature-level imbalance caused by random sampling of FPN was addressed. The CARAFE operator in BFPCAR allowed the up-sampling operation to expand the receptive field and focus on the semantic features of multi-scale objects. 256 output features of BFPCAR enter the RPN, and the 3 × 3 convolution is performed on the feature map using the sliding window method, keeping the number of channels features unchanged. Then the obtained new feature maps of 256 channels go into two fully connected layers to complete anchor and foreground and background predictions. The RPN structure generates region proposals, and bounding box regression corrects the anchors to obtain accurate proposals for object detection tasks. In this paper, we presented our improved framework in detail in two parts, and the framework design is shown in Figure 3.

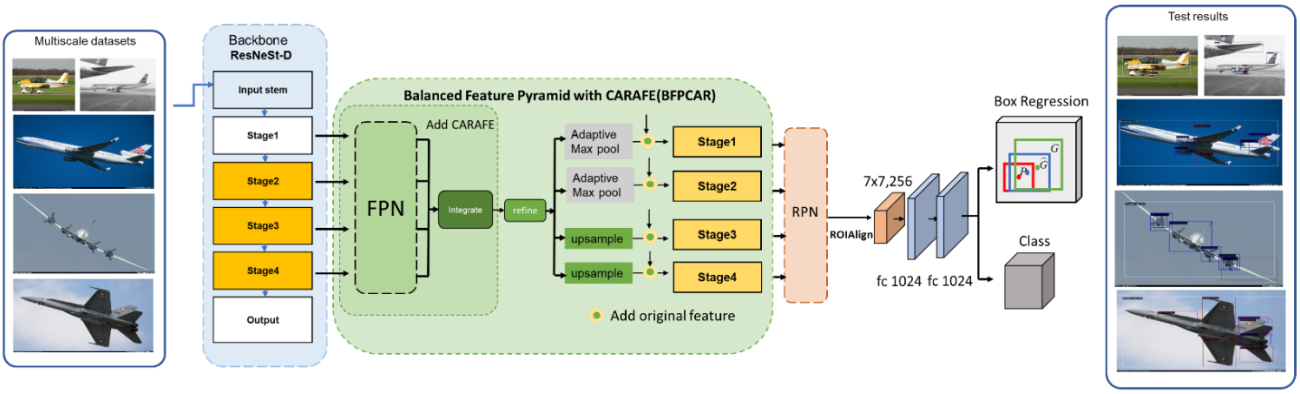

Figure 3. Multi-scale information augmentation framework (Contains ResNeSt-D and BFPCAR based on Faster R-CNN).

3.1. Optimization of Backbone

Traditional Faster R-CNN generally uses ResNet-50 or VGG 16 as its backbone. Since the ResNet-50 model has a relatively small size and a relatively deep network, it is considered a good choice as a backbone for object detection [29,31,32].

ResNet-50 performs down-sampling with a stride of two only in the first convolution of each stage. This results in the loss of 3/4 of the multi-scale information. To retain more information, we designed a new backbone module (ResNeSt-D) by combining ResNet-50 and ResNeSt. By using a new multi-path residual block, this module can retain both the input and output information of ResNet.

As in ResNeSt blocks [30], the features were split into cardinal groups determined by the cardinality K according to the channel dimension C. Then, a soft attention mechanism, Split-Attention within a cardinal group, was introduced by fusing via an element-wise summation across multiple splits to form a cardinal group. Finally, the cardinal group was stitched according to the channel dimension C. ResNeSt enables attention across feature maps and integrates the channel-wise attention with multi-path network representation to reduce information loss by stacking scattered attention in a multi-path approach.

We supplemented the Deformable Convolutional Network (DCN v2) to minimize information loss, make the receptive field more suitable for the geometric changes of images and reduce the interference of irrelevant information [33]. Thus, the information richness of the model was enhanced. This DCN supplementation did not change the calculation operation for Stage 2 to Stage 4, but only modified the convolution operation. An ordinary 3×3 convolution consists of nine positions of $R = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$. The output calculation operation is expressed as follows:

$$y_{p0} = \sum_{p_n \in R} w_{pn} \cdot (x_{p0} + x_{pn}) \tag{1}$$

where x_{p0} represents the center coordinate of the convolution, and y_{p0} represents the ordinary convolution output. The deformable operation was performed to add an offset convolution, ΔPn, for training in the convolution area and learning through an offset field that outputs 2N channels [34]. At the same time, the weight Δm_k of the sampling part was added to the 3×3 offset field, which increased the degree of freedom of learning features during training and reduced the invalid extracted information. The learning step is expressed as follows:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{2}$$

Attention was performed across feature groups in combination with DCN v2, which prevents the excessive loss of small object information when the dimension of 1×1

convolution is increased and decreased, and a large amount of attention was focused on multi-scale objects. The backbone module is shown in Figure 4.

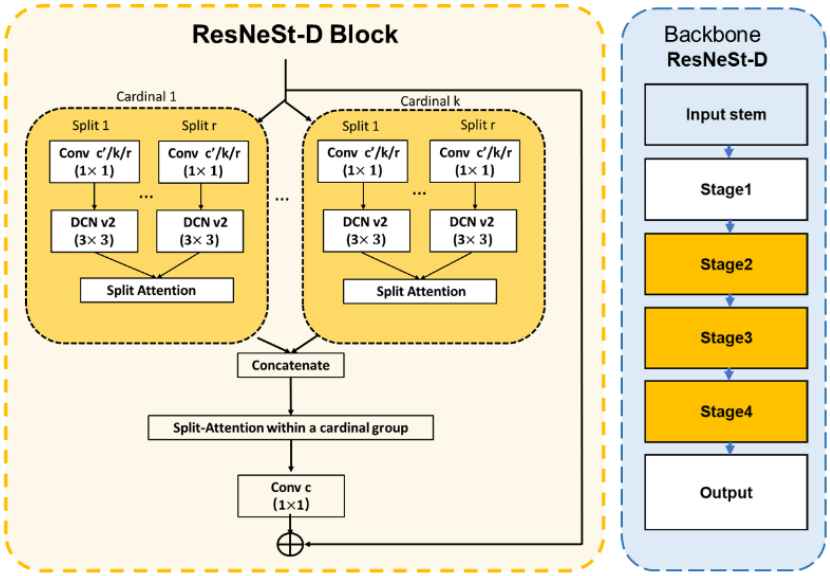

Figure 4. Backbone ResNeSt-D.

3.2. Multi-Scale Feature Fusion

In object detection, much richer semantic information can be obtained from shallow to deep convolutional neural networks, but the feature maps get smaller. As a standard component in object detection for feature fusion, FPN can fuse low-resolution feature maps with strong semantic information and high-resolution feature maps with weak semantic information. Therefore, the use of traditional Faster R-CNN in combination with FPN has been a common approach for multi-scale object detection [35].

The nearest neighbor interpolation in FPN determines the up-sampling kernel by pixel position instead of the semantic information of the feature map, which leads to a small receptive field and underutilization of the surrounding information. To overcome this problem, we added the lightweight up-sampling operator CARAFE [36], which can make full use of the object information and have a high correlation between the up-sampling kernel and the semantic information. CARAFE contains a Kernel Prediction Module and a Content-aware reassembly module. When receiving a feature map of H × W × C (where H, W, and C represent the height, width, and number of channels of the feature map, respectively). The 1 × 1 convolution is used by the up-sampling prediction module to compress the number of channels to reduce parameters and FLOPs. The up-sampling kernel size was set to kup × kup, and the up-sampling multiplicity to σ. After the input feature map was compressed with an output channel number of $\sigma^2 k_{up}^2$, the up-sampling prediction module obtained the up-sampling shape of $\sigma H \times \sigma W \times k_{up}^2$. Different channels at the same location share the same up-sampling kernel. Utilizing the Kernel Prediction Module and the Content-aware Reassembly Module, CARAFE achieves weight reduction by compressing channels and generating new feature map, which contents more semantic features in multi-scale regions.

The FPN module focused more on adjacent layers, and thus unbalanced information fusion between non-adjacent layers occurred [37]. Therefore, non-local [38] was used to refine the unbalanced fusion features to further enhance the features. Finally, four enhanced feature maps were obtained by feature separation. The BFPCAR module achieved a multi-scale fusion process, as shown in Figure 5.

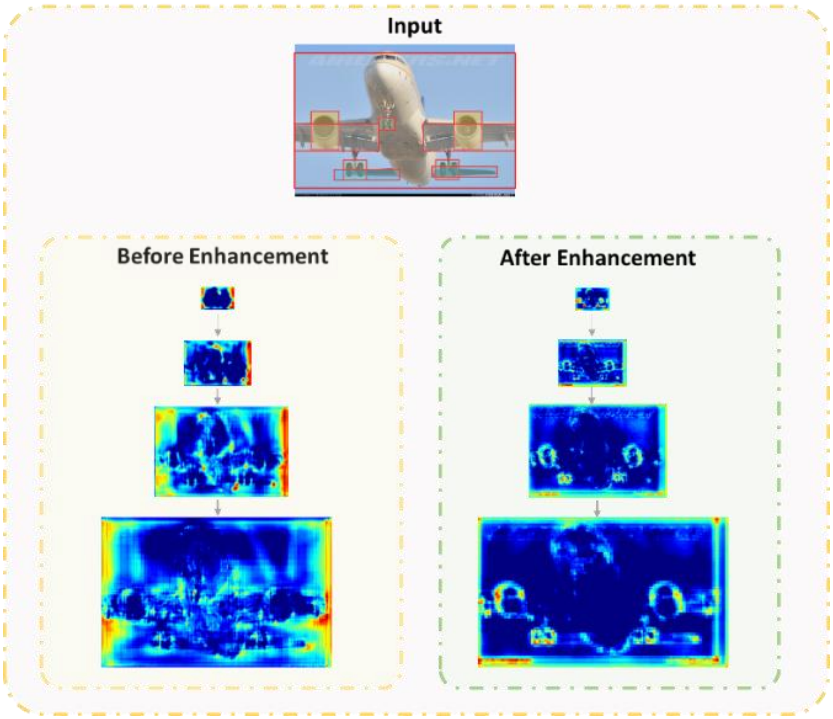

Figure 5. The performance of BFPCAR for feature enhancement.

BFPCAR addressed the information loss problem during information fusion in non-adjacent layers, and its ability to stay lightweight introduced a larger receptive field to sense the semantic features. BFPCAR divided the original images into more balanced feature maps. After multi-scale feature fusion for information enhancement, the convolutional features of the BFPCAR module were combined with the Region Proposal Network (RPN) [39]. In the head of the RPN, we set the anchor according to the statistical distribution of the AP-DATA dataset. The scales of objects in AP-DATA were counted. The results showed that the anchors obtained using the RPN had four scales (4 pixels, 8 pixels, 16 pixels, 32 pixels, and 64 pixels), and three aspect ratios (1:1, 2:5, and 6:5) can fit the ROI scales of object detection well in aircraft and critical objects. From Figure 5, it can be seen that after the Integrate and Refine operations by BFPCAR, the areas of small and medium-scale objects were more highlighted and the key parts of the aircraft were defined more clearly than before the enhancement.

4. Experiment

To verify whether the proposed model can improve the performance of multi-scale object detection, we conducted parallel experiments and ablation experiments. In the experiments, we used dual NVIDIA 1660 TI GPU for model training and testing. All pre-trained models have been made publicly available. The evaluation metric IoU was set to 0.5. For each image, we sampled 512 ROIs with a positive-to-negative ratio of 1:3. The weight decay was set to 0:0001, and the momentum was set to 0:9.

4.1. Experimental Results

We tested Faster R-CNN (FRC) models with different multi-scale modules and introduced VFNet, RetinaNet, Cascade R-CNN, TridentNet, YOLOF, and YOLOX, which perform well in the detection of multi-scale objects.

In our experiments, we used average precision (AP_S, AP_M, AP_L) and average recall (AR_S, AR_M, AR_L) to represent the detection results of objects of each scale, and AP was used to represent the average precision of the model. Each experiment was performed four times, and AP was expressed as the average and the standard deviation. The subscript

symbols S, M and L stand for small, medium and large objects respectively. Recall denoted the ratio of detected object TP to all objects that should be detected TP+FN, and AR was used to represent the average recall.

The experimental results are shown in Table 2. For large objects, the combination of FRC and the RegNetX backbone delivered the highest AP (0.897). In the ablation experiment, the combination of Faster R-CNN and the BFPCAR module and MS-IAF delivered the same AP_L, which was 0.885, 1.8% higher than that by the original Faster R-CNN. The results demonstrated that the AP of large objects was relatively high and had little room to improve, which could be attributed to the rich information in these objects. For medium objects, the combination of Faster R-CNN and the ResNeSt-D module delivered the highest AP_M, which was 0.694, 11.2% higher than the original Faster R-CNN. The results demonstrated that the ResNeSt module substantially increased the detection accuracy of medium objects by performing attention across feature groups and making the receptive field more suitable for the object. For small objects, the combination of Faster R-CNN and the ResNeSt-D module delivered the highest AP_S, which was 0.436, 9.8% higher than the original Faster R-CNN and slightly higher than Sparse R-CNN. Comprehensive assessment of overall performance, MS-IAF delivered the highest AP, which was 0.884, 2.67% higher than the original model. Our two new modules substantially improved the AP for targets of different scales compared to the original Faster R-CNN model, and MS-IAF delivered a higher AP than the other models in parallel experiments. The experiments demonstrated that MS-IAF can effectively overcome the complex backgrounds and obscured conditions in multi-scale objects and improved the detection accuracy.

Table 2. Experiment results (Comparative experiments of AP-DATA under different models).

Model	Backbone	Small		Medium		Large		AP	FLOPs (GFLOPS)	Params (M)
		AP_S	AR_S	AP_M	AR_M	AP_L	AR_L			
Sparse R-CNN [40]	ResNet-50	0.435	0.657	0.631	0.868	0.864	0.944	0.865 ± 0.013	149.9	105.95
HRNet [19]	ResNet-50	0.416	0.523	0.651	0.828	0.871	0.963	0.866 ± 0.012	174.25	27.1
Cascade R-CNN [41]	ResNet-50	0.416	0.488	0.654	0.840	0.877	0.943	0.864 ± 0.09	195.82	69.84
SSD 300 [42]	VGG-16	0.248	0.448	0.561	0.861	0.872	0.960	0.840 ± 0.002	386.25	34.31
SSD 512 [42]	VGG-16	0.393	0.635	0.601	0.923	0.860	0.966	0.838 ± 0.005	344.86	24.98
VFNet [43]	ResNet-50	0.367	0.580	0.653	0.942	0.855	0.988	0.867 ± 0.011	189.17	32.49
YOLOX [44]	Darknet53	0.433	0.617	0.648	0.877	0.835	0.900	0.871 ± 0.016	33.31	8.94
YOLOF [45]	ResNet-50	0.376	0.453	0.643	0.901	0.873	0.981	0.868 ± 0.013	98.26	42.16
RetinaNet [46]	Resnet-50	0.405	0.613	0.613	0.908	0.846	0.980	0.855 ± 0.016	206.13	36.19
Faster R C-NN+PAFPN [47]	ResNet-50	0.399	0.479	0.663	0.843	0.903	0.960	0.870 ± 0.009	218.58	44.68
TridentNet [26]	ResNet-50	0.380	0.503	0.666	0.861	0.865	0.949	0.880 ± 0.007	822.13	32.8
FRC [30]	ResNeSt	0.415	0.494	0.642	0.848	0.872	0.963	0.871 ± 0.004	220.17	43.03
FRC [48]	RegNetX	0.394	0.499	0.665	0.853	0.897	0.960	0.876 ± 0.006	170.41	31.49
Original FRC	ResNet-50	0.397	0.442	0.624	0.826	0.869	0.936	0.861 ± 0.014	193.8	41.14
FRC +BFPCAR	ResNet-50	0.409	0.486	0.641	0.850	0.885	0.954	0.873 ± 0.009	197.81	47.01
FRC	ResNeSt-D	0.436	0.521	0.694	0.869	0.879	0.954	0.881 ± 0.011	194.16	46.52
MS-IAF	ResNeSt-D	0.434	0.507	0.644	0.885	0.885	0.961	0.884 ± 0.009	197.12	52.12

The good experimental results were attributed to the following several advantages of MS-IAF, stacking the scattered attention by a multi-path approach, more features can be retained, and the use of deformable convolution can improve the adaptability to object receptive fields. the use of non-local implementation of cross-layer extraction improved the imbalanced feature fusion of non-adjacent layers of FPN, which in turn improved the detection accuracy of the model.

In this paper, the number of FLOPs refers to the number of multiply accumulation operations and measures the complexity of an algorithm or model. Params refers to the total number of parameters to be trained in the network model [49]. As shown in Table 2, YOLOX has the smallest FLOPs and Params, but it does not provide satisfactory detection accuracy in AP-DATA. As shown in Figure 6, MS-IAF achieved the highest accuracy without causing any excessive increase in FLOPs and Params. The results showed that MS-IAF achieved high detection accuracy for multi-scale objects without causing any excessive increase

in FLOPs and Params, which realized a good trade-off between memory and accuracy. Although TridentNet achieved a high AP, the excessive FLOPs imposed an additional burden on the model deployment.

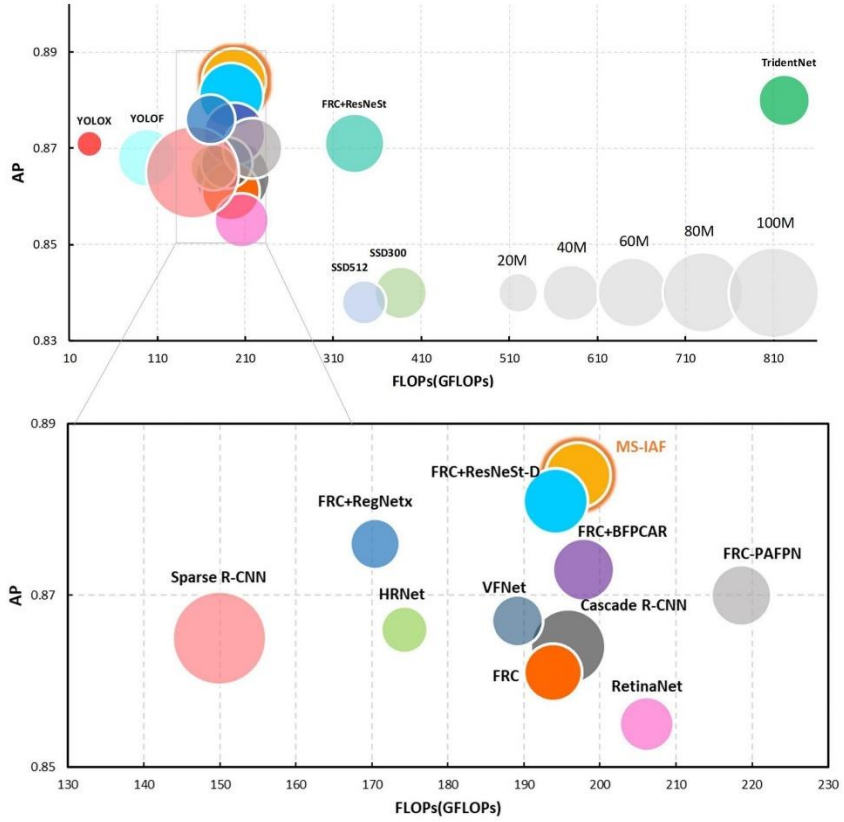

Figure 6. AP vs. FLOPs, size ∝ Params. (The vertical axis represents the average AP, the horizontal axis represents the FLOPs of the model, and the size of the circle represents the parameters of the model).

In order to observe the performance of original FRC and MS-IAF under different training data, we conducted the comparison experiments with 10%, 20%, 30%, 40%, 50%, 75%,100% of the total 6300 samples in the original training set, and the other 700 samples are used as the validation set. The difference in performance between original FRC and MS-IAF are shown in Figure 7 (AP is expressed as the average value of four experiments).

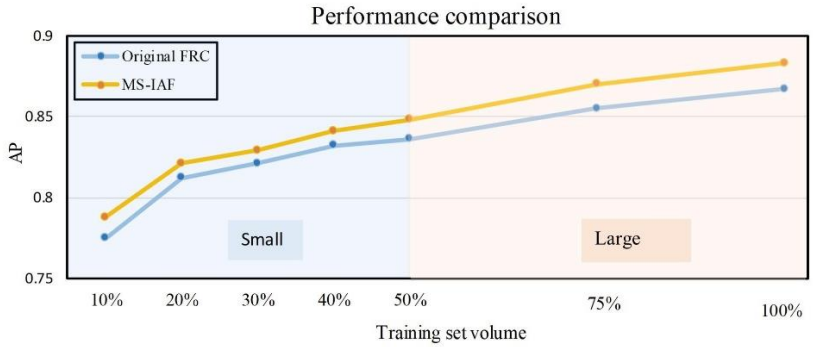

Figure 7. Performance difference between original FRC and MS-IAF models under different training set (100% represents all these 6300 samples are used for training the model).

The results in Figure 7 confirm that MS-IAF outperforms the original FRC in terms of detection ability and maintains a high AP after training on both small and large datasets. Which also demonstrated that MS-IAF has more excellent detection ability for different important objects of aircraft.

The experimental results are shown in Figure 8. For small and medium objects like engines, tail planes and landing gears, the detection results of the original model were often ignored or incorrectly detected, but MS-IAF delivered better detection results for these objects.

Figure 8. Comparison of test results on AP-DATA dataset.

4.2. Robustness Analysis

To verify the robustness of our model, we conducted experiments in the public remote sensing dataset RSOD [13]. The ratio of training to validation sets in the dataset was 5:1. The results of the original model and our model are compared in Figure 9. Our model can detect small and medium-sized aircraft objects in remote sensing images more accurately, while the original model often suffered from false detections and even missed detections, which shows it failed to find aircraft objects obscured in remote sensing images.

Figure 9. Comparison of test results on RSOD dataset.

To verify the anti-interference capability of the model, we conducted anti-interference experiments. We used the Imgaug [50] method to verify five pollution variables in the same dataset, light, fog, cloud, snow, and stains. Each pollution was divided into four different severity levels, as shown in Figure 10. With the same model setup, the trained model with

unpolluted remote sensing images was used to test 20 remote sensing images with different pollutions. The AP results of our experiments are shown in Table 3.

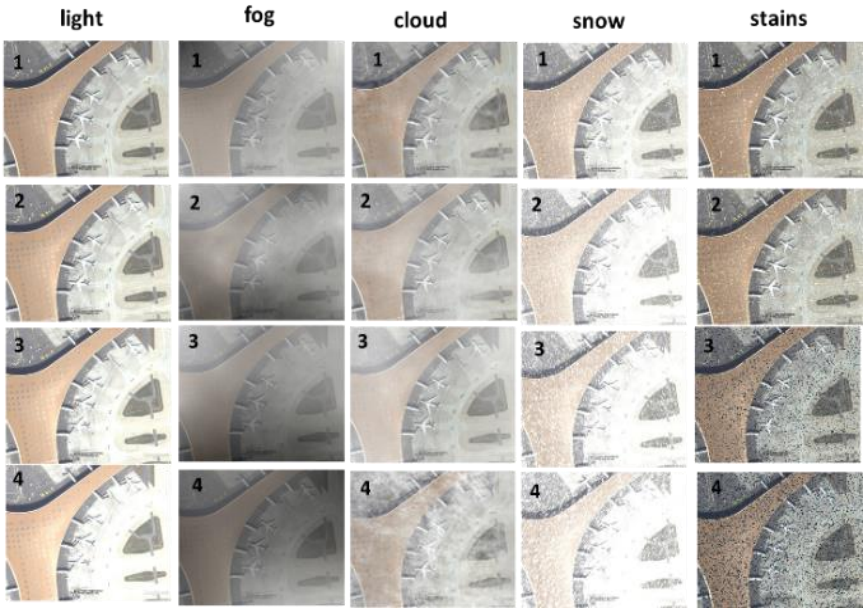

Figure 10. Pollutions with different levels of severity (Imgaug for image processing).

Table 3. Test AP results of pollutions with different levels of severity.

Model	Light:1	Fog:1	Cloud:1	Snow:1	Stains:1
FRC	0.890	0.882	0.888	0.880	0.897
HRNet	0.896	0.887	0.887	0.886	0.892
Cascade R-CNN	0.901	0.903	0.898	0.900	0.891
RetinaNet	0.903	0.903	0.900	0.892	0.900
TridentNet	0.906	0.904	0.899	0.903	0.906
MS-IAF	0.911	0.907	0.914	0.906	0.907
Model	**Light:2**	**Fog:2**	**Cloud:2**	**Snow:2**	**Stains:2**
FRC FPN	0.862	0.866	0.871	0.861	0.876
HRNet	0.870	0.866	0.874	0.866	0.873
Cascade R-CNN	0.872	0.876	0.878	0.863	0.874
RetinaNet	0.880	0.881	0.879	0.869	0.872
TridentNet	0.886	0.893	0.894	0.871	0.902
MS-IAF	0.907	0.907	0.904	0.871	0.908
Model	**Light:3**	**Fog:3**	**Cloud:3**	**Snow:3**	**Stains:3**
FRC FPN	0.837	0.859	0.846	0.759	0.729
HRNet	0.844	0.861	0.851	0.820	0.731
Cascade R-CNN	0.849	0.86	0.857	0.822	0.733
RetinaNet	0.851	0.866	0.859	0.839	0.739
TridentNet	0.853	0.870	0.856	0.833	0.741
MS-IAF	0.862	0.874	0.860	0.844	0.744
Model	**Light:4**	**Fog:4**	**Cloud:4**	**Snow:4**	**Stains:4**
FRC FPN	0.824	0.855	0.773	0.704	0.663
HRNet	0.829	0.859	0.779	0.706	0.659
Cascade R-CNN	0.833	0.858	0.786	0.711	0.667
RetinaNet	0.836	0.863	0.787	0.715	0.672
TridentNet	0.836	0.866	0.774	0.710	0.677
MS-IAF	0.839	0.869	0.788	0.721	0.679

The results from parallel experiments showed that MS-IAF outperformed the other models in terms of interference resistance. The different pollutions in the environments adversely affected the detection accuracy, causing some features to be obscured and making the model less capable of sensing features. From the variations in the detection results of our model at different severity levels, we found that our model exhibited the best robustness for fog pollution, with a 4.19% variation in the AP from severity 1 to 5. Its robustness was relatively good for light pollution, with a 7.9% variation. Its robustness was poor for cloud, snow, and stains pollutions, with 13.8%, 20.4%, and 25.1% variations, respectively. These findings indicated that fog and light pollution had the least effect on the multi-scale object features in remote sensing images but still obscured some feature information.

From the results of the robustness experiments, it was again verified that MS-IAF achieved higher accuracy for the detection of multi-scale aircraft objects in remote sensing images.

5. Conclusions

In this paper, we contributed a new multi-scale dataset AP-DATA and a multi-scale information augmentation framework MS-IAF comprising two new modules, ResNeSt-D and BFPCAR.

Most of the current multi-scale datasets contain independent objects and lack mixed annotations and complex environments, so we opensourced the AP-DATA dataset (including mixed annotations of whole aircrafts and their key parts and severe environmental interferences and obscured conditions). A new MS-IAF was proposed, based on a backbone module and a multi-scale feature fusion module. Backbone ResNeSt-D stacked scattered attention and made the receptive field more suitable for the object. The multi-scale feature fusion module BFPCAR enhanced the fusion capability and semantic features of multi-scale objects and overcame the problem of unbalanced feature extraction from non-adjacent layers. Our experiments improved the detection accuracy for multi-scale objects in complex backgrounds, making it easier to detect obscured objects. In the AP-DATA dataset, the AP of MS-IAF was 2.67% higher than that by the original Faster R-CNN. We conducted robustness experiments using aircraft datasets from ROSD remote sensing images, and the model showed significant improvements in the interference resistance.

Considering the flexibility of ResNeSt-D and BFPCAR modules, we will apply these two modules in the detection of other objects, but further improvements should be made to adapt them to other models.

Author Contributions: Conceptualization, J.L.; methodology, J.L. and P.S.; software, J.L.,W.L. and X.W; validation, P.S., Y.Z. and Q.F.; formal analysis, J.L. and X.W.; investigation, Q.F. and W.L.; resources, Y.Z.; data curation, J.L. and L.L.; writing—original draft preparation, J.L.; writing—review and editing, P.S., W.L. and Y.Z.; visualization, J.L.; supervision, L.L. and P.S.; project administration, Q.F.; funding acquisition, Y.Z. 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 (Grant No.61873307), the Hebei Natural Science Foundation (Grant No. F2020501040, F2021203070, F2022501031), the Fundamental Research Funds for the Central Universities under Grant N2123004, the Administration of Central Funds Guiding the Local Science and Technology Development (Grant No. 206Z1702G).

Data Availability Statement: The data presented in this study are available in the link: https://github.com/dlj0214/AP-DATA--MS-IAF (accessed on 17 April 2022).

Acknowledgments: All individuals included in this section have consented to the acknowledgment.

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

References

1. Liu, Y.; Zhang, X.Y.; Bian, J.W.; Zhang, L.; Cheng, M.M. SAMNet: Stereoscopically Attentive Multi-Scale Network for Lightweight Salient Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 3804–3814. [CrossRef] [PubMed]
2. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-Scale Spatial and Channel-Wise Attention for Improving Object Detection in Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 681–685. [CrossRef]
3. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
4. Ammour, N.; Alhichri, H.; Bazi, Y.; Benjdira, B.; Alajlan, N.; Zuair, M. Deep Learning Approach for Car Detection in UAV Imagery. *Remote Sens.* **2017**, *9*, 312. [CrossRef]
5. Zhang, Z.; Liu, Y.; Liu, T.; Lin, Z.; Wang, S. DAGN: A Real-Time UAV Remote Sensing Image Vehicle Detection Framework. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1884–1888. [CrossRef]
6. Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A Survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 91–124. [CrossRef]
7. Zhang, Q.; Liu, H. Multi-Scale Defect Detection of Printed Circuit Board Based on Feature Pyramid Network. In Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2021, Dalian, China, 28–30 June 2021; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2021; pp. 911–914.
8. Jiang, Y.; Han, S.; Bai, Y. Building and Infrastructure Defect Detection and Visualization Using Drone and Deep Learning Technologies. *J. Perform. Constr. Facil.* **2021**, *35*, 1–15. [CrossRef]
9. Liu, Y.; Yeoh, J.K.W.; Chua, D.K.H. Deep Learning–Based Enhancement of Motion Blurred UAV Concrete Crack Images. *J. Comput. Civ. Eng.* **2020**, *34*, 04020028. [CrossRef]
10. Muzammul, M.; Li, X. A Survey on Deep Domain Adaptation and Tiny Object Detection Challenges, Techniques and Datasets. *arXiv* **2021**, arXiv:2107.07927.
11. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE Computer Society: Washington, DC, USA, 2018; pp. 3974–3983.
12. Cheng, G.; Han, J.; Zhou, P.; Li, K. Multi-Class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
13. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
14. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; Volume 2016, pp. 2110–2118.
15. Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; Zhao, L. Infrared Small Target Detection Utilizing the Multiscale Relative Local Contrast Measure. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 612–616. [CrossRef]
16. Dong, Z.; Wang, M.; Wang, Y.; Zhu, Y.; Zhang, Z. Object Detection in High Resolution Remote Sensing Imagery Based on Convolutional Neural Networks with Suitable Object Scale Features. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2104–2114. [CrossRef]
17. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
18. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 3349–3364. [CrossRef] [PubMed]
19. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.
20. Tan, M.; Pang, R.; Le, Q. v EfficientDet: Scalable and Efficient Object Detection. In Proceeding of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
21. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. AugFPN: Improving Multi-Scale Feature Learning for Object Detection. In Proceeding of the Conference on Computer Vision and Pattern Recognition 2020 (CVPR) , Seattle, WA, USA, 14–19 June 2020; pp. 12595–12604.
22. Cao, J.; Chen, Q.; Guo, J.; Shi, R. Attention-Guided Context Feature Pyramid Network for Object Detection. *arXiv* **2020**, arXiv:2005.11475.
23. Zhang, T.; Zhang, X.; Shi, J.; Wei, S.; Wang, J.; Li, J. Balanced Feature Pyramid Network for Ship Detection in Synthetic Aperture Radar Images. In Proceedings of the IEEE National Radar Conference—Proceedings, Washington, DC, USA, 28–30 April 2020; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2020; Volume 2020.
24. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient Multi-Scale Training. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; pp. 1–8.
25. 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, 18–23 June 2018.
26. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6054–6063.

27. Tirin, M.; Marc, Z. Neural Mechanisms of Selective Visual Attention. *Annu Rev Psychol.* **2017**, *68*, 47–72.
28. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for Small Object Detection. *arXiv* **2019**, arXiv:1902.07296.
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–13. [CrossRef]
30. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
31. Fang, F.; Li, L.; Zhu, H.; Lim, J.H. Combining Faster R-CNN and Model-Driven Clustering for Elongated Object Detection. *IEEE Trans. Image Process.* **2020**, *29*, 2052–2065. [CrossRef]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016.
33. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable Convnets V2: More Deformable, Better Results. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; IEEE Computer Society: Washington, DC, USA, 2019; Volume 2019, pp. 9300–9308.
34. Zhao, Y.; Li, J.; Zhang, Q.; Lian, C.; Shan, P.; Yu, C.; Jiang, Z.; Qiu, Z. Simultaneous Detection of Defects in Electrical Connectors Based on Improved Convolutional Neural Network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–10. [CrossRef]
35. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, Hawaii, USA, 21–26 July 2017.
36. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware Reassembly of Features. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2019; Volume 2019, pp. 3007–3016.
37. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
38. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017.
39. Tang, X.; Zhang, H.; Ma, J.; Zhang, X.; Jiao, L. Supervised Adaptive-RPN Network for Object Detection in Remote Sensing Images. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Waikoloa, HI, USA, 26 September–2 October 2020; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2020; pp. 2647–2650.
40. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-End Object Detection with Learnable Proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 14–19 June 2022.
41. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2017, Honolulu, Hawaii, USA, 21–26 July 2017.
42. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 1–17.
43. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-Aware Dense Object Detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8514–8523.
44. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
45. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-Level Feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, Nashville, TN, USA, 20–25 June 2021.
46. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
47. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018.
48. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing Network Design Spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020.
49. Köpüklü, O.; Kose, N.; Gunduz, A.; Rigoll, G. Resource Efficient 3D Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1–10.
50. Cao, J.; Zhang, J.; Huang, W. Traffic Sign Detection and Recognition Using Multi-Scale Fusion and Prime Sample Attention. *IEEE Access* **2021**, *9*, 3579–3591. [CrossRef]