electronics

MDPI

Article

Few-Shot Air Object Detection Network

Wei Cai, Xin Wang *, Xinhao Jiang, Zhiyong Yang, Xingyu Di and Weijie Gao

The Third Faculty of Xi'an Research Institute of High Technology, Xi'an 710064, China; xhtu807@outlook.com (W.C.); jiangxinhao2020@outlook.com (X.J.); yangzhiyong302@outlook.com (Z.Y.); dixingyu807@outlook.com (X.D.); gaoweijie331@outlook.com (W.G.)
* Correspondence: wangxin9550@outlook.com

Abstract: Focusing on the problem of low detection precision caused by the few-shot and multi-scale characteristics of air objects, we propose a few-shot air object detection network (FADNet). We first use a transformer as the backbone network of the model and then build a multi-scale attention mechanism (MAM) to deeply fuse the W- and H-dimension features extracted from the channel dimension and the local and global features extracted from the spatial dimension with the object features to improve the network's performance when detecting air objects. Second, the neck network is innovated based on the path aggregation network (PANet), resulting in an improved path aggregation network (IPANet). Our proposed network reduces the information lost during feature transfer by introducing a jump connection, utilizes sparse connection convolution, strengthens feature extraction abilities at all scales, and improves the discriminative properties of air object features at all scales. Finally, we propose a multi-scale regional proposal network (MRPN) that can establish multiple RPNs based on the scale types of the output features, utilizing adaptive convolutions to effectively extract object features at each scale and enhancing the ability to process multi-scale information. The experimental results showed that our proposed method exhibits good performance and generalization, especially in the 1-, 2-, 3-, 5-, and 10-shot experiments, with average accuracies of 33.2%, 36.8%, 43.3%, 47.2%, and 60.4%, respectively. The FADNet solves the problems posed by the few-shot characteristics and multi-scale characteristics of air objects, as well as improving the detection capabilities of the air object detection model.

Keywords: few-shot object detection; air objects; multi-scale; deep learning

MSC: 68T45

Citation: Cai, W.; Wang, X.; Jiang, X.; Yang, Z.; Di, X.; Gao, W. Few-Shot Air Object Detection Network. *Electronics* **2023**, *12*, 4133. https://doi.org/10.3390/electronics12194133

Academic Editors: Antonio Fernández-Caballero and Byung-Gyu Kim

Received: 28 August 2023
Revised: 24 September 2023
Accepted: 2 October 2023
Published: 4 October 2023

1. Introduction

In the modern military field, air objects such as stealth fighters and unmanned aerial vehicles are combat weapons that affect the success or failure of high-tech warfare. As a means of undertaking aerial reconnaissance to combat the threat of the "visual center", these weapons can take advantage of surprise and high precision to break through the defender's three-dimensional defense network. The facilities, personnel, and equipment are crucial to implementing a "surgical" type of precision strike, paralyzing the overall combat system, and weakening the defender's combat capability. However, as high-value military targets, air objects have obvious non-cooperative and few-shot characteristics. The large amount of labeled data makes achieving accurate detection impossible. Therefore, when carrying out research on air object detection methods under few-shot conditions, realizing the accurate detection of air targets is vital to protecting key components, enhancing early warning capabilities on the battlefield, and improving the combat defense system. In addition, such research provides a reference for the accurate detection of other important non-cooperative military targets.

Few-shot object detection can be roughly divided into two types: one based on meta-learning and the other based on fine-tuning [1]. For the meta-learning type, the class-

independent parameters of the model are trained on a per-task basis to obtain a specialized meta-model. A small number of training samples are mapped to the new-class detection model to achieve the new-class detection task. For example, Yan et al. [2] introduced meta-learning ideas on the basis of mask regional convolutional neural networks (R-CNNs). They used support branches to obtain category attention vectors and fused them with query-image-extracted features to obtain new predictive features for object classification and localization. Kang et al. [3] used You Only Look Once version 2 (YOLOV2) as the basic framework and embedded meta-feature learners and feature-reweighting modules to obtain meta-features that could be generalized to new categories, enabling the detector to adapt to new categories quickly.

In contrast, fine-tuning-based methods solve few-shot problems through pre-training and fine-tuning. They apply a general supervised training approach, which minimizes the regularization losses of the pre-trained model during the fine-tuning stage using model gradient optimization methods to adapt to the detection of new categories. Gao et al. [4] proposed a multi-domain adversarial variational Bayesian inference method that minimized the inter-domain difference between the conditional distributions of the features of the base class and those of the new class. In addition, Cao et al. [5] constructed a compact new-class feature space based on fine-tuning to improve the new class's detection performance.

Meta-learning methods play a major role in few-shot object detection research. Generally considered promising in object detection, they can perform well with a small number of sample inputs for specific tasks [6,7]. For example, Chen et al. [8] solved the uncertainty representation problem based on meta-learning using a dual-awareness attention mechanism. Similarly, Perez-Rua et al. [9] addressed the open accommodation problem for new categories by obtaining new-class category feature vectors through meta-learning. However, such techniques still show slight shortcomings in two aspects. First, the complexity of the meta-learning model increases the risk of overfitting the model's parameters to the training base class [1], and second, dimensional meta-learning may not converge during the training iterations [10].

Therefore, from this perspective, fine-tuning-based learning has more advantages in terms of its universality and simplicity compared with meta-learning. Several fine-tuning-based studies have recently reported competitive results. For instance, Fan et al. [11] conducted an analysis based on the fine-tuning method and proposed a "Bias-Balanced RPN" and a secondary detector to eliminate the bias brought about by the base class during pre-training while ensuring that the class-independent knowledge is not forgotten. In other research, Kaul et al. [12] introduced a pseudo-labeling method based on the fine-tuning method, which increases the number of training samples for the new class by obtaining high-quality pseudo-labeled data, reducing the problem of too few samples in the new class and improving the model's detection ability. Compared with the uniqueness and unfamiliarity of the meta-model, the fine-tuning-based method has stronger plasticity in terms of optimization technology, loss function, data enhancement, and architecture. Although the meta-learning method has stronger adaptability, in particular in the field of few-shot object detection, the vanishing gradient problem and its overall complexity limit its calculation steps, while the fine-tuning-based method encounters no such difficulties [1].

However, the few-shot object detection method based on fine-tuning focuses on general-purpose objects in natural scenes and still exhibits some deficiencies in the identification and localization of air objects. Compared with natural scenes, the few-shot characteristics of air objects introduce more prominent problems. Unlike natural object detection, which only categorizes an object within a large class, the detection of air objects needs to accurately determine an object's specific model. Moreover, the air object's own feature recognition is low, making the model more difficult to decipher. A stronger feature extraction ability is needed to obtain more information about the object from a small number of samples so that the air object is effectively detected. In addition, because air objects fly in all weather conditions and samples are difficult to obtain, the precise sample shooting angle, as well as the time and location, cannot be determined, resulting in a wide

distribution of sample scales when detecting air objects. Furthermore, the number of air objects at each scale is small, and the few existing shot detection networks may be difficult to adapt. The discriminative nature of object features at different scales and the processing capability of multi-scale features cannot be guaranteed, and a detection model similar to universal object detection undertaken in natural scenes is challenging to design to achieve the accurate detection of airborne objects at all scales.

To address the above issues, we propose a few-shot air object detection network (FADNet). Starting from the few-shot and multi-scale perspectives, the network structure enhances the model's capability for detecting air objects. First, a multi-scale attention mechanism (MAM) is introduced after designing the backbone network to extract object features from both the spatial and channel aspects. This further aggregates the local contextual features and global features to improve the object information extraction capability of the network. Second, the feature pyramid of the neck network is improved by adding jump connections based on the path aggregation network (PANet) [13], and sparsely connected convolution is added to the multi-scale output. The number of corresponding convolution groups is set for the outputs of different scales to improve the discriminative power of the features at each scale. Lastly, we integrate a multi-scale regional proposal network (MRPN) that is designed based on the multi-scale characteristics of the features, changing the previous mode of multiple inputs and single outputs. MRPNs are built based on multi-scale outputs, and adaptive convolution is introduced in the front end to process features at different scales effectively and enhance the object recognition ability of the detection network.

The main contributions to this article are as follows:

1. FADNet is proposed to solve the problem of the low precision rate of air object detection under the influence of few-shot characteristics and multi-scale properties, improving the detection capability of the network;
2. A MAM was designed to realize the deep fusion of object features from both the spatial and channel dimensions, and more effective information about object features was extracted;
3. Based on multi-scale characteristics, first, the feature pyramid structure was improved, a jump connection was added to PANet [13], and sparsely connected convolution was introduced to the outputs of each scale, which improved the discriminative properties of the features of each scale. Second, a multi-scale regional candidate network was constructed to adaptively extract feature information for different scale outputs, and a multi-input and multi-output model was established to utilize the multi-scale features effectively;
4. The designed algorithm was experimentally validated on the general PASCAL VOC dataset and our self-developed few-shot military air object dataset, achieving good results.

2. Related Work

This section reviews the existing deep learning object detection algorithms, few-shot learning algorithms, and few-shot object detection algorithms related to this article.

First, we discuss object detection based on deep learning. Currently, object detection algorithms are divided into two main categories: two-stage and single-stage. During the detection process, two-stage algorithms first create region suggestion boxes, distinguishing between the background and foreground, and then perform classification and localization regression operations on each suggestion box. In 2014, Girshick et al. [14] proposed the two-stage algorithmic regional convolutional neural network (R-CNN) model for the first time. However, because of its cumbersome algorithm steps and slow calculations, researchers proposed a fast R-CNN [15] and a faster R-CNN [16] to improve precision and reduce calculation speed. At present, two-stage algorithms are widely applied in fields such as unmanned driving, military detection [17,18], facial recognition, and industrial detection, yielding good results.

Compared with two-stage algorithms, single-stage algorithms can directly predict, locate, and classify feature maps. Redmin et al. [19] first proposed the You Only Look Once (YOLO) algorithm in 2016, which has since undergone version updates in the YOLO series [20–22], gradually becoming an important framework for one-step object detection. The single-shot detector (SSD) algorithm [23] draws on the advantages of the two-stage algorithm and integrates the design concept of the faster R-CNN Chung-guyok algorithm into the single-stage algorithm. The deconvolutional single-shot detector (DSSD) algorithm [24] cited Resnet-101 [25] as a feature extraction network on this basis, exhibiting improved detection performance. Compared with two-stage algorithms, single-stage algorithms are simpler to implement and faster to train, but because of their lack of RPNs, their overall precision is inferior to that of two-stage algorithms.

The next area of interest relates to few-shot learning. Few-shot learning aims to use a small number of samples to acquire new knowledge. The central premise of this method is to accurately transfer knowledge from a base-class training model to a new class. The existing few-shot learning methods can be roughly divided into three categories. The first constitutes optimization-based methods, such as model-agnostic meta-learning [26] (MAML), which learn through well-initialized rules. In a relatively short period of time and using MAML as a basis, Jamal et al. [27] developed and proposed task-agnostic meta-learning (TAML) to solve the problem of meta-learner bias. The second method is based on metric learning, which obtains the generalized metric space of a category through learning to perform subsequent similarity measurement operations. Karlinsky et al. [28] introduced multimodal distribution into metric learning to achieve end-to-end training of backbone network parameters and embedded spatial distribution. Wang et al. [29] utilized global vectors for word representation encoding to embed label information into feature maps, achieving feature enhancement of the data. The third method is based on parameter generation [30]. Unlike the other methods, this method obtains a superior network model by pre-training and fine-tuning the class-related parameters in the second stage to achieve better adaptation to new tasks. Sun et al. [31] and Liu et al. [32] integrated the MAML method into model fine-tuning, achieving algorithm improvements and enhancing the generalization performance of the algorithm.

The literature also covers few-shot object detection. Similar to few-shot object classification, most few-shot object detection methods currently use two-stage training, namely, a pre-training stage and a fine-tuning stage. However, this method is different from few-shot learning in that it must not only recognize the object in the sample but also locate its specific positions based on the background, which is more difficult to achieve. In order to improve the detection accuracy of few-shot objects, meta-R-CNN [2] introduces meta-learning into an R-CNN, which does not extract feature map information from a holistic perspective. Instead, it focuses on the features of each region of interest (ROI). Fan et al. [33] designed an aggregation model called an Attention RPN based on the meta-learning network model, which measures the similarity between the support set features and the query set features from three perspectives—global, local, and cross-correlation—helping the detector to better distinguish different categories.

Li et al. [34] proposed a category marginal reconstruction method to transform the under-shot object detection problem into an under-shot object classification problem. This was carried out by introducing a fully connected layer at the end of the detector to decouple the classification and regression feature contradictions. Meanwhile, the category boundary loss is added to the feature learning to achieve the marginal space between the new class and the base class and improve the new-class detection level. Yin et al. [35] improved the meta-learning method to study the under-photographed object detection environment for incremental learning. They proposed a hyper-network-based under-photographed object detection method, which solves the difficulties encountered in incremental learning. New categories can be learned sequentially and incrementally without additional training, which improves the effectiveness of the model's detection. Similarly, Zhang et al. [6] introduced a novel inter-class correlation meta-learning strategy to achieve the robust and efficient

detection of underphotographed objects by extracting and utilizing the correlations of different categories. This inter-class correlation method can focus on multiple support categories simultaneously, reducing the misclassification of similar samples and enhancing the generalization ability of new-class samples. Qiao et al. [36] improved the fine-tuning method, analyzing it from a multi-task and multi-stage perspective. They proposed a fast decoupling method, which decouples the feed-forward network and gradient updating through the introduction of the gradient decoupling layer; they also redefined the forward and backward operations. Meanwhile, an offline classification module was added to the detection back end, which realized classification correction through extra scores and improved the ability of category judgment.

The authors of [37] proposed a multi-scale positive sample refinement (MPSR) model for few-shot object detection. This generates multi-scale samples through data augmentation and establishes a fast R-CNN branch to alleviate the problem of insufficient samples. Furthermore, Khandelwal et al. [38] improved the generalization ability and detection performance of few-shot object detection by calculating the semantic similarities between the new and base classes and transferring the regression and classification weights to the new class. Sun et al. [39] mixed the new and base classes to form a fine-tuning dataset to reduce the differences in the features between the classes. Zhang et al. [40] proposed the Cooperative Region Proposal Network (CoRPN) to solve the problem of foreground-background imbalances exacerbated by insufficient sample data, increase the number of foreground classifiers, and avoid losing more pre-selection boxes.

In the above studies, by building a new network structure and improving the fine-tuning-based or meta-learning methods, few-shot object detection alleviates the problem of having insufficient samples of new classes during detection, solves the multi-task, multi-stage coupling contradiction in detection, and reshapes the feature space of the new and base classes. This improves the accuracy of detecting new classes and the network model's overall detection capability. However, the above methods focus on the detection of natural objects and lack relevance for the detection of military targets with few shots, such as air objects. Particularly significant is the lack of attention these methods pay to the multi-scale characteristics of air objects, while the scale problem is considered the core of object detection [41], which seriously restricts the detection capabilities of air objects under the conditions of few samples. In contrast, our proposed network, which is based on the characteristics of air objects, such as their few-shot and multi-scale nature, makes targeted modifications to the backbone network, neck network, detector, etc., achieving efficient multi-scale feature extraction and feature processing and improving the detection performance of few-shot military targets such as air objects.

3. Methods

This paper proposes FADNet to build a network model for few-shot and multi-scale situations. By improving the ability to extract object features, the discriminative power of features at each scale, and the utilization of multi-scale features, air object detection performance is enhanced. The network structure is shown in Figure 1, comprising a backbone network, neck network, and detector. The backbone network includes an air object image input backbone network transformer [40], and the transformer's feature extraction network is composed of four main parts. First, the input image is divided into four blocks and $56 \times 56 \times 48$ feature vectors, which sequentially output different levels of feature vectors across four stages to achieve multi-scale feature extraction of the object. After being processed by the MAM, the output multi-scale features are input into the neck network, the IPANet, which carries out the deep extraction and fusion of object features at each scale, and then the output results of each scale are input into the detector, which comprises ROI Align, HEAD, and the MRPN, consisting of multiple RPN branches. The features of each scale are input into the corresponding region candidate network and ROI Align to form the candidate frame, and the final classification results and location results are output by HEAD.

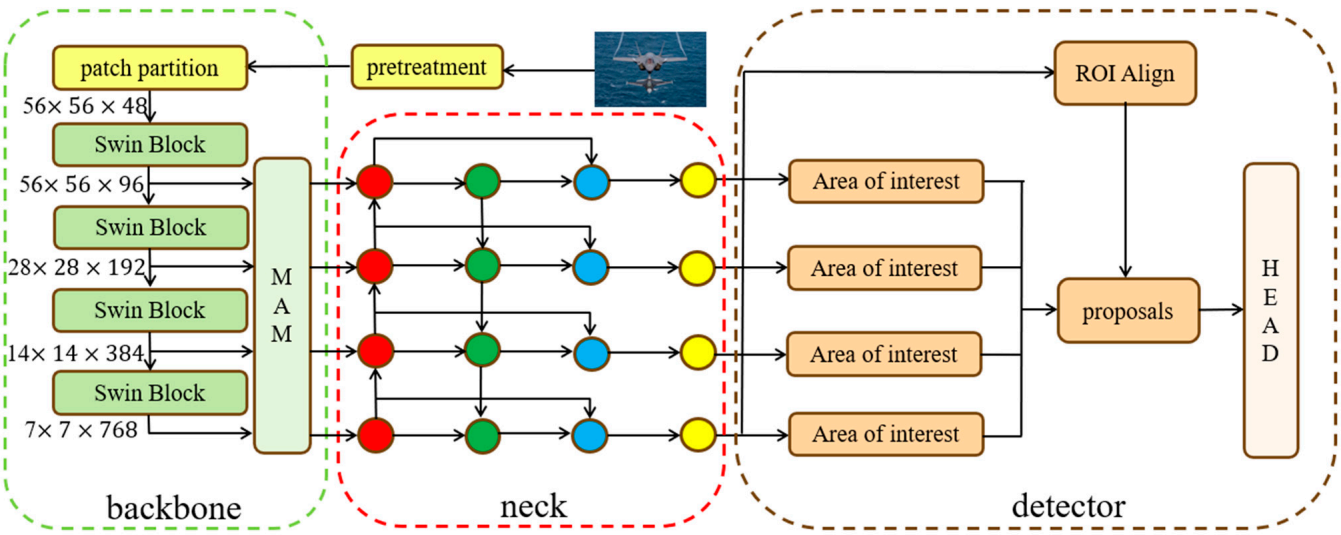

Figure 1. Overview of our proposed FADNet. The FADNet network structure consists of a backbone network consisting of a transformer network and MAM, a neck network with IPANet as the main body, and a detector consisting of multiple RPNs and HEAD.

The design premise and specific implementations of the MAM, IPANet, and MRPN are discussed in detail below.

3.1. MAM

In order to extract more air object feature information and enhance the detection performance of air object models, we designed the MAM after the backbone network. It deeply fuses the spatial W- and H-scale information and the channel local and global information with the original features to realize the multi-scale attention focus on the input features in the spatial and channel dimensions and enhance the object feature extraction capability. The MAM structure is divided into left and right parts, as shown in Figure 2. The left side constitutes the spatial attention module, which performs average pooling operations along the width W and the length H of the input feature x to $C \times H \times 1$, and it obtains the dimensional feature x_W and the $C \times 1 \times W$ dimensional feature x_H. The formulas are as follows:

$$x_W = W - AugPool(x), \tag{1}$$

$$x_H = H - AugPool(x), \tag{2}$$

where $W - AugPool$ and $H - AugPool$ are the average pooling operations of W and H, respectively.

We add x_W and x_H to the channel dimension C, and the output feature x' is input with a 1×1 convolution for channel mixing, which reduces the number of channels to C_1 and allows for the effective interaction of the feature information of each channel. The output-fused features x'' of the $C_1 \times H \times W$ dimension are separated into the $C_1 \times H \times 1$ dimension feature $x_W{}'$ and the $C_1 \times 1 \times W$ dimension feature $x_H{}'$ through the BN layer and ReLU function processing. The formulas are as follows:

$$x' = concat(x_W, x_H), \tag{3}$$

$$[x_W{}', x_H{}'] = BN(relu(Conv_1(x'))), \tag{4}$$

where *concat* represents the feature concatenation of the channel dimensions and $Conv_1$ represents a convolution with a kernel size of 1.

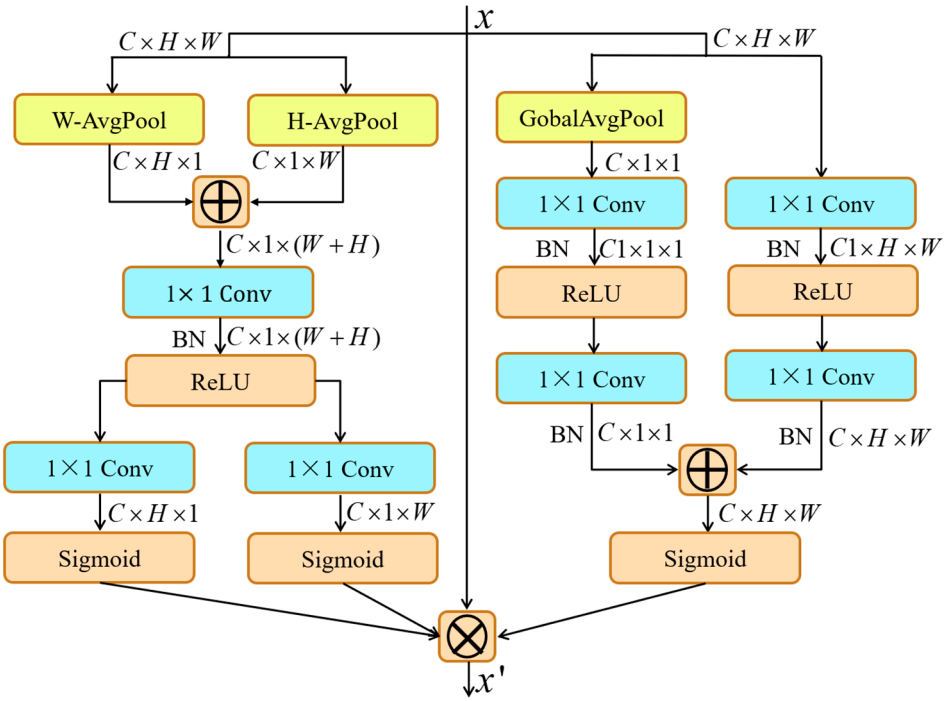

Figure 2. The MAM structure module, consisting of left and right segmentation of space and channels. Effective extraction of W and H scale features, as well as local and global features, and fusion of the extracted feature information with the original features can be realized.

$x_W{}'$ and $x_H{}'$ perform 1×1 convolutions, with the number of channels restored to C, yielding the spatial attention features $x_W{}''$ and $x_H{}''$ for the dimensions W and H, respectively, through the nonlinear operation of the sigmoid function. The formulas are as follows:

$$x_W{}'' = sigmoid(Conv_1(x_W{}'), \tag{5}$$

$$x_H{}'' = sigmoid(Conv_1(x_H{}'), \tag{6}$$

On the right is the channel attention module, which performs channel global average pooling on the input feature x to obtain the $C \times 1 \times 1$ dimensional features x_c. The x_c input is a 1×1 convolution, and the number of channels is compressed to C_1. The output $C_1 \times 1 \times 1$ dimensional features x_{c1} are processed by the BN layer and the ReLU function. The output features are then processed via a 1×1 convolution, and the number of channels is restored to C, resulting in the global attention feature x_G. The formulas are as follows:

$$x_{c1} = Conv_1(GlobalAugPool(x)), \tag{7}$$

$$x_G = Conv_1(relu(BN(x_{c1}))), \tag{8}$$

where $GlobalAugPool$ represents the global channel average pooling operation.

The other branch directly performs a 1×1 convolution operation on the input features. The output $C_1 \times H \times W$ dimensional features are then subjected to the BN layer and the ReLU function. Then, a 1×1 convolution is performed again to output the local attention features x_P. The outputs x_P and x_G are normalized by the BN layer, and then feature addition is performed in the channel dimension. The sigmoid nonlinear function is input to obtain the $C \times H \times W$ dimensional channel attention feature $x_c{}'$. The formulas are as follows:

$$x_P = Conv_1(relu(BN(Conv_1(x)))), \tag{9}$$

$$x_c{}' = sigmoid(add(BN(x_p, x_G))), \tag{10}$$

where *add* represents the feature addition operation for the channel dimension.

Finally, the spatial attention features x_W'' and x_H'' are output by the spatial attention module. The channel attention feature x_c' is output by the channel attention module, and they are multiplied by the input feature x to obtain the multi-scale attention module output feature x'. The formula is as follows:

$$x' = x \otimes x_c' \otimes x_W'' \otimes x_H'' \tag{11}$$

3.2. IPANet

PANet [13], proposed by Shun et al., includes multiple improvements to the mask R-CNN. It adds a bottom-up path to the back end of the original feature pyramid, and it uses adaptive feature pooling to incorporate full-fusion operations to address the information loss in the long path of the feature pyramid. However, PANet [13] has not completely solved this problem, especially for air objects; their few-shot characteristics and cross-scale problems mean that PANet's [13] discriminative information for object features at each scale is not sufficiently strong. As a result, we propose an improved feature pyramid based on PANet [13], as shown in Figure 3.

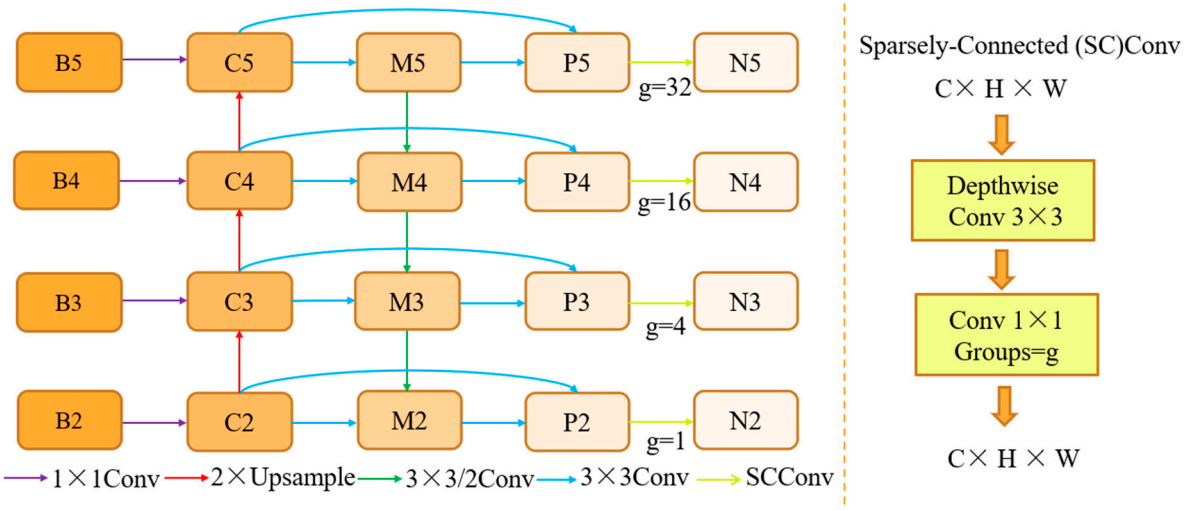

Figure 3. Schematic of the structure of IPANet, which adds jump-connected and sparse-connected convolution to PANet to enhance the extraction of features at different scales.

In Figure 3, B2, B3, B4, and B5 represent multi-scale input features with feature dimensions of $56 \times 56 \times 96$, $28 \times 28 \times 192$, $14 \times 14 \times 384$, and $7 \times 7 \times 768$, respectively. The steps of the feature pyramid can be divided into three stages. In the first stage, the input features perform 1×1 convolutions and bottom-up path up-sampling operations, strengthening the process of transmitting high-level semantic information to the low-level features. The feature channel dimension is unified to 256, and the output features are C2, C3, C4, and C5. In the second stage, the output features undergo 3×3 convolutions and bottom-up path down-sampling operations, further extracting the multi-scale information. At the same time, this strengthens the upward transmission of the low-level, strong positioning of the features to the multi-scale information, outputting the features M2, M3, M4, and M5. In the third stage, M2, M3, M4, and M5 are convolutionally processed into the output features P2, P3, P4, and P5, respectively. The formulas for the three stages are as follows:

$$C_i = Conv_1(B_i + Upsample(C_{i-1})) \quad (2 \le i \le 5, C_1 = 0), \tag{12}$$

$$M_i = Conv_3(C_i + Conv_{3/2}(M_{i-1})) \quad (2 \le i \le 5, M_1 = 0), \tag{13}$$

$$P_i = Conv_3(M_i) \quad (2 \le i \le 5), \tag{14}$$

On this basis, we propose two points of improvement. First, in the transmission of information at various scales, jump connection paths are added so that the output layer not only effectively fuses high-level and low-level feature information through the up and down paths but also retains the unmerged information of the original nodes. This reduces the loss of information during the transmission process. The formula after adding a skip connection is as follows:

$$P_i = Conv_1(C_i + Conv_3(M_i)) \quad (2 \leq i \leq 5), \tag{15}$$

Second, we add sparsely connected convolutions to the back ends of features P2, P3, P4, and P5 to further extract and fuse the multi-scale information. The specific structure of the sparsely connected convolution is shown on the right side of Figure 3, and it comprises deep 3×3 convolutions and groups of 1×1 convolutions. The deep 3×3 convolutions are used to extract information at various scales, while the groups of 1×1 convolutions are used to enhance the information fusion between the channels. In features P2, P3, P4, and P5 at various scales, because of the gradual decrease in feature levels and the gradual decrease in information interaction among the channels, the number of groups increases gradually when performing groups of 1×1 convolution operations to promote the fusion of the low-level feature information. The formulas are as follows:

$$N_2 = Conv_1^i(Conv_3(P_2)) \quad (i = 1), \tag{16}$$

$$N_3 = Conv_1^i(Conv_3(P_3)) \quad (i = 4), \tag{17}$$

$$N_4 = Conv_1^i(Conv_3(P_4)) \quad (i = 16), \tag{18}$$

$$N_5 = Conv_1^i(Conv_3(P_5)) \quad (i = 32), \tag{19}$$

where $Conv_1^i$ is a group of 1×1 convolutions and i is the number of groups.

3.3. MRPN

We contend that, for air objects, the existing structure of regional candidate networks is too simple to effectively handle the relevant information at various scales, especially in few-shot situations where the requirements for object feature selection vary from scale to scale. A single regional candidate network can constrain the information at various scales and have adverse effects on the final object recognition and positioning.

To address this issue, we improved the existing regional candidate network by building a multi-scale regional candidate network. The structure diagram is shown in Figure 4. We split a single regional candidate network into multiple networks to adapt to the feature requirements of the different scales of information and avoid conflicts caused by mixed information at the different scales. Each sub-region candidate network consists of a front-end feature extraction section and a back-end classification and localization section. The front-end processing part is composed of a 1×1 convolution and a self-adaption (SK) convolution in parallel; then, it is connected in series with a 3×3 convolution.

The SK convolution performs multi-scale feature extraction operations, adaptively extracting object features based on object information. First, the SK convolution extracts the input object feature information using a 3×3 convolution and a 5×5 convolution, concatenates the extracted information in the channel dimension, and performs global maximum pooling processing. After passing through two fully connected layers, one-dimensional feature data of the channel are obtained. Second, the obtained data are input into the softmax function, and 3×3 and 5×5 convolution feature information weights are output. The proportion of each convolution in the adaptive convolution of the candidate network in the scale region is then clarified and multiplied by the extracted

feature information of each convolution. The output result is concatenated in the channel dimension to output the adaptive convolution result. The *SK* convolution formulas are as follows:

$$U_2' = Conv_3(V) \otimes \text{Softmax}(FC_2(GobalAvgPool(Conv_3(V)))), \tag{20}$$

$$U_2' = U_3' = Conv_5(V) \otimes \text{Softmax}(FC_2(GobalAvgPool(Conv_5(V)))), \tag{21}$$

$$V1 = U_2' \oplus U_3', \tag{22}$$

where $Conv_3$ and $Conv_5$ represent the 3×3 and 5×5 convolution operations, respectively, and FC_2 represents two fully connected layer operations.

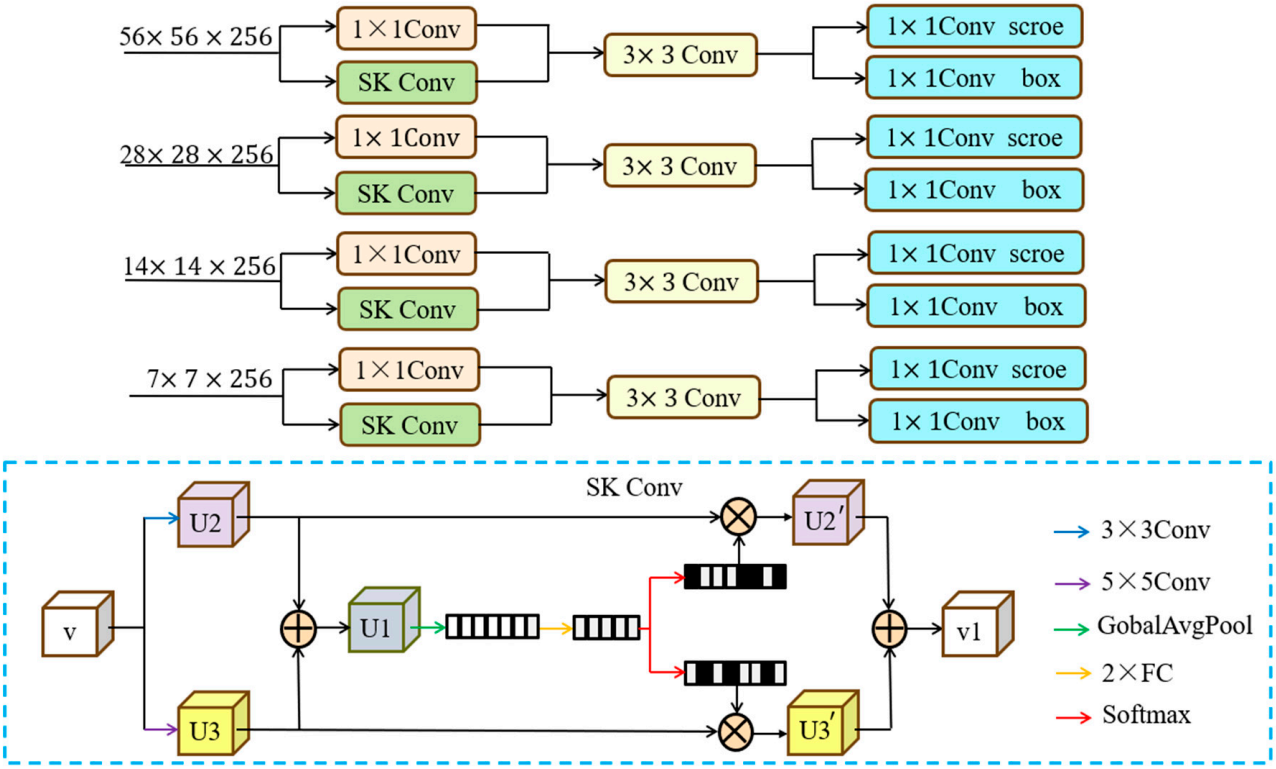

Figure 4. Structure of MRPN. We improve the single RPN structure by transforming it into multiple RPNs corresponding to the output feature maps at each scale; meanwhile, we add adaptive convolution to each single RPN to enhance the ability to process features at each scale.

After the *SK* convolution processing, the output result is multiplied by a 1×1 convolution output result, and the 3×3 convolutions that are input into the object features are extracted and input into the back-end classification and positioning part. After the classification 1×1 and the positioning 1×1 convolution processing, the preliminary position and foreground information of the object are obtained, fused to form an object candidate box, and input into the detection head network. Thus, we obtain the recognition results and precise positions of the air object.

4. Experimental Results and Analysis

4.1. Experimental Setup

The hardware platform configuration for the experimental training phase is shown in Table 1. This study used the Pytorch deep learning development framework for the experiments.

Table 1. The hardware platforms for model training. Our model training hardware platforms are divided into six categories: GPU, CPU, GPU memory size, computing platform, operating systems, and CPU (test).

Name	Related Configuration
GPU	NVIDIA Quadro GV100
CPU	Inter Xeon Silver 4210/128 G
GPU memory size	32 G
Operating system	Win10
Computing platform	CUDA10.2
CPU (test)	Inter Core i7 10700/16 G

4.2. Few-Shot Dataset of Military Air Objects

Our few-shot dataset included five types of military air attack objects: F35, Su57, MQ9, RQ4, and B2. The dataset was divided into training, validation, and testing sets. The training set included one, two, three, five, and ten photos of the five object types, according to the requirements of the few-shot tasks, and one, two, three, five, and ten shots were used for model training. At the same time, ten photos were provided for each of the five object types to form a validation set to assist in the model training. In addition, we provided five images from each of the five object categories, totaling twenty-five images, to form a test set for testing the performance of the model. Finally, the image labeling software LabelImg 1.6.0 was used to label the sample data in the training, validation, and test sets using a dataset label format similar to the PASCAL VOC data label format.

4.3. Evaluating Indicator

The definitions of detection precision and object recall in deep learning are shown in Formulas (23) and (24), respectively, as follows:

$$Precision = \frac{N_t}{N_t + N_f}, \tag{23}$$

$$Recall = \frac{N_t}{N_r}, \tag{24}$$

where N_t is the number of real objects detected by the algorithm, N_f is the number of false objects detected by the algorithm, and N_r is the number of real weak objects that actually exist in the image. The average precision AP is a combination of detection precision and object recall. According to the calculation method in [2], the confidence threshold was set to 0.5 to evaluate the detection performance of the detection model for a single category. The average AP of the detected categories was used to evaluate the overall performance of the detection model. The expression is shown in Formula (25):

$$mAP = \sum_{i}^{n} AP, \tag{25}$$

where n represents the total number of categories (normally $0 < i \leq n$). The higher the values of AP and mAP, the better the detection performance of the model, and vice versa.

4.4. Implementation Details

We used FADNet as our network model for the network implementation. In the training phase of the base class, we used 15 classes of objects in the PASCAL VOC dataset as the base-class dataset, except for birds, buses, cows, motorbikes, and sofas. We used motorbikes and sofas as the base-class dataset. The training process applied the SGD optimizer with 15,000 iterations, a learning rate of 0.02, a batch size of 16, a momentum of 0.9, and a weight decay of 0.0001. In the fine-tuning phase, the learning rate was 0.001;

the iterations of the 1-, 2-, 3-, 5-, and 10-shot tasks were 3000, 6000, 9000, and 15,000, respectively; and the batch size, momentum, and weight decay were unchanged.

4.5. Analysis of the Results of the Air Object Comparison Experiments

We used TFA/fc [7], TFA/cos [7], Attention RPN [35], TIP [42], DCNet [43], MPSR [36], LVC [12], and our algorithm to detect the military air objects in the few-shot dataset constructed in this study. A comparison of the detection results is shown in Table 2. Clearly, the algorithm proposed in this paper exhibited the strongest detection ability, especially in the three-shot task, showing a significant improvement in detection performance compared with the other algorithms. Compared with the suboptimal algorithm MPSR, the overall performance increased by an average of 1.1%, effectively improving the detection precision for air objects in various shot tasks. Figure 5 shows the decoupling-based algorithm we designed, which was combined with four AP@0.5 combinations of TFA/fc [7], TFA/cos [7], Attention RPN [35], TIP [42], DCNet [43], MPSR [36], and LVC [12]. A comparison of the visual output results of the network models discussed above is shown in Figure 5.

Table 2. Results for the comparison of our method with the remaining seven few-shot object detection methods on the air object-based datasets for 1-, 2-, 3-, 5-, and 10-shot tasks, where the results in bold are the optimal results.

Methods/Shot	*AP* @0.5				
	1	**2**	**3**	**5**	**10**
TFA/cos [7]	17.2	23.5	23.1	23.4	29.2
TFA/fc [7]	25.6	29.4	27.4	37.1	44.2
Attention RPN [34]	23.8	29.2	29.1	39.4	54.3
TIP [42]	24.4	30.1	33.5	39.1	54.6
DCNet [43]	25.7	30.3	35.4	39.7	54.8
LVC [12]	26.4	30.8	37.5	41.2	55.1
MPSR [35]	27.4	31.7	38.8	41.7	55.3
Ours	**33.2**	**36.8**	**43.3**	**47.2**	**60.4**

4.6. Air Object Ablation Experiment

The algorithm designed in this study is proposed using FADNet, based on the multi-scale problem of air objects and constructed from the MAM, IPANet, and MRPN. To evaluate the degree of optimization of the algorithm's performance according to different module combinations and improvements, we designed ablation experiments. Table 3 shows the results of the ablation experiments, which were validated on a few-shot dataset of air objects under the same experimental conditions.

Table 3. Experimental results of our proposed FADNet method for the ablation of military-type air objects.

Baseline	Transformer	MAM	I-PANet	M-RPN	*mAP* @0.5				
					1 Shot	**2 Shots**	**3 Shots**	**5 Shots**	**10 Shots**
√	√				25.6	29.4	27.4	37.1	44.2
√	√				30.5	34.2	32.5	42.4	49.5
√	√	√			31.3	35.7	36.7	43.6	54.1
√	√	√	√		32.6	36.2	40.4	44.5	56.8
√	√	√	√	√	33.2	36.8	43.3	47.2	60.4

The experimental results indicated that the different combinations had positive impacts on the overall performance of the model, with the baseline models scoring 25.6, 29.4, 27.4, 37.1, and 44.2 for the 1-, 2-, 3-, 5-, and 10-shot tasks, respectively. After replacing the backbone network resnet-101 with the transformer network, the shot tasks increased by an

average of 5.1 percentage points compared with the baseline model, improving its ability to process multi-scale information significantly. After adding the MAM, effective feature fusion was achieved for the feature information in the channel and spatial dimensions, and effective aggregation of the local and global features was achieved in the channel dimension. The performance of each shot task increased to varying degrees, especially at 3 and 10 shots, which increased by 4.3 and 4.6 percentage points, respectively. In response to the characteristics of the air object, we improved the original PANet network after the MAM, further integrating the multi-scale features of the air object. With this improvement, the detection performance increased most significantly for 10 shots by 2.7 percentage points. To further improve the processing ability of the multi-scale features and to solve the multi-scale problem of air objects, we built an MRPN to refine each shot task, especially for the 5- and 10-shot tasks, which achieved rapid increases of 3.8 and 3.7 percentage points, respectively, indicating a stronger detection precision of the model.

(**a**) (**b**) (**c**) (**d**)

Figure 5. Visualization results for our proposed method, FADNet, compared with Attention RPN, MPSR, and the reference map on the air object 10-shot task test set. (**a**) Attention RPN; (**b**) MPSR; (**c**) ours; (**d**) reference map.

4.7. Analysis of the Detection Results for the PASCAL VOC Dataset

We used TFA/fc [7], TFA/cos [7], Attention RPN [35], TIP [42], DCNet [43], MPSR [36], LVC [12], FORD+BL [44], and our algorithm for few-shot object detection on the PASCAL VOC dataset. A total of 5 types of objects—birds, buses, cows, motorbikes, and sofas—were identified as new-class objects, while the remaining 15 out of 20 were identified as the base classes. The detection results are shown in Table 4; the optimal results in the 1-, 2-, 3-, 5-, and 10-shot tasks are given in bold. The results of our proposed method in the 1- and 2-shot tasks are lower than those of the FORD+BL algorithm and, thus, suboptimal, but our method performed better than the other algorithms. Meanwhile, in the 3-, 5-, and 10-shot tasks, the proposed method yielded optimal results, performing the best of all algorithms, with the highest performance achieved for the 3-shot task, 3.3 percentage points greater than that of the next best performer. The smallest performance improvement occurred with the 10-shot task, with an increase of 1.4 percentage points.

Table 4. Results for the comparison of our method with the remaining eight few-shot object detection methods on the PASCAL VOC-based dataset's 1-, 2-, 3-, 5-, and 10-shot tasks, where the bolded numbers are the optimal results.

Methods/Shot	***AP* @0.5**				
	1	**2**	**3**	**5**	**10**
Attention RPN [35]	35	36	39.1	51.7	55.7
TFA/fc [7]	36.8	39.1	43.6	55.7	57
TFA/cos [7]	39.8	36.1	44.7	55.7	56
TIP [42]	27.7	36.5	43.3	50.2	59.6
DCNet [43]	33.9	37.4	43.7	51.1	59.6
LVC [12]	36.0	40.1	48.6	57.0	59.9
FORD+BL [44]	**46.3**	**54.2**	49.9	56.3	61.8
MPSR [36]	41.7	42.5	51.4	55.2	61.8
Ours	43.6	46.4	**54.7**	**57.8**	**63.2**

Compared with the results for the air object dataset, the results of our proposed method for the PASCAL VOC dataset show a smaller advantage. This is because our proposed method is based on the characteristics of air objects; the dataset is more specific to air objects. Unlike air objects, which exhibit less individual variability, natural objects dominate the PASCAL VOC dataset with a larger number of categories (20 in total) and exhibit more variability in various classes. This results in more 1- and 2-shot tasks when the number of samples is small. The effectiveness of our method decreases, and although our method outperforms as the number of samples rises (i.e., in the 3-, 5-, and 10-shot tasks), the resultant superiority of our method for the PASCAL VOC dataset is still lower than that for the air object dataset. However, the overall dominance of our method for the PASCAL VOC dataset, especially when the shot number is large, demonstrates the effectiveness and generalization ability of our method. A comparison of the visual output results of the network models discussed above is shown in Figure 6.

Figure 6. Visualization results for our proposed method FADNet versus TFA/cos, MPSR, and reference graphs on the 10-shot task from the PASCAL VOC test set. (**a**) TFA/cos; (**b**) MPSR; (**c**) ours; (**d**) reference map.

5. Conclusions

In this study, we aimed to solve the detection problems introduced by the few-shot and multi-scale characteristics of air objects. On this basis, we propose FANet, which realizes the deep fusion of the object features with the original features in the spatial dimension and the channel dimension by using our multi-scale attention mechanism, enhancing the extraction ability for the object features. Meanwhile, we improve the PANet network and propose IPANet, which introduces jump and sparse connection convolutions to enhance the discriminative power of the neck network for features at all scales. In addition, we propose a multi-scale regional candidate network that establishes multiple regional candidate networks based on the output feature scale and improves the extraction level of multi-scale object features through adaptive convolution. Finally, by conducting comparative experiments on two datasets, we subjectively and objectively compared our proposed method with mainstream methods. The results show that our method improves the ability to detect air objects. By analyzing the results using the PASCAL VOC data, we found that the proposed method performs well in natural object detection overall. However, in the case of very few samples, such as in 1- and 2-shot tasks, slight deficiencies were noted, and the detection accuracy was not optimal. Therefore, improving the detection performance for natural objects and enhancing the generalization ability of FADNet is a promising direction.

Author Contributions: Conceptualization, X.W.; validation, Z.Y.; formal analysis, X.W.; data curation, X.J.; writing—review and editing, X.D.; visualization, X.J.; supervision, W.C.; resources, W.G.; project administration, W.G. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to acknowledge the National Defense Science and Technology 173 Program Technical Field Fund Project (grant no. 2021-JCJQ-JJ-0871) for funding their experiments.

Data Availability Statement: Data related to the current study are available from the corresponding authors upon reasonable request. The code used during the study is available from the corresponding authors upon request.

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

References

1. Demirel, B.; Baran, O.B.; Cinbis, R.G. Meta-tuning Loss Functions and Data Augmentation for Few-shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7339–7349.
2. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9577–9586.
3. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8420–8429.
4. Gao, Z.; Guo, S.; Xu, C.; Zhang, J.; Gong, M.; Del Ser, J.; Li, S. Multi-domain Adversarial Variational Bayesian Inference for Domain Generalization. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, 1. [CrossRef]
5. Cao, Y.; Wang, J.; Jin, Y.; Wu, T.; Chen, K.; Liu, Z.; Lin, D. Few-shot object detection via association and discrimination. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16570–16581.
6. Zhang, G.; Luo, Z.; Cui, K.; Lu, S.; Xing, E.P. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, 1–12. [CrossRef] [PubMed]
7. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv* **2020**, arXiv:2003.06957.
8. Chen, T.I.; Liu, Y.C.; Su, H.T.; Chang, Y.C.; Lin, Y.H.; Yeh, J.F.; Chen, W.C.; Hsu, W. Dual-awareness attention for few-shot object detection. *IEEE Trans. Multimed.* **2023**, *25*, 291–301. [CrossRef]
9. Perez-Rua, J.M.; Zhu, X.; Hospedales, T.M.; Xiang, T. Incremental few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 13846–13855.
10. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
11. Fan, Z.; Ma, Y.; Li, Z.; Sun, J. Generalized few-shot object detection without forgetting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 4527–4536.
12. Kaul, P.; Xie, W.; Zisserman, A. Label, verify, correct: A simple few-shot object detection method. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 14237–14247.
13. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, ON, USA, 24–29 June 2014; pp. 580–587.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE international Conference on Computer Vision, Boston, MA, USA, 8–10 June 2015; pp. 1440–1448.
16. 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]
17. Jiang, X.; Cai, W.; Yang, Z.; Xu, P.; Jiang, B. IARet: A lightweight multiscale infrared aerocraft recognition algorithm. *Arab. J. Sci. Eng.* **2022**, *47*, 2289–2303. [CrossRef]
18. Jiang, X.; Cai, W.; Ding, Y.; Wang, X.; Yang, Z.; Di, X.; Gao, W. Camouflaged Object Detection Based on Ternary Cascade Perception. *Remote Sens.* **2023**, *15*, 1188. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
20. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
21. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
22. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
23. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; pp. 21–37.
24. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
25. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.

26. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
27. Jamal, M.A.; Qi, G.J. Task agnostic meta-learning for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 11719–11727.
28. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-based metric learning for classification and few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5197–5206.
29. Wang, P.; Liu, L.; Shen, C.; Huang, Z.; Van Den Hengel, A.; Tao Shen, H. Multi-attention network for one shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2721–2729.
30. Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J.B.; Isola, P. Rethinking few-shot image classification: A good embedding is all you need? In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIV 16; pp. 266–282.
31. Sun, Q.; Liu, Y.; Chua, T.S.; Schiele, B. Meta-transfer learning for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 403–412.
32. Liu, Y.; Sun, Q.; Liu, A.A.; Su, Y.; Schiele, B.; Chua, T.S. Lcc: Learning to customize and combine neural networks for few-shot learning. *arXiv* **2019**, arXiv:1904.08479.
33. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 4013–4022.
34. Li, B.; Yang, B.; Liu, C.; Liu, F.; Ji, R.; Ye, Q. Beyond max-margin: Class margin equilibrium for few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 7363–7372.
35. Yin, L.; Perez-Rua, J.M.; Liang, K.J. Sylph: A hypernetwork framework for incremental few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 9035–9045.
36. Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; Zhang, C. Defrcn: Decoupled faster r-cnn for few-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 8681–8690.
37. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-scale positive sample refinement for few-shot object detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XVI 16; pp. 456–472.
38. Khandelwal, S.; Goyal, R.; Sigal, L. Unit: Unified knowledge transfer for any-shot object detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 5951–5961.
39. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. Fsce: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 7352–7362.
40. Zhang, W.; Wang, Y.X.; Forsyth, D.A. Cooperating RPN's Improve Few-Shot Object Detection. *arXiv* **2021**, arXiv:2011.10142.
41. Zhou, P.; Ni, B.; Geng, C.; Hu, J.; Xu, Y. Scale-transferrable object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–22 June 2018; pp. 528–537.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
43. Li, A.; Li, Z. Transformation invariant few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 3094–3102.
44. Vu, A.K.N.; Nguyen, N.D.; Nguyen, K.D.; Nguyen, V.T.; Ngo, T.D.; Do, T.T.; Nguyen, T.V. Few-shot object detection via baby learning. *Image Vis. Comput.* **2022**, *120*, 104398. [CrossRef]