Article

Aircraft Target Interpretation Based on SAR images

Xing Wang , Wen Hong * , Yunqing Liu , Dongmei Hu , Ping Xin

Posted Date: 10 August 2023

doi: 10.20944/preprints202308.0800.v1

Keywords: YOLOv5s, attention mechanism, decoupled detection head, K-means++, aircraft detection, SAR image.

doi:10.20944/preprints202308.0800.v1

Article

Aircraft Target Interpretation Based on SAR Images

Xing Wang [1,2], Wen Hong [1,*], Yunqing Liu [1], Dongmei Hu [2] and Ping Xin [2]

[1] College of Electrical and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China; xingwang_1983@126.com (X.W.); mzlyq@cust.edu.cn (Y.L.)
[2] College of Electrical and Information Engineering, Beihua University, Jilin 132013, China; m13844809391@163.com (D.H.); xinping@beihua.edu.cn (P.X.)
* Correspondence: m15699554195@163.com

Abstract: Synthetic Aperture Radar (SAR) is an active sensor that uses microwave for sense, it is unrestricted by weather and illumination conditions, and it can observe targets all day and weather. Aircraft targets are important monitoring objects in military and civilian fields, and how to efficiently detect and recognize aircraft targets is an important topic in the field of SAR image interpretation. Based on the features of SAR images, such as complex background, high resolution, and multi-scale, we proposed an improved method based on YOLOv5s. Firstly, this paper proposed the structure of the multi-scale receptive field and channel attention fusion, which is applied to the shallow layer of the backbone of YOLOv5s, it can adjust the weights of the multi-scale receptive field during the training process to enhance the extraction ability of feature information. Secondly, we proposed four decoupled detection heads to replace the original part in YOLOv5s, which can improve the efficiency and accuracy of SAR image interpretation for small targets. Thirdly, in the case of the limited amount of SAR images, this paper proposed multi methods of data augmentation, which can enhance the diversity and generalization of the network. Fourthly, this paper proposed the K-means++ to replace the original K-means to improve the network convergence speed and detection accuracy. Finally, Experiments demonstrate that the improved YOLOv5s can enhance the accuracy of SAR image interpretation by 9.3%, and the accuracy of small targets is improved more obviously, reaching 13.1%.

Keywords: YOLOv5s; attention mechanism; decoupled detection head; K-means++, aircraft detection; SAR image

1. Introduction

SAR has been developed since the 1950s. Due to unique imaging features, which have been widely used in both military and civilian fields [1,2]. In the military aspect, SAR can be used for global strategic reconnaissance all day, which can play a crucial role in the victory of the war. In the civil aspect, SAR plays an active role in mineral resource detection, disaster detection, and prevention. Due to the unique performance of SAR, the interpretation technology of SAR images has been taken seriously by various countries. Aircraft is a crucial target in both military and civilian fields, which is very important to realize the rapid detection and recognition of aircraft targets in SAR images. The detection and recognition method of vehicle and ship targets have formed a relatively mature system, but the development of aircraft targets is still relatively backward. Due to two factors, Firstly, the characteristics of aircraft targets are more complex than those of vehicle and ship targets, there are many parts with different scattering characteristics in the fuselage, and the brightness of each part of the SAR image varies. Secondly, the quantity of available SAR images is limited for aircraft targets, and the acquisition cost is high. These factors limit the development of detection and recognition algorithms for aircraft targets of SAR images.

The traditional interpretation process of SAR images contains three stages. Firstly, all the suspicious objects need to be detected, and the universal algorithm contains the constant false alarm rate (CFAR). On this basis, scholars have proposed a variety of improved CFARs, such as cell average constant false alarm rate (CA-CFAR), smallest of cell average constant false alarm rate (SOCA-CFAR), greatest of cell average constant false alarm rate (GOCA-CFAR), ordered statistic constant false alarm rate (OS-CFAR), and variability index constant false alarm rate (VI-CFAR). Secondly, the detector can take advantage of the characteristics of targets to eliminate false alarms, and the common characteristics include contour, size, texture, scatter center, and so on [3–6]. Finally, the recognizer can classify the object of targets, and the universal algorithm contains template-based methods, mathematical model-based methods, and machine learning-based methods [7,8]. Based on these, scholars have proposed some improved methods [9–11]. Although these methods can shorten detection time and improve detection precision to a certain extent, the traditional interpretation methods are difficult to satisfy the needs of detection speed and accuracy, and their generalization ability is poor, which is greatly affected by the datasets.

In recent years, driven by artificial intelligence technology, deep learning algorithms have been applied to the following fields, face recognition, natural language recognition, vehicle detection, and so on. All of these applications have achieved fruitful results. The deep learning algorithm is composed of a deep network structure, which can extract deeply abstract semantic information. The detection and recognition performance based on the deep learning algorithm has already exceeded that of the traditional interpretation method. And more and more excellent algorithms have been proposed. The existing deep learning algorithm for detection and recognition mainly divides into one-stage detection and two-stage detection, which are distinguished by whether candidate boxes are predicted. The two-stage detection includes regions with convolutional neural network (R-CNN) [12], Fast R-CNN [13], Faster R-CNN [14], and Mask R-CNN [15]. They are using the region proposal network (RPN) to generate candidate boxes, then perform regression and classification of candidate boxes. The one-stage detector includes SSD [16], Retina-Net [17], YOLO [18–20]. The algorithms of the YOLO series are generally faster than others, and it has a good performance for small objects, so it has been widely applied in computer vision and pattern recognition field recently. Many scholars have conducted research based on the YOLOv5 version, and many improved methods have also been proposed. Zheng et al. [21] proposed a multi-scale detection network RebarNet with an embedded attention mechanism based on YOLOv5 to solve the problem of missed and false detection in dense small object detection. Hou et al. [22] proposed integrating coordinate attention (CA) into the backbone of the deep network, enhancing the expression ability of features and the precision of detection and recognition. Tan et al. [23] proposed adding one more detection branch based on the YOLOv5 structure, which can improve the efficiency of detection and recognition for small targets. Yuan et al. [24] proposed the feature fusion method of multi-scale adaptive, which can retain more useful feature information. It can improve the precision of detection and recognition for multi-scale targets.

SAR images of aircraft usually cover multi-scale targets, which contain lots of background noise. The detection and recognition accuracy of existing algorithms has great development potential, especially for small aircraft targets. This paper focused on the problem, of how to enhance the precision of small aircraft targes and how to suppress the background noise from scattering spots of ground-interfering objects. An improved YOLOv5s is proposed for the above-mentioned problems. There are four primary differences with the original YOLOv5s. Firstly, we proposed the multi-scale receptive field and channel attention (MRFCA) structure based on SENet [25]. MRFCA can integrate into the backbone of YOLOv5s, can change the adaptively receptive field for multi-scale targets, and capture more relevant information and critical features. Secondly, an additional detector is introduced into YOLOv5s, and all the detectors adopt decoupled operations. The new four decoupled detection heads (4DDH) structure can improve detectability for multi-scale targets, and enhance detection precision for small targets. Thirdly, we proposed the multi-method of data augmentation fusion together, it can enhance the diversity of datasets and generalization ability of the model, and prevent overfitting [26]. Finally, this paper proposed the K-means++ to replace the original K-means to improve the network convergence speed and detection accuracy. The experiment shows that the mean average precision (mAP) based on improved YOLOv5s is better than the result based on Faster-RCNN, Retina-Net, SSD, YOLOv3, and the original YOLOv5s algorithm.

2. Related Work

This section includes three parts. Part 1 describes the YOLOv5s network. Part 2 introduces the SENet network. Part 3 introduces the Inception network [27].

2.1. YOLOv5s Network

YOLOv5 was proposed in June 2020 and provided four models with different parameter sizes, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5s is the lightest model, it runs fast, but the precision is lower than the other three models. YOLOv5s consists primarily of input, backbone, neck, and head. YOLOv5s architecture is shown in Figure 1. Our improvement method is based on YOLOv5s.

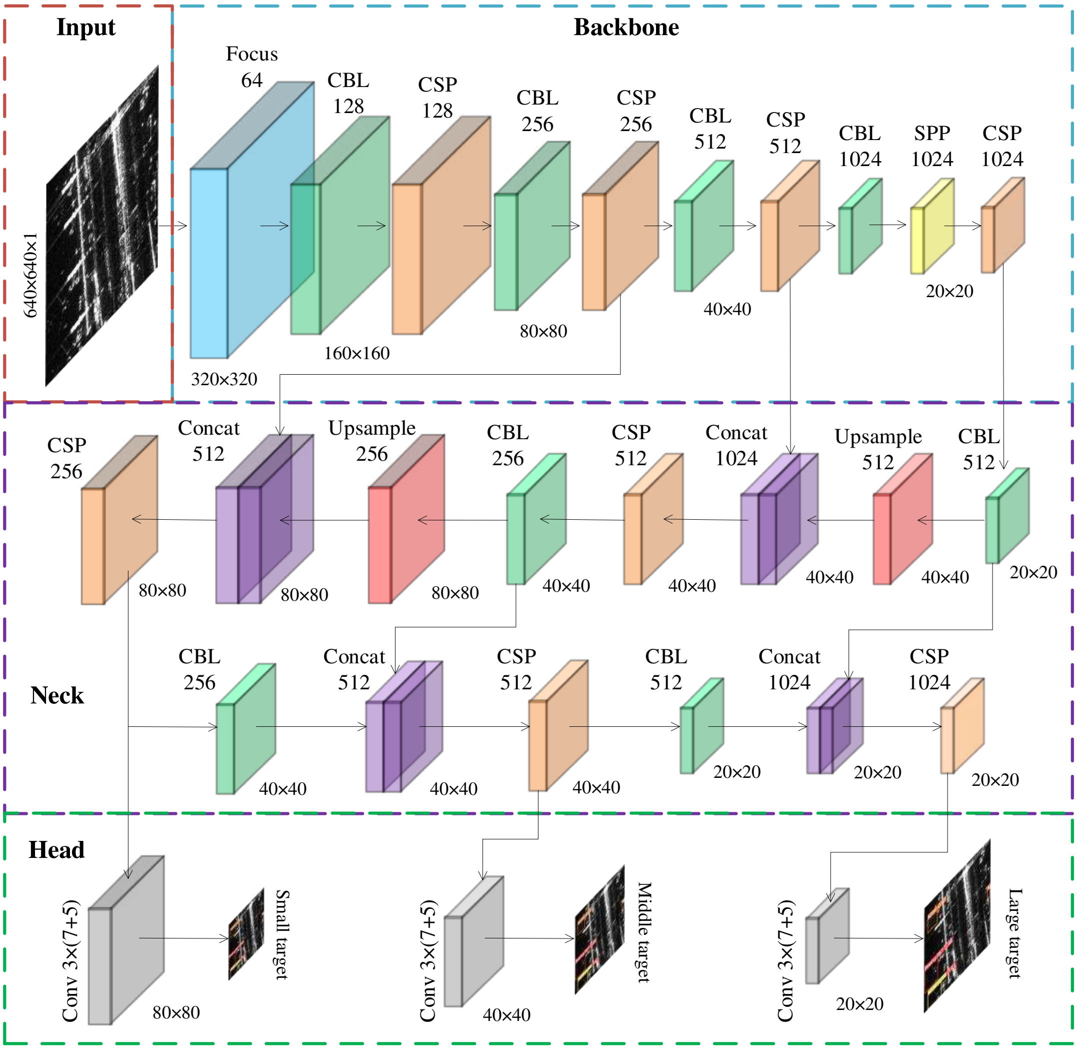

Figure 1. YOLOv5s network architecture.

The input section is responsible for data augmentation. The data augmentation method can generate new images by processing input images, which can enhance the diversity of the datasets and generalization ability, enrich the background information of input images, and improve the convergence speed of the network. The backbone section is used to extract image features. It contains Focus, CBL, CSP, and SPP structures. Focus can increase the feature dimension of the channel by slicing input images, and it can realize down-sampling without information loss; CBL can further extract the image feature by convolution, batch normalization, and leaky-ReLU operations; CSP can reduce information loss during feature transmission, improve learning ability and reduce the amount of computation; SPP can convert the feature map of arbitrary size into the feature vector of fixed size so that the network can input images of arbitrary scale. The neck section is used for feature fusion of different reception fields. It adopts PANET structure [28], which is designed based on the feature pyramid network (FPN) structure [29], it integrates down-top and top-down bidirectional information transmission paths, and it merges the feature map of different reception fields that are extracted from the backbone section, each feature layer can get rich semantic and detailed information, it can enhance feature expression ability of the network. The head section config three detectors, which are used to predict classes, confidence, and anchor box of targets of different scales.

2.2. Channel Attention Mechanism

To improve the feature expression ability of the convolutional neural network, some kinds of attention mechanisms have been proposed, such as SENet, CBAM [30], and SKNet [31]. The visual saliency of SAR images is different from that of natural images. Aircraft targets of SAR images show strong scattering spots and complex background noise. The network can more easily capture the feature region with significant contribution by integration attention mechanism. Different weights can automatically assign to each channel of the feature map by network learning, which can improve the expression ability of network features. SENet is the winner of the ImageNet2017 classification competition. The module is shown in Figure 2. It can learn the relationship between different channels, keep the attention of the network to channel information, evaluate the score of each channel, and obtain weights of each channel adaptively.

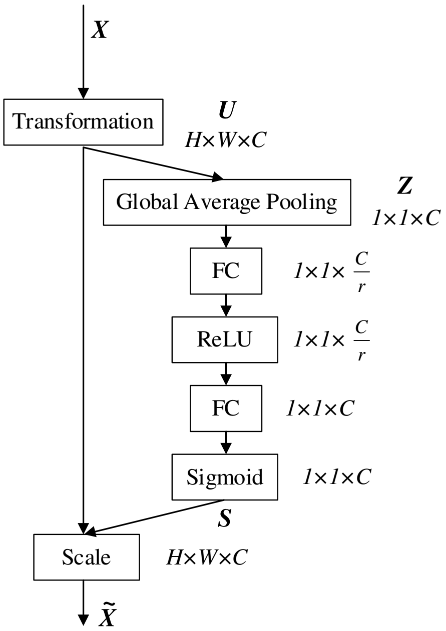

Figure 2. SENet module.

where *X* is the input feature map, and *U* ∈ *R* (*H*×*W*×*C*) is obtained based on the transformation of *X*, such as convolution, inception, or residual operation. *U* is squeezed along the spatial dimension by global average pooling operation to obtain *Z* ∈ *R* (*1*×*1*×C). Each spatial dimension of the feature map compresses into one pixel, which has a global receptive field of each spatial dimension. The output Z_c is expressed as follows:

$$Z_c = F_{sq}(U_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \qquad (1)$$

where U_c represents the *c*-th channel of the feature map, $U_c(i,j)$ represents the pixel of the *i*-th row and *j*-th column of the U_c, and F_{sq} represents the global average pooling operation. To reduce the complexity of the module, a fully connected (FC) operation is used to reduce the channel dimension to obtain *R* (*1*×*1*×*C*/*r*), the length of the channel dimension is compressed by r times, then perform ReLU activation function, and perform FC operation again to restore the channel dimension *R*

doi:10.20944/preprints202308.0800.v1

(*1×1×C*), finally, it is activated by sigmoid function to generate the weights for each channel, The output S of the sigmoid function is expressed as follows:

$$S = F_{ex}(Z, W) = \sigma[W_2\delta(W_1 Z)] \tag{2}$$

where F_{ex} is the feature excitation and reweight function, W_1 and W_2 are two FC operations, but the dimension of their input and output channel is different, δ is the ReLU activation function, σ is the sigmoid activation function, and S is the weight matrix for each channel. The output $\widetilde{X_c}$ is expressed as follows:

$$\widetilde{X_c} = F_{scale}(U_c, S_c) = S_c \cdot U_c \tag{3}$$

where S_c is the weight of *c*-th channel, and $\widetilde{X_c}$ is the final output with the weight information of each channel. After the integration of the SENet module, the extraction ability of useful features is enhanced.

2.3. Inception Network

The inception network is a milestone in the development of convolutional neural networks. Before, the network often increased the depth to achieve better performance, but in this way, the model is easier to overfit and brings lots of computation. Inception is used for solving the problem of convolutional layer stacking on the premise of ensuring the quality of the model, it sets up multiple channels in the same layer, which is obtained by operation with convolution kernels of different sizes, and it expands the width of the network. Inception can avoid redundant computation, reduce the number of parameters, and extract high dimensional features of SAR images more effectively. Inception can enhance the generalization ability and feature expression ability of the network, and the feature learning ability of the convolutional neural network is greatly improved.

The inception network performs multiple convolution operations with different scale kernels at the input feature map in parallel to get four branches. All the branches are concatenated into one feature map in the output section. Figure 3 shows the architecture of the inception network. It contains three convolution kernels with different scales and one maximum pooling layer. Each branch performs its convolutional operations, which can extract much richer features and reduce computational complexity. Inception can fuse the different scale features of the SAR image, obtain receptive fields of different scales, and improve the performance of the network.

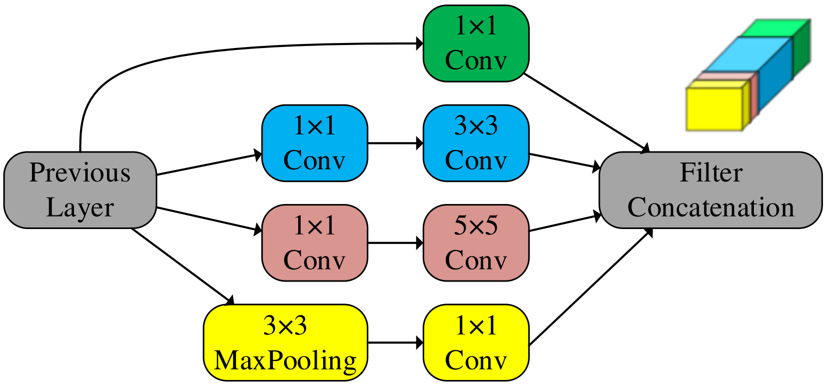

Figure 3. Inception network.

3. Method

Compared with traditional optical images, SAR images are complex, diverse, and contain lots of background noise. The precision of detection and recognition is not satisfactory when using the YOLOv5s module directly, which always leads to the loss of small objects targets. The following four improved methods are proposed based on the YOLOv5s module to solve this problem.

(1) MRFCA module integrates into the backbone of YOLOv5s, which can change the adaptively receptive field to improve feature exaction ability.
(2) 4DDH structure replaces the original part of YOLOv5s to improve the detection and recognition ability for small objects.
(3) Flip, scaling, and mosaic [32] data augmentation method fuse together to enhance the diversity of datasets and generalization of the model.
(4) This paper adopted the K-means++ to replace the original K-means algorithm to improve the network convergence speed and detection accuracy.

3.1. Multi-scale Receptive Field and Channel Attention Fusion (MRFCA)

There are significant differences in the size of aircraft targets in our datasets, the YOLOv5s algorithm has a weak feature extraction ability for multi-scale targets, especially in the case of complex backgrounds and lots of noise. Moreover, the YOLOv5s algorithm is more likely to miss small aircraft targets in the detection and recognition process. This paper proposed an MRFCA module, which combines the inception network and channel attention mechanism, it has better stability of detection and recognition for the targets with different scales, and it can adaptively adjust the weights of different receptive fields according to the input information.

The inception network can expand the width of the network. We proposed to create three branches to extract the feature information of different receptive fields, each branch uses convolution kernels of different sizes for operation. But once the training is done, the parameters of the inception network are fixed, and the weights of different receptive filed are not dynamically adjusted. We proposed to incorporate the channel attention mechanism into the inception network. It takes into account not only the relationship between channels but also the importance of the different receptive fields. With different images input, the weights are dynamically adjusted. MRFCA network is shown in Figure 4. It contains split, fuse, and select.

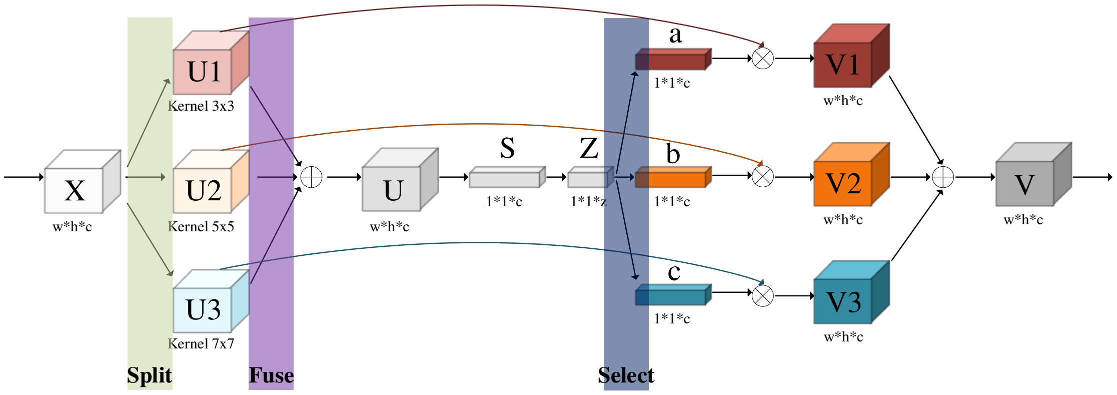

Figure 4. MRFCA network.

where $X \in R\ (h \times w \times c)$ is the input feature map, X is split into 3 branches $U1$, $U2$, and $U3$, they are separately convolved operations with kernels of different sizes. $U1 \in R\ (h \times w \times c)$ is the convolution of X and kernel (3×3), $U2 \in R\ (h \times w \times c)$ is the convolution of X and kernel (5×5), $U3 \in R\ (h \times w \times c)$ is the convolution of X and kernel (7×7). To improve the calculation speed and reduce the number of parameters, kernel (5×5) and kernel (7×7) adopt dilated convolution method [35]. The fuse operation is an element-wise summation of $U1$, $U2$, and $U3$, $U \in R\ (h \times w \times c)$ is the feature map of fusion, it contains the feature information with different receptive filed, the operation is expressed as

$$U = U1 + U2 + U3 \tag{4}$$

where $S \in \mathrm{R}\ (1 \times 1 \times c)$ is the feature vector of the global average pooling operation at the dimensions w and h of U, the dimension of each channel is composed into one pixel, which has a global receptive field of each channel. To further improve the efficiency of the network and reduce the number of parameters, S carries out cross-channel information interaction through the FC operation to get $Z \in$

R (*1×1×z*), the dimension is squeezed to *z*. At select section, *Z* performs three FC operations respectively to obtain three new feature vectors of dimension *c*, then performs softmax function to predict the weights of the different receptive fields, the weight vectors {*a*, *b*, *c* ∈ *R* (*1×1×c*)} are expressed as follows:

$$a = \frac{e^{AZ}}{e^{AZ}+e^{BZ}+e^{CZ}},\ b = \frac{e^{BZ}}{e^{AZ}+e^{BZ}+e^{CZ}},\ c = \frac{e^{CZ}}{e^{AZ}+e^{BZ}+e^{CZ}} \tag{5}$$

where *e* is the natural logarithm, *A*, *B* and *C* ∈ *R* (*c×z*) is two dimensions operational matrix, which values are constantly corrected as the input image changes. The feature map of final output *V* is expressed as follows:

$$V = V1 + V2 + V3 = aU1 + bU2 + cU3 \tag{6}$$

MRFCA is plug and play module, which integrates behind the first *C2* layer in the backbone of YOLOv5s, it can greatly improve the feature extraction ability of YOLOv5s for multi-scale targets.

3.2. Four Decoupled Detection Heads (4DDH)

The backbone of YOLOv5s uses the three-scale feature map {C_3, C_4, C_5} of 8 times, 16 times, and 32 times down-sampling to predict the targets of the small, medium, and large scales respectively. There are some small targets in our dataset. The experiment demonstrates that the three-scale feature maps are difficult to detect these small targets. To extract more abundant feature information and improve the detection precision for small targets, this paper proposes the four-scale feature map {C_2, C_3, C_4, C_5}. After the fusion of feature information in the neck section of YOLOv5s, new four-scale feature maps {P_2, P_3, P_4, P_5} are generated as the input of YOLOv5s head to complete the detection and recognition for aircraft targets in the SAR images.

At the YOLOv5s head section, the channel dimension of the input feature map is squeezed by convolution operation first. The squeezed feature map contains anchor coordinates, confidence, and aircraft category information, which is used to complete the task of detection and recognition. The two tasks share one feature map, but the two tasks focus on different feature information, the one feature map is usually not suitable for two tasks. The images from the GF-3 satellite have a large area of complex background and interference from various types of objects, this makes the task of identifying small targets more challenging. To further enhance the extraction ability of feature information and improve the precision, the detector adopts decoupled processing to complete the detection and recognition, 4DDH network is shown in Figure 5.

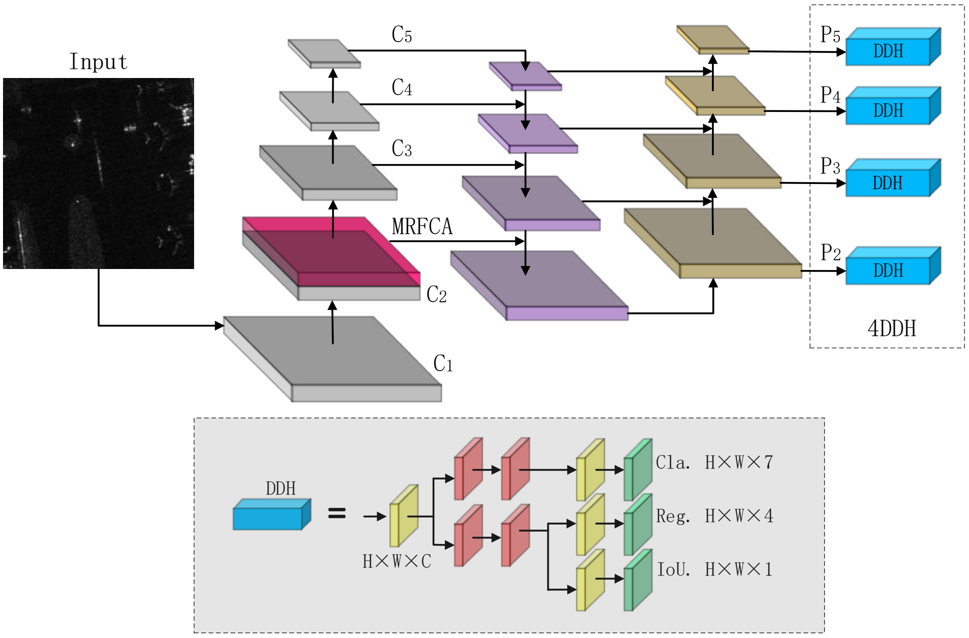

Figure 5. 4DDH network.

DDH squeezes the input channel dimension by *1×1* convolution first, which is used to decrease the number of parameters and increase operation speed, then the output is divided into two branches. The top branch is responsible for the recognition task, and first extracts the feature information through two *3×3* convolution operations, and another *1×1* convolution operation is used to adjust the channel dimension of the feature map to the number of categories of predicted aircraft, there are seven types of aircraft in our datasets. The bottom branch is responsible for the detection task, and first extracts the feature information through two *3×3* convolution operations, the feature map is divided into two more, the top one is used to predict the anchor coordinates by *1×1* convolution operation, another is used to predict the confidence by *1×1* convolution operation. 4DDH can improve the efficiency and precision of detection and recognition, especially for small targets.

3.3. Data Augmentation Method

Our dataset has only 2000 SAR images and 6556 aircraft instances. The diversity of aircraft targets in SAR images is limited. The proportion of small targets is small, the detection and recognition precision for small targets is low. This paper proposed the following three data augmentation methods to resolve the problem: Flip, Scaling, and Mosaic.

Flip and scaling are basic data augmentation methods, the input images are transformed geometrically to generate new images, and they are a simple and effective method to expand datasets. Mosaic obtains any four images from the dataset, they are randomly cropped, rotated, and scaled to synthesize an image. It can enrich the background of the SAR images, and increase the batch size and sample diversity in a single training process, the synthetic image is shown in Figure 6. Large samples are randomly scaled into small samples, increasing the number of small targets, Mosaic is very suitable for the datasets with the small number of small targets, it greatly improves the convergence speed of the model, and enhance the precision of detection and recognition.

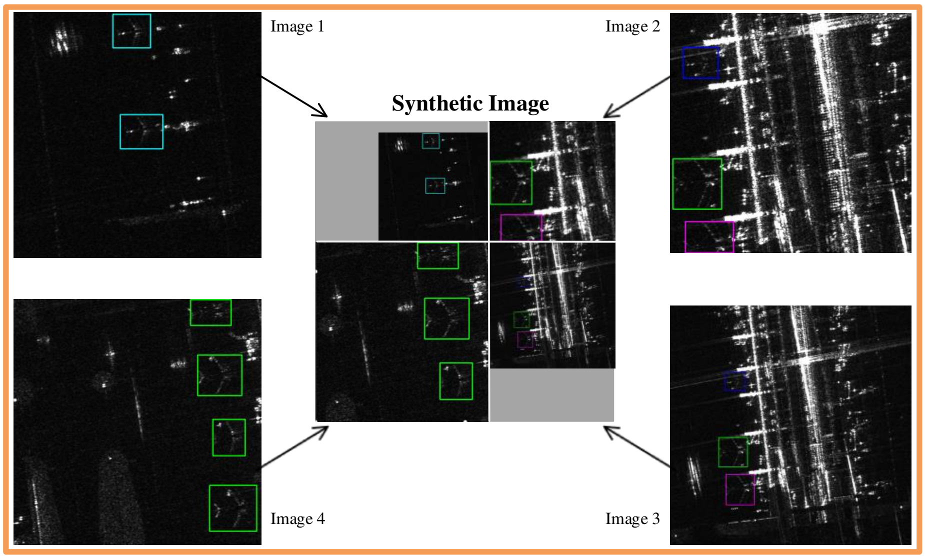

Figure 6. Mosaic synthetic image.

3.4. Optimization Method of Adaptive Anchor Box

In the process of network inference, an appropriately sized initial anchor box helps to accelerate network convergence. The network outputs some prediction boxes based on the initial anchor boxes. Compare the predicted box with the true box, calculate the difference between the two bounding boxes, then update and iterate the network parameters in reverse. The YOLOv5s uses the K-means algorithm to cluster anchor boxes in the dataset and obtain k clustering anchor boxes. The detailed steps are as follows:

doi:10.20944/preprints202308.0800.v1

- Randomly selected k initial clustering anchor boxes $C_k(w_k, h_k)$ from the dataset. (w is the width of the anchor box, h is the height of the anchor box)
- Calculate the distance from all bounding boxes $B_i(w_i, h_i)$ to $C_k(w_k, h_k)$ as follows: (i represents the total number of all bounding boxes)

$$Distance(B_i, C_k) = 1 - IoU(B_i, C_k) \tag{7}$$

- Classify bounding boxes $B_i(w_i, h_i)$ into the relevant clusters based on the principle of nearest distance. Finally, classify all bounding boxes into k clusters.
- Recalculate the new clustering center boxes $C_k(w_k, h_k)$, then repeat steps two, three, and four until the clustering anchor boxes remain unchanged.

When selecting k initial clustering anchor boxes, the distance between anchor boxes should be as large as possible. Based on this concept, the paper proposed the K-means++ algorithm to optimize step one in the K-means algorithm. First, randomly selected the first clustering anchor box, and calculated the distance from all bounding boxes to their nearest clustering anchor box to further calculate the probability of becoming a new clustering anchor box. Second, selected a new clustering anchor box by the roulette wheel method at non-clustering samples. The further away the sample is from the original clustering anchor boxes, the more likely it is to become the next clustering center. Repeat this process until k cluster anchor boxes are selected. Because this paper adopted the 4DDH method, the value of k is 12. The distribution of initial anchor boxes is shown in Figure 7. There are a total of 12 color regions, which represent different clusters. There are a total of 12 golden stars, which represent the size of the center anchor box of each cluster.

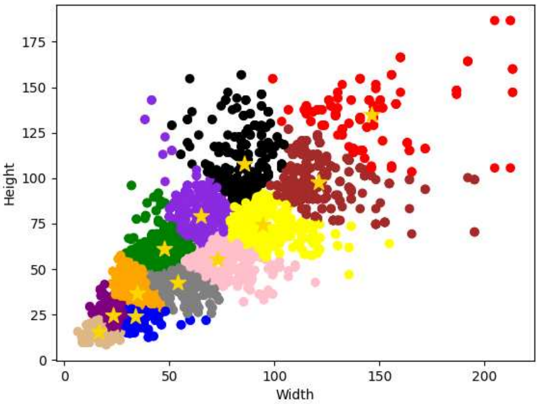

Figure 7. The distribution of 12 initial anchor boxes in the SAR image dataset.

Table 1 shows the detailed size of the initial anchor boxes between the original YOLOv5s and our improved YOLOv5s. P2, P3, P4, and P5 represent the detection layers.

Table 1. The initial anchor sizes between the original YOLOv5s and our improved YOLOv5s.

Detection Map Level	P2	P3	P4	P5
Original YOLOv5s	-	(10, 13) (16, 30) (33, 23)	(30, 61) (62, 45) (59, 119)	(116, 90) (156, 198) (373, 326)
Improved YOLOv5s	(16, 15) (23, 25) (34, 24)	(35, 37) (47, 61) (54, 43)	(65, 79) (73, 55) (85, 108)	(94, 74) (121, 98) (146, 135)

4. Experimental Result and Analysis

4.1. Experimental Environment

The SAR images of aircraft target in my datasets are from the GF-3 satellite, and the datasets used in this paper is from the 2021 GF challenge on Automated High-Resolution Earth Observation Image Interpretation [36]. The scene includes multi-temporal images of several common airports around the world. The datasets contain 2000 SAR images ranging in size from 600×600 pixels to 2048 ×2048 pixels and 6556 aircraft instances, the resolution is 1m, there are seven different types of aircraft, as shown in Figure 8. The datasets are divided into the training set, validation set, and test set according to the proportion of 8:1:1.

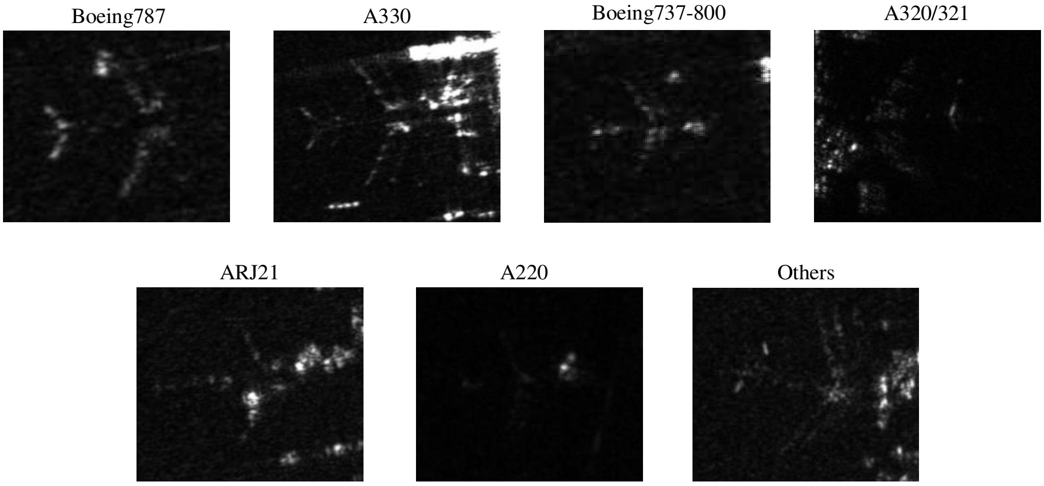

Figure 8. Examples of 7 types of aircraft.

The deep learning program runs on a 64-bit Windows 10 computer system, and Table 2 shows the software environment configuration. The number of training iterations is 200, the batch size is 32, and the initial learning rate is 0.01.

Table 2. Computer software environment configuration.

Parameter	Configuration
CPU	Inter(R) Core(TM) i7-7820X CPU @ 3.60 GHz
GPU	NVIDIA TITAN Xp
Accelerator	CUDA 10.2
Architecture	Pytorch 1.9
Language	Python 3.8

4.2. Experimental Evaluation

To accurately evaluate the performance of the network, this paper uses the average precision (*AP*) and the mean average precision (*mAP*) as the evaluation indicator. *AP* can reflect the detection performance of a single target category, and *mAP* can reflect the comprehensive detection performance of all categories. *AP* is calculated by precision and recall. The precision and recall are expressed as follows:

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

where *TP* represents a positive sample that predicts to be a positive example, *FP* represents a negative sample that predicts to be a positive example, *FN* represents a positive sample that predicts to be a negative example. At a specific Intersection over Union (*IoU*) threshold, the *PR* curve is drawn with recall (*R*) as the horizontal axis and precision (*P*) as the vertical axis. *AP* and *mAP* are expressed as follows:

$$AP = \int_0^1 P(R)dR \tag{10}$$

$$mAP = \frac{1}{\mathrm{C}}\sum AP(C) \tag{11}$$

where *C* represents the total number of the target category. This paper adopts mAP_{50} (*IoU* is 50%), mAP_{75} (*IoU* is 75%), and $mAP_{50\sim95}$ (*IoU* is from 50% to 95%, step is 5%) to evaluate the precision of detection and recognition.

4.3. Experiment Analysis

To verify the performance of the improved algorithm based on YOLOv5s, we select six kinds of object detection algorithms based on deep learning for comparative analysis, Faster R-CNN, Retina-Net, SSD, YOLOv3, and YOLOv5s. Table 3 shows the *mAP*% of related algorithms based on my datasets.

Table 3. Comparative analysis based on different network model.

Method	Backbone	mAP_{50}	mAP_{75}	$mAP_{50\sim95}$	S-Target $mAP_{50\sim95}$	L-Target $mAP_{50\sim95}$
Faster R-CNN	ResNet-50	85.9	70.3	55.7	46.1	59.8
Retina-Net	ResNet-50	81.2	66.2	52.2	43.2	57.2
SSD	VGG-16	80.5	65.3	51.8	41.8	56.9
YOLOv3	DarKnet-53	84.4	70.3	58.3	49.0	62.6
YOLOv5s	CSPDarknet53	85.1	71.6	61.0	50.5	65.9
Ours	Improved	91.4	79.3	70.3	63.6	72.4

where S-Target represents small targets whose pixel size is less than 60 × 60, our datasets contains 1956 small targets, L-Target represents large targets whose pixel size is more than 60×60, and our datasets contains 4600 large targets. From the results of our experiments, the *mAPs* of our algorithm are higher than that of the compared algorithm. The $mAP_{50\sim95}$ increases significantly, especially for small targets, up 13.1% compared with YOLOv5s. Experiments demonstrate that the improved YOLOv5s has stronger detection and recognition capabilities. Compared with other advanced algorithms, it shows certain advantages. It is more suitable for aircraft target detection and recognition tasks in remote sensing SAR images.

4.4. Ablation Experiments

To verify the effectiveness of MRFCA, we placed it after the feature layer of the backbone of YOLOv5s. Considering that the MRFCA module will affect the size of model parameters and inference speed to some extent, this paper only introduced one MRFCA module, which is placed after the C2, C3, C4, or C5 feature layer. Their respective recognition accuracy is shown in Table 4. Placing it after the C2 feature layer can achieve the best recognition accuracy, the $mAP_{50\sim95}$ has increased by 6.5%, and the $mAP_{50\sim95}$ for S-Target has increased by 9.1%. This paper proposed to place the MRFCA module after the C2 feature layer.

Table 4. Comparison of recognition accuracy of MRFCA modules.

C2	C3	C4	C5	mAP_{50}	mAP_{75}	$mAP_{50\sim950}$	S-Target $mAP_{50\sim95}$	L-Target $mAP_{50\sim95}$
×	×	×	×	85.1	71.6	61.0	50.5	65.9
√	×	×	×	89.3	76.2	67.5	59.6	69.9
×	√	×	×	89.0	75.8	66.9	59.0	69.6
×	×	√	×	88.6	75.3	66.2	58.2	69.2
×	×	×	√	88.1	74.8	65.3	57.0	68.8

To further prove that the proposed method improves the detection and recognition ability of YOLOv5s, we set up several groups of model comparison experiments. The first group is the prototype of YOLOv5s, the other experiments incorporated methods of MRFCA, 4DDH, data augmentation, and K-means++ respectively, all other structures are the same as YOLOv5s. The experimental results are shown in Table 5.

Table 5. Ablation experiment results.

Index	FS	FSM	4DDH	MRFCA	K-mean++	mAP_{50}	mAP_{75}	$mAP_{50\sim950}$	S-Target $mAP_{50\sim95}$	L-Target $mAP_{50\sim95}$
YOLOv5s	×	×	×	×	×	85.1	71.6	61.0	50.5	65.9
YOLOv5s+	√	×	×	×	×	86.5	72.5	61.6	51.2	66.4
YOLOv5s+	×	√	×	×	×	87.3	73.5	63.2	53.8	67.3
YOLOv5s+	×	√	√	×	×	88.1	74.5	65.3	57.1	68.2
YOLOv5s+	×	√	×	√	×	89.3	76.2	67.5	59.6	69.9
YOLOv5s+	×	√	√	√	×	90.6	78.6	69.7	63.1	72.0
YOLOv5s+	×	√	×	×	√	88.1	74.2	63.8	54.3	67.7
YOLOv5s+	×	√	√	√	√	91.4	79.3	70.3	63.6	72.4

where FS represents the fusion of the data augmentation method of flip and scaling, and FSM represents the fusion of the data augmentation method of flip, scaling, and mosaic. After integrating the FSM method, the $mAP_{50\sim95}$ has increased by 2.2%, and the $mAP_{50\sim95}$ for S-Target has increased by 3.3%, after integrating the FSM and 4DDH methods, the $mAP_{50\sim95}$ has increased by 4.3%, and the $mAP_{50\sim95}$ for S-Target has increased by 6.6%, after integrating the FSM and MRFCA methods, the $mAP_{50\sim95}$ has increased by 6.5%, and the $mAP_{50\sim95}$ for S-Target has increased by 9.1%, after integrating the FSM, 4DDH and MRFCA methods, the $mAP_{50\sim95}$ has increased by 8.7%, and the $mAP_{50\sim95}$ for S-Target has increased by 12.6%, after integrating the FSM, 4DDH, MRFCA, and K-mean++ methods, the $mAP_{50\sim95}$ has increased by 9.3%, and the $mAP_{50\sim95}$ for S-Target has increased by 13.1%. As can be seen from the ablation experiment results, the detection precision in each index has improved after optimizing, especially for small targets, and the improvement effect is more obvious. The YOLOv5s algorithm tends to miss some small aircraft targets on special occasions, the improved YOLOv5s algorithm can solve the problem very well. Figure 9 shows the comparison result between YOLOv5s and improved YOLOv5s.

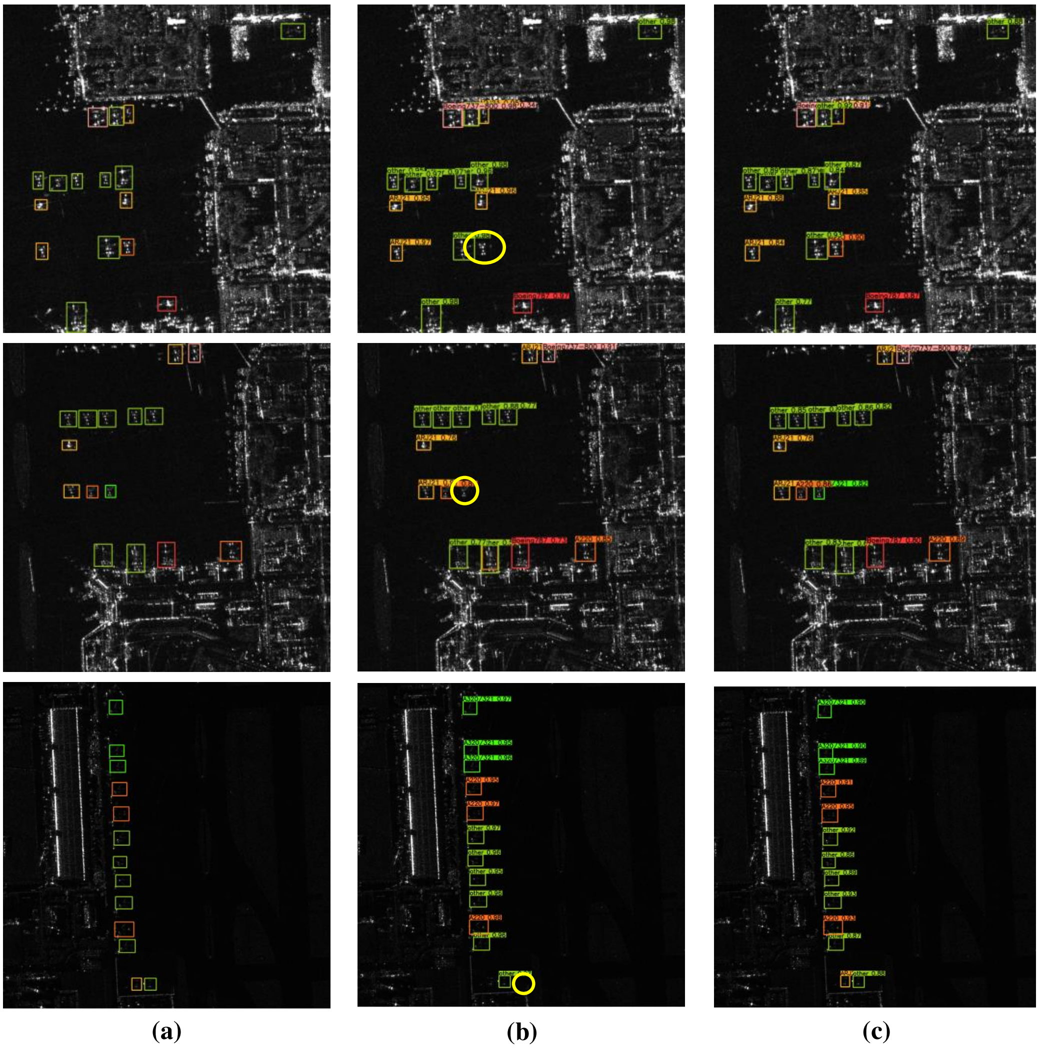

Figure 9. Comparison results of small targets recognition between the original YOLOv5s and the improved YOLOv5s. (**a**) Bounding boxes of the SAR images in the dataset. (**b**) Results for the original YOLOv5s. (**c**) Results for the improved YOLOv5s.

Compare Figure 9b and Figure 9c, the YOLOv5s algorithm misses three aircraft targets when the target is small and the background is complex, but the improved YOLOv5s algorithm has a better extraction ability of image features, there is not the case of missed detection, it has better detection and recognition performance for aircraft targets in SAR images.

5. Conclusions

It is of great significance for the accurate detection and recognition of aircraft targets in high-resolution remote sensing SAR images, an improved YOLOv5s is proposed in this paper. MRFCA network is proposed for improving the detection ability for multi-scale targets, the weights of different receptive fields are dynamically adjusted by adaptive learning, reasonably allocating the proportion of feature information of multi-scale. This paper proposed the 4DDH network, four detection branches can improve the detection ability to small targets, and decoupled detection head can effectively avoid the conflict of different feature information concerned in detection and recognition tasks, strengthening the capabilities of detection and recognition. This paper integrated the multi-method of data augmentation, which can enhance the diversity of datasets and generalization of the model, and improve the training speed of the network. This paper used the K-mean++ method, which can improve the network convergence speed and detection accuracy.

Experiments show that the improved YOLOv5s significantly improve the performance of SAR image interpretation, especially for small aircraft targets.

In the case of extremely small targets and complex backgrounds, missing detections and false detections still exist. In addition, the detection speed of improved YOLOv5s is not very ideal, further improvements are needed.

Author Contributions: Conceptualization, W.H. and X.W.; methodology, Y.L. and X.W.; software, X.W.; validation, D.H.; resources, Y.L. and P.X.; data curation, W.H.; writing—original draft preparation, X.W.; writing—review and editing, W.H. and Y.L.; supervision, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the key research and development projects of the Ministry of Science and Technology of China (Grant No. 2022YFC2203901), Science and Technology Development Plan Project of Jilin Province, China (Grand No. 20230201099GX), Department of Education Science and Technology Research Project of Jilin Province, China (Grand No. JJKH20220053KJ).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this research can be obtained from the 2021 Gaofen challenge on Automated High-Resolution Earth Observation Image Interpretation. Available online: http://gaofen-challenge.com (accessed on 1 October 2021).

Acknowledgments: The authors would like to thank the anonymous reviewers.

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

References

1. MOREIRA, A.; Prats-Iraola, P.; Younis, M.; Krieger, G.; Papathanassiou, KP. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine* **2013**, *1*, 6–43. https://doi.org/10.1109/MGRS.2013.2248301.
2. Cui, Y.; Zhou, G.; Yang, J.; Yamaguchi On the iterative censoring for target detection in SAR images. *IEEE Geoscience and Remote Sensing Letters* **2011**, *8*, 641-645. https://doi.org/10.1109/LGRS.2010.2098434.
3. Ao, W.; Xu, F.; Li, Y.; Wang, H. Detection and discrimination of ship targets in complex background from spaceborne ALOS-2 SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2018**, *11*, 536–550. https://doi.org/10.1109/JSTARS.2017.2787573.
4. Gao, G.; Ouyang, K.; Zhou, S.; Luo, Y.; Liang, S. Scheme of parameter estimation for generalized gamma distribution and its application to ship detection in SAR images. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 1812–1832. https://doi.org/10.1109/TGRS.2016.2634862.
5. Leng, X.; Ji, K.; Zhou, S.; Xing, X. Ship detection based on complex signal kurtosis in single-channel SAR imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2019**, *57*, 6447–6461. https://doi.org/10.1109/TGRS.2019.2906054.
6. Wang, X.; Chen, C. Ship detection for complex background SAR images based on a multiscale variance weighted image entropy method. *IEEE Geoscience and Remote Sensing Letters* **2017**, *14*, 184–187. https://doi.org/10.119/LGRS.2016.2633548.
7. Zhang, X.; Tan, Z.; Wang, Y. SAR target recognition based on multi-feature multiple representation classifier fusion. *Journal of Radars* **2017**, *6*, 492–502. https://doi.org/10.12000/JR17078.
8. Cheng, J.; Li, L.; Wang, X. SAR target recognition under the framework of sparse representation. *Journal of University of Electronic Science and Technology of China* **2014**, *43*, 524–529. https://doi.org/10.3969/j.issn.1001-0548.2014.04.009.
9. Wu, Q.; Sun, H.; Sun, X.; Zhang, D.; Fu, K. Aircraft recognition in high-resolution optical satellite remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 112–116. https://doi.org/10.3969/j.issn.1001-0548.2014.04.009.
10. Ge, L.; Xian, S.; Fu, K.; Wang, H. Interactive geospatial object extraction in high resolution remote sensing images using shape-based global minimization active contour model. *Pattern Recognit Lett.* **2013**, *34*, 1186–1195. https://doi.org/10.1016/j.patrec.2013.03.03.

11. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens*. **2015**, *36*, 618–644. https://doi.org/10.1080/01431161.2014.999881.
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.
13. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
14. 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. https://doi.org/10.1109/TPAMI.2016.2577031.
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969. https://doi.org/10.1109/TPAMI.2018.2844175.
16. Berg, A.C.; Fu, C.; Szegedy, C.; Anguelov, D.; Erhan, D.; Reed, S.; Liu, W. SSD: single shot multi-box detector. *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016; pp. 21–37. https://doi.org/ 10.1007/978-3-319-46448-0_2.
17. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, 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. https://doi.org/10.1109/TPAMI.2018.2858826.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.
19. Farhadi, A.; Redmon, J. Yolov3, An incremental improvement. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4, Optimal Speed and Accuracy of Object Detection. arXiv 2020, arXiv:2004.10934. https://doi.org/10.48550/ arXiv:2004.10934.
21. Zheng, Y.; Zhou, G.; Lu, B. A Multi-Scale Rebar Detection Network with an Embedded Attention Mechanism. *Appl. Sci*. 2023, 13, 8233. https://doi.org/10.3390/app13148233.
22. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717. https://doi.org/10.48550/arXiv.2103.02907.
23. Tan, S.; Yan, J.; Jiang, Z.; Huang, L. Approach for improving YOLOv5 network with application to remote sensing target detection. *Journal of Applied Remote Sensing* **2021**, *15*, 036512. https://doi.org/10.1117/1.JRS.15.036512.
24. Yu, X.; Wu, S.; Lu, X.; Gao, G. Adaptive weighted multiscale feature fusion for small drone object detection. *Journal of Applied Remote Sensing* **2022**, *16*, 034517. https://doi.org/10.1117/1.JRS.16.034517.
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
26. Kumar, T.; Mileo, A.; Brennan, R.; Bendechache, M. "Image Data Augmentation Approaches: A Comprehensive Survey and Future directions. *arXiv* **2023**, arXiv:2301.02830v4. https://doi.org/10.48550/arXiv. 2301.02830v4.
27. Li, Z.; Li, C.; Deng, L.; Fan, Y.; Xiao, X.; Ma, H.; Qin, J.; Zhu, L. Improved AlexNet with Inception-V4 for Plant Disease Diagnosis. *Comput. Intell. Neurosci*. **2022**, *2022*, 5862600. https://doi.org/10.1155/2022/5862600.
28. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. https://doi.org/10.48550/arXiv.1803.01534.
29. Lin, Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2117–2125. https://doi.org/10.48550/arXiv.1612.03144.

30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018. https://doi.org/10.48550/arXiv.1807.06521.
31. Yang, J.; Wang, W.; Li, X.; Hu, X.; Selective kernel networks. In proceeding of the 2019 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach, CA, USA, 16-20 June 2019; pp. 510-519. https://doi.org/10.1109/CVPR.2019.00060.
32. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934.
33. He, D.; Liu, J.; Xiong, H.; Lu, Z. Individual Identification of Dairy Cows Based on Improved YOLO v3. *Trans. Chin. Soc. Agric. Mach*. **2020**, *51*, 250–260
34. Goicovich, I.; Olivares, P.; Román, C.; Vázquez, A.; Poupon, C.; Mangin, J.F.; Guevara, P.; Hernández, C. Fiber Clustering Acceleration With a Modified Kmeans++ Algorithm Using Data Parallelism. *Front. Neuroinform*. **2021**, *15*, 727859. https://doi.org/https://doi.org/10.3389/fninf.2021.727859.
35. Dhivyaa, C.R.; Kandasamy, N.; Rajendran, S. Integration of dilated convolution with residual dense block network and multi-level feature detection network for cassava plant leaf disease identification *Concurrency and Computation Practice and Experience* **2022,** *34*, 1-19.
36. 2021 Gaofen challenge on Automated High-Resolution Earth Observation Image Interpretation. Available online: http://gaofen-challenge.com (accessed on 1 October 2021).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.