electronics

Article

Aircraft Detection and Fine-Grained Recognition Based on High-Resolution Remote Sensing Images

Qinghe Guan, Ying Liu *, Lei Chen, Shuang Zhao and Guandian Li

College of Electronical and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
* Correspondence: liuying02@cust.edu.cn; Tel.: +86-135-7868-2539

Abstract: In order to realize the detection and recognition of specific types of an aircraft in remote sensing images, this paper proposes an algorithm called Fine-grained S^2ANet (FS^2ANet) based on the improved Single-shot Alignment Network (S^2ANet) for remote sensing aircraft object detection and fine-grained recognition. Firstly, to address the imbalanced number of instances of various aircrafts in the dataset, we perform data augmentation on some remote sensing images using flip and color space transformation methods. Secondly, this paper selects ResNet101 as the backbone, combines space-to-depth (SPD) to improve the FPN structure, constructs the FPN-SPD module, and builds the aircraft fine feature focusing module (AF^3M) in the detection head of the network, which reduces the loss of fine-grained information in the process of feature extraction, enhances the extraction capability of the network for fine aircraft features, and improves the detection accuracy of remote sensing micro aircraft objects. Finally, we use the SkewIoU based on Kalman filtering (KFIoU) as the algorithm's regression loss function, improving the algorithm's convergence speed and the object boxes' regression accuracy. The experimental results of the detection and fine-grained recognition of 11 types of remote sensing aircraft objects such as Boeing 737, A321, and C919 using the FS^2ANet algorithm show that the $mAP_{0.5}$ of FS^2ANet is 46.82%, which is 3.87% higher than S^2ANet, and it can apply to the field of remote sensing aircraft object detection and fine-grained recognition.

Keywords: aircraft object detection; fine-grained recognition; attention module; data augmentation; SPD; KFIoU

Citation: Guan, Q.; Liu, Y.; Chen, L.; Zhao, S.; Li, G. Aircraft Detection and Fine-Grained Recognition Based on High-Resolution Remote Sensing Images. *Electronics* **2023**, *12*, 3146. https://doi.org/10.3390/electronics12143146

Academic Editor: Byung Cheol Song

Received: 7 July 2023
Accepted: 17 July 2023
Published: 20 July 2023

1. Introduction

Remote sensing image object detection is one of the research hotspots in the field of optical remote sensing image processing, among which the research of remote sensing aircraft object detection has tremendous significance in aerospace monitoring, military detection, aircraft accident prevention, and UAV detection. Also, with the continuous maturity of remote sensing satellite technology and the continuous upgrading of optical equipment, more and more high-resolution remote sensing aircraft images can be obtained, making it possible to detect remote sensing aircraft objects in a wide range without interruption and with a high accuracy. With the progress of aircraft object detection technology for industrial, civil, and military remote sensing images and the development of deep learning object detection algorithms, the detection accuracy and speed of remote sensing aircraft objects are continuously improved; however, there are problems such as small objects, arbitrary object angles, similar shapes of various types of an aircraft, and slight differences between classes for specific models of remote sensing aircraft objects. Therefore, how to accurately, efficiently, and quickly detect remote sensing aircraft objects and fine-grained recognition is a challenging research direction.

In the last decade, many scholars have used a Convolutional Neural Network (CNN) [1] to improve a large number of object detection networks in the field of remote sensing object detection due to its excellent feature extraction ability [2–4]. Compared with traditional

remote sensing aircraft detection methods, deep learning methods are faster, more accurate, and have a better generalization ability, and gradually become the preferred choice for remote sensing aircraft object detection. Kai-Ming He proposed the method of Anchor boxes and multi-scale feature extraction and designed Faster R-CNN [5]. Liu et al. designed the single-stage object detection algorithm SSD [6] with a better detection speed and accuracy. In the same year, Joseph Redmon et al. proposed another single-stage detection algorithm, YOLO [7], and many remote sensing aircraft detection methods have emerged with their improvements based on these algorithms. For aircraft rotation object detection in remote sensing images, in 2017, Siyu Wang et al. improved LeNet-5, adding a saliency prediction step to the network to improve the coarse localization capability of the network [8]. Pang et al. reconfigured the network module, improved small objects' detection accuracy, and proposed R^2CNN [9]. Jian Ding et al. designed RRoI Align based on the horizontal region of interest Align and proposed ROI Transformer [10] to rotate the sampled points to the corresponding coordinates in the feature map to complete the detection for rotated objects. Xue Yang et al. proposed the R^3Det [11] network using a stepwise regression method from coarse-grained to fine-grained to detect objects quickly and accurately, using horizontal anchor points to obtain detection objects and then refining and correcting to obtain an oriented bounding box. Regarding ReDet [12], an invariant rotational network was added to the detector to extract the rotational invariant features of the remote sensing object and RiROI Align was proposed to extract the invariant features. Jiaming Han et al. proposed the S^2ANet [13] algorithm, with the overall model using RetinaNet [14] as the backbone network. The model uses the FPN for feature fusion, and at the detection head, the model uses the FAM and ODM modules for feature alignment, regression, and classification, significantly improving the detection accuracy. The SASM [15] algorithm proposed by Liping Hou et al. divides the overall oriented bounding box object detection into a shape-adaptive selection (SAS) module and a shape-adaptive measurement (SAM) module, which improves the accuracy of rotating object detection.

The above studies for remote sensing aircraft object detection have a good detection accuracy and speed, but there are some problems:

1. Some studies only focus on detecting aircraft objects and do not fine-identify specific aircraft types, which has a poor applicability in some scenarios.
2. The differences between various aircraft types in remote sensing images are very subtle, making it challenging to classify specific aircraft types accurately.
3. In high-resolution remote sensing images, the high complexity of the object background causes many difficulties in detecting and classifying an aircraft.
4. Because of the imbalanced instances of various aircrafts, the number of instances of various aircraft types in the existing dataset shows a long-tailed distribution.

In order to solve the above problems, this paper proposes a remote sensing aircraft object detection and fine-grained recognition algorithm called FS^2ANet, and the main innovations of this algorithm are as follows:

1. Introduces an attention mechanism in the detection head, combines multi-scale feature fusion, and constructs the AF^3M, which makes feature aggregation occur in the channel and space between the fine-grained features of different aircraft types and increases the semantic feature distance between different aircraft types.
2. Optimizes the regression loss function by replacing the original Smooth L1 [16] loss with the KFIoU [17] loss, which is more effective than the oriented bounding box IoU of Smooth L1.
3. Introduces the SPD [18] module in the Neck to improve the traditional FPN [19] and construct the FPN-SPD structure with three input layers and five output layers, and replaces the convolutional layer with SPD in the downsampling stage, followed by BN and ReLU to reduce the fine-grained information loss and improve the detection accuracy of small objects.
4. In this paper, we perform data augmentation on some remote sensing images to alleviate the accuracy loss caused by the imbalanced number of instances and replace

the RetinaNet backbone with ResNet101 [20] to extract aircraft features further and improve the detection accuracy.

2. Related Work

2.1. S^{2}ANet

S^{2}ANet is mainly applied in the field of remote sensing image object detection. The network consists of the backbone, FPN, Feature Alignment Module (FAM), and Orientation Detection Module (ODM), which solves the problems of an extensive distribution range and the arbitrary orientation of remote sensing objects.

S^{2}ANet selects RetinaNet as the backbone network and replaces the original horizontal box regression with oriented bounding box regression while the rest settings remain unchanged. After RetinaNet extracts features of the remote sensing image, the backbone sends the future maps to the Neck for feature fusion. In FPN, the strong semantic features in the top layer benefit object classification, and the high-resolution features in the bottom layer benefit object localization. The model gradually expands the receptive field and fuses the feature map, which enables it to better obtain the overall information and improves the feature extraction ability of the network. Then, the feature map enters the Head to generate high-quality rotated anchors with FAM and align the anchors and features through alignment convolution, after which the ODM generates orientation-sensitive features beneficial for subsequent classification and regression. Figure 1 shows the overall structure of S^{2}ANet.

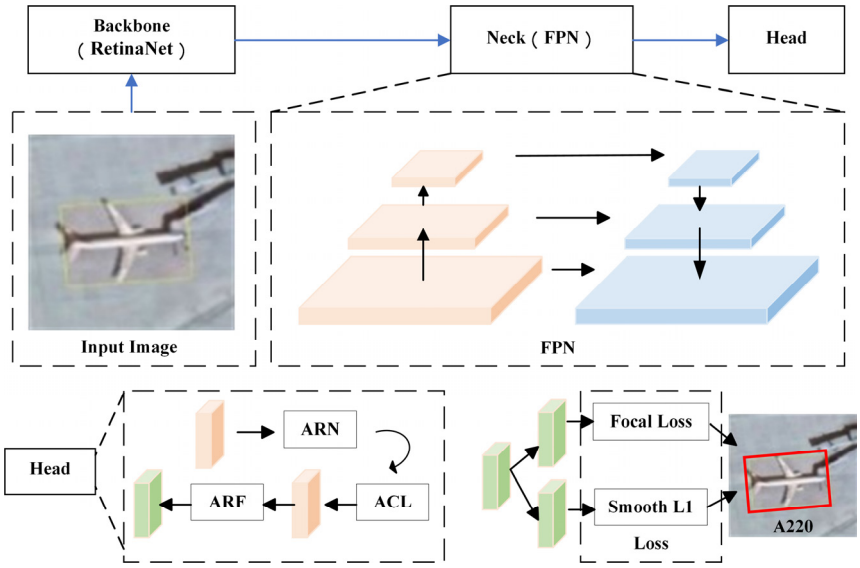

Figure 1. Structure diagram of S^{2}ANet.

2.2. Attention Mechanisms

The attention mechanism mainly solves the region of interest problem of the feature map, which makes the region of interest well-focused to a certain extent. In computer vision, the common attention mechanisms generally divide into the channel and spatial attention mechanisms. Since SENet [21] was proposed in 2017, the attention mechanism has rapidly become a hot spot in the field of computer vision to improve the accuracy of models. ECANet [22] removes the fully connected layer based on SENet, eliminates the information loss caused by a dimensionality reduction in the channel, and enhances the cross-channel interaction. The CBAM [23] attention mechanism consists of the channel attention module (CAM) and spatial attention module (SAM), focusing the features from two dimensions to improve the model performance. Firstly, the feature map passes through the CAM, and after max pooling and average pooling, the channel weights are obtained

through MLP and summation. Then, the feature map passes through the SAM, and after max pooling and average pooling, the spatial weights are obtained through convolution.

2.3. Regression Loss

The regression loss of S^{2}ANet is $smooth_{L1}$ Before $smooth_{L1}$, there are two classic losses, L_1 and L_2. When the difference between the predicted box and the genuine box is small, the loss can be close to zero, and the gradient gets smaller and smaller. On the contrary, the loss value cannot grow too fast, which requires the gradient not to grow too fast. Therefore, by comparing L_1 and L_2 loss, a $smooth_{L1}$ loss combining them is proposed. The formulas for the three are given below.

$$L_1(x) = |x| \tag{1}$$

$$L_2(x) = x^2 \tag{2}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{3}$$

According to the formula, the $smooth_{L1}$ loss is a small value, and the gradient also slowly becomes smaller (gradually to zero); it is an enormous value, and the gradient also reaches the maximum value of 1, solving the problem that the L_1 loss function fluctuates around the stable value. L_2 loss is an enormous value appearing in a gradient explosion.

2.4. Space-to-Depth

In the original convolutional network, each convolutional downsampling brings a loss of fine-grained information and inefficient feature extraction, which makes some small objects poorly classified and detected in classification and detection tasks. In order to obtain smaller feature maps without a loss of accuracy, SPD splits the feature map with the size of S × S × C into feature maps with the size of S × S × 1 × 1 × C along the channel direction and draws on the focus operation to assign the 1 × 1 × C feature maps according to the downsampling factor Scale to finally obtain the feature maps with the size of (S/Scale) × (S/Scale) × (Scale × 2C).

3. Algorithm Improvement

3.1. AF^{3}M

In this paper, we combine multi-scale feature fusion and construct the AF^{3}M module to highlight the delicate features of various types of an aircraft and improve the feature extraction capability of the network in complex remote sensing contexts.

First, the S^{2}ANet network directly feeds the feature maps into the anchor refinement network (ARN) after passing FPN in the Neck. ARN can subdivide into two branches, one for classification and one for regression, which is a lightweight network, and the classification branch separates different anchors by category. In contrast, the regression branch refines the anchors for regression to generate high-quality regression boxes. Figure 2 shows the specific framework of ARN.

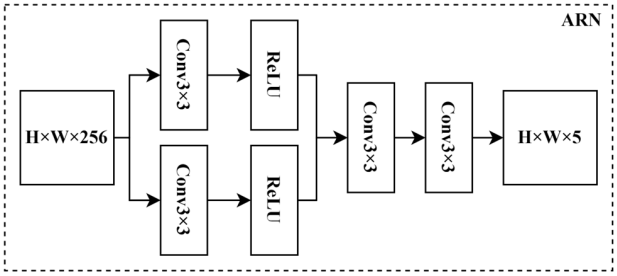

Figure 2. Frame diagram of ARN.

To further enhance the aircraft position and fine-grained semantic information, improve the accuracy of remote sensing aircraft object detection and recognition, and avoid interference from high-level coarse-grained position information to low-level fine-grained position information, this paper designs the AF^{3}M before the ARN module to improve the accuracy of the spatial position information of the feature map. Firstly, the feature maps are obtained using convolution with 3 × 3, 5 × 5, and 7 × 7 kernels, respectively. We use two 3 × 3 kernels to replace the 5 × 5 kernel and three 3 × 3 kernels to replace the 7 × 7 kernel to reduce computational complexity and increase the nonlinearity degree. Different scale feature maps are added and passed through maximum and average pooling. Then, feature maps pass through the MLP and sigmoid to obtain channel weights. The weights are multiplied with input feature maps and sent to the spatial attention module for maximum and average pooling. Unlike channel attention, the spatial attention module concatenates the future maps after the pooling layer. After passing through a 7 × 7 convolution and multiplying it with the output feature maps of the channel attention module, the final feature map is obtained and connected to the input feature map via residual connections. Figure 3 shows the position of the AF^{3}M module in the Head, and Figure 4 shows its specific structure.

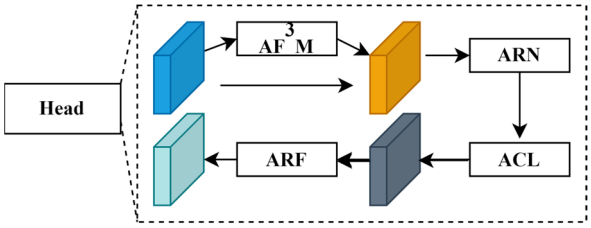

Figure 3. Location map of AF^{3}M.

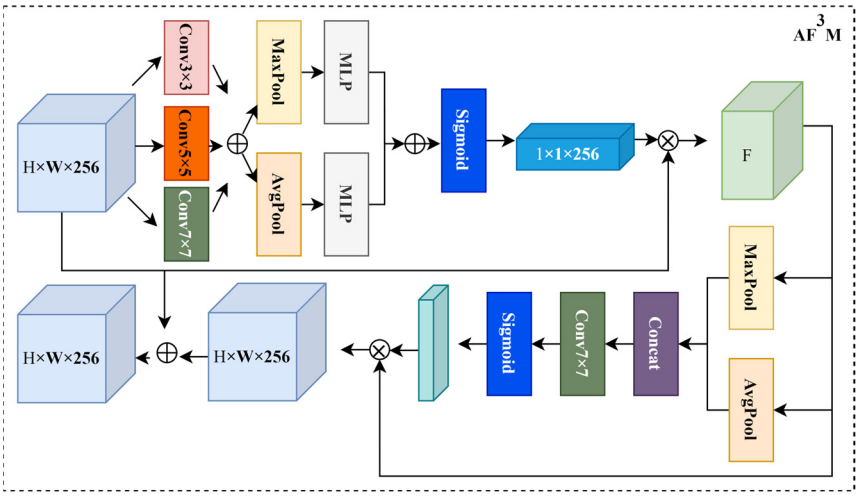

Figure 4. Frame diagram of AF^{3}M.

Attention mechanisms are effective in most networks, allowing feature maps to highlight feature information in channel, spatial, or hybrid dimensions, and the effectiveness of different attention mechanisms varies in specific networks. In the experiments of remote sensing aircraft object detection and fine-grained recognition, this paper compares different attention mechanisms and the AF^{3}M to study the impact of different improvements on experimental accuracy, and Table 1 shows the results.

Table 1. Comparative experimental results of AF^{3}M.

Algorithm	$\mathbf{mAP_{0.5}}$%	$\mathbf{AP_{max}}$%
S^{2}ANet	42.92	86.3
S^{2}ANet+SE	43.12	81.3
S^{2}ANet+CBAM	43.65	83
S^{2}ANet+ECA	43.55	83.6
S^{2}ANet+AF^{3}M (Ours)	**44.21**	81.3

As can be seen from Table 1, SE, CBAM, ECA, and AF^{3}M can all improve the model accuracy to different degrees, by 0.2%, 0.73%, 0.63%, and 1.29%, respectively, and AF^{3}M works best under the experimental conditions in this paper. Comparing AP_{max}, it can be found that after adding the attention mechanism, AP_{max} is reduced, which indicates that the accuracy of the most easily detected and identified aircraft among 11 types of an aircraft decreases, and the overall accuracy improves, which improves the accuracy of other small objects that are difficult to detect and identify to some extent.

3.2. Optimize Regression Loss

The loss function is crucial in remote sensing aircraft object detection and fine-grained recognition. A better regression loss can make the predicted oriented bounding box closer to the actual box, and the regression process is more consistent with the fine-grained characteristics of aircraft objects.

KFIoU Loss is based on the Gaussian function and Kalman filter, which can build a loss similar to the actual rotation IoU. First, any oriented bounding box with any dimension has a corresponding transformation relationship with the Gaussian function, which is defined as (x, y, h, w, θ) in two-dimensional space, and the oriented bounding box is transformed into the Gaussian distribution $g(\mu, \Sigma)$, where

$$\Sigma = R\Lambda R^T, \mu = (x, y, (z))^T \tag{4}$$

R is the rotation matrix, Λ is the diagonal matrix of the eigenvalue, and their formula is

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \Lambda = \begin{pmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{pmatrix} \tag{5}$$

Once the conversion formula between the oriented bounding box and the Gaussian function is known, first, convert the prediction box and the actual box into the Gaussian function, and then use the center loss L_c to make the centers of the two Gaussian functions gradually close until they coincide. The formula of L_c is as follows:

$$L_c = \sum_{i \in (x,y)} l_n(t_i, t_i') \tag{6}$$

Multiplying the two Gaussian distributions obtained and using Kalman filtering can obtain the Gaussian distribution function of the overlapping region:

$$\alpha g_{kf}(\mu, \Sigma) = g_1(\mu_1, \Sigma_1) g_2(\mu_2, \Sigma_2) \tag{7}$$

$$\mu = \mu_1 + K(\mu_2 - \mu_1), \Sigma = \Sigma_1 - K\Sigma_1 \tag{8}$$

$$K = \Sigma_1(\Sigma_1 + \Sigma_2)^{-1} \tag{9}$$

Finally, transform the Gaussian function of the overlapping area into an oriented bounding box, and calculate the $KFIoU$ according to the area formula of the oriented bounding box:

$$\nu_B(\Sigma) = 2^n \sqrt{\Pi eig(\Sigma)} = 2^n \cdot |\Sigma|^{\frac{1}{2}} \tag{10}$$

$$KFIoU = \frac{\nu_{B3}(\Sigma)}{\nu_{B1}(\Sigma) + \nu_{B2}(\Sigma) - \nu_{B3}(\Sigma)} \tag{11}$$

$\nu_B(\Sigma)$ is the area of the oriented bounding box and shows that when the center of the oriented bounding box coincides, the loss is only related to the covariance Σ_1 and Σ_2 of the two Gaussian distributions. This way, the calculated loss value will have a maximum threshold, and the loss will be within a specific range. Once the upper bound is found, it can be extended to [0, 1] for easy calculation.

$KFIoU$ does not directly calculate the IoU of two oriented bounding boxes. Instead, $KFIoU$ converts the two oriented bounding boxes into two Gaussian functions and calculates the overlapping area of two Gaussian functions, and then converts the overlapping area into an oriented bounding box to use the relationship between the three oriented bounding boxes to calculate the IoU. This method solves the problems of boundary discontinuity and the cross area close to the square. It can be extended to high-dimensional fields to calculate the IoU, effectively improving the algorithm's convergence rate and reducing training fluctuations. Figure 5 shows the loss changes in training with and without $KFIoU$, where Iters represents the sampling points within each epoch. Although the calculation formulas of the two are different and the regression loss values are different, it can be seen from Figure 5 that the rate of convergence of the algorithm has significantly improved after using the $KFIoU$; after the 10th Iter, the algorithm began to converge, and the loss value did not fluctuate significantly. Instead of not using $KFIoU$, the algorithm needs to converge after the 35th Iter and the loss value has significant fluctuations.

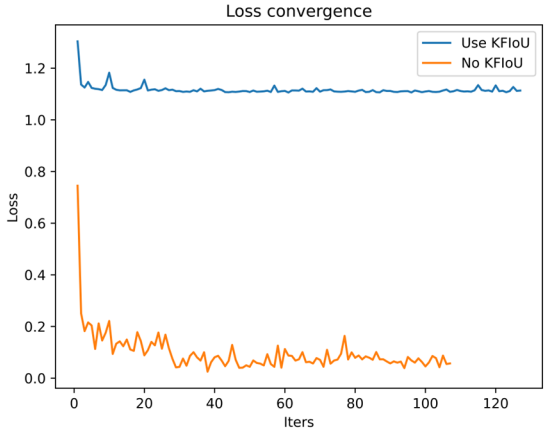

Figure 5. Diagram of loss convergence.

3.3. FPN-SPD

In S^2ANet, feature maps emerge from the backbone and are sent to the Neck (FPN) for feature fusion. The FPN constructs a top-down upsampling network, which is fused with the feature map from the downsampling process during feature extraction to obtain better multi-scale semantic features. With the increase in the depth of the Convolutional Neural Network, the deep feature map with more image semantic information is suitable for detecting large objects, and the shallow, high-resolution feature map with more object positioning information is suitable for detecting small objects. With the gradual downsampling, the receptive field of the feature map is gradually expanded, which can better obtain the global information and then merge with the feature map after upsampling, significantly improving the feature extraction ability of the network.

After improvement, this paper takes the feature maps of the backbone's Stage 2, Stage 3, and Stage 4 as the input of FPN-SPD, and the downsampling rate is 8, 16, and 32 times. Regarding these feature maps through 1×1 convolution, lower resolution feature maps are upsampled and fused with the previous feature maps and obtain three-layer-output feature maps using 3×3 convolution. Two additional output layers are obtained by downsampling the input feature maps (32 times) through the S module, with downsampling rates of 64 and 128 times. Figure 6 shows the S module and the overall FPN-SPD structure. In the S module, the input feature map with a downsampling of 32 times is first segmented by the SPD with Scale = 2 and then, after 1×1 convolution, adjusts the number of channels and obtains the output feature map through the BN layer and ReLU activation function.

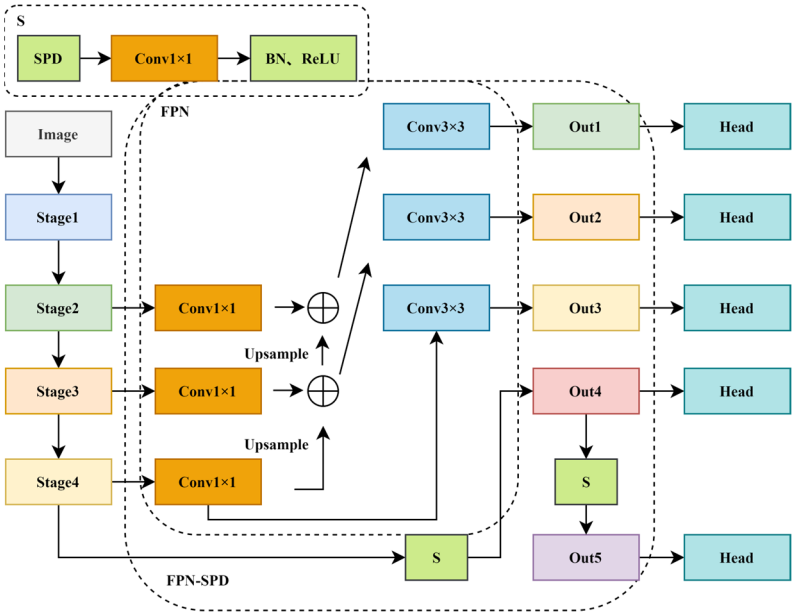

Figure 6. Frame diagram of FPN-SPD.

3.4. Selection of Activation Function

The activation function is usually used to enhance the nonlinear ability of the network. The effects of different activation functions are different for different network structures, use occasions, and datasets. Some activation functions will better affect some problems, such as a gradient disappearance, gradient explosion, and slow convergence rate. Therefore, it is essential to select a suitable activation function. The adaptive activation function can automatically adjust the parameters [24] during the training process so that the network can converge faster and improve stability. This section compares the convergence and accuracy of the algorithms that continue to use the ReLU activation function in the S^2ANet algorithm, use the adaptive ReLU activation function–Average Biased ReLU, and use the adaptive Tanh activation function–Penalized hyperbolic tanh.

The calculation formulas of ReLU [25], Average Biased ReLU [26], and Penalized hyperbolic tanh [27] are

$$\mathrm{ReLU}(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{12}$$

$$AB\mathrm{ReLU}(x) = \begin{cases} x - \beta & x - \beta \geq 0 \\ 0 & Otherwise \end{cases} \tag{13}$$

$$PH\tanh(x) = \begin{cases} \tanh(x) & x \geq 0 \\ \alpha\tanh(x) & x < 0 \end{cases} \tag{14}$$

After using the three activation functions, respectively, Figure 7 shows the training results of the FS^{2}ANet algorithm. ReLU, Average Biased ReLU, and the Penalized hyperbolic

tanh activation function have the same impact on the convergence of model training, and there is no apparent difference in the accuracy of the models. Therefore, this paper follows the ReLU activation function used by the S^2ANet algorithm.

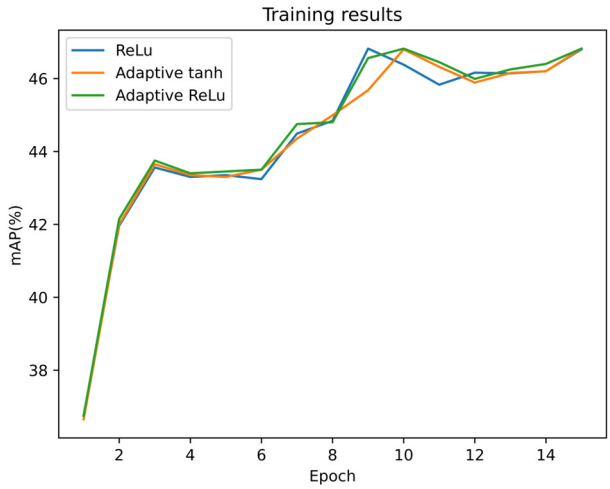

Figure 7. Training results of different activation functions.

4. Experiment and Result Analysis

4.1. Experimental Environment

This article's remote sensing aircraft object detection and fine-grained recognition experiment was conducted on a Windows 10 system (64-bit), using an Intel i7-11700 8-core CPU and NVIDIA GeForce RTX 3080Ti to build a training platform. The deep learning framework is mmrotate, and the CUDA version is 11.1. The experimental optimizer uses SGD, lr = 0.0025, momentum = 0.9, and weight_Decay = 0.0001, as well as trains 100 epochs as a whole, and if convergence occurs earlier, ends training earlier.

4.2. Experimental Indicators

In the same experimental environment, the evaluation indicators used in the comparison and ablation experiments of various algorithms include mAP, Precision, and Recall. Firstly, calculate the P and R:

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

Average Precision (AP) is the area enclosed by the PR performance curve for a specific category, which can be calculated by the P and R corresponding to the threshold of that category. Based on various Average Precisions and setting the IoU threshold to 0.5, $mAP_{0.5}$ can be calculated. The calculation formula is as follows:

$$AP = \int_0^1 P(R)dR \tag{17}$$

$$mAP_{0.5} = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{18}$$

where N is the number of categories, and AP_i is the average accuracy of each category.

4.3. Image Enhancement

This article's remote sensing aircraft dataset is sourced from the high-resolution fine-grained recognition challenge FAIR1M [28] remote sensing dataset. There are 11 types of an

aircraft, namely Boeing 737, Boeing 777, Boeing 747, Boeing 787, A321, A220, A330, A350, C919, ARJ21, and another aircraft. The unbalanced number of instances can easily affect the accuracy of aircraft recognition. The number of aircraft instances in FAIR1M shows a long-tailed distribution, as shown in Figure 8a; the number of ARJ21 and C919 instances is tiny. This issue needs to be addressed, along with minimizing the potential impact of introducing additional aircraft instances. This article uses methods such as flipping, rotation, color space transformation, contrast transformation, etc., to enhance the data of the images where ARJ21 and C919 are located and fill excess pixels with black pixels to generate new images. This alleviates the problem of significant differences in the number of instances and expands a certain number of datasets. The training involved 14,022 remote sensing aircraft images, a training set including 9922 images, a validation set including 2700 images, and a test set including 1400 images.

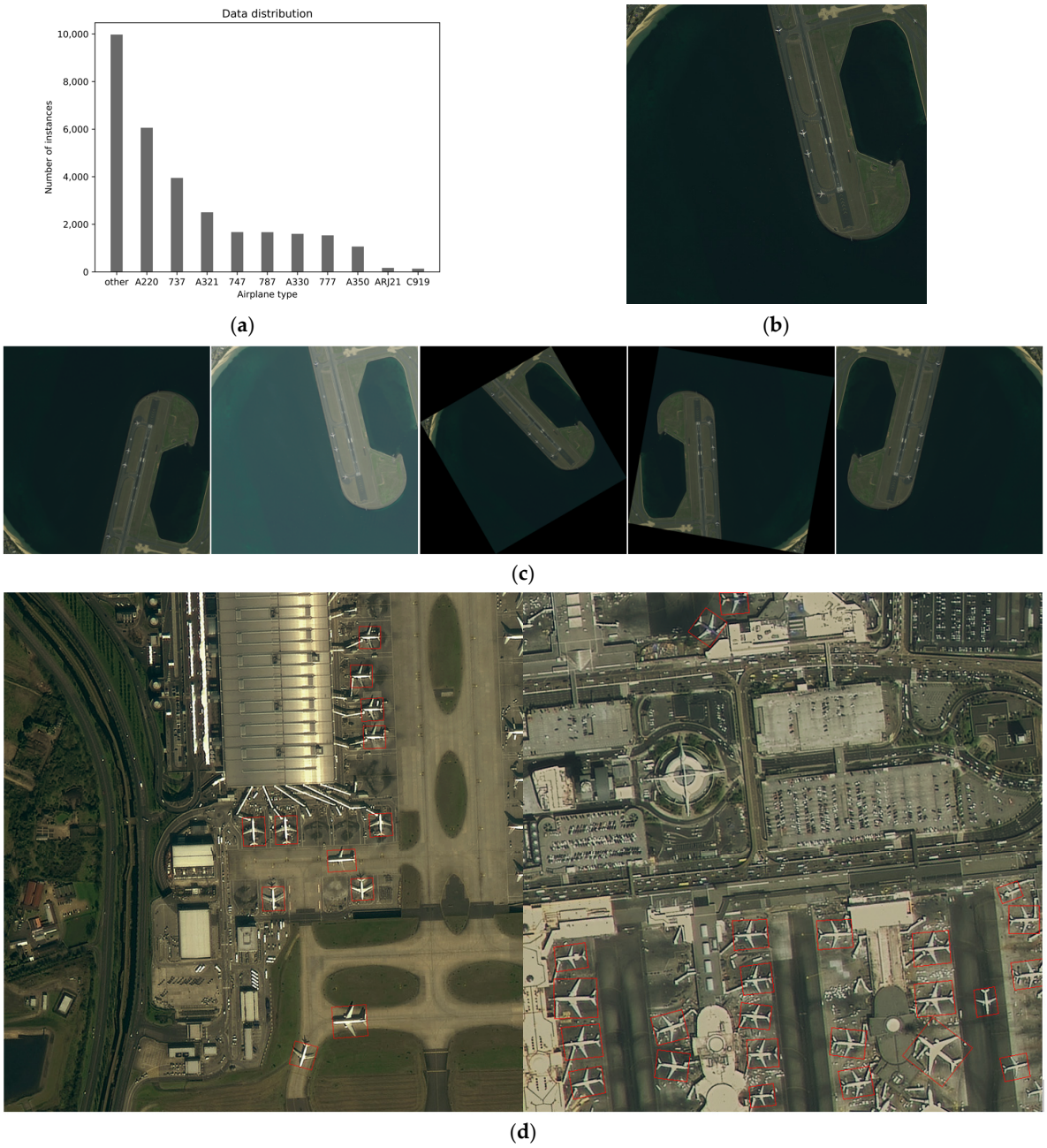

Figure 8. FAIR1M dataset and data augmentations: (**a**) Instance Distribution Map of FAIR1M, (**b**) example of FAIR1M, (**c**) the process of data augmentation, and (**d**) example of dataset annotation.

Considering that the aircraft object size in the dataset is tiny, the image background is complex and variable, and the image size is large, this study divided the remote sensing images in the dataset into 1024 × 1024 before training to reduce the impact of image zooming on the loss of various aircraft detail features. Figure 8b shows the original image of the FAIR1M dataset, Figure 8c depicts the data augmentation process, and Figure 8d presents the annotation of the dataset.

4.4. Comparative Experiment

4.4.1. Performance on the FAIR1M Dataset

Different models have a different detection and recognition accuracy of various aircraft types in the FAIR1M dataset. In order to verify the effectiveness of the improved algorithm in this paper, we compare the FS^{2}ANet model with excellent models in remote sensing rotated object detection, such as Roi-Transformer, SASM, ReDet, R^{3}Det, Faster-Rcnn, Rotated RetinaNet, GWD, and S^{2}ANet, in the experiment. The comparative experimental results are shown in Table 2, which mainly compares the mAP$_{0.5}$, AP, and Recall of each algorithm.

Table 2. Results of the comparison experiment.

Algorithm	mAP$_{0.5}$/%	AP-Max	AP-Min	Recall-Max	Recall-Min
Roi-Transformer	42.27	0.756	0.01	0.944	0.143
SASM	33.33	0.638	0.004	0.985	0.679
ReDet	41.99	0.706	0.053	0.949	0.179
R^{3}Det	41.82	0.839	0.003	0.989	0.643
Faster-Rcnn	41.76	0.836	0.028	0.967	0.132
Rotated RetinaNet	38.31	0.776	0.002	0.988	0.679
GWD	42.28	0.816	0.002	0.983	0.679
S^{2}ANet	42.95	0.863	0.003	0.978	0.679
FS^{2}ANet (Ours)	**46.82**	0.849	0.131	0.971	0.357

Figure 9 shows the growth of mAP with the epoch in training. The horizontal axis represents the training epoch, and the vertical axis represents the mAP. The shorter the horizontal axis, the higher the vertical axis, indicating a faster convergence speed and better model accuracy. Observing Table 2, we can know the following:

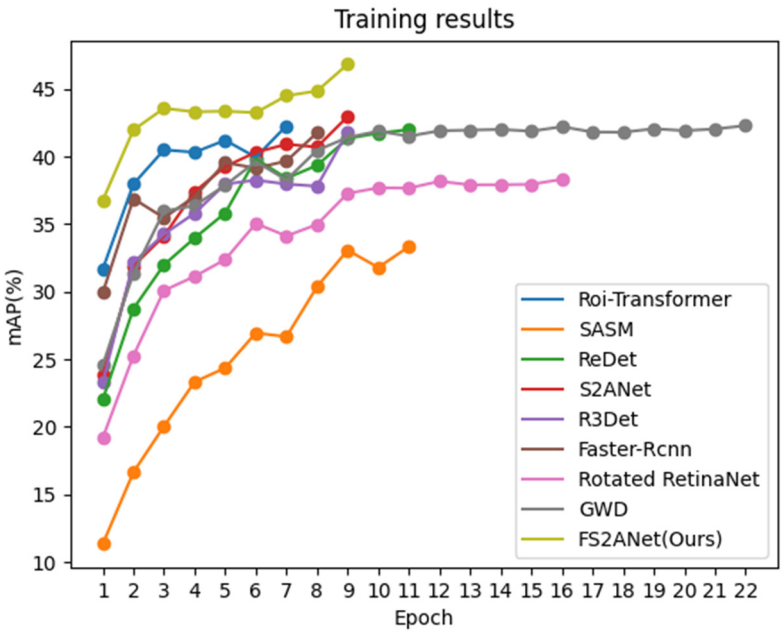

Figure 9. Training results of the comparison experiment.

The $mAP_{0.5}$ of the FS^{2}ANet model is 46.82%, which is 4.55%, 13.49%, 4.83%, 5%, 5.06%, 8.51%, 4.54%, and 3.87% higher than Roi Transformer, SASM, ReDet, R^{3}Det, Faster-Rcnn, Rotated RetinaNet, GWD, and S^{2}ANet, respectively. The FS^{2}ANet model has a significant improvement in accuracy.

Regarding AP, the AP_{max} can represent the model's preference for aircraft categories that are easy to detect and classify, while the AP_{min} can represent the model's handling of aircraft categories that are difficult to detect and classify. The AP_{max} of FS^{2}ANet is slightly lower than S^{2}ANet and higher than the AP_{max} of other models; this indicates that the model in this paper has a good accuracy for the easy detection and classification of aircraft categories, and the model has the highest AP_{min}; this indicates that the improved module in this paper significantly improved in preserving fine-grained information about an aircraft and extracting fine features of various aircrafts. Compared to the AP_{max} decrease, the AP_{min}'s improvement is higher. It has more excellent value for practical applications and optimization for the overall model.

It can be seen from Figure 9 that the FS^{2}ANet proposed in this paper converges faster in the training process. After the ninth epoch, the mAP tends to be stable, and the overall process is flat. There is no significant training fluctuation like Roi-Transformer, ReDet, and SASM. The line chart of FS^{2}ANet is closest to the upper left corner, which indicates that the model training speed is fast and the accuracy is high, and it has a good improvement compared with other models.

Figure 10 shows the changes in loss convergence and error convergence during the training process of the FS^{2}ANet algorithm to demonstrate the convergence of the improved algorithm. It can be seen from Figure 10 that the improved algorithm has a faster rate of convergence, especially regarding the regression loss, which tends to converge after the third epoch. Due to the tiny size of various aircraft targets and the tiny class differences, the classification loss only converges after the ninth epoch. Finally, the classification loss converges around 0.01, regression loss converges around 1.1, and error loss converges around 2.2.

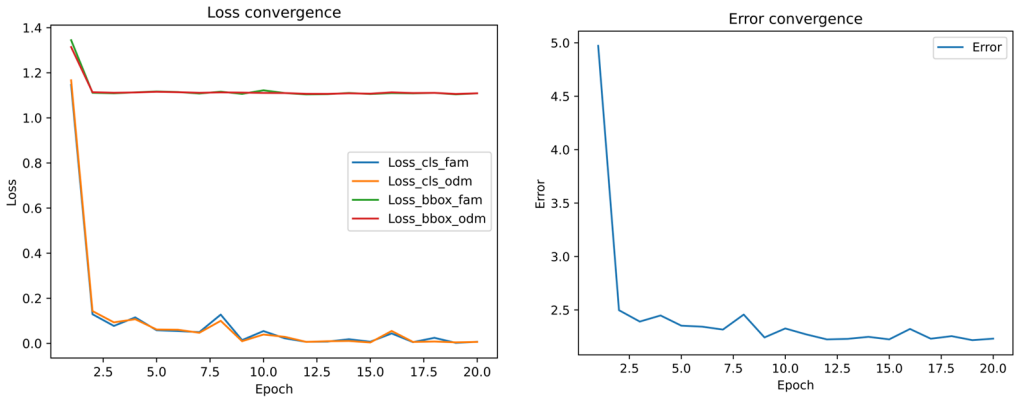

Figure 10. Loss convergence and error convergence.

Table 3 shows the results of using the FS^{2}ANet model to detect various aircraft types; the detection and recognition accuracy of various aircraft types varies, mainly due to the imbalance in the number of aircraft instances; the differences in fine-grained features of various aircrafts is slight, with a varying complexity of the remote sensing background and many small objects. Regarding FS^{2}ANet, modules such as AF^{3}M and FPN-SPD were added to extract fine-grained aircraft features, reduce the loss of fine-grained features, and improve the accuracy of Boeing 737, Boeing 777, Boeing 787, and A220. In contrast, data augmentation reduced the long tail effect. However, data augmentation will introduce other aircraft instances, so it can only partially solve the problem of instance imbalance, such as improving the accuracy of C919.

Table 3. Training results for various types of aircrafts in the dataset.

Class	S^2ANet		FS^2ANet	
	Recall	AP	Recall	AP
Boeing737	0.97	**0.386**	0.948	**0.427**
Boeing747	0.976	0.863	0.971	0.849
Boeing777	0.968	0.141	0.845	0.219
Boeing787	0.978	0.463	0.947	0.524
ARJ21	0.839	**0.11**	0.552	**0.131**
C919	0.679	**0.003**	0.357	**0.189**
A220	0.977	**0.446**	0.969	**0.471**
A321	0.97	0.608	0.944	0.584
A330	0.958	**0.393**	0.874	**0.452**
A350	0.925	0.583	0.919	0.576
Other airplane	0.953	0.728	0.940	0.727

4.4.2. Performance on the DOTA Dataset

In order to ensure the generalization performance of the network proposed in this article and the effectiveness of each module, we added another comparative experiment on the DOTA dataset [29]. Table 4 shows the experimental results, and we can see that the $mAP_{0.5}$ of FS^2ANet is still the highest, indicating that the network has a good generalization performance compared with the accuracy of Roi Transformer, SASM, ReDet, R^3Det, Faster-Rcnn, Rotated RetinaNet, GWD, and S^2ANet. Figure 11 shows the detection results applied to the DOTA dataset.

The effectiveness of FS^2ANet for small target detection can also be seen in Figure 11. The DOTA dataset has 15 types of targets: a plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer field, and swimming pool. Table 5 shows the AP and Recall for each category.

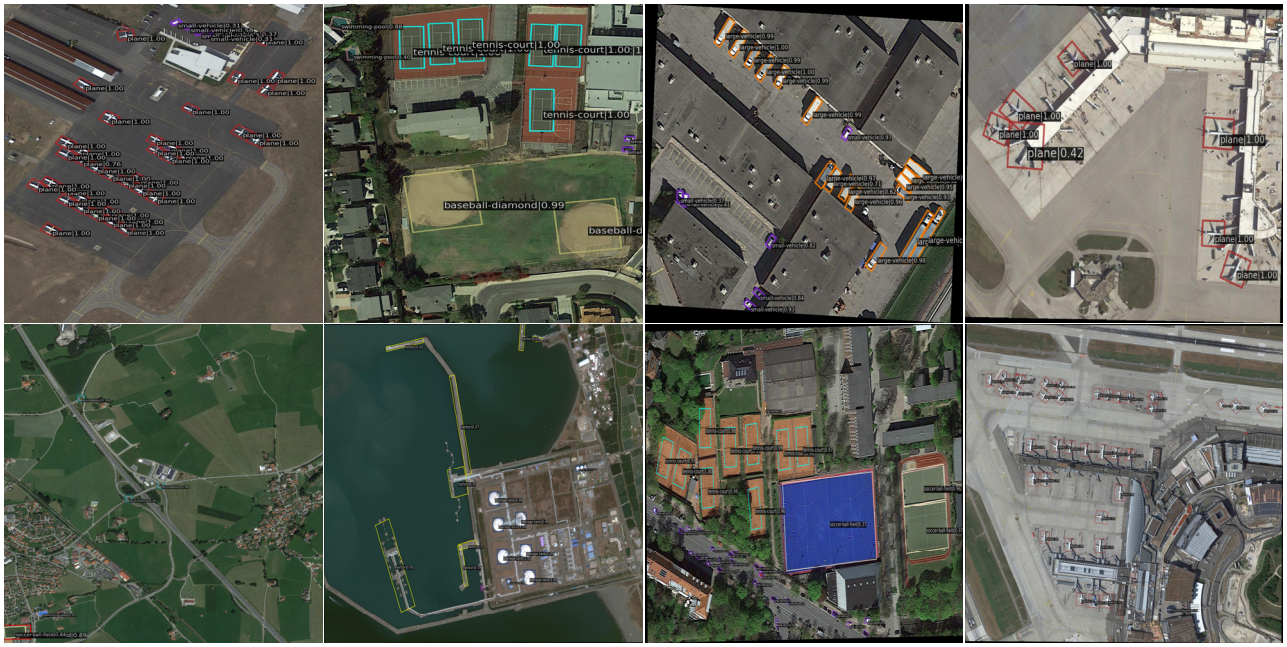

Figure 11. Detection results of FS^2ANet in the DOTA dataset.

Table 4. Training results in the DOTA dataset.

Algorithm	Backbone	Size	$mAP_{0.5}$/%
Roi-Transformer	ResNet50	1024 × 1024	77.83
SASM	ResNet50	1024 × 1024	77.65
ReDet	ResNet50	1024 × 1024	78.21
R^3Det	ResNet50	1024 × 1024	76.50
Faster-Rcnn	ResNet50	1024 × 1024	77.40
Rotated RetinaNet	ResNet50	1024 × 1024	76.62
GWD	ResNet50	1024 × 1024	77.87
S^2ANet	ResNet50	1024 × 1024	77.86
FS^2ANet (Ours)	ResNet50	1024 × 1024	**78.40**

Table 5. Training results for each category in the DOTA dataset.

Class	Gts	Dets	Recall	AP
plane	4449	10,617	0.955	0.905
ship	18,537	39,330	0.955	0.895
storage tank	4740	14,108	0.822	0.780
baseball diamond	358	3277	0.936	0.842
tennis court	1512	6159	0.959	0.908
basketball court	266	4006	0.962	0.885
ground track field	212	3075	0.863	0.721
harbor	4167	24,890	0.864	0.779
bridge	785	10,254	0.783	0.622
large vehicle	8819	47,535	0.942	0.867
small vehicle	10,579	71,087	0.864	0.717
helicopter	122	6040	0.869	0.700
roundabout	275	2752	0.909	0.795
soccer field	251	3754	0.837	0.670
swimming pool	732	5639	0.816	0.674

4.5. Ablation Experiment

In order to solve the above problems in remote sensing aircraft object detection and fine-grained recognition, the method proposed in this paper mainly has four parts of improvement: building the AF^3M module in the detection head, modifying the regression loss function to KFIoU loss, building the FPN-SPD module, using data enhancement to alleviate the long tail effect, and replacing the backbone with ResNet101.The ablation experiment mainly verified the effectiveness of these four parts and their impact on the $mAP_{0.5}$ of the model. Table 6 shows the ablation experimental results.

Table 6. Results of ablation study.

Algorithm	Data Augmentation	ResNet101	AF^3M	KFIoU	FPN-SPD	$mAP_{0.5}$ (%)
S^2ANet						42.95
S^2ANet+Data enhancement	√					45.3 (+2.35)
S^2ANet+ResNet101		√				44.1 (+1.15)
S^2ANet+AF^3M			√			44.21 (+1.26)
S^2ANet+KFIoU				√		43.46 (+0.51)
S^2ANet+FPN-SPD					√	43.53 (+0.58)
S^2ANet+AF^3M+FPN-SPD			√		√	44.52 (+1.57)
S^2ANet+AF^3M+KFIoU			√	√		44.55 (1.6)
S^2ANet+KFIoU+FPN-SPD				√	√	43.81 (+0.86)
S^2ANet+AF^3M+KFIoU+FPN-SPD			√	√	√	44.8 (+1.85)
FS^2ANet (Ours)	√	√	√	√	√	**46.82 (+3.87)**

In this table, √ means that the corresponding method was adopted.

From the above table, we can see that when using data augmentation, ResNet101, AF^{3}M, KFIoU, and FPN-SPD alone, mAP$_{0.5}$ increased by 2.35%, 1.15%, 1.26%, 0.51%, and 0.58%, respectively. All modules can make the model improve to a certain extent. The combination of AF^{3}M, KFIoU, and FPN-SPD increased by 1.57%, 1.6%, and 0.86% for mAP$_{0.5}$, respectively, verifying the effectiveness of each module. Finally, this study used all the improvements to construct the FS^{2}ANet model. The experiment showed that the FS^{2}ANet model had the best performance—the mAP$_{0.5}$ reaching 46.82%, which was 3.87% higher than S^{2}ANet. Figure 12 shows the detection results of the FS^{2}ANet model and the label of the aircraft object.

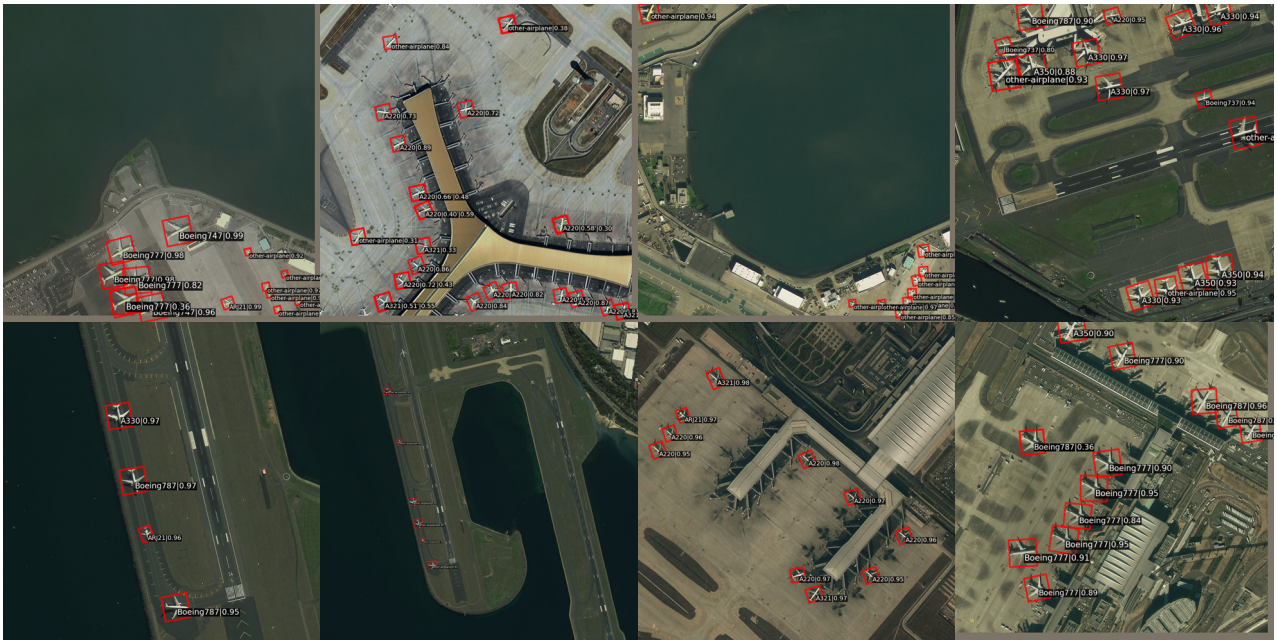

Figure 12. Detection and recognition results of FS^{2}ANet.

Using ResNet101 with deeper network layers can capture deeper aircraft features, which helps the algorithm detect various aircraft targets in complex remote sensing backgrounds. Moreover, with deeper network layers, the category information between different types of aircrafts is more abstract, and the semantic feature spacing between different types of aircrafts is more extensive, which is conducive to classification between different types of aircrafts, thereby improving the accuracy of the aircraft target detection and fine-grained recognition of the algorithm. The detection and recognition results show that small aircraft objects' detection and recognition accuracy has significantly improved. Some aircraft objects have missed detection or insufficient recognition accuracy due to the small pixel proportion, and the accuracy of the rotated angle of the aircraft object box needs to be improved. Some oriented bounding boxes have tilted angles, but the model still has a good aircraft object detection and fine-grained recognition performance in complex remote sensing backgrounds.

5. Discussion

The novelty of this study mainly lies in the construction of the AF^{3}M and FPN-SPD modules, which can effectively extract object features and reduce the loss of fine-grained information. In addition, to verify the effectiveness and generalization performance of the FS^{2}ANet network, experiments were conducted on the FAIR1M and DOTA datasets. The two sets of comparative experiments in Section 4.4 show that the algorithm proposed in this paper achieved the best mAP in small object detection.

From the experimental results in Section 4.5, the optimization of each module and network structure designed in this article improved the detection accuracy of the network

to a certain extent. The AF^{3}M module can enable the network to obtain better object features and improve the network's ability to detect small objects. The FPN-SPD module adopts a multi-output layer structure, and the network can use the multi-layer output to detect objects of different scales, which is conducive to improving the detection accuracy of small objects; in the S module, it can also reduce the loss of object information. This paper introduces KFIoU as a regression loss function to measure the similarity between the genuine box and the predicted box, which improves the regression accuracy of object detection. Data augmentation also significantly improved the detection accuracy of the network for C919 and ARJ21. However, this article did not consider other remote sensing scenarios like cloudy, foggy, and dark conditions. As this article only focused on improving detection and recognition accuracy and did not consider an increased computational cost, based on existing modules, the increase in computational complexity is insignificant and can be applied in practical remote sensing object detection and fine-grained recognition. In addition, aircraft target detection and fine-grained recognition algorithms based on high-resolution remote sensing images still face many limitations and challenges in the future. For example, obtaining and annotating many high-resolution remote sensing aircraft images while ensuring a balanced number of aircraft instances in each category is a significant challenge. Moreover, the higher the resolution of remote sensing images, the more detailed information and pixels specific to the target for object detection and classification; this is beneficial for detection and classification, but designing algorithms with a higher input resolution and faster speed is also a considerable challenge. A faster hardware computing speed can drive larger models. It is also a good direction to improve the adaptive activation function to speed up the rate of convergence and detection accuracy [30] of larger models in the future, and we will design a network structure that is faster and more accurate than FS^{2}ANet for complex remote sensing scenarios.

6. Conclusions

This article proposes an FS^{2}ANet algorithm based on the current pain points of remote sensing aircraft object detection and fine-grained recognition. This network designs and adds the AF^{3}M and FPN-SPD based on S^{2}ANet to extract fine-grained features and reduce information loss, replaces the regression loss function with KFIoU to reduce the regression box loss, and its data enhancement alleviates the long tail effect and uses a deeper backbone to extract remote sensing aircraft features. According to experimental results on the FAIR1M and DOTA datasets, the FS^{2}ANet algorithm achieved the best detection and recognition performance compared to other methods, effectively improving the accuracy of remote sensing aircraft target detection and fine-grained recognition. It can achieve the detection and fine-grained recognition of 11 types of remote sensing aircraft targets, such as Boeing 737, A321, and C919. In future work, we will strive to address the issues and challenges mentioned in the Discussion and design high-resolution models to improve the accuracy and speed of aircraft target detection.

Author Contributions: Conceptualization, Q.G. and Y.L.; methodology, Q.G. and Y.L.; software, Q.G. and G.L.; formal analysis, L.C.; writing—original draft preparation, Q.G.; writing—review and editing, L.C. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Department of Science and Technology of Jilin Province, China (20210201130GX, 20230203028SF).

Data Availability Statement: All datasets used for training and evaluating the performance of our proposed approach are publicly available and can be accessed from [28,29].

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

Abbreviations

S^2ANet	Single-shot Alignment Network
FS^2ANet	Fine-grained S^2ANet
AF^3M	Aircraft fine feature focusing module
SPD	Space-to-depth
FPN	Feature pyramid network
KFIoU	SkewIoU based on Kalman filtering
CNN	Convolutional Neural Network
SA-S	Shape adaptive selection
SA-M	Shape adaptive measurement
FAM	Feature Alignment Module
ODM	Orientation Detection Module
CAM	Channel attention module
SAM	Spatial attention module
ARN	Anchor refinement network
AP	Average precision
IoU	Intersection over union
R	Recall
P	Precision
YOLO	You only look once
SSD	Single-shot multi-box detector
mAP	Mean average precision
MLP	Multi-layer perceptron
BN	Batch Normalization
SGD	Stochastic gradient descent

References

1. Lecun, Y.; Bottou, L. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
2. Ju, M.; Niu, B.; Jin, S.; Liu, Z. SuperDet: An Efficient Single-Shot Network for Vehicle Detection in Remote Sensing Images. *Electronics* **2023**, *12*, 1312. [CrossRef]
3. Guo, J.; Wang, Z.; Zhang, S. FESSD: Feature Enhancement Single Shot MultiBox Detector Algorithm for Remote Sensing Image Target Detection. *Electronics* **2023**, *12*, 946. [CrossRef]
4. Yu, L.; Zhou, X.; Wang, L.; Zhang, J. Boundary-Aware Salient Object Detection in Optical Remote-Sensing Images. *Electronics* **2022**, *11*, 4200. [CrossRef]
5. 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]
6. 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; Springer Nature Switzerland: Cham, Switzerland, 2016; pp. 21–37.
7. 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.
8. Wang, S.; Gao, X.; Sun, H.; Zheng, X.; Sun, X. Aircraft object detection method based on CNN for high-resolution SAR images. *J. Radar* **2017**, *6*, 195–203.
9. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *arXiv* **2017**, arXiv:1706.09579.
10. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
11. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *arXiv* **2019**, arXiv:1908.05612. [CrossRef]
12. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021, Nashville, TN, USA, 20–25 June 2021.
13. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *arXiv* **2020**, arXiv:2008.09397. [CrossRef]
14. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
15. Hou, L.; Lu, K.; Xue, J.; Li, Y. Shape-adaptive selection and measurement for oriented object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 923–932.

16. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
17. Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; Tian, Q. The kfiou loss for rotated object detection. *arXiv* **2022**, arXiv:2201.12558.
18. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, 19–23 September 2022; Springer Nature Switzerland: Cham, Switzerland, 2023; pp. 443–459.
19. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
21. 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.
22. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
23. 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; pp. 3–19.
24. Jagtap, A.D.; Karniadakis, G.E. How important are activation functions in regression and classification? A survey, performance comparison, and future directions. *arXiv* **2022**, arXiv:2209.02681. [CrossRef]
25. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323, Workshop and Conference Proceedings.
26. Dubey, S.R.; Chakraborty, S. Average biased relu based cnn descriptor for improved face retrieval. *Multimed. Tools Appl.* **2021**, *80*, 23181–23206. [CrossRef]
27. Eger, S.; Youssef, P.; Gurevych, I. Is it time to swish? comparing deep learning activation functions across nlp tasks. *arXiv* **2019**, arXiv:1901.02671.
28. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
29. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
30. Jagtap, A.D.; Karniadakis, G.E. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *J. Comput. Phys.* **2019**, *404*, 109136.