applied sciences

MDPI

Article

Aircraft Rotation Detection in Remote Sensing Image Based on Multi-Feature Fusion and Rotation-Aware Anchor

Feifan Tang [1], Wei Wang [1,*], Jian Li [1], Jiang Cao [1], Deli Chen [1], Xin Jiang [1], Huifang Xu [2] and Yanling Du [2]

[1] School of Computer Science, Fudan University, 220 Handan Rd., Shanghai 200433, China; 19110240048@fudan.edu.cn (F.T.); cbrdcbrd@163.com (J.L.); caojiang128@163.com (J.C.); cdl_1@TOM.com (D.C.); beita1986@163.com (X.J.)

[2] College of Information and Technology, Shanghai Ocean University, No.999, Huchenghuan Rd., Shanghai 201306, China; xuhuifang_bb@163.com (H.X.); yldu@shou.edu.cn (Y.D.)

* Correspondence: weiwang1@fudan.edu.cn

Abstract: Due to the variations of aircraft types, sizes, orientations, and complexity of remote sensing images, it is still difficult to effectively obtain accurate position and type by aircraft detection, which plays an important role in intelligent air transportation and digital battlefield. Current aircraft detection methods often use horizontal detectors, which produce significant redundancy, nesting, and overlap of detection areas and negatively affect the detection performance. To address these difficulties, a framework based on RetinaNet that combines a multi-feature fusion module and a rotating anchors generation mechanism is proposed. Firstly, the multi-feature fusion module mainly realizes feature fusion in two ways. One is to extract multi-scale features by the feature pyramid, and the other is to obtain corner features for each layer of feature map, thereby enriching the feature expression of aircraft. Then, we add a rotating anchor generation mechanism in the middle of the framework to realize the arbitrary orientation detection of aircraft. In the last, the framework connects two sub-networks, one for classifying anchor boxes and the other for regressing anchor boxes to ground-truth aircraft boxes. Compared with state-of-the-art methods by conducting comprehensive experiments on a publicly available dataset to validate the proposed method performance of aircraft detection. The detection precision (P) of proposed method achieves 97.06% on the public dataset, which demonstrates the effectiveness of the proposed method.

Keywords: multi-feature fusion; rotating anchor; harris corner detection; retinanet; oriented aircraft detection

Citation: Tang, F.; Wang, W.; Li, J.; Cao, J.; Chen, D.; Jiang, X.; Xu, H.; Du, Y. Aircraft Rotation Detection in Remote Sensing Image Based on Multi-Feature Fusion and Rotation-Aware Anchor. *Appl. Sci.* **2022**, *12*, 1291. https://doi.org/10.3390/app12031291

Academic Editors: Daniel Ortiz-Arroyo and Petar Durdevic Løhndorf

Received: 15 November 2021
Accepted: 21 January 2022
Published: 26 January 2022

1. Introduction

Object detection is one of the fundamental tasks in computer vision and has attached wide attention due to the success of deep learning [1]. With the development of sensor and aerospace technology, large volume and high-resolution remote sensing images have accumulated. Object detection in remote sensing images has become a research hotspot [2–4]. Currently, aircraft are a common means for transportation and warfare. Therefore, aircraft detection in remote sensing images plays a great significance in civilian air traffic management and military intelligence [5], especially detecting accurate aircraft position and size information, and has crucial practical significance and military value for such as enemy's operational readiness analysis and regional situation assessment.

Aircraft detection consists of two parts: locating object regions and classifying object in the candidate regions. In order to meet the requirements of rapid and accurate aircraft detection, it is necessary to get invariant features and train a classifier with a strong generalized capability. Extensive research has focused on feature design and feature expression to get good aircraft detection results. It usually constructs rotation-invariant and scale-invariant features based on object shape, texture, and geometric features and collaborates with the general classifiers, e.g., SVM, Bayes, KNN, and other classifiers.

Most methods realize object detection by extracting excellent features, such as Harris corner feature [6,7], Histogram of oriented gradient (HOG) [8], scale-invariant feature transform (SIFT) [9,10], and Fourier descriptor [11,12]. These methods have achieved good performance in the field of object detection, but the aircraft scales and shapes in the optical remote sensing images are complex and diverse, and the manual features are difficult to adapt to the diverse remote sensing images. The methods based on manual-designed features are laborious and lack generalization ability, which leads to the disadvantage that the accuracy cannot meet the requirements of real applications.

Recently, with the great success of deep learning, much progress has been made in object detection and recognition. Since 2012, Hinton proposed a convolutional neural network (CNN) model called AlexNet, which won the championship of 2012 ImageNet Challenge [13]; many CNN-based deep learning models have mushroomed. The popular detection models can be divided into two categories: two-stage and one-stage object detectors. The two-stage methods, such as R-CNN, Fast R-CNN, and Faster R-CNN [14–16], mainly consist of region proposals generation followed with feature extraction from these regions, and then classifiers and regressors used for classification and regression in the second stage. The one-stage method has obvious advantages in efficiency, especially for the large scale of remote sensing images [17]. OverFeat [18] is one of the first one-stage detectors based on CNN. It performs object detection in a multiscale sliding window fusion via single forward propagation through the CNN. Compared with region-based methods, Redmon et al. [19] proposed YOLO, a unified detector, casting object detection as a regression problem from image pixels to spatially separated bounding boxes and associated class probabilities. To preserve real-time speed without sacrificing too much detection accuracy, Liu et al. [20] proposed single shot multibox detector (SSD). The work solves the class imbalance problem by proposing RetinaNet with Focal loss [21] and further improves the accuracy of one-stage detector. Following their works, Hu et al. [22] used the saliency detection algorithm to reduce the number of proposal boxes. They obtained the object position information using the saliency algorithm based on the background priori, and then used a deep CNN to determine the category and fine-tune bounding boxes of the objects. Wu et al. [23] used Edgeboxes algorithm to generate region proposals, and then used CNN to perform feature extraction and classification. Similarly, Yang et al. [4] proposed a "region proposal-classification-accurate object localization" framework for detecting objects in remote sensing images. However, all the above methods have the problem of redundancy in candidate regions. In order to solve those problems, Liu et al. [5] proposed an aircraft detection method based on corner clustering and CNN, which used the mean-shift clustering algorithm to generate candidate regions for the corner points detected in the binary image, and then utilized CNN to determine the target category in the candidate region. Compared to traditional methods, these approaches based on CNN models could learn the essential features of the object from a large amount of data, without the need to design the features manually, and the detection accuracy has been significantly improved.

However, aircraft detection in remote sensing images faces emerging challenges because of the diversity of orientation, scales, and appearance (color, texture, size, shape, etc.). These problems are very serious in remote sensing images. First, similar to the objects in natural images, the aircraft in remote sensing images also have a multi-scale problem. Huge difference is caused by two aspects, one is that the same objects in remote sensing images may have different sizes due to the varying spatial resolutions of sensors, and the other is the diverse scales of aircrafts [24]. In summary, aircraft in remote sensing images vary greatly in scale and orientation, resulting in a low detection accuracy (high rate of missing detection or false alarm) and region redundancy, especially nesting and overlap in densely arranged region.

To address these challenges mentioned above, an aircraft detection framework based on a multi-feature fusion module and a rotating anchor generation mechanism is proposed in this paper. First, multi-scale features and local corner features for enhancing the ability

of feature expression are considered in our multi-feature fusion module. Among them, multi-scale features are extracted by the feature pyramid network, and local corner features that can well express the salience of aircraft are extracted on every layer output of each layer of FPN, finally the feature map group is constructed. Then, the rotating anchors are generated by multiple aspect ratios, angles and scales. Finally, aircraft detection with accurate location and size information is achieved by classification and regression.

The rest of this paper is organized as follows. The proposed method is described in Section 2. Section 3 illustrates the datasets, details of implementation, evaluation criteria, and experiment results. Finally, Section 4 concludes this paper with a discussion of the experimental results.

2. Methods

In this paper, we build an aircraft detection framework for multi-scale and arbitrary-oriented aircraft in remote sensing images, as illustrated in Figure 1. Our contributions consist of following major parts.

(1) We propose a rotated aircraft detection framework based on RetinaNet. The framework integrates multi-feature fusion and rotating anchors generation mechanism, which can efficiently detect arbitrary-oriented aircrafts and fit oriented bounding boxes.
(2) Multi-feature fusion. Multi-scale features are extracted from a single resolution input image by feature pyramid network (FPN) with a top-down pathway and lateral connections, which includes high semantic features, as well as basis spatial detail features. In addition, in order to enrich the feature expression, Harris corner features are extracted from binary feature maps of different scales in feature pyramid layers and together fused into the feature map group.
(3) Rotating anchors generation mechanism. To build a rotation-aware aircraft detector, we adopt five parameters (x, y, w, h, and θ) to describe object bounding box (OBB), and design multiple rotating anchors with different aspect ratios, angles, and scales, which can fit OBBs well and provide accurate positions of arbitrary-oriented aircrafts.

The overview framework of oriented aircraft detection is shown as Figure 1. First, the high semantic feature is extracted by ResNet-101, and the Harris corner feature is extracted from binary remote sensing images. In order to obtain sufficiently well expression of the aircrafts multi-scale characteristics, we use FPN to extract multi-levels of features that include high semantic features and low detail features (much information about location, scale and shapes). At the same time, multi-feature fusion also integrates with Harris corner, which can well describe the aircraft local salient feature and enriches the feature expression. Moreover, we design rotating anchor generation mechanism to improve the sensitivity of location and the fitting of OBBs. Finally, oriented aircraft detection is achieved by classification and regression. More details are provided in the following subsections.

2.1. Backbone Network

We use the one-stage object detection network RetinaNet [21] as the basic network and construct a multi-scale rotation anchor generation mechanism with multi-scale, multi-aspect ratio, and multi-rotation angle to realize automatic oriented detection of aircraft. The model structure is shown in Figure 1. It consists of four parts: Residual Network (ResNet) for initial extraction of aircraft features [25], Feature Pyramid Network (FPN) for multi-level features fusion and local salient feature integration [26], multi-scale rotating anchor generation, and sub-network for classification and location.

In the initial stage of feature extraction, linear convolutional neural networks are generally used to extract features, such as AlexNet [27], GoogLeNet [28], and VGGNet (visual geometry group) [29]. These networks often increase the depth of the network to improve the ability to express features. When the network reaches a certain depth, the accuracy of training will tend to be flat. If we increase the number of layers again, the optimization effect will become worse. Therefore, we use the ResNet network to initially

extract the features of aircrafts, by introducing residual mapping and jump connections in the network and adding the feature information of the previous residual block to the next residual block. It effectively avoids the disappearance of gradients and loss of feature information caused by the excessive depth of the network.

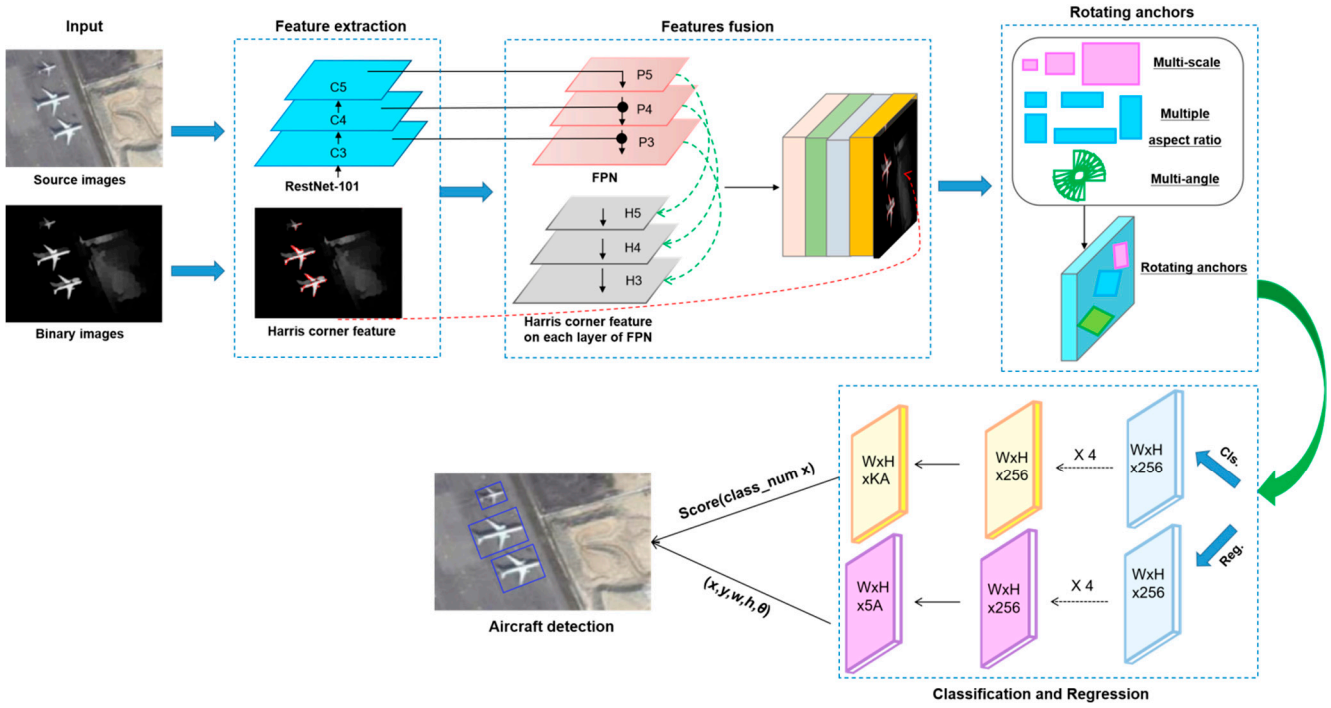

Figure 1. Overview framework of multi-feature fusion with rotating anchors generation mechanism for oriented aircraft detection. The proposed framework based on FPN, fuses multi-scale features from different CNN layers and integrates Harris corner features which extracted from binarized feature maps of FPN.

RetinaNet is a simple one-stage object detector, named for its dense sampling of object locations in an input image [21]. The large class imbalance encountered during training of dense detectors overwhelms the cross-entropy loss. Easily classified negatives comprise the majority of the loss and dominate the gradient. The Focal Loss is designed to address the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes during training.

Specifically, a modulating factor $(1 - p_t)^\gamma$ is added to the cross entropy loss, with tunable focusing parameter $\gamma \geq 0$. The focal loss can be shown as:

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t), \tag{1}$$

2.2. Multi-Feature Fusion

A corner is one of the most important local features in object detection owing to its strong invariance to rotation, scale and illumination variation. In this paper, corners are extracted by the Harris operator [6], which is a popular corner detection algorithm that can be divided into the following steps.

(1) Calculate the gray gradient in the X direction (denoted as I_x) and Y (denoted as I_y) direction of the image I, then filter the ${I_x}^2$, ${I_x}^2 \times {I_y}^2$, ${I_y}^2$ with a Gaussian window.
(2) The covariance matrix M of the object pixel is denoted as below:

$$M = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \tag{2}$$

(3) The value of corner response R of each pixel in the image can be calculated by:

$$R = Det(M) - k(Tr(M))^2, \quad (3)$$

where $Det(M)$ represents the matrix determinant of M, $Tr(M)$ is the trace of M, the k is empirical constant which generally take the value 0.04–0.06. The local non-maximum suppression of R is carried out. That is, retain the points whose values are the local maximum and greater than the threshold T, reset the values of the other points to zero. These non-zero points are corners. Threshold T is generally 10^{-2} times of the global maximum response value, and the smaller the T value is, the more corner points are retained. As shown in Figure 2, the measure R with a large positive value indicates the presence of a corner; R with a large negative value indicates an edge; a small absolute value $|R|$ indicates a flat area.

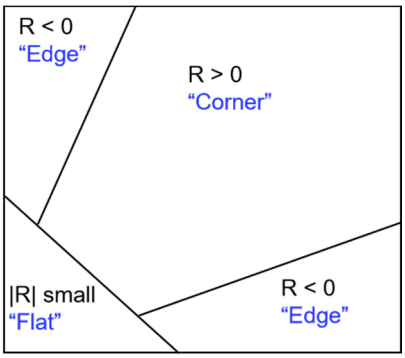

Figure 2. The relation of the image classification with the value of R.

Different levels of features provide more information for accurate object detection. For the diversity scales of the aircraft, the feature pyramid network fuses the low-level and high-level features, providing rich spatial location and high-level semantic information. The feature pyramid network integrates the Harris corner features from different layers, which well expresses the aircraft salient feature for accurate aircraft detection. The network architecture consists of two parts: a bottom-up network, a top-down network and horizontal connection, as shown in Figure 1. The bottom-up feed forward network {C1, C2, C3, C4, C5} is composed of the output of the last convolutional layer in each residual block in the ResNet network. Since C1 and C2 need to take up a lot of memory, then remove them, and the step size of the convolutional layer is set to {8, 16, 32} respectively. The feature map generated by each layer of the network has the characteristics of resolution decreasing layer by layer and semantic information enhancing while location information weakens layer by layer. Therefore, in the top-down network, up-sampling and level connection are used to generate features {P3, P4, P5} containing rich information. P5 is obtained by using a 1×1 convolutional layer to reduce the channel of C5, and then adding a 3×3 convolution to obtain the first layer network P5 of the FPN. P4 is obtained by fusing the rough feature map obtained from C4 and the feature map obtained from P5 and adding 3×3 convolution to eliminate the aliasing effect of up sampling, so as with P3. In contrast to the traditional FPN, the model introduces two more layers with small scales in order to improve the detection accuracy of obvious objects with significant structures, which are demoted as P6 and P7. P6 is obtained by performing convolution with a 3×3 kernel and the step size of 2, taking C5 as the input. P7 is obtained by adding RELU activation and a step size of 2, 3×3 convolution operation based on P6. The FPN structure can effectively use a single image to construct multi-scale feature maps so that the feature map output from each layer of the pyramid has strong semantic information, which provides more abundant spatial location and high-level semantic feature information for the detection of multi-scale and arbitrary orientation.

At the same time, we fuse the corner features to enhance the feature expression in multi-feature fusion module, which is shown in Figure 1. The fusion process includes two parts: one is the corner feature extracted from the binarization of the original image, the other is the corner features obtained from each layer feature maps of the feature pyramid network based on binarization. For example, the feature map P3 is binarized, and then the Harris corner features are extracted to obtain H3, which is the same process for H4, H5, H6, and H7. In the end, the corner features are fused into the feature map group together with the multi-scale feature maps.

2.3. Rotating Anchor Generation Mechanism

The input of the multi-scale rotation anchor is the output of FPN, and the rotation area network is constructed based on the area generation network and the fully convolutional neural network. The network first generates a multi-scale rotation anchor composed of scale, aspect ratio, and angle, and puts it into two sub-networks of classification and regression. The classification sub-network predicts the probability of the target, and the regression sub-network regresses the offset between the boundary box and the true value box. Finally, the non-maximum suppression operation and backward processing operation are performed to obtain the final detection result.

In the multi-scale rotation anchor generation mechanism, three parameters are used to control the anchor, namely, scale, aspect ratio, and angle. To save the calculation cost and make the anchor cover the aircraft as much as possible, the basic scale of the initial anchor in {P3, P4, P5, P6, P7} layers is set as {32, 64, 128, 256, 512}, the scale of each layer is set as $\{2^0, 2^{1/3}, 2^{2/3}\}$ to control the size of the anchor, and the aspect ratio is set as {1, 1/2, 2,1/3, 3, 1/5, 5} to control the proportion of the anchor. At the same time, five different angles are set, namely $\{-15^\circ, -30^\circ, -45^\circ, -60^\circ, -75^\circ\}$, to realize the anchor rotation to different directions. The multi-scale rotating anchor generation strategy can well solve the problem of detection frame nesting and overlapping in the horizontal detection of aircraft with diverse scales and arbitrary orientation. In addition, assuming that each feature point of the feature map will generate A anchors, the H*W feature map will generate A*H*W anchors.

In the prediction stage of the sub-network, the traditional horizontal bounding box (dashed box) is usually represented by x_{min}, y_{min}, x_{max}, and y_{max}; based on these four parameters, the bounding box regression transformation is performed to obtain the predicted bounding box, but it cannot well cover the aircraft in different directions. Therefore, the representation of rotating bounding box (solid wireframe) is defined, which is not only controlled by scale and aspect ratio, but also by direction. Five variables (x, y, w, h, and θ) are used to uniquely determine a bounding box in any direction, as shown in Figure 3. Where x and y represent the coordinates of the center point, respectively, θ represents the angle between the horizontal axis and the first side of the rectangle, and the range of the angle is $[-90^\circ, 0^\circ]$.

Based on the above five parameters, the classification sub-network predicts the probability (KA) of the object in the anchor at each spatial position, where A represents the number of anchors, and K represents the category. The regression sub-network outputs the position information (5A) of each anchor box based on the five parameters defined above and performs regression transformation based on the offset between the anchor box, and the truth box calculates the intersection over union (IOU) and gets the rotating anchor box, which has the largest IOU with the truth box. The position information of the target in the predicted bounding box includes the coordinates, width, height, and angle of the center point. Based on the obtained position information, the center, scale, and radius of the aircraft in the prediction box can be calculated, which lays a foundation for the next processing.

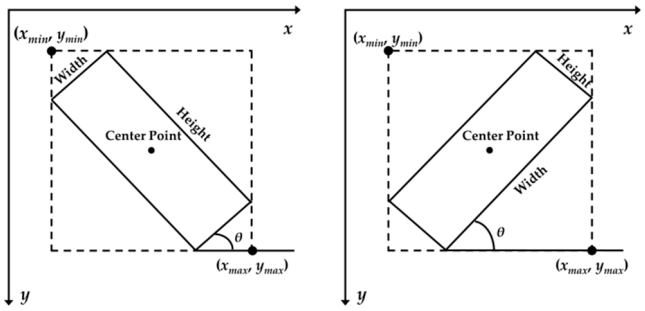

Figure 3. Horizontal bounding box and rotation bounding box.

3. Results

In this section, we conduct ablation experiments to explore the proposed aircraft detector and demonstrate the effectiveness of our methods. First, we describe the dataset for experiments and the implementation details of our method. Second, several groups of ablation experiments are conducted to evaluate the performance of the proposed modifications, and a detailed analysis is conducted on the aircraft detector. Finally, we present the comparisons with existing object detection algorithms and demonstrate that our method is a competitive aircraft detection framework for remote sensing images.

3.1. Dataset

We select the aircraft dataset from DOTA [30] and UCAS-AOD [31] to train and test the proposed model. Rather than the horizontal box in previous datasets, each instance is annotated with an oriented bounding box, which provides more accurate location and size of the object.

The DOTA image size, ranging from around 800 × 800 to 4000 × 4000 pixels, contains aircrafts that exhibit a great variety of scales, orientations, and shapes. Specifically, the DOTA dataset maintains a large number of small instances and provides large objects over 500 pixels as well. In short, DOTA provides a reliable benchmark for studying multi-scale and arbitrary-oriented object detection in aerial images. We selected 79 images containing 3509 objects from DOTA.

In UCAS-AOD, we selected 1000 aircraft images containing 7482 objects. All objects in UCAS-AOD were labeled with both oriented bounding boxes. In our experiments, we randomly divided the training and test set by 8:2.

3.2. Evaluation Metrics

Three common evaluating metrics, Precision (P), Recall (R), and F1 value, are used to measure the accuracy of the proposed method. The number of image frames processed per second (FPS) is the measurement of the method speed performance. The Precision metric represents the ratio of detection that are true positives, and the Recall metric means the ratio of positives that are detected accurately. The Precision and Recall metrics can be formulated as follows:

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

where, TP, FP, and FN indicate the number of true positive, false positive, and false negative, respectively.

The F1 value is calculated as (6):

$$F1 = \frac{2 * (\mathrm{Precision} * \mathrm{Recall})}{\mathrm{Precision} + \mathrm{Recall}}, \tag{6}$$

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

The experiment builds and trains the model based on TensorFlow framework. The model training process is carried out in the environment of two GPUs (Tesla p100-pcie). Based on the constructed aircraft training set, image samples and annotation files are put into the model for training, and the model super-parameters are adjusted according to experience and experimental results to test the best performance of the model.

In this experiment, the IoU is set to 0.5, the model learning rate is initialized to 0.0005, batch size is initialized to 2 images, weight decay and momentum are initialized to 0.0001 and 0.9, respectively, and the maximum value of iteration is set to 500,000. In each iteration, the input image is subjected to a complete feed forward calculation and back propagation in the network, and the training parameters are updated until the return loss and classification loss of the network become convergent. The convergence curve of the proposed method of training is shown in Figure 4. The total loss is the sum of classification loss and regression loss.

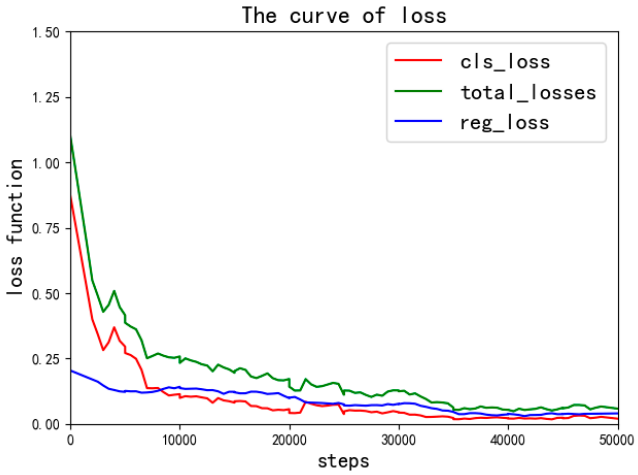

Figure 4. Curves of classification loss (red line), regression loss (blue line), and total loss (green line).

As shown in Figure 4, we can find that the loss rate is basically stable when the number of iterations reaches 50,000 and the total loss rate is only reduced about 1% when the number of iterations is increased to 50,000. Therefore, we can use the model of 50,000 iterations to detect aircraft in the real-time application.

In order to verify the performance of the feature fusion module and the rotation anchor generating module, we performed an ablation study of the proposed method. At the same time, we use 50% cross validation to obtain an average value. Table 1 shows that rotating anchors can improve performance by about 7% (0.9641 versus 0.8894 as shown in Table 1), which is significant. We believe that the main reason is that the rotated anchor can fit the target well and reduce the number of negative samples, which are effective for improving performance. In addition, by fusing the multi-scale feature maps with local corner features the feature fusion module increases F1 performance by 8.12%, compared with RetinaNet (88.94%), reaching 97.06%. This phenomenon reflects the feature fusion module enhance the expression of the feature.

Table 1. Ablative study of each component in our method.

Backbone	Feature Fusion Module	Rotation Anchor	Recall	Precision	F1
ResNet-50	×	×	0.8616	0.8556	0.8584
ResNet-101	×	×	0.9072	0.9016	0.8894
ResNet-50	×	√	0.9633	0.9534	0.9592
ResNet-101	×	√	0.9750	0.9552	0.9641
ResNet-50	√	√	0.9691	0.9569	0.9601
ResNet-101	√	√	0.9863	0.9583	0.9706

In addition, we compare the performance of the state-of-the-art deep learning networks in aircraft detection, including two-stage networks of Faster RCNN and FPN, and one-stage networks of RetinaNet, YOLOv4, SSD and R3Det [1]. Among them, R3Det is the one-stage method with the best performance at present for rotated detection. The two-stage methods need to generate object candidate proposals in the first stage, then classify and locate the object in the second stage. They can achieve high accuracy while taking a long time to train the network. The one-stage object detection methods directly convert the detection problem into a regression problem, which greatly improves the detection speed. It can be seen from Table 2 that the aircraft detection accuracy of FPN is the best, 90.84%, which is 1.6% and 0.68% higher than Faster RCNN and RetinaNet, respectively. The F1 value of RetinaNet is the highest, 88.94%, which is 6.3% and 4.26% higher than Faster RCNN and FPN, respectively. At the same time, according to the FPS value in Table 2 and the model training and testing time in Table 3, the efficiency of RetinaNet is the best. Experimental analysis shows that RetinaNet has better stability when it achieves better accuracy. Therefore, we use RetinaNet as the backbone network to study the effectiveness of the proposed method.

The effects of the rotating anchors generation module and multi-feature fusion module on the method are verified and the results are shown in Table 2. Compared with the optimal HBB detection method (Faster RCNN, FPN and RetinaNet), the aircraft detection performance of our proposed OBB is improved by about 8% (8.12%), although there is a slight decrease (about 1.62 fps) in speed performance.

The compared results between R3Det and proposed method show R3Det has the best detection performance (F1 = 97.49%), which is 0.43% higher than the proposed method. However, the feature refinement module FRM in R3Det requires a lot of computing resources. Table 2 shows that the speed performance decrease about 2.35 fps. Through the analysis of experimental comparison, it is shown that the proposed method can achieve the considerable detection performance as R3Det while with better performance in speed.

Table 2. Comparison of aircraft detection performance of deep learning methods.

Methods	Backbone Network	R	P	F1	FPS
Faster RCNN	Resnet-101	76.95%	89.24%	82.64%	9.11
FPN	Resnet-101	79.30%	90.84%	84.68%	9.71
RetinaNet	Resnet-101	90.72%	90.16%	88.94%	13.0
SSD	VGG-16	84.09%	90.17%	86.40%	28.04
YoloV4	CSPDarknet53	86.10%	96.03%	91.08%	41.02
R3Det	Resnet-50	97.01%	95.96%	96.12%	9.27
	Resnet-101	98.70%	96.39%	97.49%	9.03
RetinaNet + Rotating Anchor	Resnet-50	96.33%	95.34%	95.92%	12.79
	Resnet-101	97.50%	95.52%	96.41%	12.12
Proposed Method	Resnet-50	96.91%	95.69%	96.01%	11.85
	Resnet-101	98.63%	95.83%	97.06%	11.38

Table 3. Training time and test time for aircraft detection methods based on deep learning.

Method	Faster RCNN	FPN	RetinaNet	R3Det	RetinaNet + Rotating Anchor	Proposed Method
Train	0.46 s	1.35 s	0.18 s	0.48 s	0.43 s	0.44 s
Test	0.17 s	0.20 s	0.09 s	0.16 s	0.12 s	0.13 s

In addition, the visualization results of the proposed OBB aircraft detection method in this paper and HBB detection by FPN are shown in Figure 5. First, we can see that rotation detection significantly reduces the overlap and nesting of detection boxes in the red rectangular region. What's more, compared with the results of HBB and OBB detection methods proposed in this paper, in the purple rectangular region, the redundancy of the rotation detection boxes is significantly reduced, and the rotation detection boxes fit the aircrafts better, obtaining much more accurate information of aircraft scales and locations.

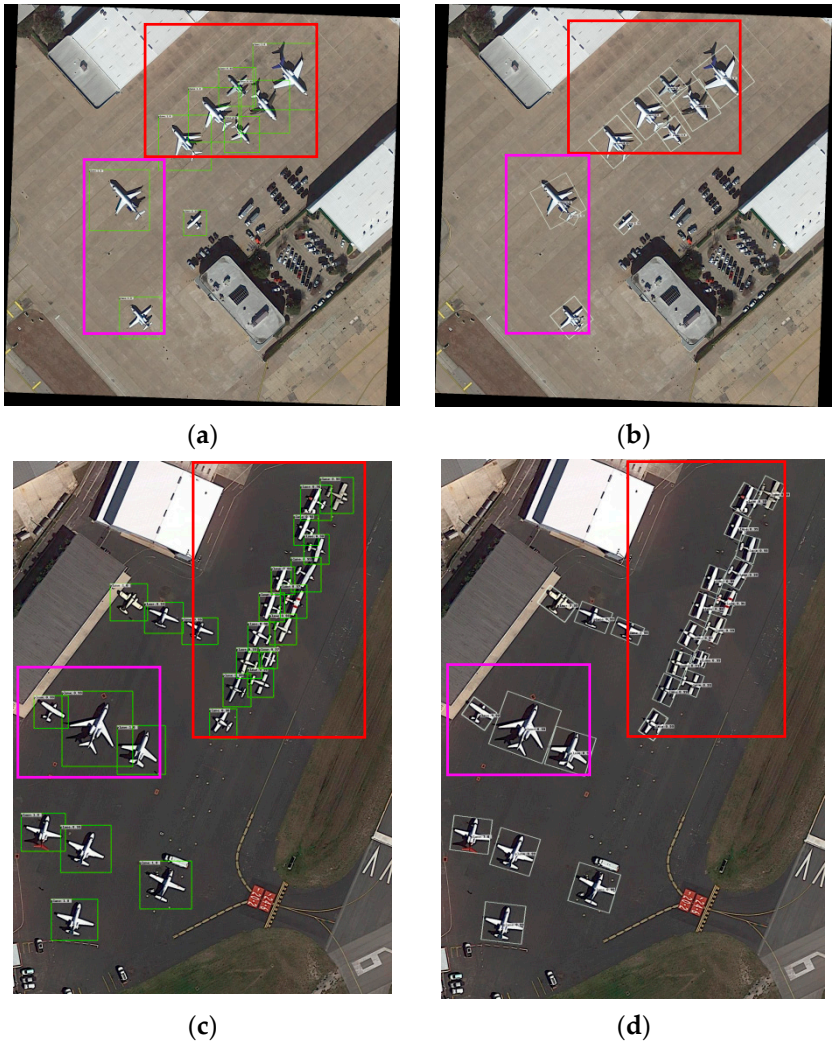

Figure 5. Visualization of aircraft detection results of proposed method on OBB in (**b**,**d**) and HBB method in (**a**,**c**).

4. Conclusions

In this paper, we propose a novel and end-to-end one-stage framework for arbitrary-oriented and multi-scale aircraft detection in remote sensing images. Aiming to improve the shortcoming of low detection accuracy caused by redundancy of object detection region, especially the overlap and nesting of detection areas dense areas, we design a

multi-feature fusion layer with different scales and feature levels and rotating anchors generation mechanism. We perform comparative experiments on two aircraft rotation detection datasets, including DOTA and UCASAOD, and demonstrate that our method achieves state-of-the-art detection accuracy with high efficiency. The experiments show the following: (1) the multi-feature fusion module learns multi-level features with semantic information and fine details, which is necessary for aircraft detection in remote sensing images; (2) the oriented anchor is more robust to aircraft in remote sensing images, and adding anchors with extra angles and scales brings improvement; and (3) the proposed methods help the network handle the issue of arbitrary-orientation and diverse-scale aircrafts in remote sensing images, which can obtain more accurate aircraft position and size information.

Author Contributions: Conceptualization, W.W. and J.L.; methodology, F.T. and D.C.; model training and validation, J.C., X.J., Y.D. and H.X.; writing—original draft preparation, F.T.; and Y.D. writing—review and editing, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China under Grants 41906179 and the local capacity building project of Shanghai Municipal Commission of science and technology under Grants 20050501900.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author wishes to thank the providers of DOTA dataset and UCAS-AOD dataset.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv* **2019**, arXiv:1908.05612.
2. Ji, F.; Ming, D.; Zeng, B.; Yu, J.; Qing, Y.; Du, T.; Zhang, X. Aircraft Detection in High Spatial Resolution Remote Sensing Images Combining Multi-Angle Features Driven and Majority Voting CNN. *Remote Sens.* **2021**, *13*, 2207. [CrossRef]
3. Wu, Z.-Z.; Wan, S.-H.; Wang, X.-F.; Tan, M.; Zou, L.; Li, X.-L.; Chen, Y. A benchmark data set for aircraft type recognition from remote sensing images. *Appl. Soft Comput.* **2020**, *89*, 106132. [CrossRef]
4. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
5. Liu, Q.; Xiang, X.; Wang, Y.; Luo, Z.; Fang, F. Aircraft detection in remote sensing image based on corner clustering and deep learning. *Eng. Appl. Artif. Intell.* **2019**, *87*, 103333. [CrossRef]
6. Harris, C.; Stephens, M. A combined corner and edge detector. In Proceedings of the Alvey Vision Conference, Manchester, UK, 31 August 1988; Volume 15, pp. 10–5244.
7. Mitra, D.; Baksi, A.; Modak, A.; Das, A.; Das, A. Machine Learning Approach for Signature Recognition by HARRIS and SURF Features Detector. *Int. J. Comput. Sci. Eng.* **2019**, *7*, 73–80. [CrossRef]
8. Liu, B.; Wu, H.; Su, W.; Zhang, W.; Sun, J. Rotation-invariant object detection using Sector-ring HOG and boosted random ferns. *Vis. Comput.* **2017**, *34*, 707–719. [CrossRef]
9. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **2003**, *31*, 3812–3814. [CrossRef] [PubMed]
10. Rashid, M.; Khan, M.A.; Sharif, M.; Raza, M.; Sarfraz, M.M.; Afza, F. Object detection and classification: A joint selection and fusion strategy of deep convolutional neural network and SIFT point features. *Multimed. Tools Appl.* **2018**, *78*, 15751–15777. [CrossRef]
11. Zahn, C.T.; Roskies, R.Z. Fourier Descriptors for Plane Closed Curves. *IEEE Trans. Comput.* **1972**, *C-21*, 269–281. [CrossRef]
12. Lin, C.-S.; Hwang, C.-L. New forms of shape invariants from elliptic fourier descriptors. *Pattern Recognit.* **1987**, *20*, 535–545. [CrossRef]
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]

14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision 2015, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
17. Chen, K.; Li, J.; Lin, W.; See, J.; Wang, J.; Duan, L.; Chen, Z.; He, C.; Zou, J. Towards accurate one-stage object detection with ap-loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 5119–5127.
18. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision 2016*; Springer: Cham, Germany, 2016; pp. 21–37.
21. 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 2017, Venice, Italy, 22–29 October 2019; pp. 2980–2988.
22. Hu, G.; Yang, Z.; Han, J.; Huang, L.; Gong, J.; Xiong, N. Aircraft detection in remote sensing images based on saliency and convolution neural network. *EURASIP J. Wirel. Commun. Netw.* **2018**, *2018*, 1–6. [CrossRef]
23. Wu, H.; Zhang, H.; Zhang, J.; Xu, F. Typical target detection in satellite images based on convolutional neural networks. In Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics 2015, Hong Kong, China, 9–12 October 2015; pp. 2956–2961.
24. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
26. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Image net classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. 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 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
31. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP) 2015, Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.