IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 7, 2021, accepted January 3, 2022, date of publication January 6, 2022, date of current version January 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140876

Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv5

SHUN LUO[1], JUAN YU[1], YUNJIANG XI[2], AND XIAO LIAO[3]

[1]School of Economics and Management, Fuzhou University, Fuzhou 350108, China
[2]School of Business Administration, South China University of Technology, Guangzhou 510641, China
[3]School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou 510521, China

Corresponding author: Juan Yu (yujuan@fzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71771054 and Grant 72171090.

ABSTRACT Dealing with the insufficient detection accuracy and speed of aircraft targets in remote sensing images under complex background, this paper proposes a new detection method, YOLOv5-Aircraft, based on the YOLOv5 network. The YOLOv5-Aircraft model is improved in 3 ways: (1) At the beginning and end of original batch normalization module, centering and scaling calibration are added to enhance the effective features and form a more stable feature distribution, which strengthens the feature extraction ability of network model. (2) The cross-entropy loss function in the confidence of the original loss function is improved to the loss function based on smoothed Kullback-Leibler divergence. (3) For reducing information loss, the CSandGlass module is designed on the backbone feature extraction network of YOLOv5 to replace the residual module. Meanwhile, low-resolution feature layers are eliminated to reduce semantic loss. Experiment results demonstrate that the YOLOv5-Aircraft model can enhance the accuracy and speed of aircraft target detection in remote sensing images while achieving easier convergence.

INDEX TERMS Remote sensing image, aircraft detection, YOLOv5, batch normalization, loss function.

I. INTRODUCTION

With the continuous development of satellite remote sensing technology, the information amount of high-resolution remote sensing images has increased sharply, and the detailed information contained in is getting more abundant. Some sensitive targets such as ships, tanks, airplanes and ports can also be clearly visible to naked eyes, for which the detection methods have become a hot spot for scholars. Aircraft play an irreplaceable role in both the civilian and military fields. Therefore, the detection method of aircraft targets in remote sensing images is of great significance.

However, the detection of aircraft targets in remote sensing images remains to be a challenging problem because it is susceptible to interference of external factors such as weather, light, shadows, etc. Besides, when there are small targets in the images with high exposure and complex background, the difficulty of aircraft detection is expected to rise.

Many solutions have been proposed to solve the above problems of target detection [1]. Traditional methods such as template matching are fast, simple and easy to implement, but they have high requirements on the target state and target size and perform badly in complex backgrounds. The machine learning methods are designed to be flexible and highly targeted, but they are solidified and have poor robustness [2]. In recent years, deep learning methods have developed rapidly. Many target detection algorithms based on CNN (Convolutional Neural Networks) have been proposed and applied to target detection in remote sensing images [3], [4]. At present, target detection methods can be classified into two main types: Two-Stage methods and One-Stage methods [5]. The Two-Stage method is a deep convolutional network based on the candidate region. It first generates possible candidate blocks containing the detection target, and then classify and correct the candidate blocks and obtain the detection frame to achieve target detection. The more common algorithms are R-CNN (Region CNN) [6], Fast R-CNN (Fast Region-Based CNN) [7] and Faster R-CNN (Faster Region-Based CNN) [8], etc. These methods have high detection accuracy, but low speed. The One-Stage method is based on the target detection of the deep convolutional network of regression calculation, which uses an end-to-end target detection method, such

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

as SSD (Single Shot MultiBox Detector) [9]–[11], YOLO series [12]–[15] and so on. These methods have a faster detection speed and can meet real-time requirements.

More specifically, scholars have done a lot of research work on target detection in remote sensing images. Reference [16] used the k-means algorithm to cluster the data set, and learned from the Densenet network idea to improve the YOLOv3 network to detect aircraft targets in remote sensing images, which greatly improves the detection accuracy. Reference [17] introduced the spatial pyramid pooling structure, transition module and residual network to improve the YOLOv3 network, and the comprehensive performance indicators for detecting ship targets in remote sensing images were greatly improved. Reference [18] used the PIIFD descriptor to process the transformation between the background and the target of different images, and verified that it had better performance in remote sensing image target detection in the geographic space environment. Reference [19] strengthened the CSP feature extraction network of the YOLOv4 network, replaced the original activation function with the Mish function and added a pyramid pooling module to reduce the scale sensitivity, and improve the detection accuracy and recall rate. Reference [20] used a multi-scale fusion method to solve the problem of small target semantic information transmission in a fully convolutional neural network. In summary, it can be seen that the deep learning methods have high application value in remote sensing image target detection.

Therefore, we tested YOLOv3, YOLOv4, and YOLOv5 on aircraft targets detection in remote sensing images. The experimental results show that the detection accuracy is high, but the detection speed is too low to meet the requirements of real-time detection. For images with complex backgrounds, the complexity of the network structure will increase the difficulty of training and reduce the detection speed. Meanwhile, overfitting is prone to occur when the amount of data is small and the network structure is too simple to effectively describe the feature of the target, which results in a decrease in detection accuracy. In the task of small target detection, traditional convolutional layers usually fail to be both accurate and real-time because it is difficult to extract the characteristics of small targets, no matter for a simple network or a complex one.

This paper presents a network model, YOLOv5-Aircraft, based on improved YOLOv5 to enhance the detection accuracy and detection speed of aircraft targets in remote sensing images. The content of this paper is arranged as follows. Chapter 2 describes the YOLOv5 target detection model. Chapter 3 explains in details how to improve the YOLOv5 model to YOLOv5-Aircraft. Chapter 4 conducts experimental analysis. Finally, Chapter 5 gives the research conclusions.

II. INTRODUCTION OF YOLOv5 DETECTION NETWORK

YOLOv5 proposed by Ultralytics LLC is an improved version based on YOLOv4. It is a one-stage detection network in terms of accuracy and detection speed [21]. After learning from the advantages of the previous version as well as other networks, YOLOv5 changes the characteristics of the previous YOLO target detection algorithm that the detection speed is faster but the accuracy is not high. YOLOv5 has improved detection accuracy and real-time performance, which not only meets the needs of real-time image detection, but also has a smaller structure. Therefore, this article uses YOLOv5 as the detection model. Its network model is divided into 4 parts, namely Input, Backbone, Neck and Prediction, and its network structure is shown in figure 1 [22].

Input includes three parts: mosaic data enhancement, adaptive anchor frame calculation and adaptive image scaling. The input terminal of YOLOv5 adopts the same mosaic data enhancement method as YOLOv4. The random clipping, random scaling and random distribution are used to splice the images. The four images are spliced, which enriches

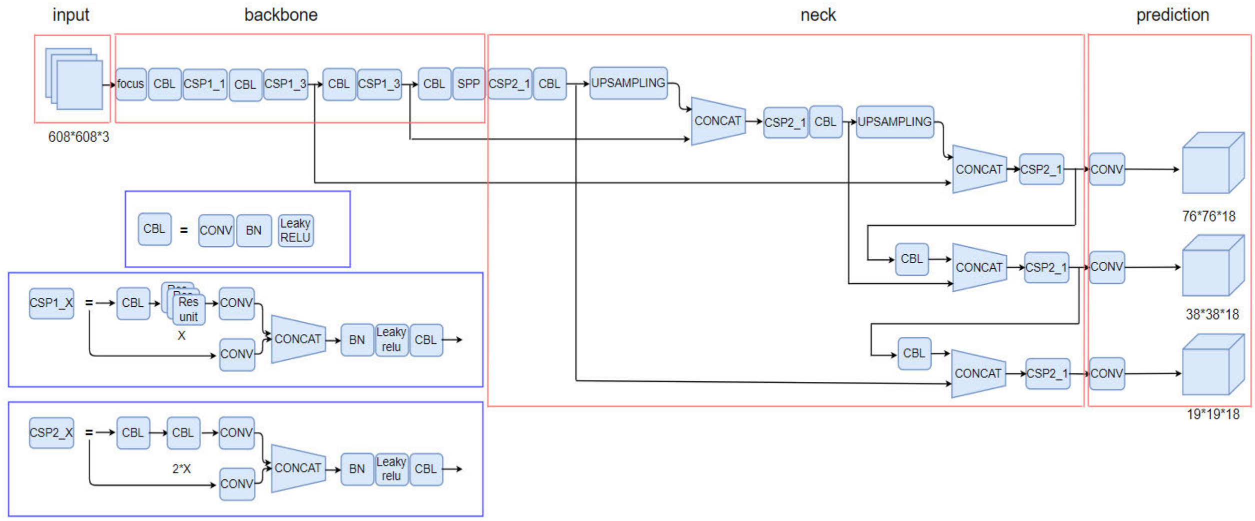


FIGURE 1. The main modules of YOLOv5 network.

the detection data set, improves the robustness of the network, reduces the calculation of GPU, and increases the universal applicability of the network; Adaptive anchor frame calculation sets the initial anchor frame for different data sets, outputs the prediction frame on the basis of the initial anchor frame, and then compares it with the real frame. After calculating the gap, it updates the network parameters reversely and iterates the network parameters continuously. The anchor frame parameters are [116,90,156,198,373,326], [30,61,62,45,59119], [10,13,16,30,33,23]. Adaptive image scaling is to scale the image to a uniform size, which has been implemented in the data preprocessing stage.

Backbone includes focus structure and CSPnet (cross stage partial network) structure. Focus slices the image of 608 × 608 × 3 to get the feature map of 304 × 304 × 12. Then, after convolution of 32 convolution kernels, the feature map of 304 × 304 × 32 is obtained, and the process is shown in figure 2.

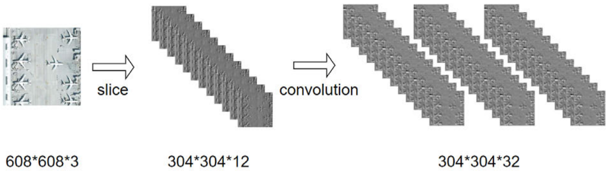


FIGURE 2. **Slicing operation.**

Neck uses FPN (feature pyramid networks) and PAN (pyramid attention network) structure, and its structure is shown in figure 3. FPN transfers and fuses high-level feature information through up sampling from top to bottom to convey strong semantic features. PAN is a bottom-up feature pyramid to convey strong positioning features. Both of them are used at the same time to enhance the ability of network feature fusion.

Prediction includes bounding box loss function and NMS (non-maximum suppression). YOLOv5 uses GIOU loss function as the loss function of bounding box, which effectively solves the problem of non coincidence of bounding boxes, and improves the speed and accuracy of prediction box regression. In the object detection and prediction stage, weighted NMS is used to enhance the ability to recognize multiple objects and occluded objects, and obtain the optimal object detection frame.

III. IMPROVEMENT OF YOLOv5

A. BATCH NORMALIZATION IMPROVEMENTS

Batch normalization (BN) has become the default component of modern neural network stability training. In BN, centering and scaling operations as well as mean and variance statistics are used for feature normalization on batch dimensions. The batch dependence of BN makes the network have stable training and better representation. However, BN inevitably ignores the representation differences between instances. In order to perform feature correction in BN, centering and scaling calibration were added at the beginning and end of the original normalization layer of BN, respectively [23]. Given input feature $X \in R^{N\times C\times H\times W}$, where N, C, H, and W are batch size, the number of channels, height, width of the input feature, respectively, the centering calibration of features is written as follows:

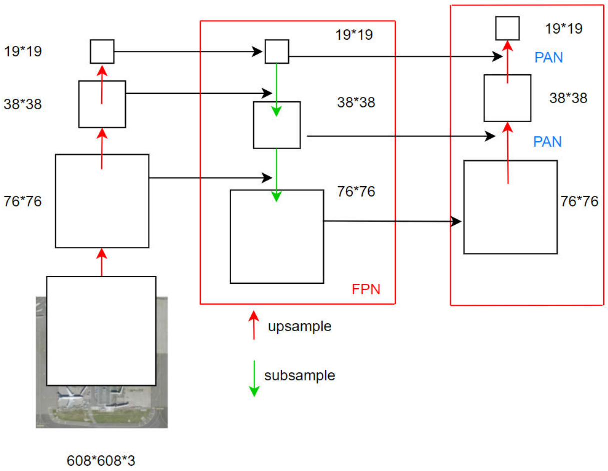


FIGURE 3. **Schematic diagram of neck's structure.**

$$X_{cm} = X + w_m \odot \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} X_{(n,c,h,w)} \tag{1}$$

where $w_m \in R^{1\times C\times 1\times 1}$ is the learnable weight vector, and its value changes with the number of network layers as shown in Figure 4(a). The value in most layers is close to 0 and its absolute value increases as the number of layers increases, because the higher the number of layers, the network has more instance-specific features. X_{cm} is the centering calibration of features. $\odot$ is the dot product operator that broadcast two features to the same shape and then conduct dot product.

Then the centered features with the centering calibration can be written as:

$$X_m = X_{cm} - E(X_{cm}) \tag{2}$$

where $E(X_{cm})$ is the mean of X_{cm}. By scaling X_m like BN, we can deduce the following formula:

$$X_s = \frac{X_m}{\sqrt{Var(X_{cm}) + \varepsilon}} \tag{3}$$

where $Var(X_{cm})$ is the variance of X_{cm}, and ε is used to avoid zero variance. Then, the scaling calibration operation is added to the original scaling operation:

$$X_{cs} = X_S \cdot R(w_v \odot K_s + w_b) \tag{4}$$

where $w_v, w_b \in R^{1\times C\times 1\times 1}$ are learnable weight vectors, as shown in Figure 4(b). Similar to x, their value tends to 0 in most layers and its absolute value increases with the increase of the number of layers. and $R()$ is the restricted function, which can be defined with multiple forms. In this work, we choose to use the Tanh function to suppress extreme values. Similar to K_m, K_s is the statistics of the instance

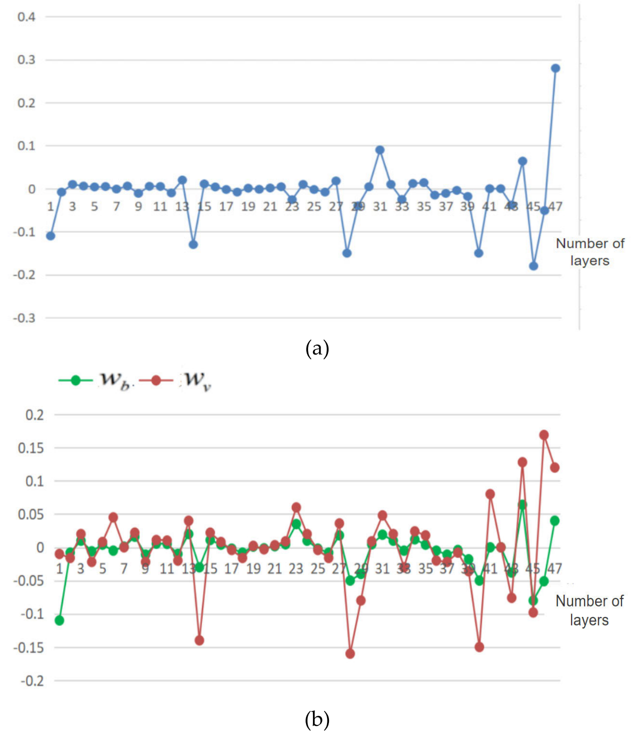


FIGURE 4. Line chart of learnable weight variables.

feature X_S, that can be set to multiple values. The restricted function $R()$ along with the w_v and w_b in Eqn. (4) suppress out-of-distribution features, making the feature distribution more stable.

Finally, the trained learnable scale factor γ and deviation factor β are linearly transformed to obtain the final representative batch normalization result Y. The affine transformation can be written as follows:

$$Y = X_{cs}\gamma + \beta \tag{5}$$

To utilize the optimization of batch normalization in existing deep learning frameworks, we add the centering and scaling calibrations at the beginning and ending of the original normalization layer of batch normalization, respectively, which enhances the effective features and forms a more stable feature distribution, and enhances the feature extraction ability of the network model. We extract the feature map in the network, and the result is shown in the figure 5. It can be seen that the feature map on the far right after the centering and scaling calibration is more significant than the original feature map output in the middle.

B. LOSS FUNCTION BASED ON KULLBACK-LEIBLER DIVERGENCE

The mean square error (MSE) is used in the target frame coordinate regression process of YOLOv5, and the cross entropy is used as the loss function of confidence and category. However, as the loss function of the target frame, the loss of MSE is more sensitive to the target frame. In order to further improve the convergence stability, this paper improves the cross-entropy loss function to the smoothed Kullback-Leibler divergence loss function when designing the loss function for confidence. KL divergence is also called relative entropy [24]. For two probability distributions P and Q of the same continuous variable, the definition of KL divergence is:

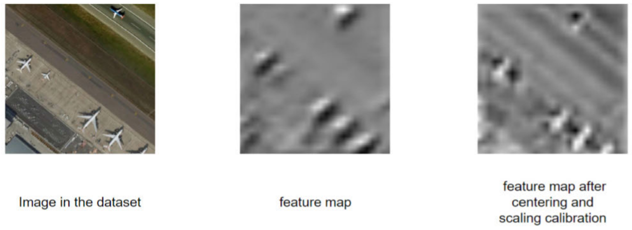


FIGURE 5. Centering and scaling calibration renderings.

$$D_{KL}(Q \parallel P) = \int Q(x)(\ln \frac{Q(x)}{P(x)})dx \tag{6}$$

In this paper, $\hat{\phi}$ is used to represent the parameter change process of minimizing the KL divergence between the predicted probability distribution and the real label distribution of n input samples. The formula is as follows:

$$\hat{\phi} = \arg\min_{\phi} \frac{1}{n} \sum D_{KL}(Q_D(x) \parallel P_\phi(x)) \tag{7}$$

where $Q_D(x)$ is the probability distribution of real label coordinates, $P_\phi(x)$ is the probability distribution of predicted coordinates, $Q_D(x)$ and $P_\phi(x)$ are defined as Gaussian distribution functions. Therefore, the boundary box regression loss function L_{KL} can be written as:

$$\begin{aligned} L_{KL} &= D_{KL}(Q_D(x) \parallel P_\phi(x)) \\ &= \int Q_D(x)(\ln \frac{Q_D(x)}{P_\phi(x)})dx \\ &= \int Q_D(x)\ln Q_D(x)dx - \int Q_D(x)\ln P_\phi(x)dx \end{aligned} \tag{8}$$

Then, according to the properties of Gaussian distribution function, the derivation is written as follows:

$$\begin{aligned} L_{KL} &= \int Q_D(x)\ln Q_D(x)dx \\ &\quad - \int Q_D(x)(-\ln(\sqrt{2\pi}\sigma) + \ln e^{\frac{-(x-x_e)^2}{2\sigma^2}})dx \\ &= \int Q_D(x)\ln Q_D(x)dx \\ &\quad + \int Q_D(x)(\frac{1}{2}\ln(2\pi\sigma^2) + \frac{(x-x_e)^2}{2\sigma^2})dx \end{aligned} \tag{9}$$

The definition of Dirac delta function is shown in (10). Since Gaussian distribution is the approximation of Dirac delta function when the standard deviation is close to 0, equation (9) is deduced according to the screening property of Dirac delta function in equation (11), and finally equation (12) is obtained. The equations are as follows:

$$\int \delta(x)dx = 1 \tag{10}$$

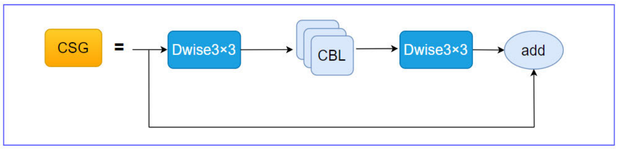


FIGURE 6. Structure diagram of CSandglass.

$$\int \delta(x - x_g)y(x)dx = y(x_g) \tag{11}$$

$$\begin{aligned} L_{KL} &= \int Q_D(x)\ln Q_D(x)dx \\ &\quad - \int Q_D(x)(-\ln(\sqrt{2\pi}\sigma) \\ &\quad + \ln e^{\frac{-(x-x_e)^2}{2\sigma^2}})dx \\ &= \int Q_D(x)\ln Q_D(x)dx \\ &\quad + \int Q_D(x)(\frac{1}{2}\ln(2\pi\sigma^2) \\ &\quad + \frac{(x-x_e)^2}{2\sigma^2})dx \end{aligned} \tag{12}$$

where x_e is a constant term. Since the constant term has no effect on the derivation, the term without parameters can be removed as follows:

$$L_{KL} \propto (\frac{\ln(\sigma^2)}{2} + \frac{(x_g - x_e)^2}{2\sigma^2}) \tag{13}$$

where x_g is a constant term. If the initial value of σ in equation (13) is large, it is easy to cause gradient explosion in the initial stage of training, which leads to the failure of convergence of the model. And The input of x in function $\ln x$ is limited in mathematical calculation. Therefore, in the prediction stage of model training, this paper sets the relationship between variables α and σ as shown in equation (14), and then brings equation (14) into equation (13) to obtain the loss function equation (15):

$$\alpha^2 = \ln \sigma^2 \tag{14}$$

$$L_{KL} \propto (\frac{\alpha^2}{2} + \frac{e^{-\alpha^2}(x_g - x_e)^2}{2}) \tag{15}$$

In order to further enhance the robustness of the model, the KL divergence loss function is smoothed. When $|x_g - x_e| > 1$, the regression loss function of the model's bounding box is:

$$L_{KL} = \mathrm{e}^{-\alpha^2}|x_g - x_e| + \frac{\alpha^2}{2} \tag{16}$$

In the process of model training, the smoothed loss function will not produce a sudden change to the noisy sample data, so as to reduce the interference in the process of back propagation, and the convergence of model is more stable.

C. IMPROVEMENT OF NETWORK STRUCTURE

The CSPNET structure in YOLOv5 divides the feature layer of the base layer into two parts and then uses a cross-stage hierarchical structure to merge the two, so that the network can achieve richer gradient combination information, but this is also more likely to cause information loss and gradient confusion. Therefore, this paper draws on the ideas of mobileneXt [25] and uses the hourglass-like module CSandGlass to replace the Res unit module in the YOLOv5 network. The structure of the CSandGlass module is shown in figure 6. Unlike the bottleneck structure with depthwise convolution in the middle, this paper moves the 3 × 3 depthwise convolution layer (Dwise) to both ends of the residual path with high-dimensional representation, and the two basic components of YOLOv5, CBL, are placed in the middle. Two depthwise convolutions can encode more spatial information, and make more gradients propagate across multiple layers, reducing information loss. Figure 7 shows the before and after comparison using the CSandGlass module. The two pictures on the left show the results of using 6 consecutive convolutions without using the CSandGlass module. It can be seen that the edge features of the aircraft are not well extracted, and the information is severely lost. The two pictures on the right are the improved feature extraction results using CSandGlass. The edge feature information of the building has been better extracted, and the background information and feature information are also more distinct.

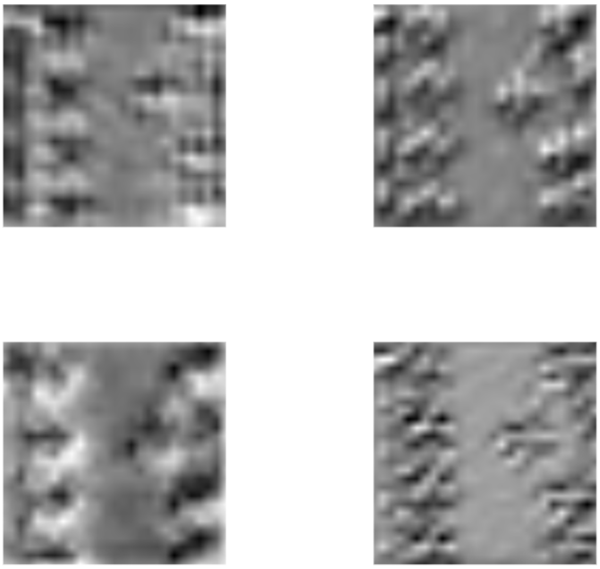


FIGURE 7. Comparison before and after improvement.

In the input of the original version of YOLOv5, the feature number of the fully connected layer behind the convolution layer is fixed, so that the size of our input image will be fixed at 608 × 608, and the sizes of the feature layer network are 19 × 19, 38 × 38, 76 × 76, respectively. The smaller the size of feature layer is, the larger the receptive field of neurons is, which means that the semantic level is richer, but the local and detail features will be lost. On the contrary, when the convolutional neural network is shallow, the receptive field becomes smaller, and the neurons in the feature map tend to be partial and detailed [20]. In order to reduce semantic loss,

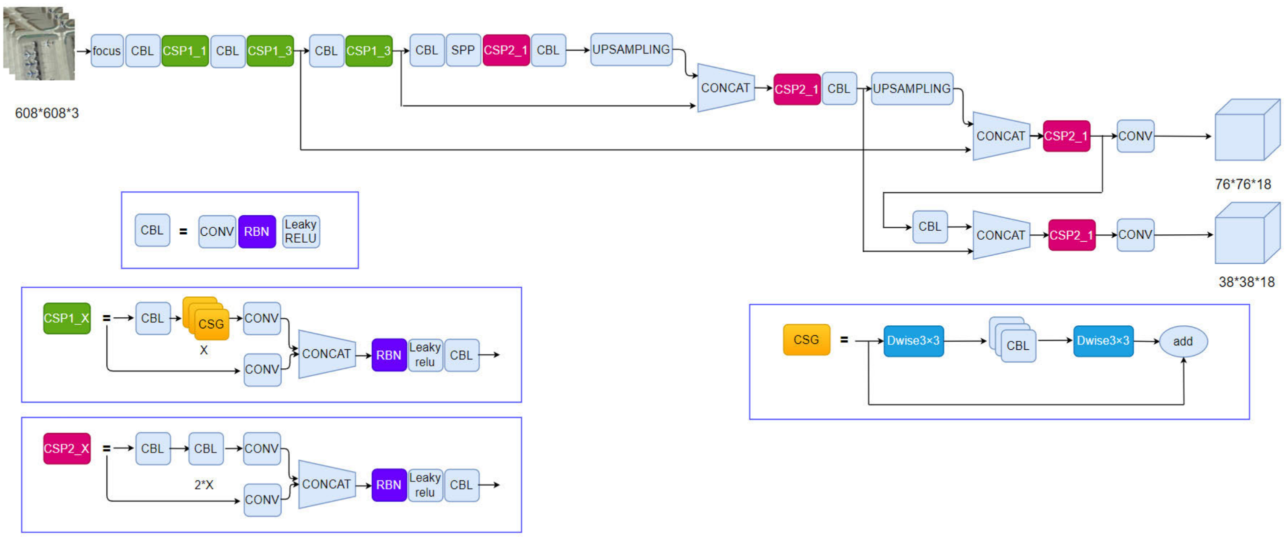

608*608*3
focus
CBL
CSP1_1
CBL
CSP1_3
CBL
CSP1_3
CBL
SPP
CSP2_1
CBL
UPSAMPLING
CONCAT
CSP2_1
CBL
UPSAMPLING
CONCAT
CSP2_1
CONV
76*76*18
CBL
CONCAT
CSP2_1
CONV
38*38*18
CBL = CONV RBN Leaky RELU
CSP1_X = CBL CSG X CONV CONV CONCAT RBN Leaky relu CBL
CSP2_X = CBL CBL 2*X CONV CONV CONCAT RBN Leaky relu CBL
CSG = Dwise3×3 CBL Dwise3×3 add

FIGURE 8. Network structure diagram of improved YOLOv5.

the 19 × 19 feature layer in the main feature extraction network is removed and the other two feature layers are retained. This not only reduces the semantic loss, but also reduces the amount of network parameters. Figure 8 shows the improved network structure of YOLOv5, where CSG is the CSandGlass module, and RBN is the improved BN module.

IV. EXPERIMENTS

A. EXPERIMENTAL ENVIRONMENT

In this paper, the deep learning platform is built in OpenCV. Test environment: NVIDIA Tesla V100, 16G GPU memory, CUDA version 10.1, cudnn version 7.6.5, and python 3.8 as the compiler language.

B. EVALUATING INDICATOR

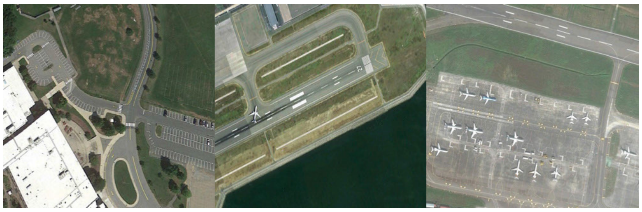
In the field of object detection, recall, precision and mAP (mean Average Precision) are usually used to evaluate the performance of object detection algorithm. Recall rate is used to describe how many samples are detected in prediction [26]. The calculation formula is as follows:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{17}$$

where R is the recall rate, TP is the number of positive samples predicted by the algorithm as positive examples, FN is the number of positive samples predicted as negative examples, that is, the number of missed detection. Precious rate is used to describe the ratio of the predicted positive examples to all the positive examples:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{18}$$

where P is the precious rate, FP is the number of individuals who predict negative examples in the sample as positive examples, that is, the object of detection errors. However, in general, it is difficult to maintain both the recall rate and the precious rate at a high level. Therefore, a parameter is needed to integrate these two parameters. The mAP is used to measure the algorithm performance of the detection network. It is suitable for single-label and multi-label image classification and calculation. The equation can be written as:



(a) (b) (c)

FIGURE 9. Main types of aircraft targets in the data set. (a) no target; (b) single target; (c) multiple targets.

$$\mathrm{m}AP = \frac{\sum_{k=1}^{N} P(k)\Delta R(k)}{C} \tag{19}$$

where N is the number of samples in the test set, $P(k)$ is the size of the precious rate when k samples are recognized at the same time, $\Delta R(k)$ is the change in the recall rate when the number of detected samples changes from $k-1$ to k, C is the number of categories in the multi-class detection task.

V. DATA PREPROCESSING

A. THE SOURCE OF DATA SET

The remote sensing images studied in this paper are from Google Earth, with 78 images in total [27]–[29]. These images are remote sensing images containing aircraft targets, including non-target images, single-target images and multi-target images. Figure 9 shows several typical remote sensing images in the data set.

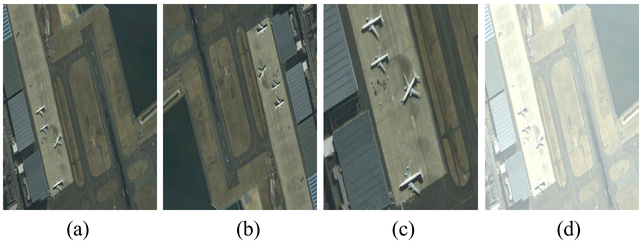


FIGURE 10. Schematic diagram of preprocessing picture. (a)Original image; (b) Original image after rotation; (c) Enlarged original image; (d) Brightened original image.

B. CONSTRUCTION OF DATA SET

The original picture size is relatively large. If the original size is used as the training data set, it will cause too many parameters. Therefore, the original aerial picture size is reduced to 608 pixels × 608 pixels by pixel transformation. And on the basis of the original image, the image is rotated, cropped, and contrasted, so that the remote sensing image has different manifestations and scales, which helps to avoid the occurrence of overfitting, thereby improving the generalization ability of the training network [14]. Figure 10 shows pictures in different forms after preprocessing.

After preprocessing, 1000 remote sensing pictures are finally obtained. Imitating the format of the VOC2007 data set, this paper uses LabelImg to mark the outer frame of the aircraft targets in these images in turn, and converts them into the XML format required for training [30]. Labelimg is an image annotation tool in deep learning, which is used to annotate the category name and location information of objects in the image.

C. RESULTS AND ANALYSIS

In order to compare YOLOv5 and the improved model proposed in this paper, the processed images will be trained with the same number of epochs. At the same time, use the YOLOv4 network to make multiple comparisons under the same conditions. For convenience of comparison, the improved model is called YOLOv5-Aircraft. Firstly, compare the loss reduction between the models. Figure 11 shows the loss graphs of the three models. The abscissa is the number of epochs, and the ordinate is the loss. The blue line, orange line and green line respectively represent three different models. It can be seen that the loss of YOLOv5-Aircraft decreases faster than that of YOLOv5 and YOLOv4, indicating that the loss of YOLOv5-Aircraft converges faster. After convergence, the loss of YOLOv5-Aircraft is closer to 0 and smoother.

While YOLOv4, YOLOv5 and YOLOv5-Aircraft are comparatively analyzed, this paper combines the test results of models such as YOLOv3 and Faster RCNN for multivariate analysis, as shown in TABLE 1.

In TABLE 1, FPS (Frame Per Second) is the detection speed, which is the number of images that the algorithm can detect per second. Analyzing the data in TABLE 1, it can be seen that the detection accuracy index of YOLOv5-Aircraft is improved compared with the original YOLOv5 network, mAP is increased by 3.74%, and the detection speed is also greatly improved by 6.93. From the comparative data, it can be seen that YOLOv5-Aircraft has improved its ability to accurately predict the location of aircraft, and its detection speed has also been greatly improved. Faster RCNN adopts two-stage detection mechanism and fine-tuned the anchor area twice, but its mAP only exceeds YOLOv3 compared to other algorithms and the detection speed is much lower than the latter.

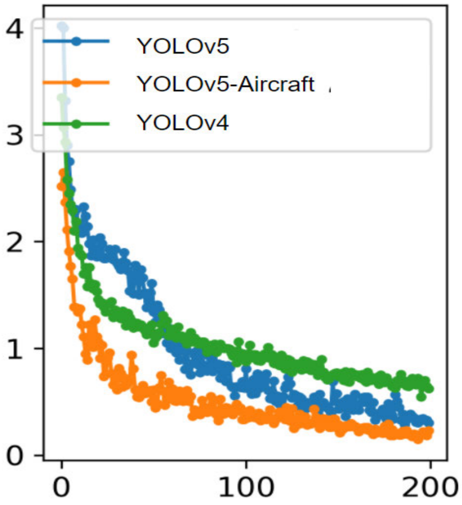


FIGURE 11. Loss curve comparison chart.

TABLE 1. Comparisons of different models.

Model	mAP	Recall	FPS
YOLOv3	73.52	78.13	32.46
YOLOv4	76.34	81.62	35.21
YOLOv5	81.51	85.12	41.92
YOLOv5-Aircraft	85.25	90.24	48.85
Faster RCNN	74.86	78.63	8.93
PP-YOLO	82.63	88.26	39.41
SSD	79.34	84.73	40.52

The improved YOLOv5 model proposed in this paper is further subjected to ablation experiments to verify its effectiveness. Some network modules are replaced, and the results are shown in TABLE 2, where CUT is the operation of removing low-resolution feature layers. From the data in the table, it can be seen that The use of RBN module, CSG module and loss function based on KL divergence all improve the accuracy and speed of detection. Removal of low-resolution feature layers improves the detection speed, but reduces the

TABLE 2. Test results of ablation.

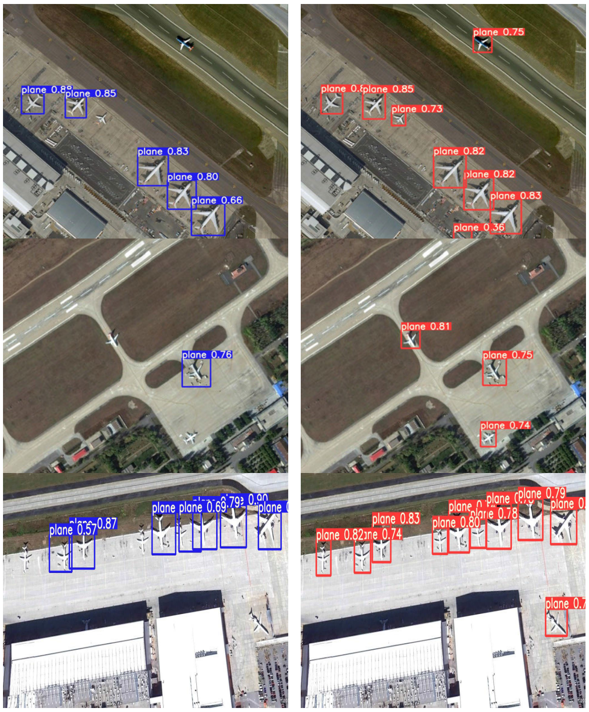
Method	mAP	Recall	FPS
YOLOv5	81.51	85.12	41.92
YOLOv5+RBN	82.72	86.87	44.21
YOLOv5+KL	83.06	87.54	41.95
YOLOv5+CSG	82.25	86.14	43.18
YOLOv5+CUT	78.98	83.42	46.76

detection accuracy. Because the RBN module strengthens the feature extraction capabilities of the network, its detection accuracy and speed have been greatly improved by 1.21% and 2.29 respectively.. The loss function based on KL divergence reduces the noise interference and improves the detection accuracy to a certain extent, but it has no obvious effect on the detection speed. The CSG module reduces the information loss in the gradient descent process, and also improves the accuracy and speed of detection to a certain extent. The network without the low-resolution feature layer has one feature layer less than the network with the low-resolution feature layer and reduces operations such as convolution and splicing, which significantly improves the network detection speed, but also reduces the detection accuracy. The shortcoming after removing the low-resolution feature layer also shows the necessity of RBN module, CSG module and the loss function based on KL divergence to improve the accuracy of model detection.

In addition, YOLOv5-Aircraft performs well when the background of the detection image is complex, and the detection results are shown in Figure 12. From the comparison of the pictures in the first two rows, we can see that YOLOv5 has missed detection of some targets. The improved YOLOv5 improves the detection accuracy of small aircraft targets, and there is no missed detection. Moreover, as shown in the third row of the picture, when the brightness of the picture is increased due to sunlight, the original algorithm missed the detection more serious, and the improved YOLOv5 can well identify all aircraft targets in the image, indicating the improved algorithm still has good recognition ability in abnormal lighting conditions.

VI. CONCLUSION

We proposed a convolutional neural network-based aircraft detection algorithm, YOLOv5-Aircraft, to detect and track aircraft targets in remote sensing images under complex backgrounds. Various types of aircraft targets in remote sensing images were evaluated and researched. Firstly, we utilised batch normalization module added centering and scaling calibration to strengthen the feature extraction ability of network model. Then the cross-entropy loss function in the confidence of the original loss function was improved to the loss function based on smoothed Kullback-Leibler divergence. Finally, the CSandGlass module was designed to replace the residual module for reducing information loss and the low-resolution feature layer of the backbone network was removed for reducing the loss of local details, which ultimately led to an efficient and accuracy aircraft targets detector, applicable in various complex environment.



(a) (b)

FIGURE 12. Comparison of detection results between YOLOv5 and improved YOLOv5. The detection results of YOLOv5 are shown on the left and the results of our method are shown on the right.

The proposed method was evaluated for remote sensing image data set from Google Earth and Vaihingen data set, including 2000 frames, and approximately 13,000 aircraft targets. YOLOv5-Aircraft was demonstrated to be able to perform in a variety of challenges including shades, lighting variations and partial visibility, and showed a major development in terms of accuracy(85.25%) and speed(48.85fps). Therefore, YOLOv5-Aircraft offered a robust aircraft target recognition algorithm in remote sensing images. However, through a large number of test experiments, it is found that factors such as light and weather still have a certain influence on the detection results. In the subsequent training process, it is necessary to collect more image data in a complex environment to improve the generalization ability of the YOLOv5-Aircraft model.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[2] H. Zhu, L. Qin, and B. Sun, "Review on parallelization of deep neural networks," *J. Chin. J. Computer.*, vol. 41, no. 8, pp. 171–191, Aug. 2018, doi: 10.11897/SP.J.1016.2018.01861.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classfication with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[5] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[10] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[11] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[16] Z. Q. Zheng, Y. Y. Liu, C. C. Pang, and G. N. Li, "Application of improved YOLOv3 in aircraft recognition of remote sensing images," *J. Electron. Opties Control.*, vol. 26, no. 4, pp. 28–32, Apr. 2019.

[17] M. Gong, Y. Y. Liu, and G. N. Li, "A ship detection method for remote sensing images based on improved YOLO-v3," *J. Electron. Opties Control.*, vol. 27, no. 5, pp. 102–107, May 2020.

[18] F. Ghorbani, H. Ebadi, and A. Sedaghat, "Geospatial target detection from high-resolution remote-sensing images based on PIIFD descriptor and salient regions," *J. Indian Soc. Remote Sens.*, vol. 47, no. 5, pp. 879–891, Jan. 2019.

[19] C. Cao, J. Wu, X. Zeng, Z. Feng, T. Wang, X. Yan, Z. Wu, Q. Wu, and Z. Huang, "Research on airplane and ship detection of aerial remote sensing images based on convolutional neural network," *Sensors*, vol. 20, no. 17, p. 4696, Aug. 2020.

[20] Y. Xu, M. Zhu, P. Xin, S. Li, M. Qi, and S. Ma, "Rapid airplane detection in remote sensing images based on multilayer feature fusion in fully convolutional neural networks," *Sensors*, vol. 18, no. 7, p. 2335, Jul. 2018.

[21] D. Wu, S. Lv, M. Jiang, and H. Song, "Using channel pruning-based YOLOv4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," *J. Comput. Electron. Agricult.*, vol. 178, no. 5, pp. 174–178, Nov. 2020, doi: 10.1016/J.COMPAG.2020.105742.

[22] S. Tan, X. Bie, G. Lu, and X. Tan, "Real-time detection for mask-wearing of personnel based on YOLOv5 network model," *J. Laser J.*, vol. 42, no. 2, pp. 147–150, Feb. 2021, doi: 10.14016/j.cnki.jgzz.2021.02.147.

[23] S.-H. Gao, Q. Han, D. Li, M.-M. Cheng, and P. Peng, "Representative batch normalization with feature calibration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021.

[24] Z. Zhao, Y. Li, Z. Zhen, Y. Zhai, K. Zhang, and W. Zhao, "Typical fittings detection method w-ith Faster R-CNN combining KL divergence and shape constraints," *J. High Voltage Eng.*, vol. 46, no. 9, pp. 3018–3026, 2020.

[25] Z. Daquan, Q. Hou, Y. Chen, J. Feng, and S. Yan, "Rethinking bottleneck structure for efficient mobile network design," 2020, *arXiv:2007.02269*.

[26] T. Hou and Y. Jiang, "Application research of improved YOLOv4 in remote sensing aircraft target detection," *J. Comput. Eng. Appl.*, 2021. [Online]. Available: https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2021&filename=JSGG202112029&uniplatform=NZKPT&v=3Qinehu7XHL1npT2x0OGQn3DtwRtWbBHcrZQ13NonaoH%25mmd2BTLh5wG6WhR2b1A9HnvU

[27] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[28] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

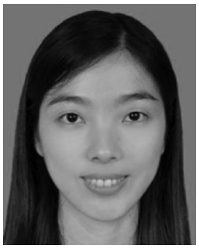
[29] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[30] Y. Yang and Y. Cao, "Detection of glass insulators in images taken by drones based on improved YOLOv3," *J. Comput. Eng. Appl.*, 2020. [Online]. Available: https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JSGG20201222009&v=UVJbamaWiqPDceB2ZkAPiN2ZMKMxsSRme8i4Hclh1iQfZbZ1R8L3gdhLJSzGvDMs and https://kns.cnki.net/kcms/detail/11.2127.TP.20201224.1102.010.html

SHUN LUO received the B.S. degree from Fuzhou University, in 2019, where he is currently pursuing the M.S. degree in information management and information system with the College of Economics and Management.

JUAN YU received the Ph.D. degree in management science and engineering from the Dalian University of Technology, Dalian, China, in 2010. She is currently an Associate Professor with the School of Economics and Management, Fuzhou University, Fuzhou, China. Her research interests include text-mining, ontology, and knowledge graphs.

YUNJIANG XI received the Ph.D. degree in management science and engineering from the Dalian University of Technology, Dalian, China, in 2007. He is currently an Associate Professor with the School of Business Administration, South China University of Technology, Guangzhou, China. His research interests include knowledge management, microblog data mining, and social media networks.



XIAO LIAO received the Ph.D. degree in management science and engineering from the Dalian University of Technology, Dalian, China, in 2015. She is currently a Lecturer with the School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou, China. Her research interests include knowledge supernetwork, knowledge management, and user innovation.

• • •