Date of publication December 15, 2022.

Research on Aircraft Image Recognition Based on Transfer Learning and Improved YOLOv5 Model

Huanyu Yang[1], Jun Wang[1], Lijun Yang[1], and Yuming Bo[1]

[1] School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China

Corresponding author: Jun Wang (e-mail: wangj1125@163.com).

ABSTRACT Effective differentiation of aircraft types using images is important for providing military combat information as well as civilian aircraft operations. Aircraft image recognition has many difficulties such as large variations of target scale, complex backgrounds, and difficult data set acquisition, which lead to the low recognition accuracy of existing models. To address the problem of low recognition accuracy caused by the above difficulties. This paper proposes the improved YOLOv5 model for the recognition of aircraft images. First, this paper designs the CSPResNet50dCA network as the backbone of the YOLOv5 model to enhance the feature extraction capability for small target aircraft in images. By introducing the coordinate attention mechanism and the CSP structure, the feature-focusing capability and the computing speed of the model are enhanced. Afterward, we use data enhancement to expand the data set and transfer learning to improve the generalization ability and convergence speed of the model, so as to improve its robustness. The experimental results show that the improved YOLOv5 model has significantly improved the recognition accuracy of aircraft targets， and significantly enhanced feature extraction ability for small target aircraft with good generalization ability.

INDEX TERMS aircraft recognition, yolov5, attention mechanism, target detection, CSPResNet50dCA

I. INTRODUCTION

In terms of the civil field, recognition and classification of aircraft targets in images can help airlines supervise and dispatch airport flights and find lost aircraft targets. In terms of the military field, the current stage of warfare has changed from the past mechanized warfare to information-based warfare, and aircraft targets, as fast and flexible airspace combat forces, have important military functions such as reconnaissance, transportation, and combat, and their dynamic can provide important military information. Therefore, the recognition of aircraft targets is not only beneficial to the development of civil aerospace enterprises but also has vital significance to the situation estimation of the military battlefield and the making of military decisions. Due to the interference of weather (fog, dust), noise, light intensity (exposure), pollution, and other factors, the internal structure and texture information of aircraft images will be affected, thus affecting the accuracy and efficiency of target detection, bringing great difficulty to the image target detection [1]. Among image detection with complex backgrounds, high target density, and different aircraft types, the detection accuracy is low, and it is easy to produce missed and false detection. Due to the disadvantages of extensive model parameters and large computation, it is difficult to meet the requirements of real-time detection when a trained model is used to solve the actual problem. Therefore, researchers hope that the accuracy and speed can be further improved when target detection is performed on aircraft images.

Currently, there are two mainstream target detection methods, one-stage and two-stage detection methods [2]. The typical algorithm of the two-stage method is Faster R-CNN. Sha [3] et al. proposed a remote sensing image aircraft target detection method based on improved Faster R-CNN, which improves the localization accuracy of multi-scale aircraft targets in remote sensing images by modifying the scale of candidate regions in RPN with the help of multi-level fusion structure and multi-scale RPN (Region Proposal Network) mechanism; Zhu [4] et al. proposed an improved ROI-pooling scheme based on bilinear interpolation for aircraft target detection based on the Faster R-CNN algorithm, which solved the region mismatch problem caused by twice quantization. The two-stage method has some advantages in accuracy performance, but it cannot achieve directional detection for aircraft targets with variable directions and is too slow in detection speed to meet the demand of real-time detection. There are also many scholars devoted to the study of one-stage detection algorithms. Wang [5] et al. proposed an algorithm for aircraft remote sensing image target detection based on SSD, using a modified deep residual network to replace the skeleton network of SSD, and designing a new feature pyramid network containing a feature perceptual field enhancement module and attention mechanism module, which makes both deep and shallow networks get structured level-

rich fusion features. The SSD algorithm and its improved improvement algorithm have achieved good results, but with the proposed YOLO series algorithm, many scholars gradually shift their research focus to the YOLO series algorithm, because this method is very fast in detection and can meet the real-time requirements. Shi [6] et al. proposed to apply the YOLOv4 algorithm to the detection of aircraft targets in remote sensing images and achieved good results. Zhang [7] et al. proposed a remote sensing aircraft target detection model based on improved YOLOv4, using the K-means++ algorithm to optimize the anchoring frame for the target samples, and fusing convolutional kernel pruning and interlayer pruning to sparsely train the convolutional kernel and batch normalized BN layer to simplify the network structure and reduce the number of parameters. Based on the results of various studies, the YOLOv5 algorithm in the YOLO series has better comprehensive detection ability than other YOLO models due to its accuracy and detection accuracy [8-10].

To address the problems of complex background of aircraft images, difficult extraction of aircraft semantic information and image features due to the presence of occlusions, and inaccurate localization of small target aircraft, this study proposes an improved YOLOv5 model for the detection of aircraft images. First, a CSPResNet50d network is designed to extract the aircraft features in the image; then a CA module with a coordinate attention mechanism is added to the network, which is used to improve the network's ability to focus on small target aircraft, ignore useless information and focus on information useful for the detection task, thus reducing the missed detection rate. Then the designed CSPResNet50dCA network is used to replace the backbone of the original YOLOv5 model. Finally, data enhancement methods and transfer learning strategies are used to expand the data set, improve the convergence speed and generalization ability of the model, and train the best YOLOv5-CSPResNet50dCA model. Through these improvements, the final model proposed in this paper significantly improves the accuracy and computing speed of aircraft detection.

II. MATERIALS AND METHODS

A. DATA ACQUISITION

There are two datasets applied in this paper, Dataset 1 is a classification dataset of aircraft images, which is to test the extraction ability of the CSPResNet50CA network proposed in this paper for aircraft features. Dataset 2 is the target detection dataset of aircraft images, which is used to verify the performance of the Yolov5-CSPResNet50dCA model proposed in this paper.

1) Dataset 1

The images and tags used in Dataset 1 are mainly collected from the FGVC-Aircraft dataset, while the aircraft images of some categories are crawled from the web by the Python crawler according to the aircraft model tags, and the aircraft data with corresponding tags are added. The whole dataset contains 10,000 aircraft images, including 70 classes of "series" (such as A330, Boeing 737, C-130, etc.). The number of differently labeled data was appropriately balanced, and 6667 images were randomly selected to form the training set, while the remaining 3333 images were used as the test set. Some of the image data are shown in Fig. 1.

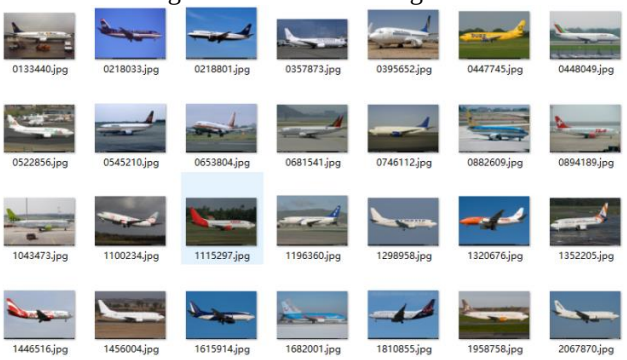


Fig. 1. Schematic diagram of some data sets

2) Dataset 2

The main body used for Dataset 2 is the Military Aircraft dataset. This dataset is used by the U.S. side for military aircraft recognition and contains 34 types of aircraft such as A10, B1, B2, B52 Be200, C130, C17, C5, E2, and EF2000. However, this dataset has insufficient data for some types, for example, it is difficult to obtain pictures of XB70-type bombers on the real battlefield due to their scarce number and few military operations conducted. To solve this problem, this paper obtained images of various types of aircraft using Python crawlers and populated the data for the types of aircraft with insufficient data to make the overall data set evenly distributed. The final composition of the dataset used in this task contains 5003 training images and 1235 test images.

B. DATA LABELING

LabelImg software was used for the labeling process, and the labeled files were all in XML format. When applied to the YOLO algorithm, the labeled files only need to perform format conversion. Part of the labeling is shown in Fig. 2.

Fig. 2. Schematic diagram of image annotation

C. DATA ENHANCEMENT

A large dataset is required to train the deep learning network. To improve the confidence of this experimental model, data augmentation of the training set is required to improve the learning effect and generalization performance of the network. This paper uses HSV, rotation, displacement, scaling,

cropping, flipping, Mix-up, and Mosaic to extend the dataset, each using random probability to determine whether the images need to be extended. The Mix-up is a simple linear transformation of the input image data that allows the images to mix between different categories. Mosaic data enhancement can improve the overall quality of the data set. Four images can be cropped at random, and the length, width, and position of the cut can be changed at random. Then they are stitched into one image as training data, and the target frame is adjusted accordingly, it enriches the background of the detected objects and facilitates the detection of small targets. in the case of extracting 4 images, each image is reduced to a different degree and the original target size is closer to the size of the small target. The effect of Mosaic enhancement is shown in Fig. 3.

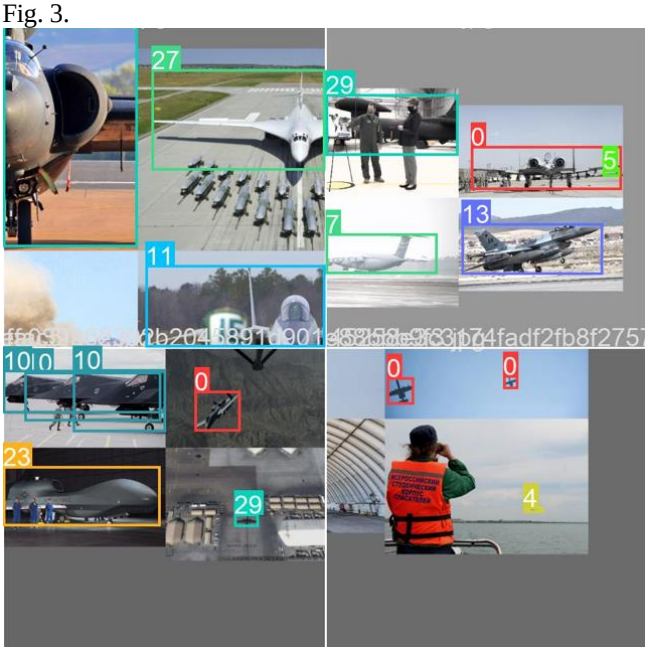


Fig. 3. Mosaic data enhancement

By using image enhancement to make up for the insufficient training samples and solve the problem of insignificant differences in image features, the training speed, generalization ability and robustness of the model are thus improved.

D. METHODOLOGY

1) YOLOv5 model

YOLOv5[11] is a very popular target detection framework, and its overall structure is shown in Figure 1. The network structure of YOLOv5 is relatively simple, and it can be roughly divided into three parts: the Backbone network for feature extraction, the Neck network for feature fusion, and the Head network for target class and location regression detection.

The input side part of YOLOv5 adaptively scales the image by automatically setting the size of the initial anchor frame and performs batch normalization of the input image size. Also, image data can be pre-processed. K-means [12] clustering of anchor frame sizes of labeled samples are used to determine the most appropriate anchor frame size at each training.

The backbone network of YOLOv5 is composed of Conv modules, CBS modules, and an SPP structure. A CBS module mainly performs convolutional operations to extract feature information from the images. The spatial pyramid pooling layer (SPP) module is introduced in the backbone network to solve the problem of the non-uniform size of input images.

The neck of YOLOv5 is mainly composed of a bottom-up Feature Pyramid Network (FPN) and a top-down Path Aggregation Network (PAN) structure. The multi-scale feature fusion of aircraft images by FPN and PAN enables the feature map to contain semantic and feature information of aircraft, ensuring accurate recognition of aircraft targets of different sizes.

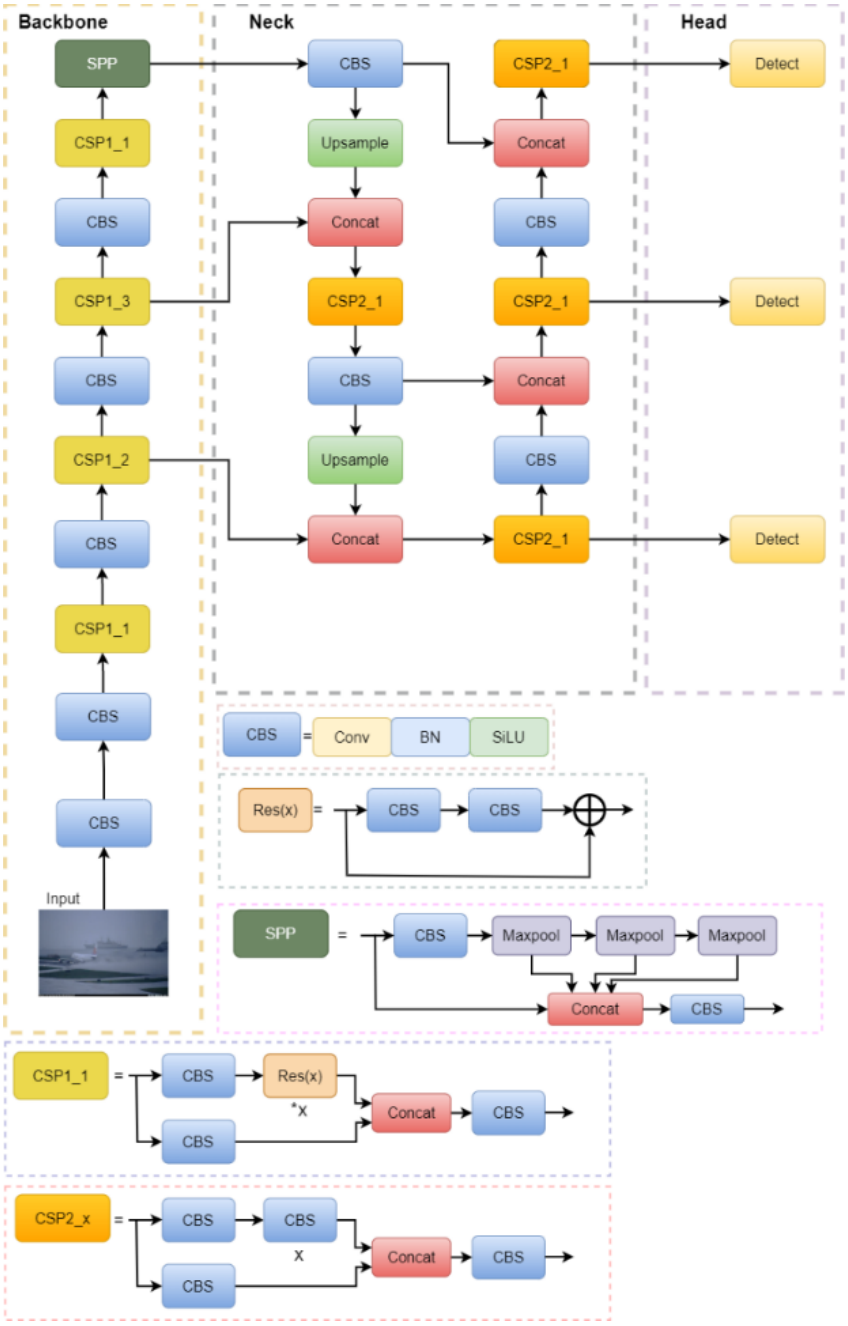


Fig. 4. Structure diagram of the original yolov5 model

2) CA attention mechanism

It has been shown that the channel attention mechanism can significantly improve the performance of image recognition networks, but the use of the channel attention mechanism can easily lead to the problem of ignoring the spatial location information in the high-level feature maps. The popular attention mechanisms include SE (Squeeze and Excitation)[13], and CBAM (Convolutional Block Attention Module)[14]. Among them, SE only considers remeasuring the importance of each channel by modeling channel relationships, while ignoring the location information and spatial structure, which are essential for generating spatially selective attention maps. CBAM encodes global spatial information by global pooling on channels, which compresses global spatial information into a single channel descriptor, and thus makes it difficult to preserve the spatial location information of objects in channels. It is therefore difficult to preserve the spatial location information of objects in the channel.

The CA module, on the other hand, considers not only the relationship between channels but also the location information in the feature space. Its essence is to encode channel relationships and long-term dependencies by precise

location information. CA decomposes attention into X-direction and Y-direction and uses one-dimensional feature encoding to obtain long-range point-space location relationships while obtaining more precise location information. Then the direction-sensitive and location-sensitive feature maps are formed by feature encoding, which enhances the representation of the target of interest by features with location information. As shown in Fig. 5, the specific operation can be divided into two steps: coordinate information embedding and coordinate attention generation.

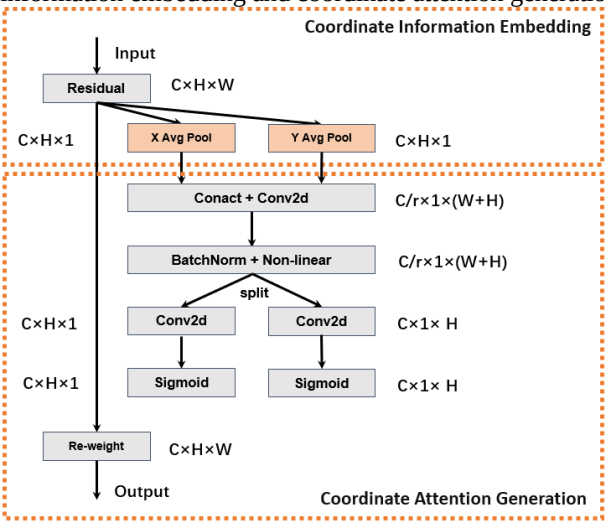


Fig. 5. Diagram of coordinate attention module

(1) Coordinate information embedding

The encoding of the spatial location of aircraft images by channel attention is usually global pooling, and low-level features with rich spatial location information are pooled to obtain high-level semantic features, but high-level features are difficult to retain global spatial location information. Thus two one-dimensional feature encodings are used to decompose the global pooling to enable greater interaction between the locations of distant points as much as possible.

If the height of the feature map is H, then its c-th channel in the vertical direction after pooling is characterized as in (1).

$$z_c^h(h) = \frac{1}{W}\sum_{0<i<W} x_c(h,i) \tag{1}$$

Similarly, if the width of the feature map is W, the output of its c-th channel in the horizontal direction can be written as (2).

$$z_c^\omega(\omega) = \frac{1}{H}\sum_{0<j<H} x_c(j,\omega) \tag{2}$$

The above 2 pooling methods operate in different directions of the same dimensional features, and their resulting aggregated features have some perceptibility of all values in both directions of the feature map. These two transformations ensure that the attention module captures the long-term dependencies of the features along one spatial direction and preserves the precise location information of the features along the other spatial direction, which helps the network to locate the information of interest more accurately.

(2) Coordinate attention generation

The feature map is decomposed and pooled from two dimensions according to the method in section (1) so that the pooled features have a larger perceptual field to make full use of the information near the foreground target of the aircraft image. It enables the distant points on the same dimensional features to retain the mutual relationship under the special pooling. To incorporate the transformed features into the neural network, the final features with weights need to be generated. Coordinate attention generation should follow the following design principles.

1. Firstly, to improve the efficiency of the overall network model, the complexity of the feature conversion should not be too high and the conversion should be as simple as possible.

2. Secondly, the conversion should retain and utilize the location information in the feature map as much as possible, which can better grasp the overall spatial location information of the image and establish the connection between distant feature points on the aircraft image to capture the region of interest.

3. Finally, the whole process can effectively bring out the key channel information features of the aircraft image and capture the relationship between channels as effectively as possible.

After information embedding, the information generation process mainly includes information fusion and convolutional transformation. Information fusion mainly stitches together all the information of different regional features, and then convolution, batch normalization, nonlinear activation, and other operations, as shown in (3).

$$f = \delta\left(F_1([z^h, z^\omega])\right) \tag{3}$$

Where $[z^h, z^\omega]$ is the stitching and fusion of two feature maps of different orientations along the spatial dimension, F_1 is the convolution operation, δ is the nonlinear activation function, and $f \in R^{\frac{C}{r}\times(H+W)}$ is the intermediate feature map where spatial information is encoded in horizontal and vertical directions, where r is the reduction rate of the regulatory dimension, and to reduce the dimensionality of the feature vector and improve the efficiency of network training, an appropriate ratio r is chosen to reduce the number of channels. The intermediate feature maps f along the x and y directions are decomposed into f^h and f^w, which correspond to the two dimensions of the horizontal and vertical directions of the feature map, respectively. The convolutional transform and nonlinear activation are performed on the two tensors, as shown in (4) and (5), respectively:

$$\boldsymbol{g}^h = \sigma\left(F_h(\boldsymbol{f}^h)\right) \tag{4}$$

$$\boldsymbol{g}^w = \sigma\left(F_w(\boldsymbol{f}^w)\right) \tag{5}$$

Of which F_h and F_w are the 1*1 convolutional change operations, σ is the Sigmoid activation function, and the outputs $\boldsymbol{g}^h$ and $\boldsymbol{g}^w$ are the attention weights of the horizontal and vertical directions of the input X, respectively.

Ultimately, the output of the feature $x_c(i,j)$, which denotes the height and width on the c-th channel of input X is i and j, after the coordinate attention module, can be expressed as (6).

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{6}$$

3) CSPResNet50dCA feature extraction network

In general, deepening the network can enhance the learning ability of the model and better extract the sample features. However, increasing the depth of the network will increase the training difficulty, and the gradient explosion or gradient disappearance will occur easily during the training process, and the increase of parameters will often make the network tend to overfit. Therefore, too-deep networks tend to degrade in accuracy rather than increase. The main structure of ResNet is the residual module, which contains two main parts, residual learning, and identity mapping channel. The residual structure is shown in Fig. 6.

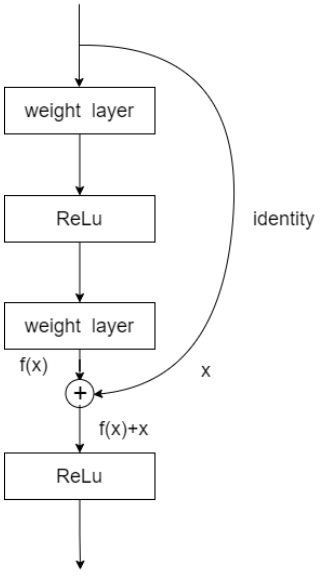


Fig. 6. Schematic diagram of the residual structure

ResNet still performs well after the layer deepening because the gradient is preserved by the identity mapping channel. According to the stacking of a different number of residual blocks, there are five ResNet structures with different depths, which are 18, 34, 50, 101, and 152. The ResNet50 structure is used in this paper.

ResNet50d is a ResNet-D network with 50 convolutional layers [15]. ResNet-D moves the down-sampling of the original ResNet residual branch to the 3×3 convolution at the back to avoid a large amount of information loss. At the same time, it leaves the down-sampling of the identity mapping part to the average pool to avoid the information loss caused by 1×1 convolution and down-sampling at the same time. The specific module structure of block1 and block2 of ResNet-D is shown in Fig. 7.

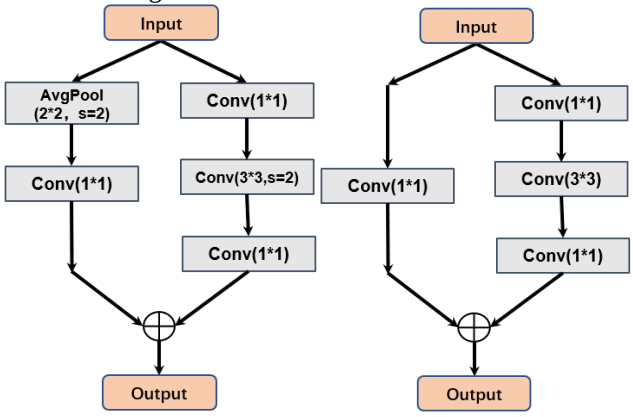


Fig. 7. Two blocks of ResNet-D

The overall structure of ResNet50d is shown in Table 1. The input image first goes through three 3×3 convolutions and one maximum pooling to change the image size to 1/4 of the original size, and then goes through stage1, stage2, stage3, and stage4 in turn, to further extract features. Each stage is composed of 1 block1 and k block2.

TABLE I. NETWORK STRUCTURE OF THE RESNET50 MODEL

Layer	Num	Kernel size	Stride	Output size	Channels
Input				640*640	3
3*Conv	1*Conv1	3*3	2	320*320	32
	2*Conv2	3*3	1	320*320	64
Max Pool	1		2	160*160	64
Stage1	1*block1		1	160*160	64
	2*block2		1	160*160	
Stage2	1*block1		2	80*80	128
	3*block2		1	80*80	
Stage3	1*block1		2	40*40	256
	5*block2		1	40*40	
Stage4	1*block1		2	20*20	512
	2*block2		1	20*20	

To further improve the learning capability of the ResNet50d network and remove the computational bottleneck, this paper nests the CSPNet[16] structure in the ResNet50d network, which is shown in Fig. 8 and can reduce the use of video memory and accelerate the inference speed of the network.

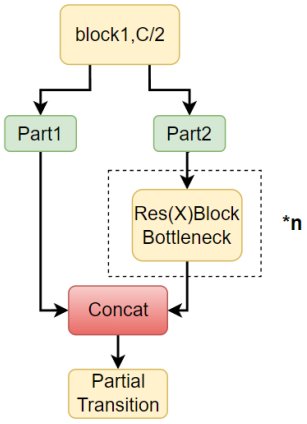


Fig. 8. Structure of CSPNet

In this paper, CSPNet is nested in stage 1, stage 2, stage 3, and stage 4 to effectively enhance the learning ability of the convolutional neural network and improve the accuracy of the model. The feature extraction network CSPResNet50d of this paper is designed through the CSP structure, as shown in Fig. 9.

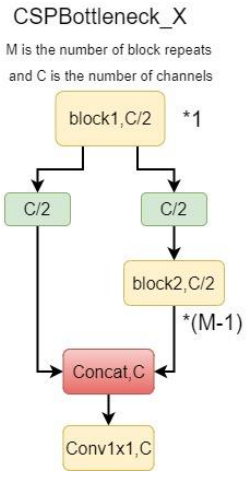


Fig. 9. Structure of CSPResNet50d

M is the number of block repeats and C is the number of channels.

The difficulty of small target detection lies in the relatively small number of available features for small targets, the high requirement for target localization accuracy, the lack of accurate position information, and the incomplete representation of features, so the coordinate attention mechanism CA module is introduced in the CSPResNet50d network.

The CSPResNet50dCA feature extraction network is designed by these tricks to achieve adequate extraction of small target features, and its overall structure is shown in Fig. 10.

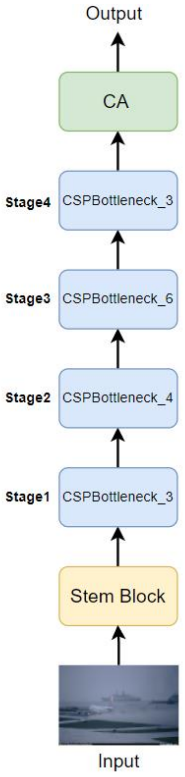


Fig. 10. The Overall structure of CSPResNet50dCA

4) Improved Algorithm of YOLOv5-CSPResNet50dCA

Realistic aircraft backgrounds are complex and there are often many occlusions. Different angles and attitudes lead to obvious disparities in the images of the same type of aircraft and the size of aircraft in different images varies greatly. In addition to that, there are many small target aircraft. These factors make it difficult to locate and identify the aircraft. To achieve accurate recognition of aircraft targets, this paper uses CSPResNet50d to replace the YOLOv5 backbone network CSPDarkNet53 and adds an attention mechanism to the CSPResNet50d network to form the CSPResNet50dCA network. The improved YOLOv5-CSPResNet50dCA algorithm can detect and identify small target aircraft more accurately.

Since the existing aircraft image datasets are limited, a good model cannot be trained with these small datasets alone, so this paper applies a transfer learning strategy to train the model [17-18]. Firstly, we construct a new transfer learning model using the deep learning model pre-trained on the large-scale ImageNet21k image dataset. Then, we set reasonable model hyperparameters, and use the weighted sum of training loss, validation loss and distance between the training set and validation set as training cost. Finally, we determine the best transfer learning model by layer-by-layer training and validation. The overall structure of the algorithm is shown in Fig. 11.

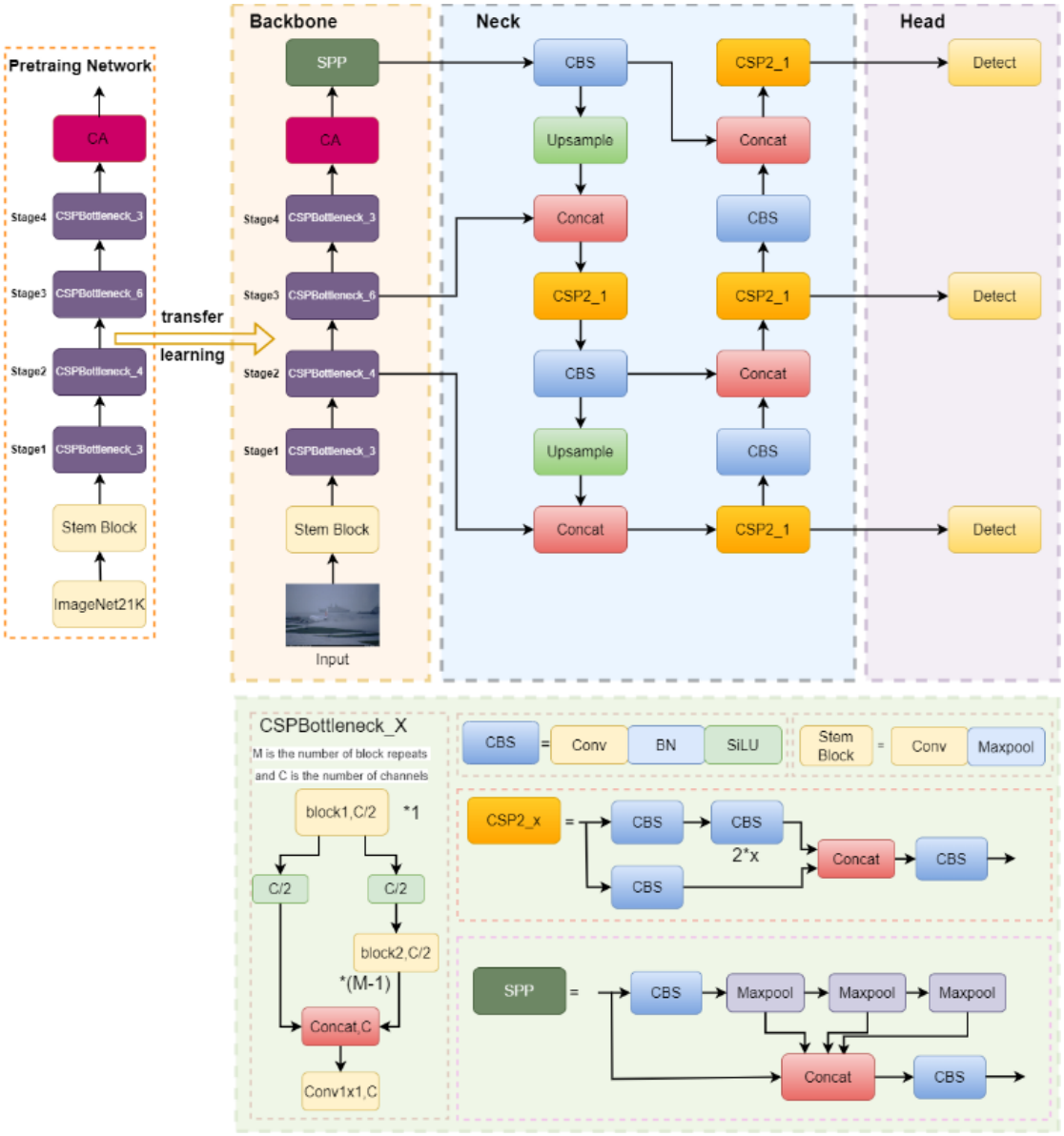


Fig. 11. The Overall structure of The Improved YOLOv5 model

III. RESULTS

A. TRAINING ENVIRONMENT AND EVALUATION INDICATORS

All experiments were run in our lab using an Intel(R) Core(TM) i7-10750H CPU (2.60GHz CPU, 16GB RAM) and an NVIDIA GeForce RTX 2070 (8G video memory). Model training and testing were done in the PyTorch framework. Both model training and testing are performed using GPUs to accelerate the computation. The experimental conditions and computer hardware information in this paper are shown in Table 2.

TABLE II. EXPERIMENTAL CONDITIONS

Experimental Environment	Details
Operating system	Windows 10
Compiler	Pycharm 2022.1.3
Programming language	Python 3.6
Deep Learning Framework	Pytorch 1.5.1
GPU model	NVIDIA GeForce RTX2070 8GB
CUDA version	12.0
Central Processing Unit	Intel(R) Core(TM) i7-10750H CPU

Detection accuracy and detection speed are important indicators for measuring model performance. The indicators

of detection accuracy include precision rate(P), recall(R), average precision (AP), and mean average precision (mAP). The calculation formulas are expressed in (7)-(10).

$$P = \frac{TP}{TP+FP} \times 100\% \tag{7}$$

$$R = \frac{TP}{TP+FN} \times 100\% \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{10}$$

where TP is the number of correctly predicted positive samples, FP is the number of falsely predicted positive samples, FN is the number of wrongly predicted negative samples, n is the number of target categories to be detected, APi is the AP of the i-th target class, N is the number of images to be detected, and t is the detection time.

B. RESULTS OF THE IMPROVED BACKBONE EXPERIMENT

To verify the feature extraction ability of the CSPResNet50dCA model, this paper conducted an aircraft classification and recognition experiment using Dataset 1, and trained the classification model using the training set before classifying aircraft models on the test set. The control group models were set up: ResNet50, Vision-Transformer-B, ResNet50-CBAM, and CSPDarknet53. The evaluation metric for this experiment is the precision rate (P).

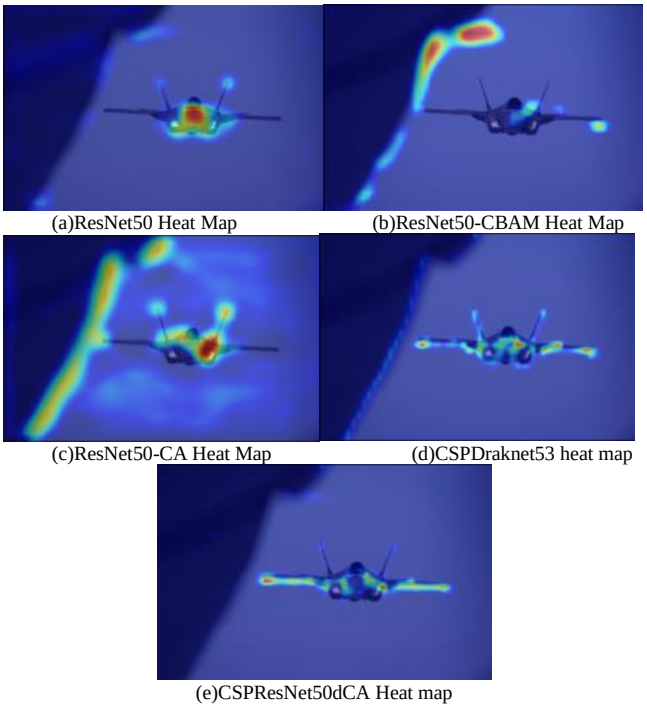
The hyperparameters in this experiment are determined by combining the setting laws in the references and local multiple experiments. Considering the hardware conditions and training time, the sample batch size is set to 16 for both testing and training. The number of runs for this experiment is set to 40 epochs, and the test interval and the print result interval are set to 1 epoch. The optimizers used in the experiments are all adaptive moment estimation (Adam) optimizers, which include exponential decay in the learning rate update strategy. In the initial stage of training, a large learning rate is set to quickly reach the vicinity of the optimal solution, and then the learning rate is gradually reduced to avoid the drastic oscillations caused by a large learning rate. The dataset for transfer learning in this paper is ImageNet21k, and the weights of the model pre-trained in the ImageNet21k dataset are migrated to the aircraft image dataset for re-training, and then the weights are fine-tuned to complete this task. The experimental results are shown in Table 3.

TABLE III. ACCURACY RATES OF DIFFERENT ALGORITHMIC MODELS

Algorithm model	precision (%)	Epoch
ResNet50	86.2	40
ResNet50-CBAM	86.4	40
Vision-Transformer-B	85.9	40
CSPDarknet53	88.3	40
CSPResNet50dCA	89.7	40

The experimental results show that the accuracy of the CSPResNet50dCA model on Dataset 1 is significantly improved compared with other models. Compared with the backbone of yolov5, CSPDarknet53, the accuracy is improved by 1.4% and has better aircraft feature extraction ability.

To verify and analyze the reason for the enhanced capability of the CA attention mechanism designed in this paper for small target aircraft features, the heat map of feature extraction of different models for the same aircraft image is drawn in Fig. 12. It can be seen by comparison that the CSPResNet50dCA network can ignore the useless background information and enhance the attention to small target aircraft profile information, so the location covered by the heat map is also more precise, and more feature information is extracted.



(a)ResNet50 Heat Map (b)ResNet50-CBAM Heat Map

(c)ResNet50-CA Heat Map (d)CSPDraknet53 heat map

(e)CSPResNet50dCA Heat map

Fig. 12. Heat map of different models

C. RESULTS OF THE ABLATION EXPERIMENT

The improved methods proposed in this study are the addition of the attention mechanism CA and the replacement of the backbone with CSPResNet50dCA. To verify the effectiveness of these improved methods, we designed ablation experiments: (1) based on the original YOLOv5 algorithm, only one improved method was added to verify the improvement effect of each improved method on the original algorithm. (2) The CA attention mechanism and improved backbone were freely combined to select the optimal detection model. To compare the performance of different models, Mean Average Precision (mAP) and single image detection time (infer time) are used as metrics in this paper.

The experiments were conducted on Dataset 2. The experimental results are shown in Table 4.

TABLE IV. RESULTS OF THE ABLATION EXPERIMENT

model	mAP@0.5	Infer time/ms
Yolov5	69.3	32.51
CSPResNet50d-Yolov5	69.6	25.68
CA-Yolov5	71.3	32.92
CSPResNet50dCA-Yolov5	73.6	25.88

The experimental results show that both adding the CA attention mechanism and replacing the backbone lead to an improvement in mAP. The CSPResNet50dCA-Yolov5 model achieves the highest 73.6% mAP. Adding the CA attention mechanism leads to a 2% improvement in mAP, indicating that the CA module improves the attention to small targets, fuses multi-scale information, and improves the detection effect of the network. Replacing the backbone with CSPResNet50d improves mAP by 0.3%, and fully fusing the residual structure and CSP structure can improve the localization effect of the model for small targets. Meanwhile, the improved model monitors 6.63ms faster than Yolov5.

D. RESULTS OF THE COMPARATIVE EXPERIMENT

To further prove the superiority of the algorithm proposed in this study, it was compared with the YOLOv5, YOLOv4, and YOLOv3 algorithms on Dataset 2. The experimental results are shown in Table 5.

TABLE V. Performance comparison of the different algorithm models.

model	mAP@0.5	Infer time/ms
Yolov3	64.2	37.88
Yolov4	66.9	42.68
Yolov5	69.3	32.51
CSPResNet50dCA-Yolov5	73.6	25.88

The mAP@0.5 of YOLOv3, YOLOv4, YOLOv5, and YOLOv5-CSPResNet50dCA can reach 64.2%, 68.1%, 69.3%,73.6%, respectively. Fig.13 shows the change curve of mAP during training. The initial accuracy rate of YOLOv5-CSPResNet50dCA was low during training, and the accuracy rate fluctuated considerably. However, the convergence speed was high, and the accuracy rate was the highest. The detection speed of YOLOv5-CSPResNet50dCA is higher than those of YOLOv3, YOLOv4, and YOLOv5.

Fig. 14 shows the comparison of the loss curves of YOLOv5 and YOLOv5-CSPResNet50dCA during the training process. The loss value of YOLOv5-CSPResNet50dCA is 0.0026, which is 0.011 lower than that of the original YOLOv5. The model is further optimized.

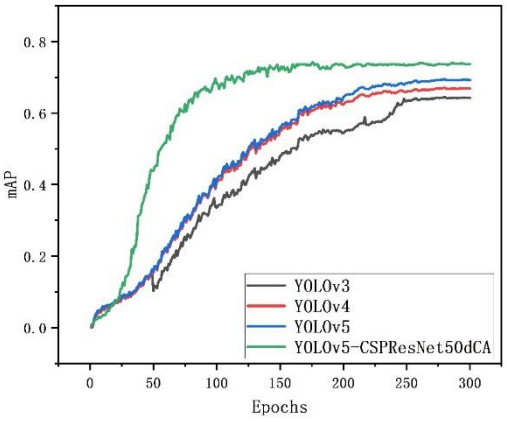


Fig. 13. mAP change curve(The gray curve represents YOLOv3, the red curve represents YOLOv4, the blue curve represents YOLOv5, and the green curve represents the model we proposed.)

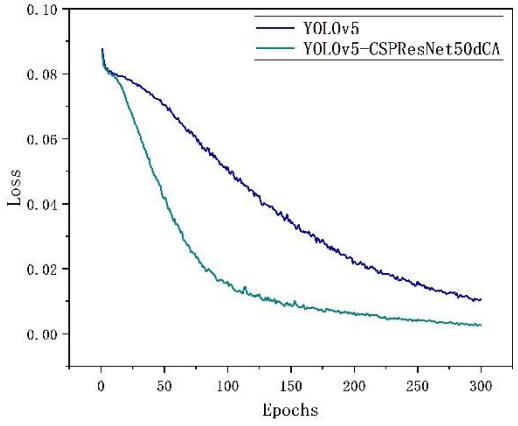


Fig. 14. Loss change curve

The YOLOv5 mosaic data enhancement can be achieved by splicing four images, as shown in Fig. 15, which considerably enriches the background of the detected object. Table 4 shows that the performance of YOLOv5-CSPResNet50dCA has been further improved compared with that of YOLOv5. The mAP of YOLOv5-CSPResNet50dCA is 73.6%, which is 9.4%, 6.7%, and 4.3% higher than those of YOLOv3, YOLOv4, and YOLOv5, respectively. The detection effect is shown in Fig. 16 and 17.

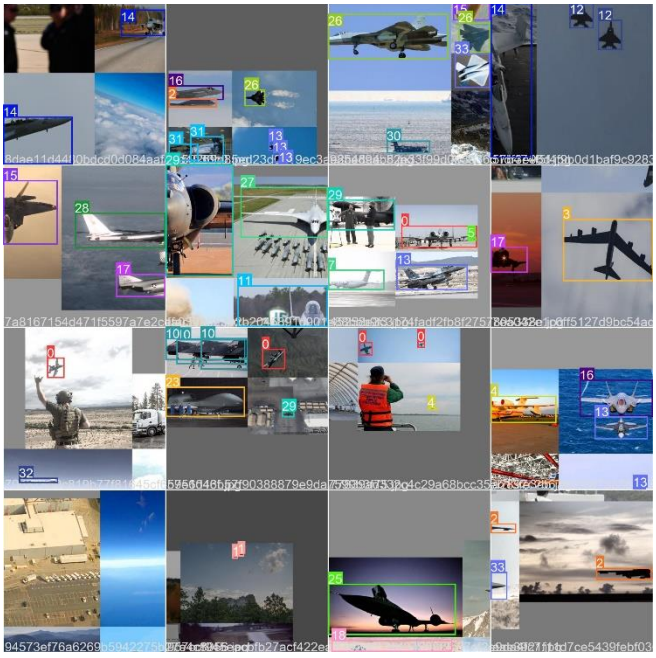


Fig. 15. Mosaic data enhancement during training

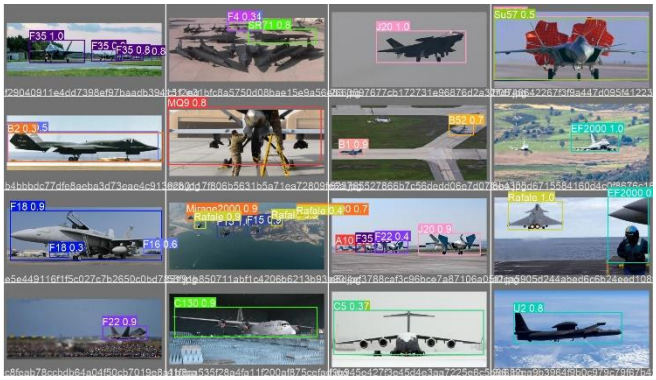


Fig. 16. Batch detection effect

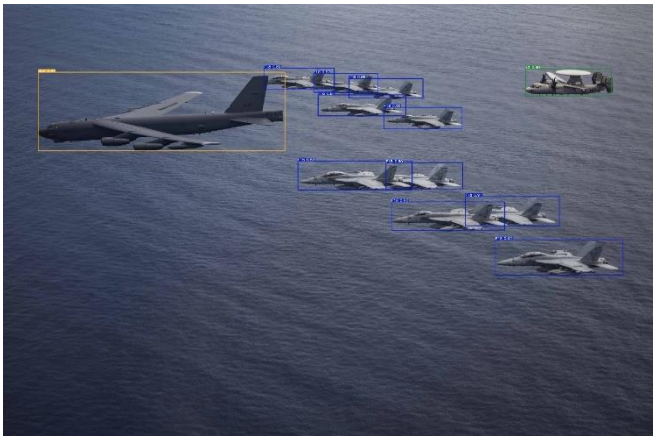


Fig. 17. Detection effect

IV. CONCLUSIONS AND FUTURE RESEARCH

The detection of aircraft targets is of great importance to both military and civilian fields. In this paper, an aircraft detection model based on the improved YOLOv5 network is established: Firstly, the original YOLOv5 backbone is replaced with the CSPResNet50d network to enhance the extraction capability of the network for aircraft features. Then the CA attention mechanism module is inserted into the model species for feature fusion to obtain more feature information of small target aircraft, and the network attention is focused on useful information, which improves the network performance at a smaller cost. Finally, the transfer learning strategy is used to reduce the computation of the model and improve the detection speed, while strengthening the generalization ability of the model. The improved model detects mAP up to 73.6%, and the detection time (infer time) of a single image is only 25.88ms, which is better than the original YOLOv5 model and other models. The improved network model not only has a high detection rate but also has a significant improvement in the recognition rate of small target aircraft. The model improvements studied in this paper are designed to maintain a balance between detection performance and detection speed to meet the demand for real-time detection of different types of aircraft. The research is also applicable to other military and civilian target detection fields to provide technical references for military decision-making and civil aviation applications.

However, the improved YOLOv5 model has limitations, for example, missed and incorrect detection of small target aircraft still exists. This is due to the difficulty in acquiring aircraft image datasets as well as the uneven quality. The poor quality of the acquired images is caused by factors such as weather, lighting, and the shooting environment. In the future, we can try to use an improved generative adversarial network to generate more images under bad weather and lighting conditions from existing aircraft images, and then use them for model training to enhance the robustness of the model and improve the accuracy rate. In addition, we can try to replace the backbone with a more lightweight model to reduce the number of model parameters.

REFERENCE

[1] H. L. Ma, R. Zhang, F. Li, et al. The application of trend line analysis method in the life assessment of pyrotechnic products[J]. Fireworks, 2013, 25(3): 53-56.

[2] W. T. Zhu, B. R. Xie, Y. Wang, et al. An overview of research on aircraft target detection techniques in optical remote sensing images [J]. Computer Science,2020,v.47(S2):175-181+192.

[3] M. M. Sha, Y. Li, and A. Li. 2022. multiscale aircraft detection in optical remote sensing imagery based on advanced Faster R-CNN. national Remote Sensing Bulletin, 26(8):1624-1635.

[4] W. T. Zhu, X. C. Lan, Y. L. Luo, B. Yue, and Y. Wang. Improved Faster R-CNN for optical remote sensing aircraft target detection[J]. Computer Science,2022,49(S1):378-383.

[5] H. T. Wang, and Z. H. Guo. Improved SSD for aircraft remote sensing image target detection[J]. Liquid Crystal and Display,2022,37(01):116-127.

[6] R. P. Shi, D. N. Jiang, and Q. Fang. Remote sensing image aircraft target detection based on YOLOv4[J]. Survey and Mapping Bulletin,2021(S1):134-138.

[7] T. Zhang, Y. H. Liu, and S. C. Li. Aircraft target detection based on improved YOLOv4 for remote sensing images[J]. Electro-Optics and Control,2022,29(12):101-105+117.

[8] Thuan. D. Evolution of Yolo Algorithm and Yolov5: The State-of-the-Art Object Detention Algorithm. Bachelor's Thesis, Oulu University of Applied Sciences, Oulu, Finland, 2021.
[9] Shao. H, Pu. J, and Mu. J. Pig-posture recognition based on computer vision: Dataset and exploration. Animals 2021, 11, 1295.
[10] Krizhevsky. A, Sutskever. I, Hinton. G.E.J.C, and Otto. A. Imagenet classification with deep convolutional neural networks. 2017, 60, 84–90. Commun. ACM 2017, 60, 84–90.
[11] W. Wu, H. Liu, L. Li, Y. Long, X. Wang, Z. Wang, J. Li, and Y. Chang. Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image.[J]. PloS one,2021,16(10).
[12] J. Z. He, D. Jiang, D. H. Zhang, J. Li, and Q. G. Fei. Interval model validation for rotor support system using K-means Bayesian method[J]. Probabilistic Engineering Mechanics,2022,70.
[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[14] Z. Zhang, and M. Wang. Convolutional Neural Network with Convolutional Block Attention Module for Finger Vein Recognition[J]. arXiv preprint arXiv:2202.06673, 2022.
[15] T. Bello, W. Fedus, X. Du, et al. Revisiting resnets: Improved training and scaling strategies[J]. Advances in Neural Information Processing Systems, 2021, 34: 22614-22627.
[16] C. Y. Wang, H. Liao, Y. H. Wu, et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2020.
[17] W. Zhang, J. Wang, and F. Lan, "Dynamic hand gesture recognition based on short-term sampling neural networks," IEEE/CAA Journal of Automatica Sinica, 2021, 8(1), 110-120.
[18] P. Zhou, G. Liu, J. Wang, Q. Wang, K. Zhang, and Z. Zhou, ZiYuan, "Lightweight unmanned aerial vehicle video object detection based on spatial-temporal correlation," International Journal of Communication Systems, 2022, 35(17), e5534.

AUTHOR BIO/IMAGE

HUANYU YANG received his bachelor's degree from North University of China in 2020. Since 2021, he has been studying for his Ph.D. degree in Control Science and engineering in Nanjing University of Science and Technology under the guidance of researcher Yuming Bo. His main research direction is identification and tracking of military targets.

JUN WANG received the B.S. degree in School of Automation in 2003, and the Ph.D. degree in control science and engineering in 2009, all from Nanjing University of Science and Technology, Nanjing, China. He is currently an Associate Professor with School of Automation, Nanjing University of Science and Technology. His research interests include vibration control, optimal standby control and estimation.

LIJUN YANG received his master's degree from Nanjing University of Science and Technology in 2022. Her research interests include fire control and intelligent control.

YUMING BO received the B.S. degree in School of Automation in 1984, and the M.S. and Ph.D. degrees in control science and engineering in 1987 and 2005, all from Nanjing University of Science and Technology, Nanjing, China. He is currently a Full Professor with School of Automation, Nanjing University of Science and Technology. His research interests include vibration control, filtering and system optimization. Prof. Bo is a member of the Chinese Association of Automation and the Vice Chairman of Jiangsu Branch. He is a standing council member of China Command and Control Society. He was granted a secondary prize of natural science from the Ministry of Education of China in 2005 and a secondary prize of technology promotion from Shandong Province in 2012.