[1]Rui Zhou
[2,*]Ming Li
[3]Shuangjie Meng
[4]Shuang Qiu
[5]Qiang Zhang

Aircraft Objection Detection Method of Airport Surface based on Improved YOLOv5

Journal of Electrical Systems

Abstract: - An aircraft object detection method on the basis of improved YOLOv5 was proposed to address the issues of large model size, high number of parameters, and inability to meet real-time monitoring requirements of aircrafts in traditional object detection. Firstly, the basic unit of ShuffleNetv2 network was optimized through replacing 3x3 convolution with 5x5 convolution and removing subsequent 1x1 convolution. Simultaneously, the original ReLU activation function was replaced with PReLU. Secondly, CBAM (Convolutional Block Attention Module) attention mechanism was developed to enhance the detection accuracy of the improved network. Finally, improved ShuffleNetv2 network was applied as the backbone structure of YOLOv5. Experimental results revealed that the parameter number of the improved YOLOv5 method introduced in this paper was decreased by 18 times, with a model size of 1.03M. Therefore, a 20.8% increase was achieved in frames per second (FPS) in GPU environments and a 234.6% increase was observed in FPS in CPU environments, while a mean average precision (mAP@0.5) of 0.99 was maintained compared with traditional YOLOv5 network. Because of the advantages of fewer parameters, faster recognition speed, higher localization accuracy, and smaller memory requirement, the developed method was found to be suitable for real-time monitoring of aircrafts in airport surface.

Keywords: Surface Surveillance, Object Detection, Improved YOLOv5, Shufflenetv2, Activation Functions, Attention Mechanism.

I. INTRODUCTION

Monitoring of aircrafts in airport is critical for the safe operation of civil aviation. Today, aircraft surveillance often involves the integration of multiple information sources, such as airport surveillance radars and multi-point positioning systems [1]. However, due to high costs, this is a challenge for regional airports. To solve this issue, video surveillance offers a cost-effective alternative, since it does not require onboard receiving equipment installation on aircraft. In addition, it can act as a supplementary monitoring tool in radar-obstructed areas.

Significant progress has been made in image-based object detection with extensive application of deep learning. YOLOv3 [2] object detection model, introduced by REDMON et al., introduced a novel method by incorporating Darknet-53 fully convolutional network. This innovative model effectively minimized the loss of low-level features, utilizing residual structures and multi-scale detection. Therefore, this enhanced accuracy while maintaining fast detection. YOLOv4 [3] algorithm developed by Bochkovskiy et al. introduced a novel method for achieving enhanced object detection performance by replacing the main network backbone and integrating spatial pyramid pooling. This method applied path aggregation network (PAN) for feature fusion. As a different method, YOLOv5 [4] proposed five different network architectures with varying depths and widths to enhance object detection performance. In addition, YOLOv5 utilized adaptive image scaling and anchor box generation to further improve detection accuracy. Ruiz-Barroso [5] developed an approach utilizing region proposal network (RPN) and optical flow maps among frames, referred to as optical flow region proposal (OFRP), to automatically identify object regions in videos. This advancement significantly enhanced the computational speed of the algorithm in both GPU and CPU environments. Yang [6] developed a lightweight real-time detection algorithm on the basis of YOLOv4. This algorithm replaced CSPnet in network structure with Ghostnet and regular convolutions were substituted with depthwise separable convolutions. In addition, a four-layer pyramid was constructed to enhance the accuracy, effectiveness and robustness of the model in aircraft target detection. Although classical algorithms have significantly enhanced the speed and accuracy of detection, excessive model parameters and large model sizes have restricted their application in real-time airport surface surveillance.

By analyzing existing algorithms, this research introduced a new method for aircraft object detection in airport surface, which addressed various challenges such as excessive model parameters and limited real-time performance. The basic units of Shufflenetv2 network were optimized and attention mechanisms were incorporated into YOLOv5 backbone network. Therefore, a lightweight aircraft object detection method was proposed with

[1] Air Traffic Management College, Civil Aviation Flight University of China, Guanghan, 618307, China
[2] Air Traffic Management College, Civil Aviation Flight University of China, Guanghan, 618307, China
[3] Air Traffic Management College, Civil Aviation Flight University of China, Guanghan, 618307, China
[4] Air Traffic Management College, Civil Aviation Flight University of China, Guanghan, 618307, China
[5] Air Traffic Management College, Civil Aviation Flight University of China, Guanghan, 618307, China
*Corresponding author: Ming Li

fewer parameters, smaller size, and higher accuracy. This method enabled real-time monitoring of aircrafts in airport environments.

II. IMPROVEMENT OF YOLOV5 ALGORITHM

A. Principles of YOLOv5

Aircraft object detection based on video images primarily involves extracting feature maps through convolutional neural networks (CNN), fusing features, and finally outputting detection results [7]. Figure 1 illustrates the developed algorithm flow.

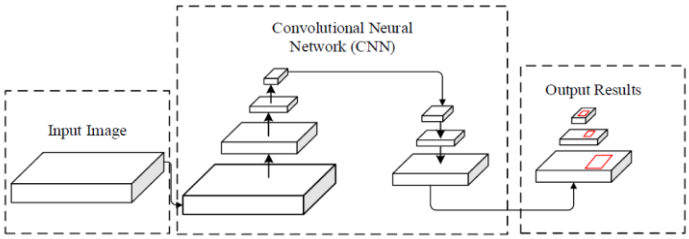


Figure 1: Algorithm Flow of Aircraft Target Detection Process

Currently, YOLOv5 is mainly adopted for object detection. As a representative one-stage end-to-end detection algorithm, YOLOv5 has five distinct models of YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x [8]. These models were designed with different sub-module depths and widths and each model achieved improved detection accuracy and increased the size of the model in the given order. YOLOv5s effectively balanced parameter number and accuracy, making it the optimal choice for aircraft detection in airport surface.

Detection procedure in YOLOv5 [9] comprised the following four primary components: YOLOv5 architecture, as shown in Figure 2, showcasing input stage, Backbone feature extraction stage, Neck feature fusion stage, and Head output stage. Before image feeding into the network, data preprocessing was conducted which included mosaic data augmentation, adaptive anchor box computation, and dynamic image scaling. In the main network, YOLOv5 utilized C3 module to obtain feature fusion and decrease computation cost. Spatial pyramid pooling – fast (SPPF) module substituted a single large pooling kernel found in SPP module with multiple smaller pooling kernels to further improve processing speed while preserving its original functionality. In Neck feature fusion part, the ideas of feature pyramid network (FPN) and PANet were introduced to exchange deep semantic information with shallow positional information for enhancing information fusion at multiple scales. In the output phase of Head network, various scales of feature maps from Neck were processed to predict and regress target boxes to form three prediction boxes for each target using non-maximum suppression for target box selection.

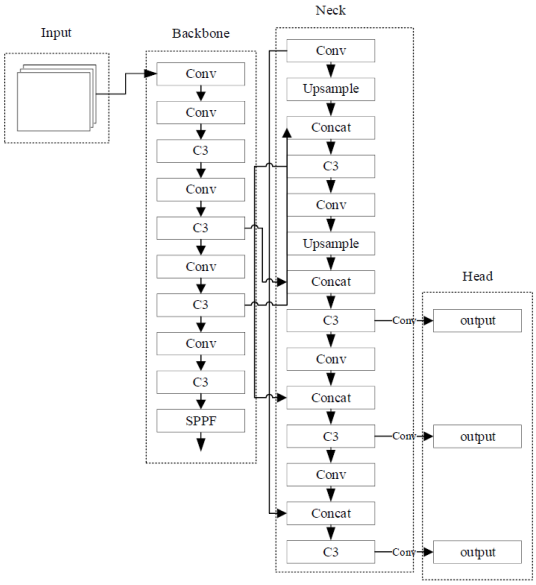


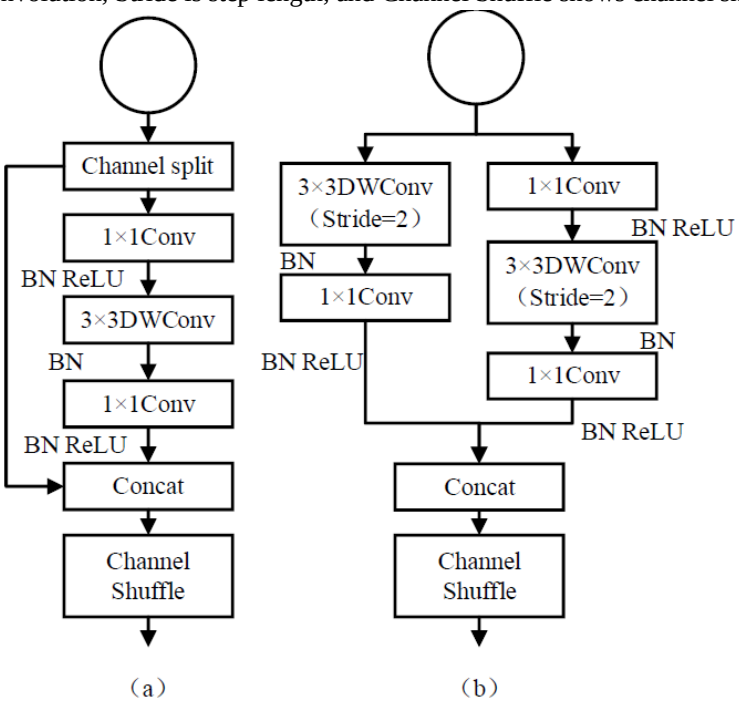
Figure 2: YOLOv5 Architecture Diagram

Because YOLOv5 backbone network had a complex structure and high computational workload, it resulted in slow algorithm execution and large model size. However, in airport surface surveillance, there were challenges due to limited computing resources, high real-time requirements, and long-term operation [10]. To address computation

resource and real-time performance requirements for airport surface monitoring, an optimization YOLOv5 approach was developed by replacing its backbone network with Shufflenetv2.

B. Principles of ShuffleNetV2 Network

Shufflenetv2 [11] served as a lightweight network structure that incorporated specialized network design and channel reordering operations. This architecture effectively reduced computational complexity and improved model efficiency, leading to smaller model sizes minimizing storage requirements. In addition, it ensured high computation speed, remarkably satisfying real-time monitoring demands. Figure 3 illustrates the basic unit of Shufflenetv2; Figure 3(a) represents the basic unit and Figure 3(b) shows downsampling unit. In these figures, Conv stands for regular convolution, BN denotes batch normalization, ReLU is activation function, DW Conv stands for depthwise convolution, Stride is step length, and Channel Shuffle shows channel shuffling [12-13].



(a) Shufflenetv2 Basic Unit, (b) Shufflenetv2 Downsampling Unit

Figure 3: Shufflenetv2 Network Diagram

However, direct application of classical Shufflenetv2 in airport surface monitoring had some limitations, which impacted its detection accuracy and real-time capability. Firstly, classical Shufflenetv2 adopted smaller convolutional kernels, which restricted the ability of network to perceive larger objects or complex details in airport surface. Since aircrafts in airports come in different sizes and possess intricate structural features, smaller receptive fields might not adequately capture target information, reducing detection accuracy. Secondly, classical Shufflenetv2 performed a 1x1 convolution operation after depthwise convolution, which increased computational burden and model complexity. Since airport surfaces are relatively static, 1x1 convolution was not necessary and introduced unnecessary computational overhead, decreasing the efficiency and real-time performance of the algorithm. In addition, the original Shufflenetv2 adopted ReLU as its activation function, potentially causing information loss and gradient vanishing when handling complex targets. These limitations decreased the discriminative ability of the network for complex objects, reducing detection accuracy and robustness.

To solve these problems, this research optimized Shufflenetv2to adapt to the need of airport surveillance tasks. Depthwise convolution kernels were expanded to 5x5 to improve the receptive field of the network. This allowed for better capturing of detailed information from complex objects in airport surfaces, improving detection accuracy and robustness. In addition, 1x1 convolution operation was removed, which simplified network structure, reduced computation load and model complexity, and improved algorithm efficiency and real-time performance. Regarding activation functions, PReLU [14] was introduced as a replacement for ReLU. PReLU was a rectified linear unit with learnable parameters that provided stronger non-linear fitting capabilities. This change was made for further enhancing network accuracy. PReLU activation function was more suitable for adapting to varying shapes and appearances of aircrafts in target detection tasks. It enhanced the feature extraction and representation capabilities of the network, therefore, enhancing aircraft detection accuracy and robustness. Figure 4 compares PReLU and

ReLU activation functions. In aircraft target detection, shapes and appearance features of aircrafts demonstrated significant variation and complexity. By utilizing PReLU activation function, network could effectively adapt to different shapes and appearances of aircraft targets. This enhanced its extraction capability and represented target features, ultimately enhancing detection accuracy and robustness. Figure 5 illustrates the fundamental unit of optimized Shufflenet where Figure 5(a) shows the basic unit and Figure 5(b) represents downsampling unit.

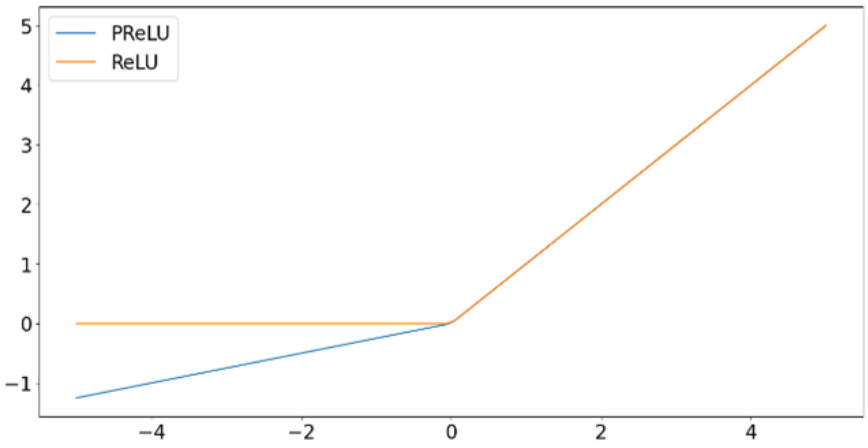
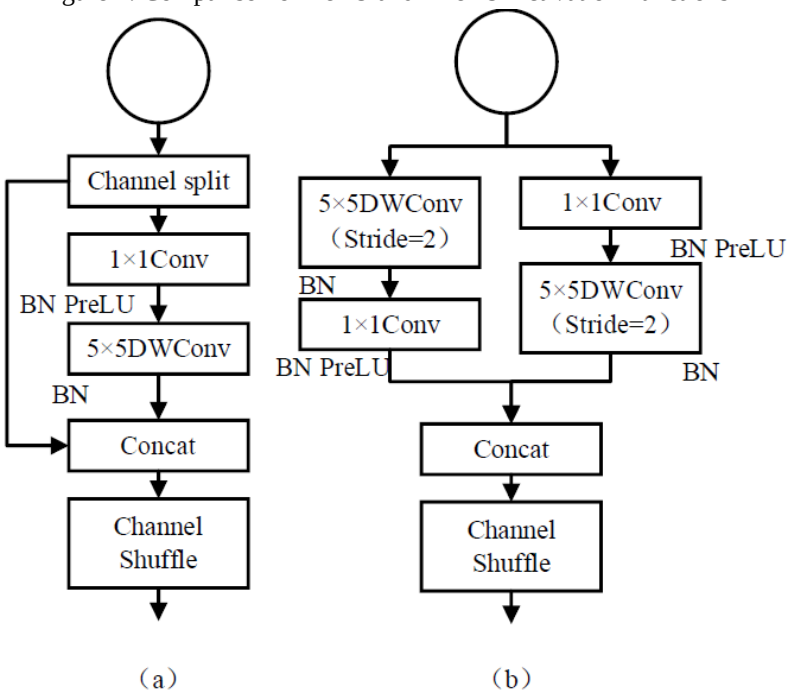


Figure 4: Comparison of ReLU and PReLU Activation Functions



(a) Improved Shufflenetv2 Basic Unit, (b) Improved Shufflenetv2 Downsampling Unit

Figure 5: Improved Shufflenetv2 Network Diagram

C. Improved YOLOv5 Algorithm

To further improve aircraft detection performance, an attention mechanism was introduced in optimized Shufflenet for improving YOLOv5 backbone network. Common attention mechanisms included squeeze-and-excitation (SE) [15], SimAM [16], efficient channel attention (ECA) [17], and CBAM [18]. CBAM attention mechanism possessed comprehensive channel and spatial attention adjustment capabilities. Unlike standalone channel attention mechanisms such as SE or spatial attention mechanisms such as ECA, CBAM concurrently assessed channel-to-channel relationships within the feature map and spatial position relationships. This comprehensive attention adjustment capability contributed to a more comprehensive enhancement of the ability of network to represent and focus on aircraft targets to further improve detection accuracy. Therefore, CBAM attention mechanism gained extensive popularity in advanced object detection and image classification tasks, achieving good performance. It had advantages in enhancing feature representation and improving object detection performance.

CBAM attention module boosted essential channels and spatial features within the feature map, which improved object detection localization accuracy and emphasized on object clustering. This helped address challenges such as false and missed detections due to overlapping objects. Therefore, CBAM attention mechanism was integrated into YOLOv5 enhancement within Neck structure, as illustrated in Figure 6.

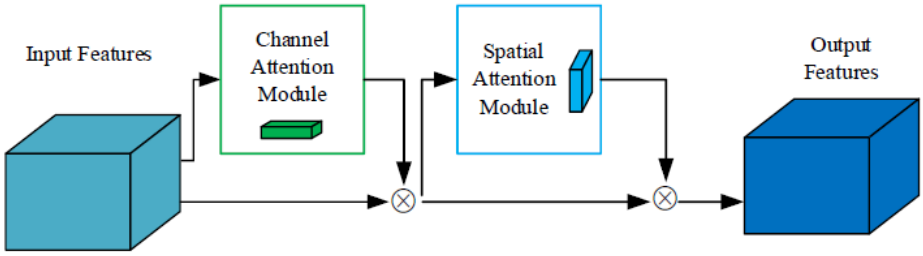


Figure 6: CBAM Attention Mechanism Module

CBAM module comprised two distinct parts: channel attention module (CAM) and spatial attention module (SAM). It enhanced the feature extraction capability and effectively boosted the detection accuracy of the network [18].

From an input feature map $\boldsymbol{F} \in \boldsymbol{R^{C\times H\times W}}$, we derive one-dimensional feature map $\boldsymbol{M_C} \in \boldsymbol{R^{C\times 1\times 1}}$ using Eq. (1), and two-dimensional feature map $\boldsymbol{M_s} \in \boldsymbol{R^{1\times H\times W}}$ using Eq. (2).

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{max}^c))) \end{aligned} \tag{1}$$

where σ represents $sigmoid$ activation function, $AvgPool(F)$ is average pooling, $MaxPool(F)$ is max pooling, and MLP denotes multi-layer perceptron operations. Also, $W_0 \in R^{C/r\times C}$ and $W_1 \in R^{C\times C/r}$ are the two sets of weights used in MLP and F_{avg}^C and F_{max}^C represent that features were obtained through average pooling and max pooling, respectively.

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times 7}([AvgPool(F);MaxPool(F)])) \\ &= \sigma(f^{7\times 7}(F_{avg}^s);(F_{max}^s)) \end{aligned} \tag{2}$$

To calculate spatial attention feature, first, average and max poolings compressed the feature map into a single layer along channel axis, resulting in two two-dimensional features of $F_{avg}^s \in R^{1\times H\times W}$ and $F_{max}^s \in R^{1\times H\times W}$. These features represented a one-dimensional channel with average and max poolings for each spatial position (H×W). Following convolution, we obtained the attention feature map $M_s(F) \in R^{1\times H\times W}$, as stated in Eq. (2), with $f^{7\times 7}$ representing a standard convolution using a 7×7 kernel

With an input feature map $\boldsymbol{F} \in \boldsymbol{R^{C\times H\times W}}$, Eqs. (3) and (4) gave the outputs of channel features and spatial features, respectively, as:

$$F' = M_c(F) \otimes F \tag{3}$$

$$F'' = M_s(F') \otimes F' \tag{4}$$

where $\otimes$ is element-wise multiplication, F' is channel feature output, and F^" is spatial feature output.

Therefore, in this research, YOLOv5 backbone structure was replaced by improved Shufflenet network to decrease the parameter number and increase the speed of the algorithm. In addition, incorporating CBAM attention mechanism enhanced both feature extraction and detection accuracy. Even by decreasing overall parameters and computational load, improved YOLOv5 structure maintained its high accuracy. Figure 7 shows the structure of improved YOLOv5.

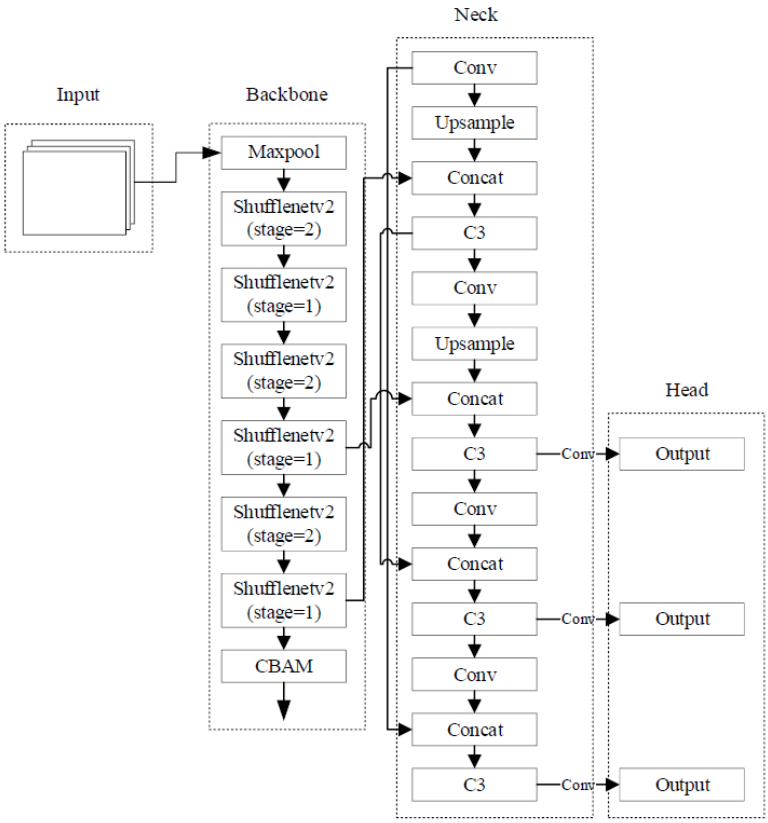


Figure 7: Improved YOLOv5 Structure Diagram

III. Experimental Results and Analysis

A. *Experimental Environment and Dataset Preparation*

To verify the developed method, the airport surface images taken in Linyi Qiyang Airport were applied as experimental data, which is a regional airport with a runway grade of 4D. The camera used was HF899_3.0mm (110-degree undistorted) and experiments were conducted on a Windows 10 system with an Intel(R) Core(TM) i5-7300HQ CPU and an NVIDIA GeForce GTX 1050Ti GPU. The simulations were performed using Torch version 1.11.0+cu113 and Python version 3.7 software.

A dataset for aircraft detection in airport surface was created by capturing 700 images under different operating conditions and time periods. To augment the dataset, the original image data was flipped horizontally and vertically, resulting in a total of 2800 images. After manual quality filtering to remove poorly captured images, the final dataset contained 2578 images. Then, the processed dataset was manually annotated using LabelImg. The dataset was divided into training, testing, and validation sets at 7:2:1 ratio, with 1805 images for training, 515 images for testing, and 258 images for validation. Figure 8 illustrates the experimental process.

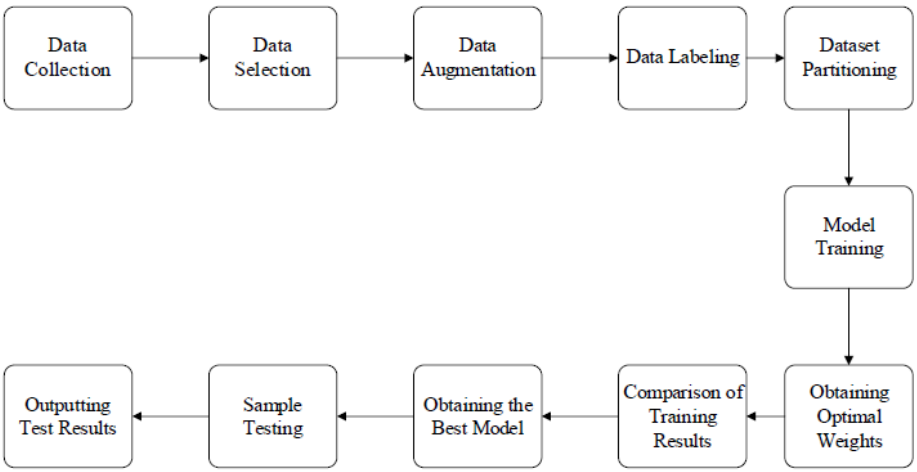


Figure 8: Experimental Flow for Aircraft Detection in Airport Surface

B. Model Training

In this experiment, the inference speed of the model was measured using frames per second (FPS) on test videos. Model complexity was analyzed by examining parameter number (Parameters), while model size was quantified in megabytes (Weight Size/MB). Evaluation of the detection performance of the developed model involved metrics such as precision (P), mean average precision (mAP), and recall (R) [19], as expressed in Eqs. (5-7).

$$Precision = \frac{T_p}{T_p + F_p} \tag{5}$$

$$Recall = \frac{T_p}{T_p + F_N} \tag{6}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{7}$$

where P is precision, R is recall, TP is true positives, FP is false positives, FN is false negatives, and AP is the area under precision-recall curve, as stated in Eq. (8).

$$AP = \int_0^1 P(R)dR \tag{8}$$

Considering the influences of both precision and recall, mAP reflected the recognition quality of the model for different classes.

C. Analysis of Experimental Results

The performances of original YOLOv5, YOLOv5 optimized by Shufflenetv2 and the improved method developed in this research and the comparative results of different network models are summarized in Table 1. As was seen from the table, our developed method substantially decreased the size and parameter number of network model while maintaining its accuracy, recall rate, and mAP values comparable to those of the original YOLOv5 network.

Table 1: Performance of Different Network Models

Model	Precision	Recall	mAP@0.5	Parameters	Weight Size/MB
YOLOv5	0.980	0.976	0.984	7022326	13.6
YOLOv5-Shufflenet	0.935	0.919	0.953	334199	0.95
YOLOv5-Improved	0.981	1	0.990	383613	1.03

Comparison results for enhanced YOLOv5 model with different introduced attention mechanisms are given in Table 2. Within these experiments, Method-1, Method-2, and Method-3 improved YOLOv5 model through the incorporation of Simam, SE, and ECA attention mechanisms, respectively. Table 2 showed that CBAM attention mechanism outperformed other attention mechanisms in terms of Precision, Recall, and mAP.

Table 2: Comparison of Different Network Models

Model	Precision	Recall	mAP@0.5	Parameters	Weight Size/MB
Method-1	0.983	0.966	0.986	375050	1.01
Method-2	0.960	0.966	0.980	506122	1.26
Method-3	0.978	0.967	0.984	375055	1.01
YOLOv5-Improved	0.981	1	0.990	383613	1.03

mAP@0.5 curves for the training processes of YOLOv5 and YOLOv5-improved are illustrated in Figure 9. It was observed that, YOLOv5-Improved attained a higher mAP value than YOLOv5 and reached a stable state at about 20 training epochs without underfitting or overfitting. However, YOLOv5 demanded about 50 training epochs to achieve stability

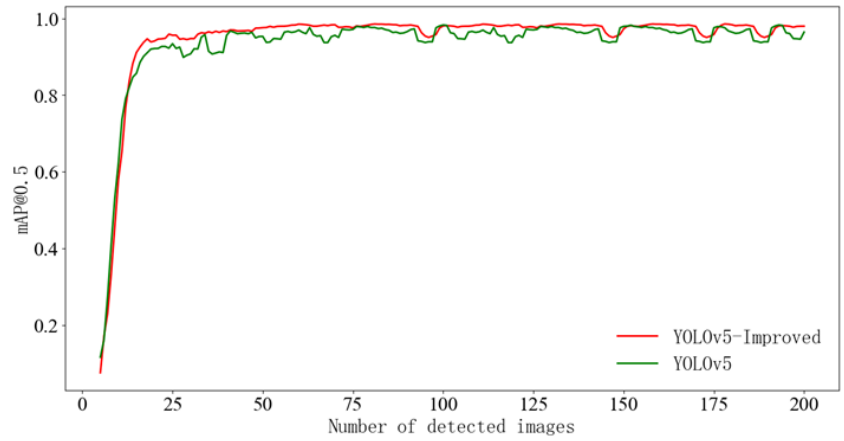


Figure 9: Comparison of mAP@0.5

Figure 10 shows Box_Loss comparison of YOLOv5 and YOLOv5-Improved. Both YOLOv5 and YOLOv5-Improved were gradually stabilized after about 100 iterations, but YOLOv5-Improved consistently demonstrated lower Box_Loss. This indicated that YOLOv5-Improved reduced boundary loss, thereby enhancing the localization capability of the model.

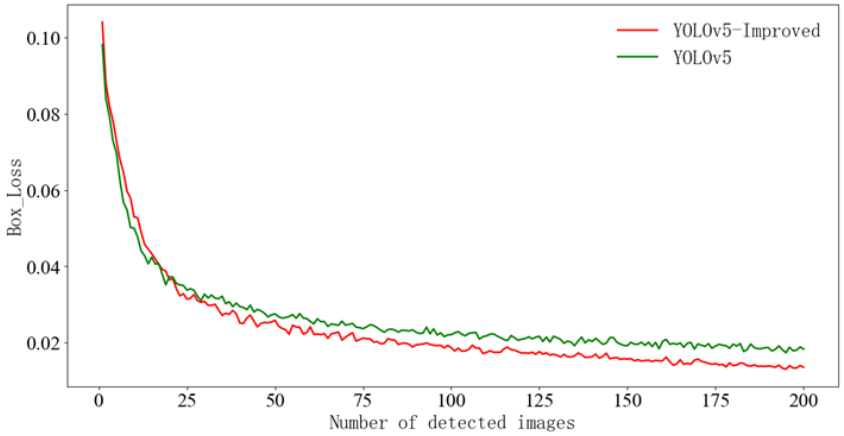


Figure 10: Comparison of Box Loss

FPS comparisons of YOLOv5-Improved and YOLOv5 on GPU and CPU environments are illustrated in Figures 11 and 12, respectively. In GPU environment, YOLOv5-Improved achieved maximum FPS of 85, minimum FPS of 70, and average of 77.8, while YOLOv5 achieved maximum FPS of 67, minimum FPS of 58, and average of 64.4. Therefore, YOLOv5-Improved was superior to YOLOv5 in terms of average FPS in GPU environment, with an improvement of 20.8%. In CPU environment, however, YOLOv5-Aircraft achieved maximum FPS of 20, minimum FPS of 14, and average of 17.57, while YOLOv5 attained maximum FPS of 6, minimum FPS of 2, and average of 5.25. Consequently, YOLOv5-Improved exhibited a significant improvement in average FPS over YOLOv5 in CPU environment, with 234.6% increase.

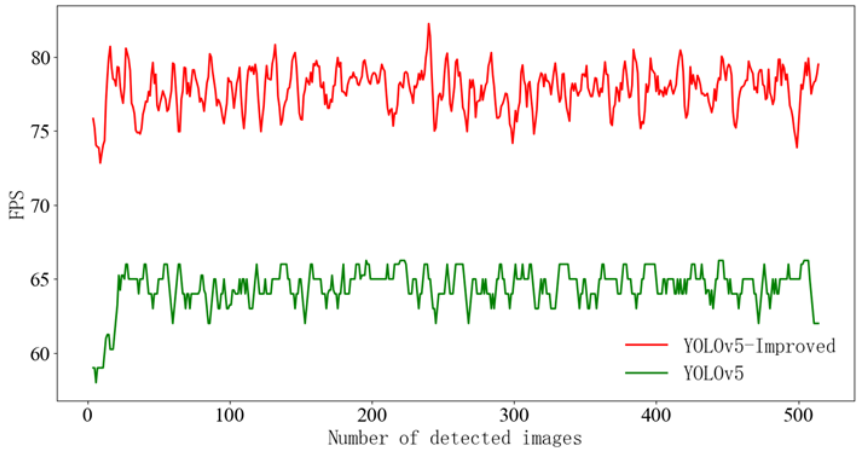


Figure 11: FPS Comparison in GPU Environment

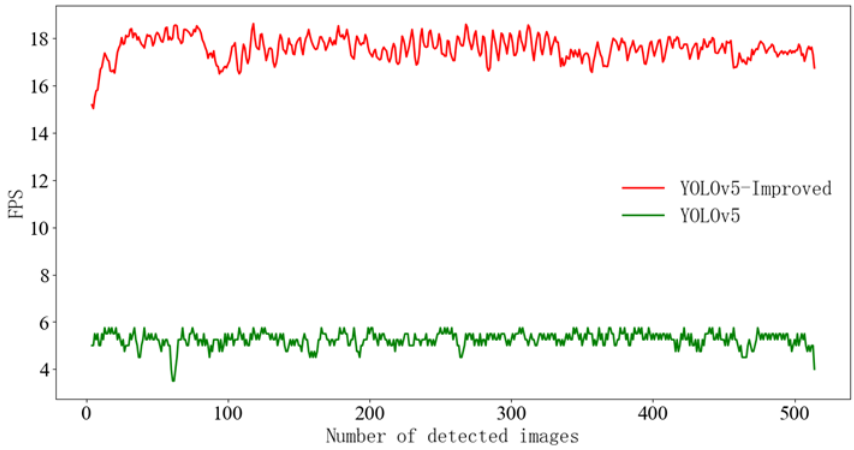


Figure 12: FPS Comparison in CPU Environment

Detection performances of YOLOv5 and YOLOv5-Aircraft models with different numbers of aircraft are shown Figures 13 and 14, respectively. It was evident that YOLOv5-Improved model demonstrated enhanced confidence levels in comparison to YOLOv5 model for different aircraft quantities.

Figure 13: Detection Performance of YOLOv5 with Different Aircraft Numbers

Figure 14: Detection Performance of YOLOv5-Aircraft with Different Aircraft Numbers

IV. Conclusions

In this research, a lightweight YOLOv5-based network was specifically designed for aircraft object detection in airport surveillance scenarios. This approach utilized an optimized Shufflenet as backbone structure for YOLOv5 and incorporated CBAM attention mechanism, resulting in the number of parameters of the model is reduced by 18 times. Model weight size was 1.03M, leading to a 20.8% increase in FPS in GPU environment and a 234.6% increase in FPS in CPU environment. However, through achieving model lightweightness, the proposed method maintained a comparable mAP@0.5 of 0.99, which was equivalent to YOLOv5 performance. This method successfully addressed issues related to large model parameter number, low real-time performance, and high hardware cost. Experimental results demonstrated significant performance improvements for YOLOv5-Aircraft in both GPU and CPU environments while maintaining high-precision detection, offering a cost-effective solution for regional airports.

References

[1] Tang H, Zhang H. Study on IoT Big Data Direction in Civil Aviation[C]//International Conference on Artificial Intelligence and Security. Cham: Springer International Publishing, 2022: 300-309.

[2] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[3] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[4] Ajayi O G, Ashi J, Guda B. Performance evaluation of YOLO v5 model for automatic crop and weed classification on UAV images[J]. Smart Agricultural Technology, 2023, 5: 100231.

[5] Ruiz-Barroso P, Castro F M, Guil N. Real-Time Unsupervised Object Localization on the Edge for Airport Video Surveillance[C]//Iberian Conference on Pattern Recognition and Image Analysis. Cham: Springer Nature Switzerland, 2023: 466-478.

[6] Yang K, Dong B, Wu Y, et al. A Lightweight Yolov4 Network-Based Approach to Airport Field Surveillance[J]. Available at SSRN 4188696.

[7] Li W, Liu J, Mei H. Lightweight convolutional neural network for aircraft small target real-time detection in Airport videos in complex scenes[J]. Scientific reports, 2022, 12(1): 14474.

[8] Souza B J, Stefenon S F, Singh G, et al. Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV[J]. International Journal of Electrical Power & Energy Systems, 2023, 148: 108982.

[9] Wu W, Liu H, Li L, et al. 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): e0259283.

[10] Dong X, Yan S, Duan C. A lightweight vehicles detection network model based on YOLOv5[J]. Engineering Applications of Artificial Intelligence, 2022, 113: 104914.

[11] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[12] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.

[13] Huang Z, Ben Y, Luo G, et al. Shuffle transformer: Rethinking spatial shuffle for vision transformer[J]. arXiv preprint arXiv:2106.03650, 2021.

[14] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[15] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[16] Yang L, Zhang R Y, Li L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]//International conference on machine learning. PMLR, 2021: 11863-11874.

[17] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.

[18] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[19] Jocher G, Chaurasia A, Stoken A, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation[J]. Zenodo, 2022.