ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

Unmanned vending counter abnormal behavior recognition based on YOLOv5

Zhiyuan Wang[1], Yan Li[2], Bibo Lu[1,*], Lishan Zhao[1], Shisong Zhu[1], Yi He[3]

[1]College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
[2]Jiaozuo Metallurgy Building Materials Senior Technical School, Jiaozuo, China
[3]Henan Zhongyuan Zhixin Technology Limited Company, Jiaozuo, China
[*]Corresponding author

Abstract: *In the open-door unmanned vending machine monitoring scenario, a detection scheme for multiple states of user's hand is designed to analyze the abnormal behavior during the user's shopping process, and a hand multi-state detection algorithm based on YOLOv5 is proposed. To improve the inference speed of the algorithm, the 3×3 convolution in YOLOv5 is replaced with RepVGG by using the idea of structural re-referencing. The accuracy of the algorithm is improved by adding CBAM attention mechanism. The model size is greatly reduced while ensuring the model recognition accuracy. The experimental results show that the algorithm in this paper can accurately identify the hand state of the user when shopping and has some practical application value.*

Keywords: *deep learning, unmanned vending cabinet, YOLOv5, CBAM, model reparameterization*

1. Introduction

With the rapid development of technologies such as mobile payment, artificial intelligence and Internet of Things, a new retail model is gradually taking shape. At present, because of the fixed track size in the widely used closed-door unmanned container, the types and sizes of goods sold are limited, the space utilization rate is low, and users cannot exchange goods, so the interaction is poor. The emerging open-door unmanned container allows a variety of goods to be sold and can be changed at any time according to the season and place, which provides more business opportunities for merchants and more choices for users, and users can choose and change items, which increases the user's shopping experience and is widely welcomed by the market and users. However, there are many hidden dangers in open-door unmanned containers. If customers exchange goods maliciously, replacing high-value goods with low-value goods will cause economic losses to businesses, replacing normal food with expired food will cause food safety accidents, and putting other items at will will affect the reputation and image of businesses. Therefore, it is necessary to analyze the behavior of users in the shopping process. Because the hand must participate in the whole shopping process, the detection of the user's hand and its state is the key link to accurately analyze the user's normal/abnormal behavior.

In the unmanned open-door unmanned container scene, hand detection will be affected by camera angle, motion blur, occlusion and other problems, which is challenging. In this paper, by analyzing the specific requirements of this scene, a variety of hand state detection schemes are designed, and based on the single-stage target detection algorithm YOLOv5 model, CBAM attention mechanism is introduced, and the RepVGG structure with re-parameterization idea is adopted to quickly and accurately identify the user's hand state.

2. Problem background and solution design

2.1. Problem background and demand analysis

As shown in Figure 1(a), the open-door unmanned container considered in this paper has double doors on the left and right, and the five-story goods storage area can store a variety of goods. The camera is located above the cabinet and used to record the shopping video of users. A touch screen industrial control integrated machine is installed between the two doors, which can provide users with services such as code scanning and shopping information viewing, and can also provide storage space and computing power as edge computing equipment.

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

The working flow chart of the open-door unmanned container is shown in Figure 1(b). When the user scans the code to open the cabinet door, the camera starts to record the video, when the cabinet door is closed, the video recording ends, and after shopping, the user's shopping video is saved.

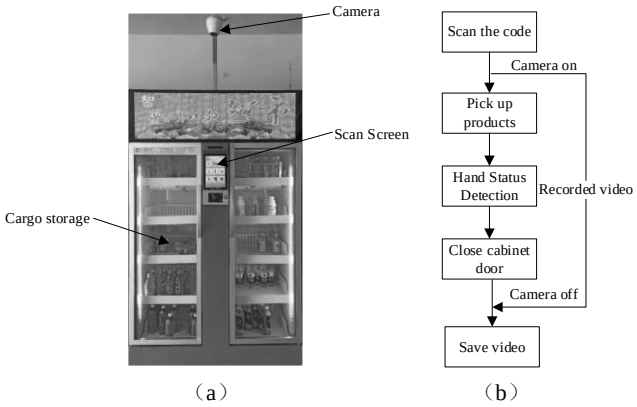


Figure 1: (a) open-door unmanned vending cabinet; (b) Container workflow diagram.

Considering the variety of open-door unmanned containers, different network environments, and the traffic cost of a large number of users uploading shopping videos to the cloud, it is necessary to put model reasoning and calculation on edge computing equipment to reduce data transmission and cost. At the same time, in order to reduce the waiting time of users, expect smaller model files and faster reasoning speed, the algorithm is lightweight and improved.

2.2. Scheme design

Users are involved in the whole shopping process, and they need to hold the unknown items in their hands when putting them into the container. According to the design purpose and scheme, the user's hand state in the whole shopping process is divided into three categories: empty-handed, hand-held goods and hand-held mobile phones. The specific reasons are as follows.

(1) Empty-handed: the state that the user has not taken the task items, including entering the container for sale empty-handed, or leaving the container for sale without taking the items. Because users must have hands to participate in shopping, hands are a necessary factor, and empty hands are taken as the detection target. This status label is set to "hand".

(2) Hand-held articles: The user holds and sells articles in the container, including taking out articles and returning articles. The detection of this state can be used for the trajectory analysis of subsequent analysis, which can analyze whether the goods are taken out of the unmanned container by the user or put other goods into the sold container by the user. If users put other items into the container for sale, they need to give an alarm to avoid food safety accidents. When taking goods or putting other items into the container, they use their hands to operate the items in the container, so they take the hand-held items as the detection target. This status label is set to "handled items".

(3) Handheld mobile phone: the state that the user holds the mobile phone. Because the shopping time of users is short, and the code scanning of mobile phones is needed during the shopping process, users hold mobile phones most of the time. Because unmanned containers do not sell mobile phones, in order to prevent the mobile phones from being mistaken for other commodities, the handheld mobile phones are also regarded as one of the detection states. This status label is set to "handled phone". Example of various states of that hand are shown in figure 2.

Figure 2: Example of hand state. (a) Hands; (b) Handheld items; (c) Handheld phone.

According to the equipment configuration and user demand of the open-door unmanned container,

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

combined with the characteristics of the shopping process, the specific implementation flow of the design scheme is shown in Figure 3. After the user finishes shopping, the hand state in the shopping video is detected and tracked to analyze the user's abnormal behavior. When the user's behavior is normal, a log is recorded. If the user's behavior is judged to be abnormal, an alarm is given to the backstage personnel, and an abnormal video is uploaded.

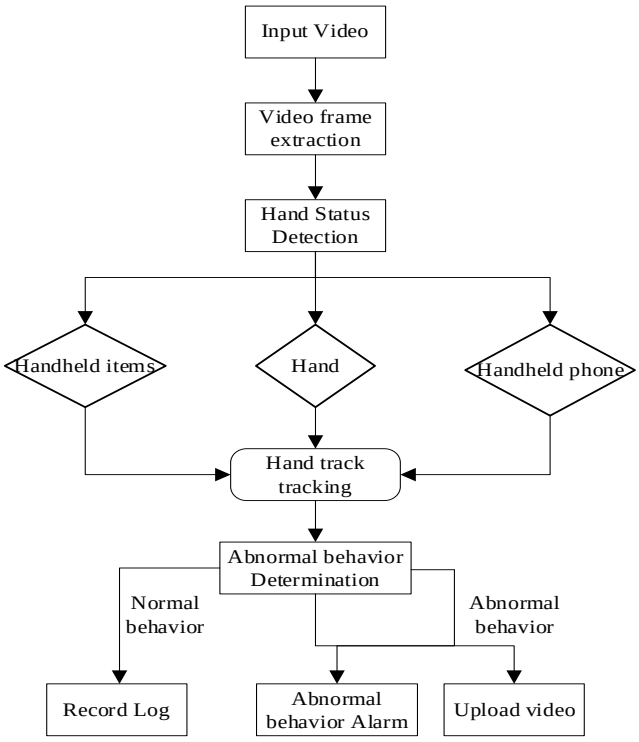


Figure 3: Flow chart of user abnormal behavior analysis scheme.

The difficulty of hand state detection lies in the mutual occlusion in the process of hand-object interaction, which makes the algorithm prone to miss detection or false detection. The feature extraction ability of the model should be further improved when designing the algorithm. Due to the different habits of users, the left and right hands of users are not detected. The algorithm is based on the terminal goal of unmanned containers, and the later deployment of the algorithm must be taken into account when designing the algorithm, so the model size and real-time detection should be guaranteed when designing the algorithm.

3. Target detection algorithm for numanned containers

3.1. YOLOv5 algorithm

YOLOv5 is a new generation target detection network with YOLO [1-4] architecture series. This network model has the advantages of high detection accuracy, lightweight and faster detection speed. On the other hand, the weight file of YOLOv5 target detection network model is smaller, which is more suitable for real-time detection on embedded devices. YOLOv5 [5] is divided into four general modules, including Input terminal, Backbone network, Neck network and Head output terminal.

YOLOv5 input uses Mosaic data enhancement operation, which is more ideal in small target detection and suitable for small target detection in this data set. The Backbone network adopts Focus [6] structure, C3 [7] structure and SPP structure. The SPP module is a spatial pyramid pool, which aims to improve the receptive field of the network by transforming any size feature map into a fixed size feature vector. GIOU_Loss [8] is used as the loss function at the output. In the post-processing process of target detection, YOLOv5 adopts weighted NMS operation for multi-target box screening. In this paper, RepVGG lightweight model and CBAM attention module are integrated into YOLOv5s network to improve network accuracy and achieve lightweight model. The improved YOLOv5s network structure is shown in Figure 4.

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

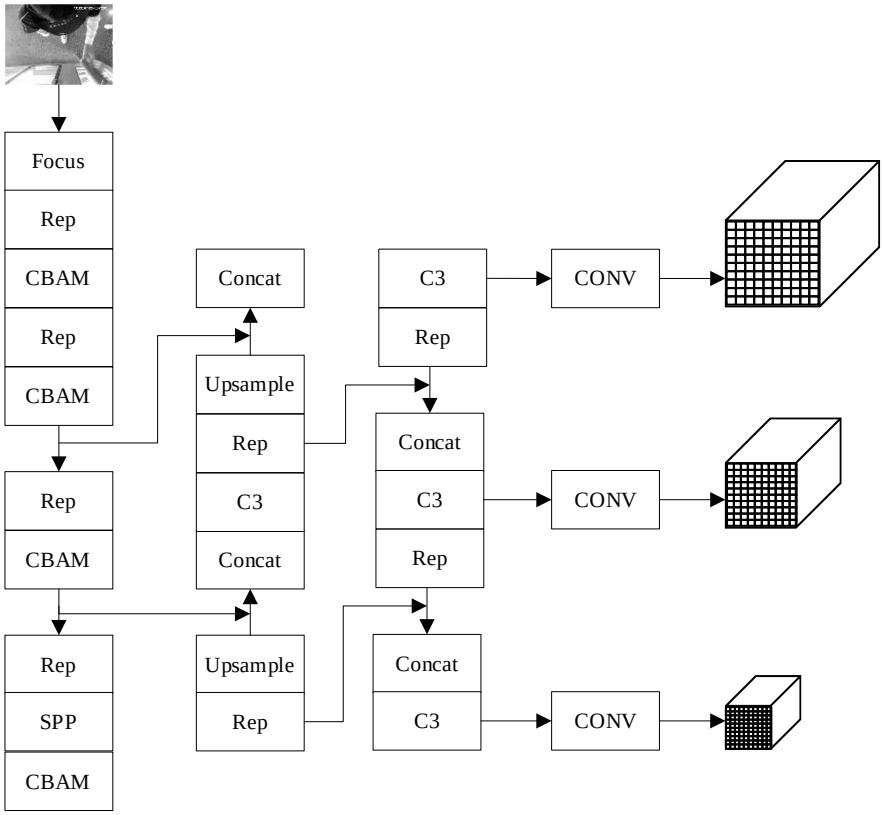


Figure 4: Improved YOLOv5s network structure.

3.2. RepVGG structure

Considering the deployment of the model on unmanned cabinets, the algorithm needs to take into account both the model size and the real-time detection. In this paper, the relatively lightweight YOLOv5 algorithm is adopted as the basic model. In target detection, YOLOv5 algorithm achieves the balance between detection accuracy and lightweight. However, when the algorithm is deployed on unmanned cabinets, the weight and reasoning speed of the model still cannot meet the required deployment requirements and the real-time detection of the algorithm. In order to improve the computing speed of the model, reduce the computing power requirements after the model is deployed, and meet the real-time detection, this paper integrates the idea of RepVGG [8] structure re-parameter into YOLOv5 network. Optimize the Backbone network of YOLOv5, and replace the original Conv module of the Backbone network with RepVGG module.

RepVGG network consists of 3×3 convolution, 1×1 volume integral branch and Identity [9] residual difference branch. The network structure of training stage and reasoning stage is shown in Figure 5. When the number of channels is c, the receptive fields of two 3×3 convolutions are the same as those of one 5×5 convolution, and the parameters of two 3×3 convolutions are 3×3×C×2=18C, and that of one 5×5 convolution is 5×5×C=25C. In contrast, the parameters of two 3×3 convolutions are reduced by 7C. Compared with 1×1 convolution, 5×5 convolution and 7×7 convolution, 3×3 convolution has higher computational density (theoretical calculation amount/time used). In the training process, the multi-branch model is used. In the training stage, the network is composed of 3×3 convolution, 1×1 volume integral branch and Identity residual connection, and ReLU is used as the activation function. In deployment and reasoning, a parallel 1×1 volume integral branch and Identity residual difference branch are added to each 3×3 convolution, and the 1×1 convolution convolution kernel is padding into 3×3 convolution, and Identity is equivalent to the convolution layer with special weight. According to the additivity of convolution, it is added with 3×3 convolution and then convolved with the feature map. The multi-branch model in the training stage is transformed into a single-branch model, which improves the operation speed and maintains the detection accuracy.

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

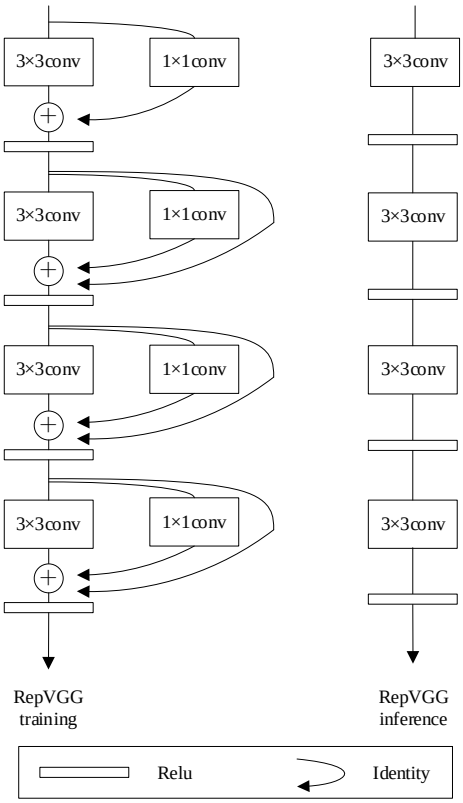


Figure 5: RepVGG structure.

3.3. YOLOv5 fusion attention mechanism

In the process of selling unmanned containers, due to the variety of goods sold and the different postures of holding the goods, the situation that the goods cover the hands is more frequent. In the detection process of YOLOv5 original algorithm, there will be cases of missing detection due to mutual occlusion of hands and objects. In order to improve the problem of missed detection caused by occlusion and improve the feature extraction ability of the algorithm, this paper integrates CBAM into YOLOv5 algorithm to realize the weight distribution of feature maps, pay attention to important features, suppress irrelevant features and improve the overall accuracy of target detection.

Convolutional attention module (CBAM) is divided into channel attention module and spatial attention module, which is a simple and effective attention module for feedforward convolutional neural networks. CBAM [10] module infers attention maps along two independent dimensions: space and channel. Firstly, the feature map is input to the channel attention module, and the corresponding attention map is output. Then, the inferred attention map is multiplied by the input feature map for adaptive feature optimization. When the output passes through the spatial attention module, the same operation is carried out, and finally the output feature map is obtained. Channel attention module applies average pooling and maximum pooling to compress the spatial dimension of feature map respectively. The spatial attention module applies average pooling and maximum pooling along the channel dimension respectively. Maximum pooling only considers the largest element, ignoring other elements in the pooled area and retaining more texture information of the image. Average pooling calculates the average value of all elements in the pooled area to retain more image background information. The structure of CBAM is shown in Figure 6.

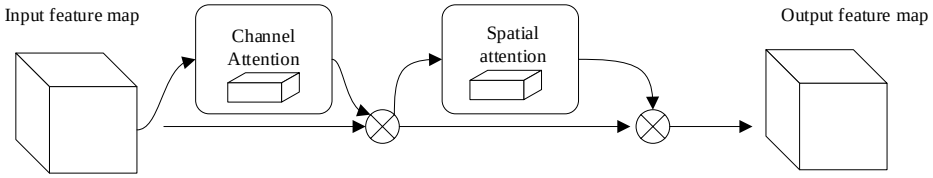


Figure 6: CBAM network structure.

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

4. Target detection algorithm for unsold containers

The experimental environment of this paper is as follows: Windows operating system, CPU is Intel(R) Core(TM) i9-10900K, GPU is NVIDIA Ge-Force RTX 3090, deep learning framework is Pytorch1.10 and CUDA10.2. In this paper, ablation and contrast experiments were carried out in the above environment.

4.1. Data set and evaluation index

Text the data sets used in this paper are all collected by ourselves. Randomly collect a number of user shopping videos on the open door vending cabinet, covering various scenes such as day, night and different light. The data set is marked with three categories, and the number distribution of categories is shown in Table 1. When training the data, the user's shopping videos are divided into single-frame images, and a total of 7970 original data sets are obtained. The data set is randomly divided into training set and test set according to the ratio of 8: 2. An example of some images is shown in Figure 7.

Table 1: Data set annotation information statistics table.

Category	Handheld items	Hand	Handheld phone
Number of tags	4769	10650	4820
Proportion	23.6%	52.6%	23.8%

Figure 7: Example of data set.

In this paper, Recall, mean Average Precision (mAP), weight and FPS are used to evaluate the performance of the model. Recall is defined as the proportion of all targets detected by the model, which is used to measure the recall rate of the model. MAP is the average of average Precision, which is the average of the average precision (AP) of the model in multiple detection categories. The calculation of mAP needs two indicators: precision and Recall. The weight size is the size of the weight file trained by YOLOv5s. FPS represents the number of video frames processed by the target network per second, that is, the number of frames transmitted per second. The more frames per second, the smoother the picture displayed in the video. The PR curve made with recall as the abscissa and precision as the ordinate, the area enclosed under the PR curve is the average accuracy (AP), and the average value of all kinds of AP is the mAP. The formulas are as follows:

$$AP = \int_0^1 P(r)dr \tag{1}$$

In formula (1), P is Precision and r is Recall.

$$mAP = \frac{\sum_{n=1}^{C} AP(n)}{C} \tag{2}$$

In formula (2), C is the total number of categories for target detection.

4.2. YOLOv5s ablation experiment

In this experiment, YOLOv5s was ablated with self-collected data sets to verify the improved strategy

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

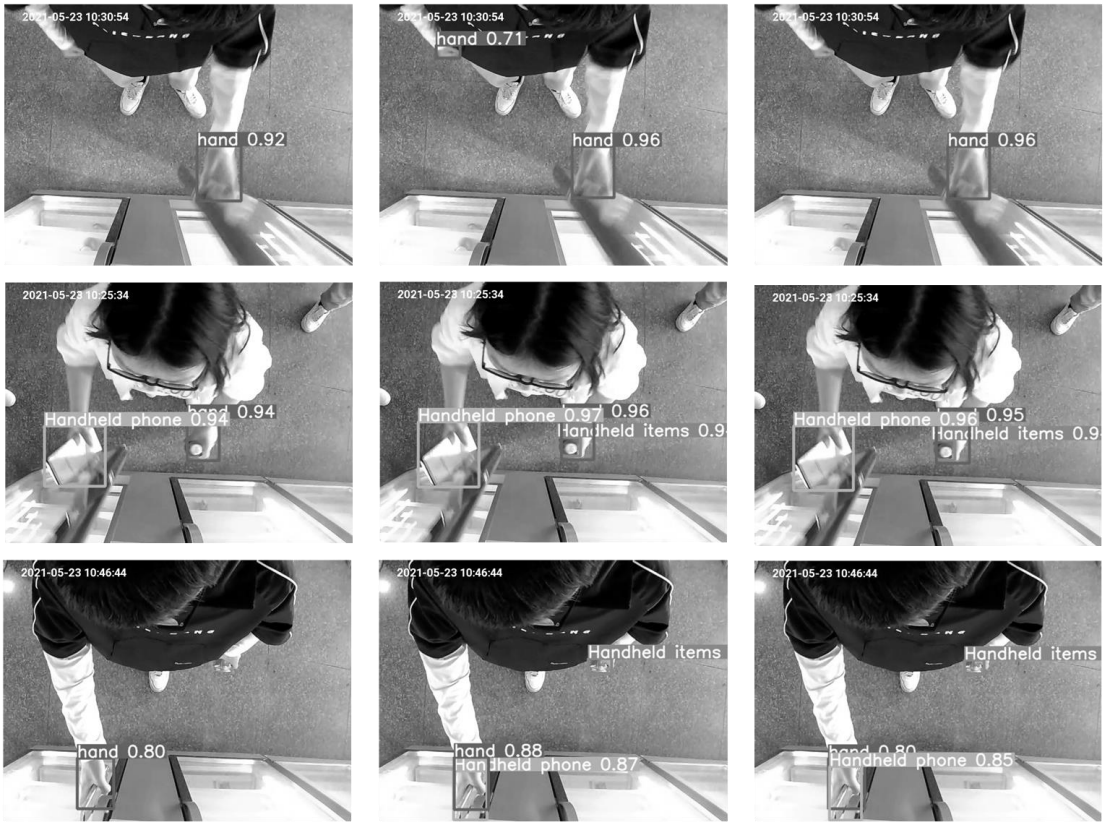
proposed in this paper. According to the experimental data and operating environment, the corresponding parameters are adjusted. The initial learning rate is set to 0.01, the momentum factor is set to 0.937, the Batch size is set to 8, and the Epochs is set to 300. The experimental results are shown in Table 2.

The first line of Table 2 shows the running results of the data set on the original YOLOv5s, with the weight of 14.4MB, FPS of 89 and mAP of 0.908. From the experimental results, it can be seen that after the integration of RepVGG network, compared with the original YOLOv5s network, the weight is reduced to 7.7MB, the FPS is increased to 126, and the mAP is increased to 0.913. After the introduction of CBAM attention module, the mAP of the model is obviously improved on the original YOLOv5s and the YOLOv5s network after adding RepVGG. The experimental results show that the improved network greatly reduces the weight, improves the reasoning speed and improves the detection accuracy.

Table 2: This caption has one line so it is centered.

RepVGG	CBAM	Recall	Weight/MB	mAP@0.5	FPS
		0.88	14.4	0.908	89
	√	0.899	14.5	0.934	81
√		0.888	7.7	0.913	126
√	√	0.882	7.8	0.92	109

In order to analyze the influence of CBAM and RepVGG on the detection results, this paper evaluates the performance of the improved model through the test set. Fig. 8 shows the comparison of the detection results of some test sets between the original algorithm and the improved algorithm. Fig. 8(a) shows the detection results of the original YOLOv5s algorithm. It can be seen that the algorithm is prone to miss detection when occlusion or small targets appear in the process of hand-object interaction. Fig. 8(b) is the detection result after introducing CBAM into YOLOv5s. For occluded and small targets, the CBAM module realizes the weight distribution of feature maps, and pays more attention to important features. In the image missed by the original algorithm, the algorithm after introducing CBAM has good detection effect on occluded and small targets. Fig. 8(c) shows the detection results after introducing CBAM and RepVGG into YOLOv5s. Compared with the original YOLO V5 S and the algorithm after introducing CBAM, the improved algorithm has greatly improved the detection effect on occluded targets and small targets.



(a)YOLOv5s *(b)YOLOv5s+CBAM* *(c)YOLOv5s+CBAM+RepVGG*

Figure 8: Comparison of ablation experimental results.

ISSN 2616-5775 Vol. 6, Issue 4: 110-117, DOI: 10.25236/AJCIS.2023.060415

4.3. YOLOv5s contrast experiment

In order to verify the effectiveness of the original algorithm in this paper, three target detection networks, Faster RCNN, SSD and YOLOv3, are selected for comparison, and the four networks are trained and tested on the same data set. The experimental results are shown in Table 3. On the whole, the improved YOLOv5s algorithm is superior to the other three networks in mAP, weight and FPS. Table 3 shows the comparative experimental results of YOLOv5s.

Table 3: YOLOv5s contrast experiment.

Models	mAP	Weight/MB	FPS
Faster RCNN	0.922	108.3	27
SSD	0.844	91.6	54
YOLOv3	0.892	66.3	75
YOLOv5s	0.908	14.4	89
Ours	0.920	7.8	109

5. Conclusion

In this paper, based on the YOLOv5 target detection algorithm, we detect the hand states of unmanned vending machine users during the shopping process in order to analyze the abnormal behavior of users more intelligently. By incorporating the attention module and the lightweight improvement of the model, the detection accuracy of the model is improved on the basis of the original algorithm, and the inference speed of the model is increased; combined with the mAP of the network, the model size and the inference time, the algorithm can accurately identify the user's hand state in the unmanned kiosk scenario.

This paper detects the hand states of users in the open-door unmanned vending cabinet scenario to support the next step of analyzing the abnormal behavior of users during the shopping process. The next stage of research focuses on: (1) the specific setting of abnormal behavior; (2) the extraction of hand motion trajectory; (3) the analysis of the abnormal behavior of the user during the shopping process based on the hand state and motion trajectory.

References

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].
[2] Redmon J, Farhadi A. YOLO9000: better, faster, stronger; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2017 [C].
[3] Redmon J, Farhadi A. Yolov3: An incremental im-provement [J]. arXiv preprint arXiv:180402767, 2018.
[4] Bochkovskiy A, Wang C-Y, Liao H-Y M. Yolov4: Optimal speed and accuracy of object detection [J]. arXiv preprint arXiv:200410934, 2020.
[5] Ultralytics. YOLOv5 [EB/OL]. (2021-04-12) [2022-4-20]. https://github.com/ultralytics/yolov5.
[6] Deng L, Gong Y, Lu X, et al. Focus-Enhanced Scene Text Recognition with Deformable Convolutions; proceedings of the 2019 IEEE 5th International Conference on Computer and Communications (ICCC), F 6-9 Dec. 2019, 2019 [C].
[7] Wang C-Y, Liao H-Y M, Wu Y-H, et al. CSPNet: A new backbone that can enhance learning capability of CNN; pro-ceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, F, 2020 [C].
[8] Ding X, Zhang X, Ma N, et al. Repvgg: Making vgg-style convnets great again; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2021 [C].
[9] He K, Zhang X, Ren S, et al. Identity Mappings in Deep Residual Networks; proceedings of the Computer Vision – ECCV 2016, Cham, F 2016//, 2016 [C]. Springer International Publishing.
[10] 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.