2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) | 979-8-3503-2548-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/PRAI59366.2023.10332084

An Improved Soft-CBAM-YoloV5 Algorithm for Fruits and Vegetables Detection and Counting

Qian Luo
Department of Soft Engineering
South China Normal University
Foshan, Guangdong
20202031018@m.scnu.edu.cn

Zhen Zhang
Department of Soft Engineering
Huizhou University
Huizhou, Guangdong
zzsjbme@sjtu.edu.cn

Cuimei Yang
Department of Soft Engineering
Huizhou University
Huizhou, Guangdong
meikoyoung@gmail.com

Junran Lin*
Department of Soft Engineering
Huizhou University
Huizhou, Guangdong
zzhbkjdx@163.com

* Corresponding author.

***Abstract*—The quality of people's life has been improved as science and technology keep progressing, and artificial intelligence has been implemented into more and more fields these days. Among them, many researchers have paid attention to the scenes of quoting prices and settling accounts for fruits and vegetables in supermarkets and manual classification of fruits and vegetables in farmers' markets. In the traditional YOLOV5 method, when the same kind of fruits and vegetables are blocked from each other, the counting will be inaccurate, resulting in missed detection. In addition, the environment where different objects stay will affect the feature extraction of objects. In order to reduce the occurrence of missed detections as much as possible and improve the detection performance of the model, this paper proposes a new detection model counting method named Soft-CBAM-YOLOV5, by reasonably reducing the score of adjacent detection boxes and introducing the attention mechanism CBAM (Convolutional Block Attention Module) to prevent the occurrence of missed detection and strengthen the focus on the detection target to improve the feature extraction ability of the algorithm. Compared with traditional YOLOV5, the mAP value of Soft-CBAM-YOLOV5 increases from 97.05% to 97.71%. Consequently, the Soft-CBAM-YOLOV5 model has greater detection performance, robustness and lower missed detection rate.**

Keywords—Fruits and Vegetables Detection, YOLOV5, NMS, CBAM, Soft-CBAM-YOLOV5.

I. INTRODUCTION

As the largest country in terms of output of fruits and vegetables, there are countless people who buy fresh fruits and vegetables in supermarkets and farmers' markets every day in China. After the consumers have selected the commodities they need, they need to give them to the staff to input the corresponding number of the category on the scale and then weigh and label them, which is quite complex and time-consuming. With the rapid development of science and technology and the gradual rise of urban construction level, machine vision and artificial intelligence are applied more and more in urban informatization and digitization. Fruits and vegetables recognition is a typical application of machine vision information processing. However, after the application of the existing self-checkout machines, there are still a large number of people who need to queue up to weigh, resulting in the low purchase efficiency. In order to simplify the sales of fruits and vegetables in supermarkets and farmers' markets, so as to improve consumers' purchase efficiency, sales and market competitiveness, the effective way to solve this problem is to develop fully automatic AI products with the functions of fruits and vegetables classification, counting, weighing and payment, and to improve the performance of automatic recognition and classification of fruits and vegetables using machine vision.

In recent years, as the pace of people's life is getting faster and faster, many supermarkets are equipped with self-service cash registers to adapt to people's fast-paced life. However, there will still be a large number of queues in the supermarkets' fruits and vegetables zone. People desperately hope that supermarkets can change this situation and improve their purchasing efficiency. As artificial intelligence continues to develop, AI makes their dreams come true. In the field of fruits and vegetables recognition [1], researchers have been studying algorithms of fruits and vegetables recognition since the traditional vision-based fruits and vegetables classification algorithms came out. At present, the most effective way to solve the problem of fruits and vegetables recognition is to use deep learning. In this paper, YOLOV5 method is used to realize fruits and vegetables recognition. Based on Soft-NMS algorithm and CBAM attention mechanism, the missed detection and detection in different environments are optimized, so that the Soft-CBAM-YOLOV5 model can have higher accuracy and reduce the missed detection rate. This study is further compared with all the current well-known deep learning methods, which effectively proves that Soft-CBAM-YOLOV5 is more suitable for counting in fruits and vegetables recognition.

II. CONTENT OF WORK

Fruits and vegetables recognition is a fundamental research in target detection, and many research methods have been proposed in this field to solve such problems. As early as 1996, Bolle [2] et al. have developed the first fruits and vegetables recognition system Veggie-Vision, the system extracts image, size, shape and texture features of fruits and vegetables, and inputs the extracted features into the nearest neighbor classifier

for classification. In the traditional fruits and vegetables recognition algorithm, Rocha [3] et al. have also provided fruits and vegetables recognition methods based on different classifiers, but the algorithm still requires manual feature extraction.

With the continuous development of artificial intelligence, deep learning came into being. At present, the application of deep learning in target detection has been relatively mature [4], for example, accomplishment have been achieved in traffic sign detection [5], medical image detection [6], pavement defect detection [7]. Deep learning can extract image features, and achieve image classification and recognition through a lot of training. Therefore, deep learning is a very effective technology in target detection. Currently, target detection algorithms of deep learning are mainly divided into two categories. The first is multi-stage algorithm such as Mask R-CNN [8], SPP-net [9] (ROI Pooling), Fast R-CNN [10] (Selective Search + CNN + ROI), Faster R-CNN [11] (RPN + CNN + ROI), R-FCN [12], etc. The second is a single-stage algorithm, including YOLOV1 [13], YOLOV2 [14], SSD [15], YOLOV3 [16], YOLOV4 [17], YOLOV5 [18], YOLOX [19], DenseBox [20] and so on. As deep learning becoming prosperous, more researchers have begun to use deep learning to classify fruits and vegetables. In 2016, Sakai [21] et al. used deep neural network (DNN) for target classification and recognition through target extraction and learning, and used it for fruits and vegetables recognition with CNN for learning. In 2019, Dengfeng Chen used Faster R-CNN to detect fruits and vegetables [22]. According to previous researches, most of the researches are based on traditional machine vision algorithm. The accuracy of it is usually 80% to 90%, and there is still a gap in practical applications. In addition, compared with Faster R-CNN, the YOLOV5 model can achieve fast inference while ensuring detection performance, which can be more advantageous in practical applications. However, the accuracy of the model will be reduced in the case of targets of the same type occluding each other. Zhen Zhang et al. have improved YOLOV4 according to the mutual occlusion of similar objects [23]. In order to further improve the model accuracy and model detection performance, we propose an improved YOLOV5 scheme in this paper, which is called Soft-CBAM-YOLOV5.

In the traditional YOLOV5 model, there are still some problems in Non-Maximum Suppression (NMS): if an object overlaps another object in different ways, that is, when the two target boxes are close, the box with lower score will be deleted because the overlapping area is too large, resulting in the detection failure of the object and reducing the average detection rate of the algorithm which leads to missed detection [24]. The Soft-NMS algorithm reduces the score of the adjacent detection box instead of setting it to 0 directly. As long as the score of the adjacent detection box is greater than a certain threshold, the final output will contain the target object. The detection effect of the target under different background information will be affected by the background information. Although the background information is also a part of the target detection object, the problem can be better solved by introducing the CBAM attention mechanism [25].

The organization of this paper is as follows. Section 1 introduces the application of the system described in this paper and solves the corresponding problems. Section 2 introduces the progress of related research work in this field. Section 3 introduces the theory and implementation of YOLOV5, NMS, Soft-NMS, CBAM and the dataset used in experiment. Section 4 presents the comparison before and after improvement with Soft-NMS and CBAM, and the comparison of Soft-CBAM-YOLOV5 with several other deep learning methods. The conclusion is in Section 5.

III. Methods

A. The YOLOV5 Model

The YOLOV5 model is particularly well-known in the field of object detection, and is widely used in urban target detection [26], pedestrian detection [27] and other situation due to its advantages of fast response and high precision. The following describes the YOLOV5 model, and Fig. 1 is the structure of YOLOV5. Among them, we divide the YOLOV5 network structure into the following four parts: Input, Backbone, Neck network, and Prediction.

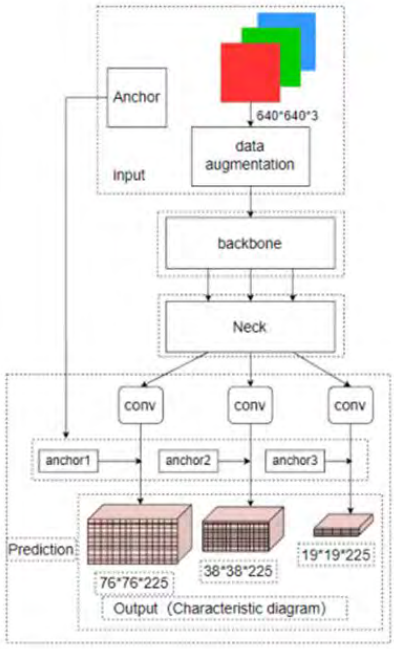

Fig. 1. The structure of YOLOV5

① Input: The input represents the input images. The input images' size of the network is 640×640, and this stage usually includes image preprocessing, that is, scaling the input images to the input size of the network, and normalize them and so on. When the network is training, the data augmentation operation of Mosaic [28] is used in this paper to improve the training speed of the model and the accuracy of the network, thereby making the model have better generalization ability.

② Backbone: The benchmark network is usually a network with excellent performance. This module is used to extract some general feature representations. YOLOV5 not only uses CSPDarknet53 structure, but also uses the Focus structure as the benchmark network.

③ Neck: The Neck is usually located in the middle of the benchmark network and the head network, and it can be used to further improve the diversity and robustness of features. The SPP module and the FPN+PAN module are used in the YOLOV5 algorithm.

④ Prediction: Prediction is the output of the Head, and the Head is used to complete the output of the target detection. For different detection algorithms, the number of branches at the output varies, usually including a classification branch and a regression branch. In YOLOV5, GIOU_Loss [29] is used to replace the Smooth L1 Loss function to further improve the detection accuracy of the algorithm.

This paper mainly focuses on Yolov5l, while Yolov5m, Yolov5s and so on are network deepening or simplifying on the basis of Yolov5l.

B. NMS Algorithm and Soft-NMS Algorithm

Non-Maximum Suppression (NMS) is used in target detection to extract target detection boxes with high confidence and suppress false detection boxes with low confidence. Generally speaking, when the analytical model is output to the target box, there will be a large number of target boxes. The specific number is determined by the number of anchors. Many duplicate boxes are located to the same target. NMS is used to remove these duplicate boxes and obtain the real target box. However, the elimination mechanism of NMS is very strict, and only the detection frame and its IOU (Intersection over Union) are considered in calculation, so it is easy to cause missed detection. Fig. 2 shows the missed detection of the target object.

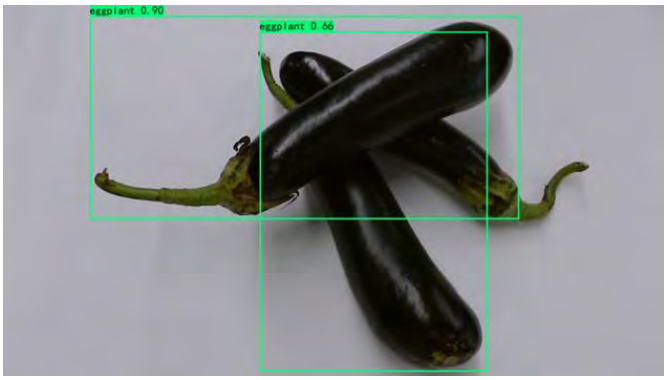

Fig. 2. Missed detection occurred when using NMS

We can obviously see from Fig. 2 that one eggplant was missed in detection. In the improvement, the key step of counting is to detect the target. When similar objects block each other, it is easy to cause missed detection. Therefore, we use Soft-NMS instead of NMS in the original YOLOV5 model to solve this problem.

From a mathematical point of view, the mechanism of NMS removing redundant detection boxes with mathematical principles is explained as follows:

$$score_i = \begin{cases} 0, IOU(M, b_i) \geq threshold\ of\ IOU \\ score_i, IOU(M, b_i) < threshold\ of\ IOU \end{cases} \quad (3\text{-}1)$$

The $score_i$ is the score of the current detection frame. In the dataset of this experiment, after a lot of tests, we found that the best threshold for IOU is 0.5.

In the course of the experiment, it was found that when the detection box with higher IOU is adjacent to the detection box with the highest score, NMS will set the score of the detection box to 0, and then delete it. In the case shown in Fig. 2, it is likely to cause missed detection. Soft-NMS can solve this problem well, and its mechanism of removing redundant detection frames is expressed as follows:

$$score_i = score_i e^{-\frac{IOU(M,b_i)^2}{\theta}} \quad (3\text{-}2)$$

It can be seen from the Gaussian function that Soft-NMS will not directly set the score of the frame to 0 when the detection frame with higher IOU is adjacent to the detection frame with the highest score. On the contrary, Soft-NMS adopts a penalty mechanism that penalizes the confidence score. We use a Gaussian function as the following weight function: $e^{-\frac{IOU(M,b_i)^2}{\theta}}$ (θ is the parameter of the weight function. After a lot of debugging, the best detection comes with θ being 0.1). The larger the area overlapping the detection box with the highest score, the more severe the score reduction of the detection box. Finally, in this way, Soft-NMS can effectively remove redundant detection boxes and reduce the occurrence of missed detection. The flowchart of the Soft-NMS algorithm is shown in Fig. 3.

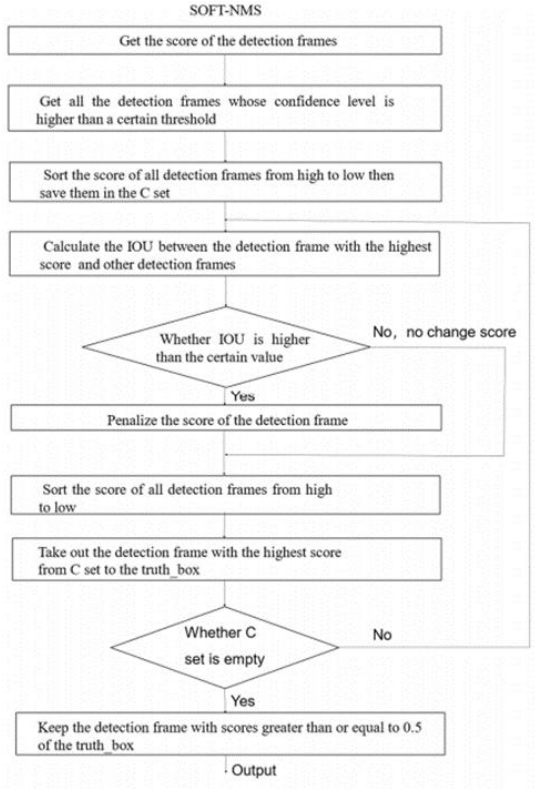

Fig. 3. Flowchart of Soft-NMS

The main ideas of Soft-NMS are as follows:

① Sort all the detection boxes in the set R according to their scores from high to low. The higher the score, the higher the probability that the detection box belongs to this category, and then the detection box H with the highest score is selected from the set R.

② Traverse all detection boxes in the set R, and calculate the IOU between each detection box and the detection box H with the highest score. Soft-NMS does not directly remove certain detection box from set R, but penalizes it accordingly to reduce the score. The larger the area overlapping the detection box with the highest score, the more severe the reduction of the detection box's score. Finally, save the detection box H into the truth_box.

③ Go back to the first step until the set R is empty. Finally, keep detection boxes with scores greater than or equal to 0.5 in truth_box as output.

After being processed by Soft-NMS, the missed detection can be improved and corrected. The results are shown in Fig. 4.

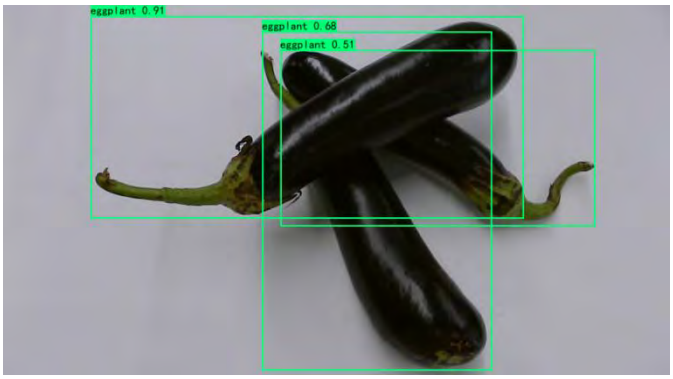

Fig. 4. No missed detection after using Soft-NMS

C. Attention mechanism of CBAM

For our visual system, we pay more attention to the main information in the image and ignore some irrelevant information. Similarly, we add the attention mechanism in the process of deep learning, so that the algorithm could judge what we focus on. For the input image, we mainly focus on the detection of fruits and vegetables, and the rest is background information, while the previous algorithm also convolved the background information during convolution, which not only increased the amount of calculation, but also slowed the operation and increased the amount of tasks, and the final prediction accuracy decreases, so we add an attention mechanism, which can not only reduce the computational load of the network, but also increase the accuracy of the prediction model. CBAM is a lightweight module that we can add to the YOLOV5 network architecture. Given a feature map, CBAM sequentially infers an attention map along two independent dimensions, namely channel and space, and then multiplies the attention map with the input feature map to perform adaptive feature refinement. During our operation, the attention mechanism can make us pay more attention to the information of fruits and vegetables, while ignoring other irrelevant information. The structure of the CBAM module is shown in Figure 5.

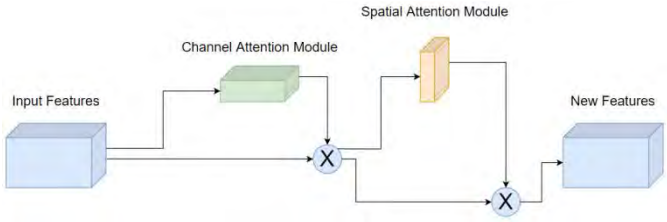

Fig. 5. Structure Diagram of CBAM Module

D. Introduction of datasets

The existing fruits and vegetables datasets (such as fruit360, Fig. 6) and the data crawled from the Internet by using the crawler technology (Fig. 7) are not in line with the actual use of the system (Intelligent settlement cashier and farmers' market), and cannot be applied to model training. The dataset studied in this paper is collected according to the actual use requirements of the system. The members of our research team used cameras to simulate the intelligent settlement scene, and photographed fruits and vegetables from different angles, different light and shadow and different types. The advantage of this is that the trained model can make predictions in different situations with better robustness and generalization ability.

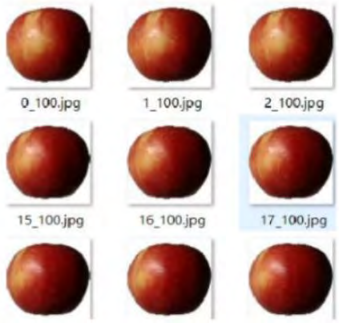

Fig. 6. The fruit360 dataset

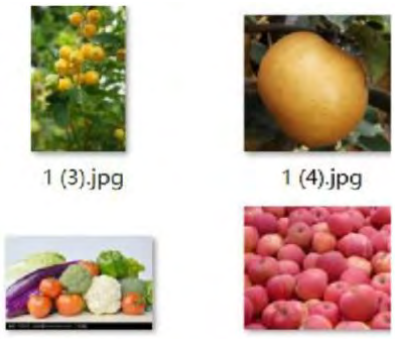

Fig. 7. Images from the Internet

The collected dataset needs to be labeled in Pascal VOC format. The size of each image is 1280x720. For the YOLOV5 model, the size of the input image is 640x640. Therefore, there will be image preprocessing at the input of the model. The processed dataset contains 6882 images, including 17 kinds of fruits and vegetables of cucumber, cauliflower, broccoli,

tomatoes, green peppers, bell peppers, eggplants, carrots, apples, and oranges. Finally, it is divided into training set, validation set and test set according to the ratio of 6:2:2.

IV. Results

To verify the validity of the Soft-CBAM-YOLOV5 model, the same prediction parameters and dataset are used for the validation of YOLOV5 and Soft-CBAM-YOLOV5. Among them, these two models use NMS and Soft-NMS respectively in terms of NMS, and use no-attention mechanism and CBAM in terms of attention mechanism respectively. In the verification process, the detection effect and performance of the model are reflected through the evaluation indicators.

The mAP value of the YOLOV5 model before improvement is 97.05%, the Precision is 95.41%, and the Recall is 97.55%. When Soft-NMS is used and CBAM is introduced, the prediction results of the Soft-CBAM-YOLOV5 model are improved, with the mAP value of 97.71%, the Precision of 98.13%, and the Recall of 95.59%.

The comparison of the two are shown in TABLE I. .

TABLE I. Comparison of before and after improvement

Model	mAP/%	Precision/%	Recall/%
Original YOLOV5	97.05	95.41	97.55
Soft-CBAM-YOLOV5	97.71	98.13	95.59

Through comparison and analysis, we can see that after using Soft-NMS and introducing CBAM, although the Recall is slightly reduced, the mAP and the Precision are improved compared to the original model. Among them, the Precision increased by 2.72%, and mAP increased by 0.66%. Since the value of Recall and Precision cannot comprehensively evaluate the effect of the algorithm, the mAP indicator is selected for evaluation. Experiments show that the mAP of Soft-CBAM-YOLOV5 is higher than that of Original YOLOV5, but the Recall is decreased. In this way, both the replacement of NMS with Soft-NMS and the introduction of CBAM in YOLOV5 are effective.

A. Comparison with State-of-the-Arts

The experiments include the following comparison methods: SSD, target detection algorithm based on YOLOV4 (abbreviated as YOLOV4), Fast R-CNN and target detection algorithm based on YOLOX (abbreviated as YOLOX [30]). All methods use the same evaluation index. It is not difficult to see that the Soft-CBAM-YOLOV5 model has improved performance compared with other algorithms. The detection results of each method on our dataset are shown in TABLE II.

TABLE II. Comparison of other deep learning methods

Model	mAP/%
SSD	91.84
YOLOV4	95.52
Fast R-CNN	96.58
YOLOX	96.36
Soft-CBAM-YOLOV5	97.71

B. Evaluation Metrics for Experiment

The metrics used to evaluate the model in this experiment include Precision value, Recall value, and mAP value. The calculation of Precision value and Recall value are expressed by formulas (4-1) and (4-2) respectively:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4\text{-}1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4\text{-}2)$$

In the above two formulas, TP means the prediction result is correctly classified as a positive sample, FP means wrongly classified as a positive sample, FN means wrongly classified as a negative sample. The mAP is the average AP of each type of target, and AP is calculated by Precision and Recall. Fig. 8 is the mAP before improvement, and Fig. 9 is the mAP after improvement. The higher the mAP, the better the prediction of the model. We can see that the mAP of improved model is better than original model.

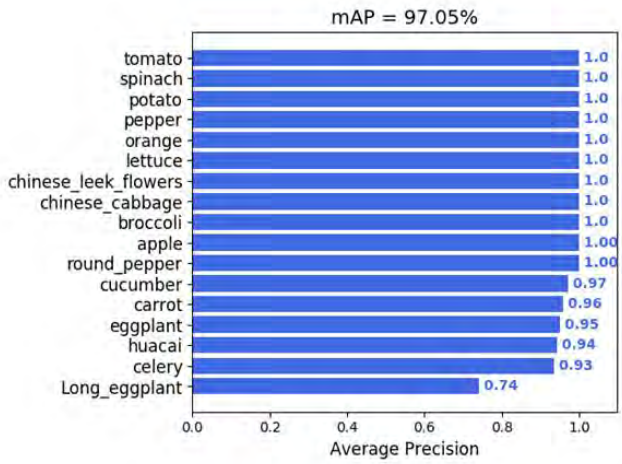

Fig. 8. The mAP value before improvement

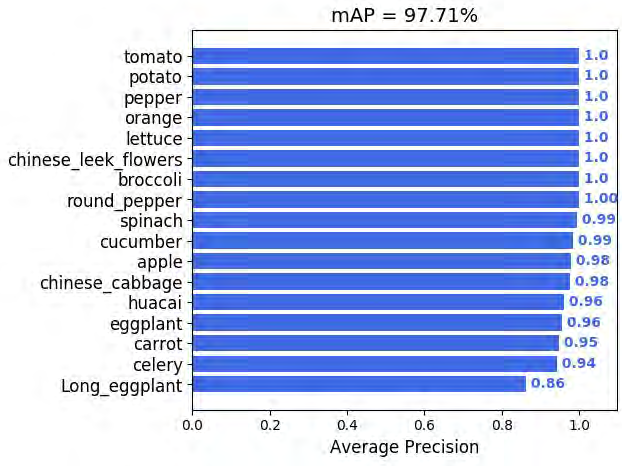

Fig. 9. The improved mAP value using Soft-NMS

V. CONCLUSIONS

Compared with other target detection models, the Soft-CBAM-YOLOV5 method proposed in this paper has better detection performance, robustness and lower missed detection rate. In the process of fruits and vegetables recognition, for the detection of similar objects occluding each other and different background information, the number of fruits and vegetables can be more accurately identified and counted, and the phenomenon of missed detection can be effectively avoided.

This paper proposes a new detection model Soft-CBAM-YOLOV5 by improving YOLOV5. Using the Soft-NMS algorithm can more accurately count fruits and vegetables, and the introduction of the CBAM attention mechanism can improve the model's ability to extract features, so as to obtain performance close to practical application requirements. The original YOLOV5 model uses the NMS algorithm to remove redundant detection boxes, while the YOLOV5 model proposed in this paper uses the Soft-NMS algorithm to penalize the scores of redundant detection boxes. After comparison and analysis, Soft-CBAM-YOLOV5 has higher accuracy and lower missed detection rate in fruits and vegetables recognition. The mAP value of Soft-CBAM-YOLOV5 is 97.71%, which is 0.66% higher than that of the YOLOV5 model. Therefore, Soft-CBAM-YOLOV5 is more suitable for counting in fruits and vegetables recognition.

REFERENCES

[1] Femling F, Olsson A, Alonso-Fernandez F. Fruit and vegetable identification using machine learning for retail applications[C]//2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE, 2018: 9-15.

[2] Bolle R M, Connell J H, Haas N, et al. Veggievision: A produce recognition system[C]//Proceedings Third IEEE Workshop on Applications of Computer Vision. WACV'96. IEEE, 1996: 244-251.

[3] Rocha A, Hauagge D C, Wainer J, et al. Automatic produce classification from images using color, texture and appearance cues[C]//2008 XXI Brazilian Symposium on Computer Graphics and Image Processing. IEEE, 2008: 3-10.

[4] Pak M, Kim S. A review of deep learning in image recognition[C]. 2017 4th international conference on computer applications and information processing technology (CAIPT). IEEE, 2017: 1-3.

[5] Wang C. Research and application of traffic sign detection and recognition based on deep learning[C]//2018 International Conference on Robots & Intelligent System (ICRIS). IEEE, 2018: 150-152.

[6] Li Z, Dong M, Wen S, et al. CLU-CNNs: Object detection for medical images[J]. Neurocomputing, 2019, 350: 53-59.

[7] Cao W, Liu Q, He Z. Review of pavement defect detection methods[J]. Ieee Access, 2020, 8: 14531-14544.

[8] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[9] Purkait P, Zhao C, Zach C. SPP-Net: Deep absolute pose regression with synthetic views[J]. arXiv preprint arXiv:1712.03452, 2017.

[10] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[11] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[12] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.

[13] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[15] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

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

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

[18] Yan B, Fan P, Lei X, et al. A real-time apple targets detection method for picking robot based on improved YOLOv5[J]. Remote Sensing, 2021, 13(9): 1619.

[19] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[20] Yu J, Jiang Y, Wang Z, et al. Unitbox: An advanced object detection network[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 516-520.

[21] Sakai Y, Oda T, Ikeda M, et al. A Vegetable Category Recognition System Using Deep Neural Network[C]// International Conference on Innovative Mobile & Internet Services in Ubiquitous Computing. IEEE, 2016.

[22] Deng C, Zhou Y, You D, et al. Intelligent fruit and vegetable settlement system based on computer vision [J]. Informatization research, 2019,45 (02): 65-70.

[23] Zhang, Z.; Xia, S.; Cai, Y. A Soft-YoloV4 for High-Performance Head Detection and Counting. Mathematics 2021, 9, 3096.

[24] Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft - NMS — Improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.

[25] 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.

[26] Jin H, Wu Y, Xu G, et al. Research on an Urban Low-Altitude Target Detection Method Based on Image Classification[J]. Electronics, 2022, 11(4): 657.

[27] Li F, Li X, Liu Q, et al. Occlusion Handling and Multi-scale Pedestrian Detection Based on Deep Learning: A Review[J]. IEEE Access, 2022.

[28] Zeng G, Yu W, Wang R, et al. Research on Mosaic Image Data Enhancement for Overlapping Ship Targets[J]. arXiv preprint arXiv:2105.05090, 2021.

[29] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 658-666.

[30] Lin, J.; Yang, C.; Lu, Y.; Cai, Y.; Zhan, H.; Zhang, Z. An Improved Soft-YOLOX for Garbage Quantity Identification. Mathematics 2022, 10, 2650. https://doi.org/10.3390/math10152650