REASEARCH ON PEAR INFLORESCENCE RECOGNITION BASED ON FUSION ATTENTION MECHANISM WITH YOLOV5
/
基于融合注意力机制的 YOLOv5 算法对梨花序的识别研究

Ye XIA [1, 2)], Xiaohui LEI [*2)], Andreas HERBST [3)], Xiaolan LYU [*2)] [1]

[1)] School of Agricultural Engineering, Jiangsu University, Zhenjiang/China;
[2)] Institute of Agricultural Facilities and Equipment, Jiangsu Academy of Agricultural Sciences / Key Laboratory of Modern Horticultural Equipment, Ministry of Agriculture and Rural Affairs, Nanjing/China;
[3)] Institute for Chemical Application Technology of JKI, Braunschweig Messeweg/Germany;
Tel: +86-025-84390082; E-mail: leixiaohui.2008@163.com

DOI: https://doi.org/10.35633/inmateh-69-01

Keywords: *agricultural machinery, pear inflorescence, object detection, YOLOv5, coordinate attention mechanism*

ABSTRACT

Thinning is an important agronomic process in pear production, thus the detection of pear inflorescence is an important technology for intelligentization of blossom thinning. In this paper, images of buds and flowers were collected under different natural conditions for model training, and the images were augmented by data augmentation methods. Model training was performed based on the YOLOv5s network with coordinate attention mechanism added to the backbone network and compared with the native YOLOv5s, YOLOv3, SSD 300, and Faster-RCNN algorithms. The mAP, F1 score and recall of the algorithm reached 93.32%, 91.10%, and 91.99%. The model size only took up 14.1 MB, and the average detection time was 27 ms, which are suitable for application in actual intelligent blossom thinning equipment.

摘要

疏花疏果是梨生产中一项重要的农艺环节，而梨树花序识别是疏花智能化过程中的一项重要技术。本文在不同的自然条件下采集了梨树花苞与花朵的图像，通过数据增强方法对图像进行了扩充。在 YOLOv5s 主干网络中增加 CA 注意力机制，将其与原生 YOLOv5s、YOLOv3、SSD 300、Faster-RCNN 算法进行对比。结果表明，改进的 YOLOv5s-CA 模型优于其他模型，其 mAP、F1 得分、Recall 分别达到了 93.32%、91.10%和 91.99%，模型大小仅为 14.1 MB，平均检测时间 27 ms，更适用于智能化疏花设备中。

INTRODUCTION

Under natural conditions, the number of blossoms in pear trees is far more than that of final fruit. Blossom thinning can save tree nutrients, avoid biennial bearing and improve fruit quality. At present, the method of pear tree thinning mainly adopts manual operation with mechanical tools, which is time-consuming with difficulty in controlling the thinning precision. Therefore, an intelligent blossom thinning method is necessary for the future development of blossom thinning equipment, especially the pear inflorescence detection algorithm based on deep learning.

With the rapid development of deep learning technology in agriculture, smart agriculture has ushered in a new era. The mainstream target detection algorithms can be divided into two-stage and one-stage algorithms according to the recognition process. Two-stage algorithms are represented by R-CNN series *(Girshick R. et al., 2013; Ren S. et al., 2016)* and SPP-Net *(He K. et al., 2014)*. This type of algorithm has a large number of parameters, and has the characteristics of high recognition accuracy but with slow recognition speed, and usually used for the tasks requiring high detection accuracy. One-stage algorithms filter the possible detection targets and return the category and classification of them. The mainstream one-stage algorithms include YOLO series *(Redmon J. et al., 2016; Bochkovskiy A. et al., 2020)* and SSD *(Wei L. et al., 2016)*, which have fast detection speed and are more suitable for deployment in embedded devices but with poor computing capacity.

[1] *Ye Xia, Master Candidate; Xiaohui Lei, Assistant Researcher; Andreas Herbst, Prof.; Xiaolan Lyu, Researcher.*

YOLO has been widely used in the field of target detection because it ensures the detection speed while taking into account the recognition accuracy. *Fan Z. et al., (2022),* compared several YOLO algorithms for the recognition of cucurbit fruits, and the results show that the YOLOv5 achieved the optimal effect, with mAP reaching 97.1% and fps reaching 90.9. *Yan L. et al., (2022),* used YOLOv4 to study the recognition of occluded orange fruits on trees, and the results showed that its detection accuracy reached 98.17%, showing a better effect on lightly occluded targets. *Jin Y., (2020),* used visual servo technology to identify the tomato in fruit picking task. The results show that the method can effectively identify and locate tomatoes during the picking process. *Wang D. et al., (2021),* used the YOLOv5 algorithm improved by the channel pruning method to detect multi-variety apple fruits. The average accuracy of the collected data set reached 95.8%, and the model size was reduced to 1.4 MB. *Farjon G. et al., (2020),* used Faster-RCNN for transfer learning and marked the flower information of apple by professional planters to realize the discrimination of different blooming intensities. The results showed that the model mAP reached 68%, and the flowering degree recognized by the model was similar to that of human discrimination in the range of 78% to 93%.

The researches above have detected the blossoms and fruits of different crops or trees with various target detection methods, but there is few related research on the detection of pear blossoms.

In this paper, pear inflorescence was taken as the research object for detection in different environments based on the improved YOLOv5s network, to provide a target positioning method for improving the operation accuracy of intelligent blossom thinning equipment in orchards.

MATERIALS AND METHODS

Image acquisition

The inflorescence data of pear trees were collected from March 10, 2022 to March 30, 2022 in Nan Jing, China. The pictures were collected in the environment of insufficient light, shadows, and overlapping occlusions. The data were captured by a smartphone, and the pictures were saved as *.jpg format with a resolution of 1920*1080. The pear trees in the data collection site were planted in horizontal trellis and the inflorescence on a single branch was used as the unit. The method of pear inflorescence data acquisition is shown in Fig. 1. A total of 1566 original pictures were collected from the sample data. The dataset was divided into 1200 training sets and 366 validation sets. Raw data was manually annotated by LabelIMG software.

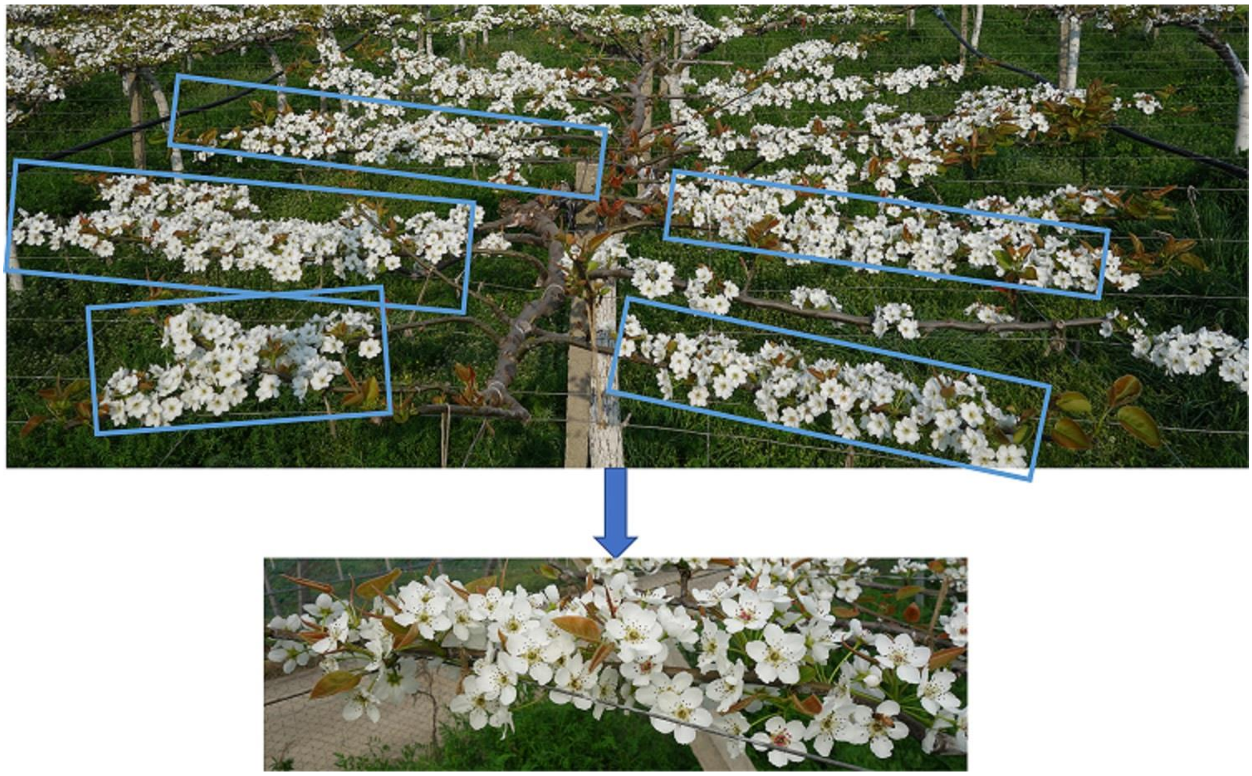

Fig. 1 - Pear inflorescence data acquisition

Data augmentation

Due to the complex shape of pear inflorescence, a data augmentation strategy was adopted for the original data set to avoid overfitting. The processing program was compiled through the OpenCV library, and the original data was transformed by changing brightness, rotation angle, adding Gaussian noise, translation, mirroring, and clipping. The specific data augmentation logic is shown in Fig. 2.

Set the total transformation times and cycle after the pictures are input. In a cycle, each factor has a 50% probability of appearing until the set transformation times are met. The specific strategy was random rotation of 0 to 70°, random crop of 2*150*150 pt, random Gaussian noise of 0.1 to 0.2, brightness adjustment from -30% to +30%, random flip horizontally, vertically and on the original spot, 100 pt to 250 pt random shift in horizontal and vertical directions. In order to save the time for repeated manual labelling, the program adopts the same position transformation method for the objects that have been labelled in the original data, so as to directly generate the expanded data containing the annotation information. In the process of augmentation, the original data were expanded by 4 times, and each picture was subjected to 6 random transformation factors to generate its own quadruple-enhanced data. The dataset after data augmentation reached 7830 images.

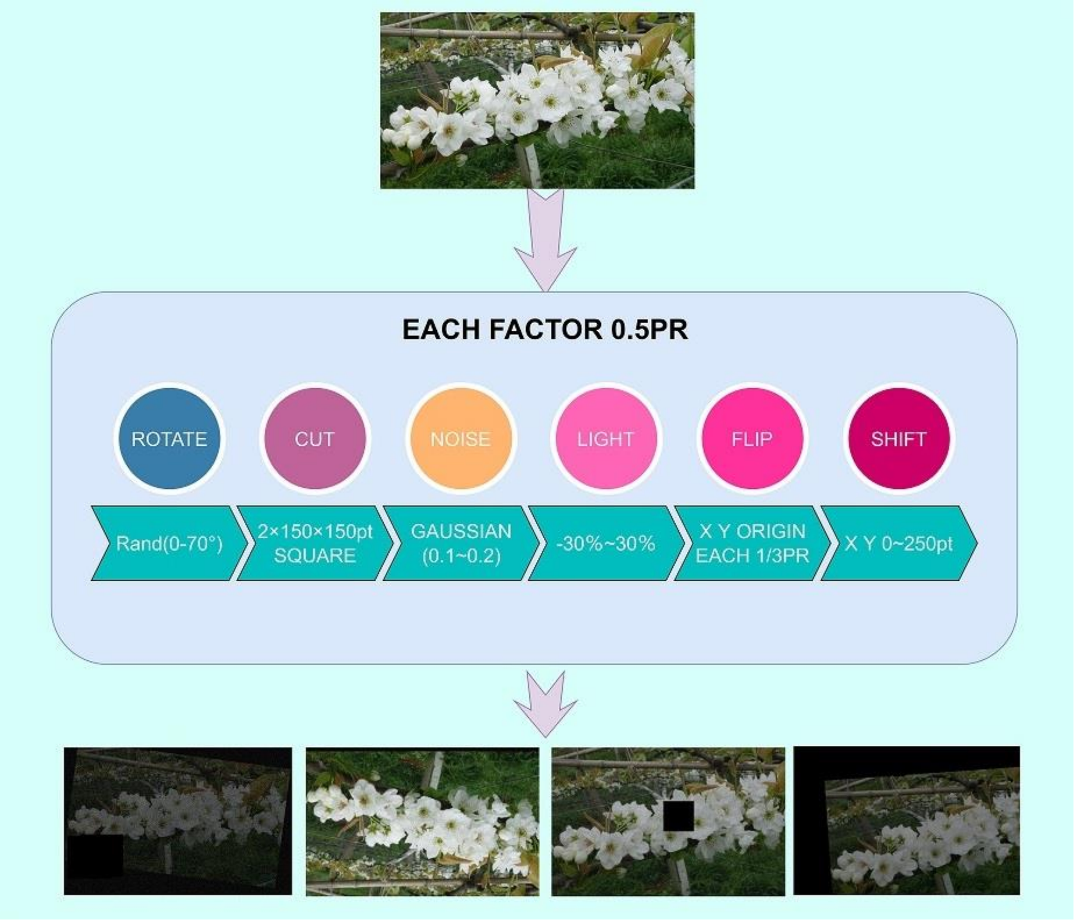

Fig. 2 - Data augmentation logic

The YOLOv5 algorithm

The YOLO algorithm transforms the object detection problem into a probabilistic regression problem by dividing the image into a finite number of anchor boxes and continuing to predict the edge box part of each anchor box. Through YOLO, the category and estimated probability of the target can be directly obtained, which greatly improves the detection speed compared with the two-stage detection network of RCNN. The standard version of YOLO can process 45 frames per second in real time, while the smaller, lighter version of YOLO can process 155 frames per second *(Shafiee M. et al., 2017)*. YOLOv5 uses the CSPDarkNet53 backbone network on the basis of the original framework. Compared with the DarkNet53 used in the previous version, CSPDarkNet53 can divide the feature map of the base layer into two parts, and merge them through the cross-stage hierarchy, which reduces the amount of calculation.

The Neck network of YOLOv5 adopts feature pyramid network (FPN) and pixel aggregation network (PAN) structure, with strong semantic and localization features. By fusing the two structures, the bidirectional aggregation of the features of different layers of the backbone network is achieved. The prediction head part makes predictions by performing probability regression judgment on the generated feature maps of different sizes using grid-based anchor boxes. The YOLOv5 algorithm organizes the network in a modular way, and the feature extraction module mainly includes convolution, batch normalization and SiLU (CBS), center and scale prediction (CSP), fast spatial pyramid pooling (SPPF). The CBS module mainly used in backbone network consists of the convolutional layer (Conv), the normalization layer (BN), and the SiLU activation function layer. The residual structure contained in it can increase the gradient value of back-propagation between layers to avoid gradient disappearing due to the deepening of the network. Finer features can be extracted without worrying about network degradation. The features of the input image are extracted by the convolution layer, normalized by the BN layer to speed up the network learning speed, and finally the features are retained and mapped by the activation function. The CSP_X consists of one CBS and X residuals, which are concatenated. SPPF first continues to pool one part of the features obtained after the maximum pooling of the small-sized feature map, and the other part is spliced with the results obtained by each pooling layer. Then the features are extracted by the CBS module and converted into fixed-size features. Let the network fuse local and global features of different sizes, and the structure diagram of each module and the entire network is shown in Fig. 3.

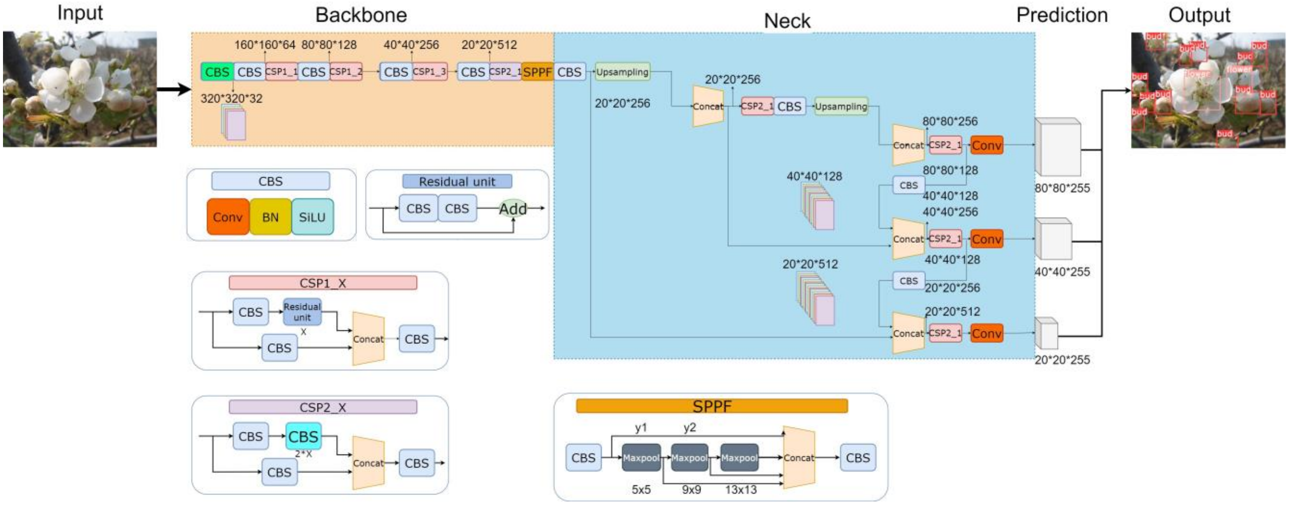

Fig. 3 – YOLOv5 network for pear inflorescence detection

The coordinate attention (CA) mechanism

The attention mechanism is a resource allocation scheme that allocates more resources to more important tasks when the computing power of the hardware platform is limited. The attention mechanism in CNN makes the network pay more attention to the information that is more critical to the current task by assigning weights to the feature maps of different parts, reducing the attention to irrelevant information or directly filtering irrelevant information, so as to improve the efficiency of the whole network processing task. Common attention mechanisms are mainly divided into two categories: spatial attention mechanism and channel attention mechanism. Typical attention mechanisms are SE-Net *(Jie H. et al., 2017)*, DRAW *(Gregor K. et al., 2015)*, which often pay attention to single-feature information, but in specific tasks, it is often necessary to take into account both attention mechanisms above in order to improve the attention to effective information.

CA mechanism *(Hou Q. et al., 2021)* is an attention mechanism embedded in mobile networks, which can transform any intermediate feature tensor in the network and output a tensor of the same size. The CA structure first divides the input feature maps into two directions by width and height, and performs global average pooling on them to obtain feature maps in both width and height directions. Then, the feature maps in the two directions of the global receptive field are spliced together for convolution, normalization, and activation operations.

Then a similar operation is performed, that is, convolve the feature map according to the original height and width to obtain a feature map with the same number of channels as the original, and then extract the attention weights in the width direction and height direction through the activation function. Finally, weighted fusion of feature maps in two directions is performed to obtain feature maps with attention weights in two directions. The CA structure is shown in Fig. 4.

For the pear inflorescence recognition task in this paper, both channel features and spatial features are meaningful, so this paper considers introducing a CA module into the network to improve the network's attention to target information. The specific way is to add the CA mechanism before the SPPF module after the last convolution operation of the backbone network part of YOLOv5s. Since the input and output feature maps of the CA module are of the same size, this addition does not change the network structure.

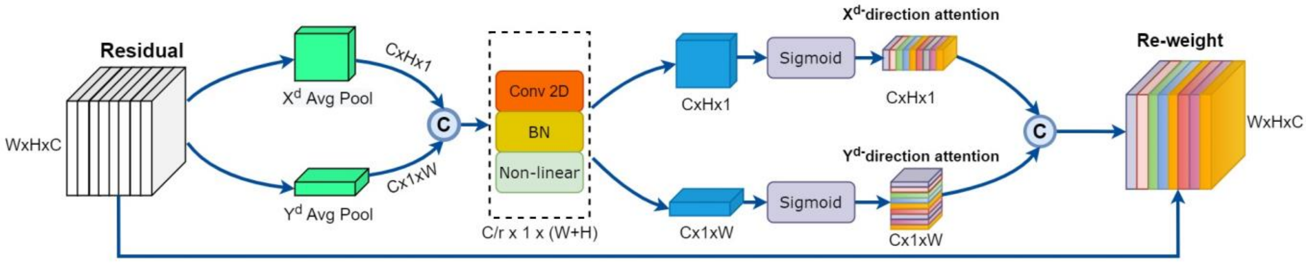

Fig. 4 – Structure of CA mechanism

Model training and evaluation indexes

The test hardware platform uses a desktop server equipped with Intel® Core E5 V3 CPU, 32 GB running memory, and 12 GB GeForce RTX 3090 GPU. The image input size is 640×640 pixels, the training epoch is set to 1000 rounds, the learning rate is initially 0.001, and hyperparameter evolution is used to dynamically adjust the learning rate through the loss rate of each round to speed up network training.

The accuracy evaluation mainly relies on the indicators of Precision, Recall, Mean Average Precision, and F1 score, which are calculated by formulas (1) to (5), respectively. In order to distinguish the recognition performance of the model, the parameter quantity, size and average run-time of the model are also important indexes for evaluation. The average run-time in this paper is obtained from running on the training platform.

$$P = \frac{TP}{TP + FP} \times 100\% \tag{1}$$

where:

P is precision, [%]; TP is true positive; FP is false Positive.

$$R = \frac{TP}{TP + FN} \times 100\% \tag{2}$$

where:

R is recall, [%]; FN is false negative.

$$AP = \int_0^1 P(R)dR \tag{3}$$

where:

AP is average precision of a class, [%].

$$mAP = \frac{\sum AP}{N(Class)} \tag{4}$$

where:

mAP is mean average precision, [%]; N is number of classes.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{5}$$

where: $F1$ is F1 Score, [%].

RESULTS AND DISCUSSION

In order to verify the superiority of the model that added CA mechanism, a comparative test between the original YOLOv5s network and the YOLOv5s network added with CA module (YOLOv5s-CA) was carried out. The performance indexes are shown in Tab. 1.

Table 1

Performance indexes of YOLOv5s and YOLOv5s-CA

Model	Class	Precision %	Recall %	F1 Score %	AP %	Parameter quantity	Average Run-Time / ms
YOLOv5s	Bud	87.33	84.16	85.72	88.54	7014617	21
	Flower	91.15	89.92	90.53	90.43		
YOLOv5s-CA	Bud	91.54	88.47	89.98	92.36	7041167	27
	Flower	94.27	93.73	94.00	94.27		

Tab. 1 shows that, the precision of YOLOv5s-CA network on the bud and flower object was increased by 4.21% and 3.12% respectively, and the recall was increased by 4.31% and 3.81% respectively, while the model parameter quantity was only increased by 0.3%. Compared with the quantity of model parameters, the precision and recall were greatly improved. In order to evaluate the contribution of CA mechanism to the network, the visual effect comparison of the two networks was carried out through gradient-weighted class activation mapping (Grad-CAM). Grad-CAM can convey network attention to the detection target through a heat map. The higher the heat value, the higher the attention of the network. Fig. 5 shows the heat map of target recognition using Grad-CAM for YOLOv5s and YOLOv5s-CA networks. The heat of the bud in Fig. 5c is higher than that in Fig. 5a, and obviously reduces in irrelevant background. For Fig. 5b and Fig. 5d, Fig. 5d removes the attention from the irrelevant background, while retaining the higher heat for the correct target and the area with high heat is better focused on the correct target area. Therefore, the network added CA module reduces attention to irrelevant information and increases the accuracy of the model. It can extract the overall feature information better than native YOLOv5.

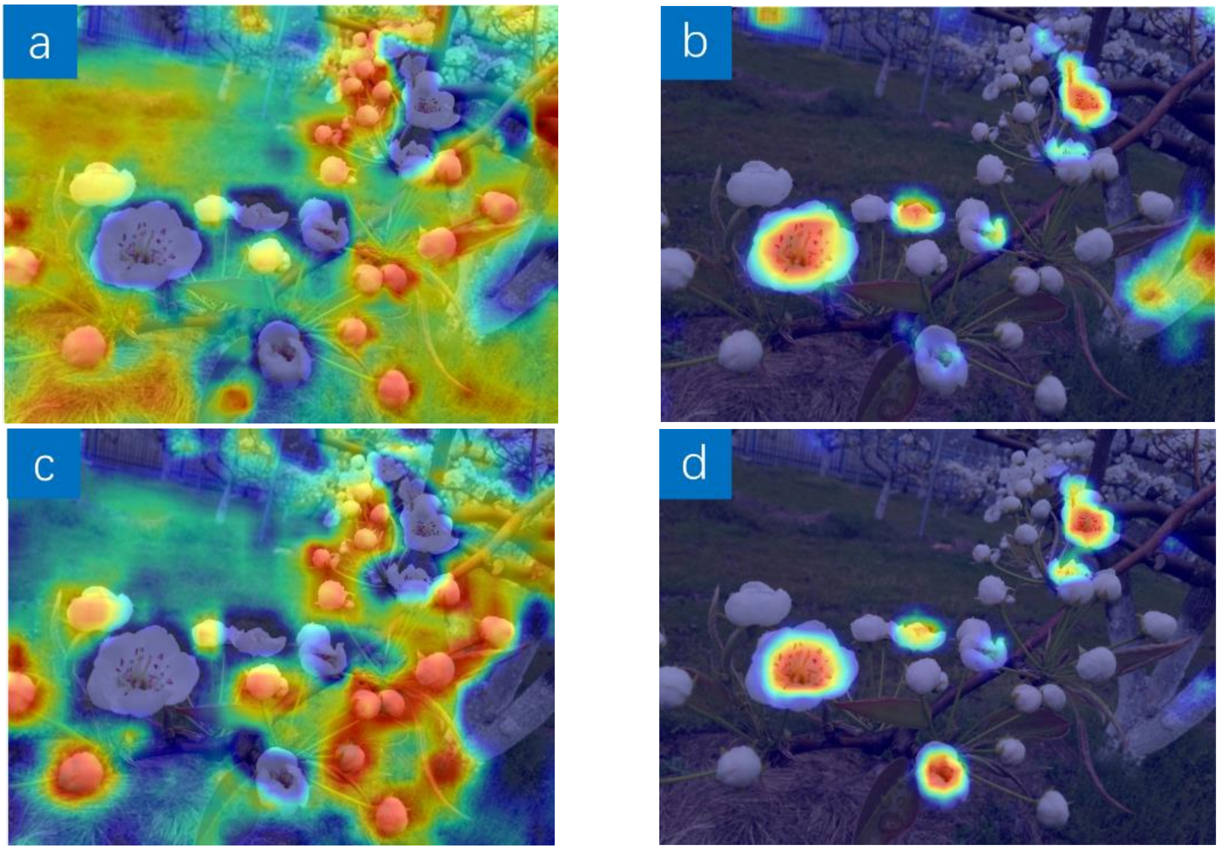

Fig. 5 – Network heat map

a, b represent the attention to buds and flowers without CA attention network;
c, d represent the attention to two types of targets after adding CA attention.

In order to compare the advantages of the model relative to other target recognition algorithms, general models SSD 300, Faster-RCNN and YOLOv3 were selected for performance comparison. In the test, the same data set and training platform were used for the 4 algorithms, and test results are shown in Tab. 2.

Table 2

Performance indexes of different detection algorithms

Algorithm	mAP %	Recall %	Model Size MB	Average Run-Time ms
SSD 300	81.53	82.76	36.71	25
Faster-RCNN	90.15	86.32	214.65	187
YOLOv3	89.71	87.46	206.36	23
YOLOV5s-CA	93.32	91.10	14.17	27

Tab. 2 shows that the mAP and recall rate of YOLOv5s-CA network is 93.32% and 91.10%, respectively, showing higher precision and recall rate than other algorithms. Compared with Faster-RCNN, the mAP and Recall of the YOLOv5-CA model were increased by 3.17% and 4.78%, respectively. At the same time, the model size was reduced by 200.48 MB and the detection speed was increased by 160 ms. Compared with SSD 300, which is also a one-stage algorithm, the mAP and recall of YOLOv5s-CA were increased by 11.79% and 8.34% respectively, the model size was reduced by 61%, and the detection time was increased by 2 ms. Compared with YOLOv3, the mAP and recall of YOLOv5s-CA was improved by 3.61% and 3.64%, the model size reduced by 192.19 MB, and the detection time increased by 4 ms. Although the detection accuracy of the two-stage algorithm was satisfactory, the detection speed was significantly lower than that of the one-stage algorithm. Therefore, two-stage algorithm is not suitable for intelligent blossom thinning equipment. In the one-stage algorithm, although YOLOv3 has better mAP and recall, its trained model is large and not suitable for actual deployment. The SSD 300 algorithm has a slight advantage in detection speed compared with YOLOv5s-CA, but it is almost negligible compared with the disadvantages of other indexes. In conclusion, YOLOv5s-CA algorithm has more advantages in actual deployment.

Pear trees are planted in relatively open orchards, and the actual detection environment is complex and has strong interference. Fig. 6 shows an example of the model recognition of some buds and flowers. From Fig. 6, it can be found that the model can effectively identify targets with different light intensities, shadows, occlusions, heterochromatic stamens, and slightly blurred targets.

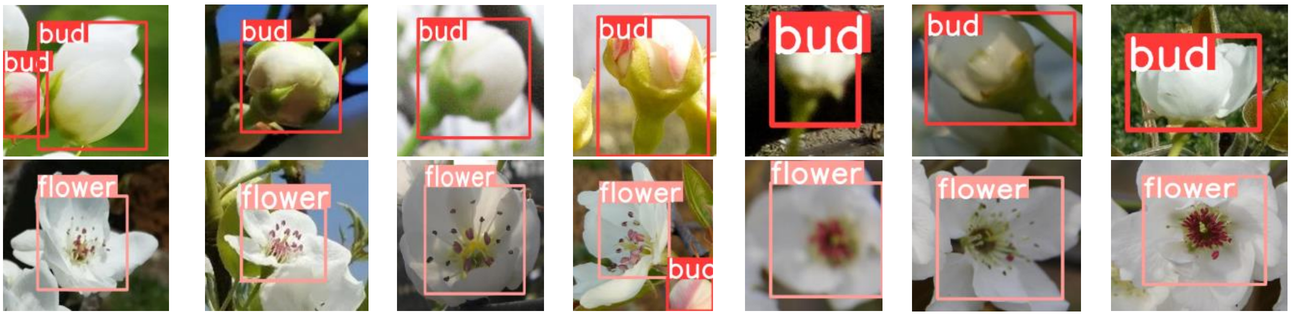

Fig. 6 – Bud and flower recognition

Although the model can work in most cases, there are many unrecognized cases in the actual application process. For example, the processing of small targets in target detection greatly affects the overall detection performance of the model. Fig. 7 shows the recognition effect of the model in this paper on small targets. When too many small objects appear, the model missed many objects. From the principle of the model, the reason is that the PAN feature fusion structure of the YOLOV5 network only fuses features of three different sizes. From the actual test effect, some targets can be detected using the feature fusion method of the original YOLOv5, but if these small targets need to be accurately detected, the current model is insufficient. In order to improve the problem of detection missing of small targets in the network, in future research, feature layers in the feature fusion network can be added according to the actual detection tasks and detection heads for small-scale targets in the detection heads of the network can also be added to improve the detection ability.

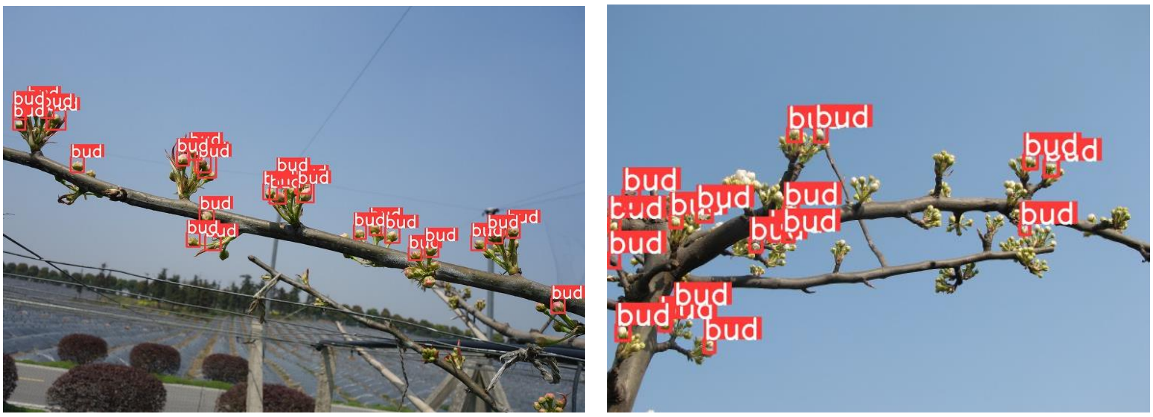

Fig. 7 – Small target detection

Similarly, the image blurring caused by camera performance also has a great impact on the target detection, as shown in Fig. 8. When the same branch is detected, the camera blurs the distant target and misses the detection. In the actual detection task, blurring will distort the characteristics of the target to a large extent, which will undermine the detection effect. In response to this problem, in this paper, Gaussian noise in the data enhancement part was added to improve the model processing capability. From the second picture of Fig. 8, it can be observed that some fuzzy targets were detected. However, due to the complexity of the specific recognition scene, there were still many virtual objects not detected. In order to improve the problem of detection missing caused by background blur, for future acquisition equipment, professional cameras without blur function can be adopted for data acquisition and detection to reduce detection missing caused by background blur.

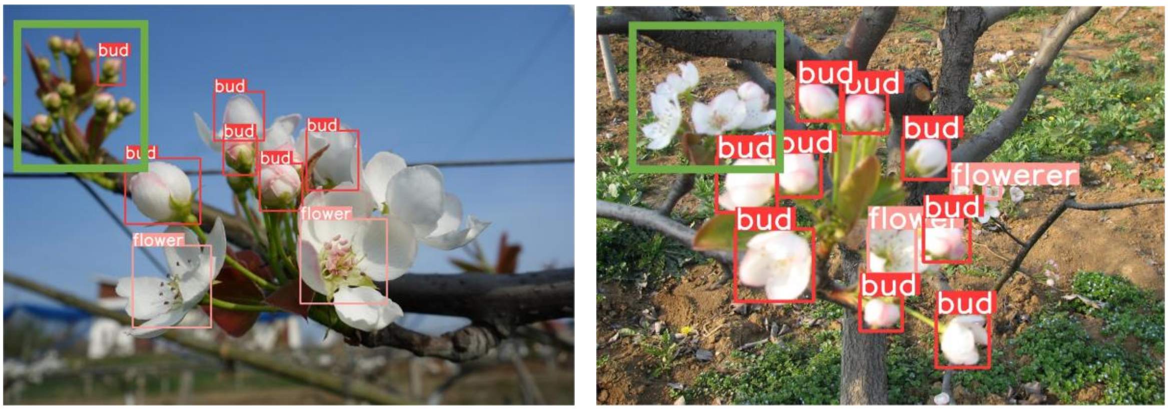

Fig. 8 – Image blurring

CONCLUSIONS

Thinning is a necessary agronomic section in pear orchard management. It can avoid biennial bearing and improve fruit quality. Intelligent blossom thinning method is necessary for the future development of blossom thinning equipment, especially the pear inflorescence detection algorithm based on deep learning. In this paper, an improved YOLOv5s model incorporating CA mechanism was proposed for identification of pear inflorescence. It can improve the network attention to the target information and improve the detection accuracy and efficiency.

(1) A total of 1566 original pictures were collected from the sample data. The pictures were captured in a real pear orchard environment with ideal conditions and possible lack of light, shadows, and overlapping occlusions. The augmentation program was compiled through the OpenCV library, and the original data were transformed by changing brightness, rotation angle, adding Gaussian noise, translation, mirroring, and clipping. The dataset after data augmentation reached 7830 images. Through data enhancement, more possible actual detection scenarios were expanded, and the generalization ability of the model was improved.

(2) The results of pear inflorescence test show that the improved YOLOv5s-CA model had a mAP of 93.32% and a recall of 91.10%. Compared with the native YOLOv5s network, the CA mechanism effectively increased the network's receptive field for the detection target, so that the network obtained higher mAP and Recall. Comparing YOLOv5s-CA with YOLOv3, SSD 300 and Faster-RCNN, the results show that the one-stage detection algorithm had better performance than the two-stage detection algorithm in the intelligent flower thinning recognition task. Among several one-stage algorithms, the YOLOV5s-CA model had better mAP and Recall. At the same time, the size of the model was only 14.1 MB, and the average detection speed was only 27 ms, which is more suitable for use in intelligent flower thinning embedded devices.

ACKNOWLEDGEMENT

The authors appreciate the financial support provided by National Natural Science Foundation of China (Grant No.32201680), China Agriculture Research System of MOF and MARA (Grant No. CARS-28-21), Jiangsu Modern Agricultural Machinery Equipment and Technology Demonstration Extension Fund (Grant No. NJ2022-14), National Science and Technology Development Program of China (Grant No. NK2022160104), Jiangsu Policy-guided Plans (Grant No. BX2019016) and Wuxi Science and technology development Fund (Grant No. N20221003).

REFERENCES

[1] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. https://doi.org/10.48550/arXiv.2004.10934

[2] Fan, Z. H. A. O., Zhang, J., Zhang, N., Zhiqiang, T. A. N., Yonghao, X. I. E., Zhang, S., ... & Mingbao, L. I. (2022). Detection of cucurbits fruits based on deep learning. *INMATEH-Agricultural Engineering*, *66*(1), pp.321-330. https://doi.org/10.35633/inmateh-66-32

[3] Farjon, G., Krikeb, O., Hillel, A. B., & Alchanatis, V. (2020). Detection and counting of flowers on apple trees for better chemical thinning decisions. *Precision Agriculture*, *21*(3), pp.503-521. https://doi.org/10.1007/s11119-019-09679-1

[4] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 580-587.

[5] Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. In *International conference on machine learning,* pp.1462-1471. PMLR.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[7] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 13713-13722.

[8] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 7132-7141.

[9] Jin, Y. (2020). Recognition technology of agricultural picking robot based on image detection technology. *INMATEH-Agricultural Engineering*, *62*(3). https://doi.org/10.35633/inmateh-62-20

[10] Li, Y., Xiao, L., Li, W., Li, H., & Liu, J. (2022). Research on recognition of occluded orange fruit on trees based on YOLOv4. *INMATEH-Agricultural Engineering, 67(2)*. https://doi.org/10.35633/inmateh-67-13

[11] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 779-788.

[12] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. https://doi.org/10.48550/arXiv.1804.02767

[13] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[14] Shaifee, M. J., Chywl, B., Li, F., & Wong, A. (2017). Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video. *Journal of Computational Vision and Imaging Systems*, *3*(1). https://doi.org/10.15353/vsnl.v3i1.171

[15] Wang, D., & He, D. (2021). Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosystems Engineering*, *210*, 271-281. https://doi.org/10.1016/j.biosystemseng.2021.08.015

[16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot MultiBox detector. *European conference on computer vision,* pp. 21-37. Springer, Cham.