Scientia Horticulturae 266 (2020) 109309

Contents lists available at ScienceDirect

Scientia Horticulturae

journal homepage: www.elsevier.com/locate/scihorti

A segmentation method for waxberry image under orchard environment

Yijie Wang[a], Jidong Lv[a,*], Liming Xu[b], Yuwan Gu[a], Ling Zou[a], Zhenghua Ma[a]

[a] School of Information Science and Engineering, Changzhou University, Changzhou, 213164, China
[b] School of Equipment Engineering, Jiangsu Urban and Rural Construction College, Changzhou, 213147, China

ARTICLE INFO

Keywords:
Neural network
Target detection
Mask RCNN
Waxberry segmentation

ABSTRACT

Regarding the identification and location of bayberries in the natural environment, the work applied the dilated convolution to the res4b module of Mask RCNN backbone network—ResNet. The method was used to realize the accurate identification and segmentation of waxberry. First, we pre-trained the D-MRCNN network transformed by the dilated convolution with the COCO dataset. Then, through the migration learning method, the representative waxberry dataset was used to train the network for the identification and segmentation of waxberry. Finally, based on the same verification sample set, the work compared the Ostu and K-Means with the deep learning segmentation networks U-net and FCN. The result showed that the algorithm in this work was optimal, with the average detection accuracy and recall rate reaching 97 % and 91 %, respectively. It has high generalization in non-structural environment and better robustness with various forms.

1. Introduction

Waxberry is a traditional Chinese specialty fruit. The global economic cultivation region is about 400,000 ha with more than 1 million tons output, of which more than 98 % comes from China (Tong et al., 2015). The picking period of waxberry is relatively short. At present, the picking of waxberry is mainly performed by hand, which is time-consuming and labor-intensive. In addition, with the aging of population and the reduction of agricultural labor, the cost of manual picking has increased, which greatly reduces the products' market competitiveness. In order to achieve timely harvesting of fruits and reduce the cost of picking, the waxberry picking robot can be introduced to achieve the rapid picking and improve the picking quality, thereby improving economic benefits.

With the development of picking robots, research scholars at home and abroad have achieved certain results (Linker, 2017; Leu et al., 2017; Henten et al., 2009; Ji et al., 2019). However, it is still unavailable for the cost-effective robots applied to large-scale production and use. A Spanish agricultural robot manufacturer launched a strawberry picking robot with 24 robotic arms— AGROBOT, but its high price and the cost of post-maintenance limit its application. At present, the research on picking robots still has a lot of room for improvement. In the operation of waxberry picking robot, the visual system is the key to guide the robot for fruit picking, whose primary task is fruit recognition. The accuracy of fruit recognition affects the success rate of fruit picking. However, due to the weather conditions in the picking season and the unstructured working environment, there are many factors affecting the accurate identification of fruits. Wherein, the natural light, fruit overlap and leaf occlusion are the main causes. Therefore, for the sake of promoting the practical and commercialization of the waxberry picking robot, it is significant to improve the waxberry visual recognition system under natural circumstances.

The visual recognition performance of the picking robot is affected by the fruit growing environment. In the unstructured growth environment, different illumination angles and multiple growth patterns of fruits are important factors affecting the recognition rate. To solve these problems, Finlayson et al. (2006) designed a vision system based on a near-infrared laser range finder. This system uses the distance and reflection information in the image to identify the fruit and reduce the influence of illumination; however its real-time performance needs to be improved. Zhao et al. (2006) applied the color sub-channel to the extraction of tomato target, which reduces the influence of illumination on the target extraction to a certain extent. Whereas the recognition rate of the method was greatly reduced for the fruits with similar color background. Meanwhile, the deep learning has developed rapidly in image recognition. Compared with machine learning, deep neural networks have shown great advantages in image recognition, and have attracted the attention of researchers in agricultural picking identification (Xiong et al., 2018; Zhao et al., 2019).

In 2015, Kuwata and Shibasaki (2015) performed the comparative experiments between Caffe's deep learning algorithm and machine learning algorithm SVR for the estimation of fruit yield. The conclusion

[*] Corresponding author.
E-mail address: vveaglevv@163.com (J. Lv).

https://doi.org/10.1016/j.scienta.2020.109309
Received 14 October 2019; Accepted 21 February 2020

showed that Caffe's deep learning algorithm can judge the yield of corn more accurately, but the SVR algorithm has poor robustness. Inkyu et al. (2016) proposed the use of multi-modal (RGB and near-infrared) fusion information and improvement of the Faster R-CNN network for the optimal fruit identification, showing that up to 7 fruits can be identified. Longsheng Fu et al. (2018) used the LeNet network model to detect several clusters of kiwi. Compared with the traditional fruit target detection method, the recognition rate is increases by 5.37 %, indicating that the neural network of dilated convolution has advantages in identifying fruits and vegetables. Xue et al. (2018) used the improved YOLOv2 network with Tiny-yolo-dense to implement the feature reuse and fusion, which improves the accuracy of detection. Kestur et al. (2018) designed a semantic segmentation deep learning network applied to mango detection and identification techniques, which can further estimate the yield. These results indicate that deep learning has broad application in picking-robot recognition system.

The algorithm in this work improved the Mask RCNN by the dilated convolution. The instance segmentation is a more difficult problem related to object detection. On the basis of object detection, it is required to segment the problematic pixel. Compared with the detection of the target fruit, the instance segmentation provides a more accurate area locating for the picking robot, which can avoid the majority losses during the picking process, especially for the fruits with weak surface, such as waxberry. Furthermore, considering the factors of light intensity and illumination angle of waxberry target, this work constructed a representative dataset.

2. Materials and methods

2.1. Experimental algorithm analysis

This section introduces the segmentation algorithm for the waxberry under natural environment. The algorithm is a deep neural network model algorithm based on the original Mask RCNN instance segmentation network. It is named the Dilated Mask RCNN in the work and abbreviated as D-MRCNN in the description below. Fig. 1 shows the flow. The dataset can be divided into two parts, including the original images and enhanced images. The pre-processed dataset is sent to the D-MRCNN network for training. After the training, the test set data is used to test and adjust the network parameters, thus obtaining the final waxberry segmentation network.

2.2. Mask RCNN

The structure of Mask RCNN (He et al., 2017) is mainly divided into three branches: shared convolution layer-backbone, RPN (Region Proposal Network), proposal classification network and mask padding-three branches (Fig. 2). Mask RCNN is a two-stage network. In the first stage, it scans the images and generates a proposal which may contain the region of detection target. The second stage is to classify the suggestion and generate the bounding box and mask. First, the features of input image are extracted by the convolutional neural network. Then, the obtained feature maps are sent to the RPN network to generate the region of interest. Afterwards, the RoI Align Layer selects the corresponding features of the RoI outputted by the RPN network on the feature map, and fixes its dimension value. Finally, the FC Layer classifies the target boxes to output the category and location of objects.

The work introduced the feature pyramid network (FPN) in Mask RCNN to extend the ResNet network. Fig. 4 shows two pyramid structures formed from C1 to C5, and from P2 to P6, respectively. The P pyramid selects the advanced features from the C pyramid, and passes them to the bottom layer. Through this process, each feature can be combined with advanced/low-level features to represent the target on multiple scales. In addition, the mask branch is a convolutional network. It takes the foreground region selected by the RoI classifier as the input region to generate their masks. Although with the low resolution of the mask, they are soft masks represented by floating-point numbers, which have more details than binary masks.

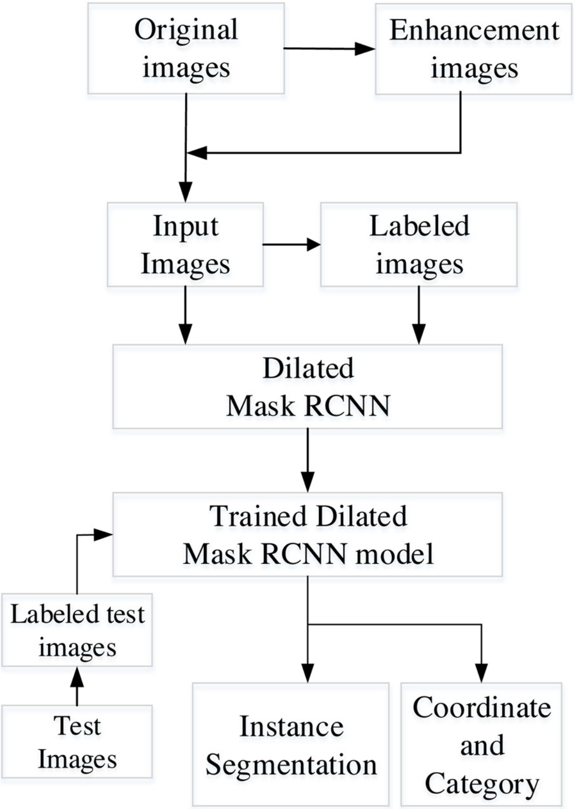


Fig. 1. The method flow chart.

2.3. Dilation convolution

The dilated convolution (Yu and Koltun, 2016) adds a cavity in the standard convolution map to increase the receptive field. Compared to the original standard convolution, the dilated convolution has an extra hyper-parameter called dilation rata, which refers to the interval index of the kernel. Fig. 3 shows the standard convolution and receptive field of dilated convolution with the convolution coefficient of 2. The dilated convolution can avoid the loss of spatial information with the increasing receptive field. The dilated convolution can expand the convolution kernel to the scale bounded by the dilation coefficient with the increasing dilation coefficient *rate* (r) based on the original convolution operation. The convolution kernel with the size of $k \times k$ changes into the convolution kernel k_d after the dilated convolution with the dilation rate of r.

$$k_d = k + (k - 1) \times (r - 1) \tag{1}$$

The dilated convolution is originally applied to the image semantic segmentation. Since the semantic segmentation requires the fine pixel-level classification of the whole image, its features are required to have specific spatial position information. After the dilated convolution, the semantic segmentation network can obtain more detailed outline information (Chen et al., 2017).

2.4. D-MRCNN

In the experiment algorithm of this work, the network performance was improved by applying the dilated convolution to several parts of the network layer in the backbone of Mask R-CNN network algorithm. Fig. 4 shows the detailed structure of the network. In the ResNet-50/101 network, after the down-sampling operation, the convolution layer

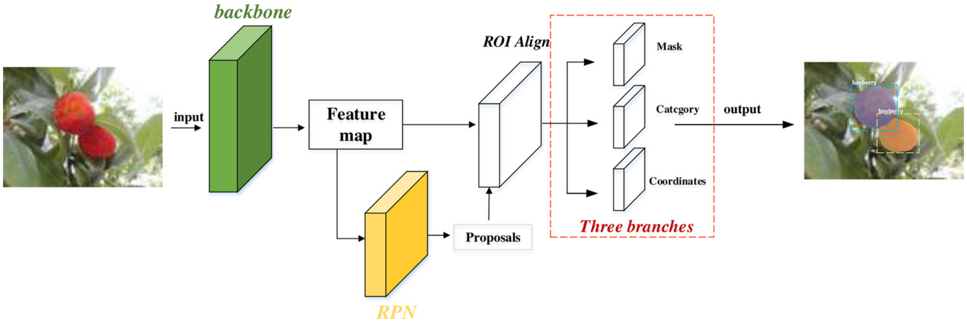


Fig. 2. Mask R-CNN.

with the step size of 2 was transformed to the one with the step size of 1. Afterwards, the dilated convolution was applied to the convolution layer with the step size of 1 and the subsequent convolution layer to ensure a reasonable balance between efficiency and precision. In this work, the experimental algorithm was used for the dilated convolution of res4b module after res4a

The algorithm of dilated convolution in this work was designed as the hybrid dilated convolution (HDC). The application of this structure was to avoid the "gridding". As described in Reference (Chen et al., 2017), if the same dilation ratio was adopted to all convolution layers, the convolution operation with cavities caused the loss of certain pixels in the feature map. The HDC structure proposed in the work set the dilation convolution rate to a series of numbers without common factor (e.g., r = 1, 2 and 3). Meanwhile, the dilation ratio was arranged in a zigzag manner which could avoid the gridding. Furthermore, the structure was simple without adding extra module.

3. Data set making and network training

3.1. Data set making

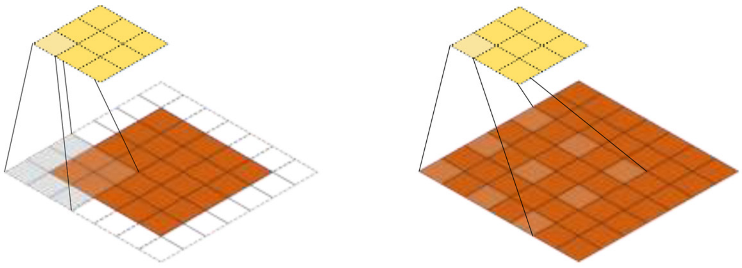
The dataset of this experiment consists of two parts—the original images collected by the camera and the amplified images after the enhancement of the original ones.

The original images were taken in the natural environment of Yangjiatou Waxberry Planting Area, Changzhou City, Jiangsu Province. The digital camera model was Canon DIGITAL IXUS 200 IS. The photographed variety was dark plum species. The maturity of the target fruit was in June whose color was mainly red and purplish black; the shooting angle was various, and the shooting distance depended on the working distance of the robotic vision sensor, ranging from 20 to 80 cm.

The collected waxberry images had the following characteristics: 1) According to the influence of light during shooting, the light conditions of the dataset were divided into two categories: under the full-lighting or the uneven illumination, including back-lighting, side-lighting and fruits in the shadow of foliage. 2) According to the different growth states of waxberry, the target fruits in the image dataset were divided into the unobstructed fruit and occluded fruit. The fruit occlusion included the multiple fruits overlapped and the fruit occluded by non-identifying targets, such as leaves or branches. Fig. 5 shows the various growth patterns of some bayberries in the dataset. The dataset of this experiment contains 1500 images in total which show various growth forms of waxberry under different light conditions.

The small training sample set leads to the over-fitting of the deep learning network, which reduces the generalization of network. In order to overcome this problem, this section of the experiment selected 1200 original images for data enhancement. Data enhancement includes image brightness enhancement/ attenuation, chroma enhancement/attenuation, contrast enhancement/attenuation and sharpness enhancement/attenuation. The brightness, chromaticity and contrast of the original image in Fig.6a were enhanced by 1.5 times. The sharpness was increased to 3 times of the original image. Meanwhile, the brightness, chromaticity, contrast and sharpness were reduced to 50, 50, 50 and 10 % of the original image, respectively (See Fig. 6b–i). In addition, in order to simulate the noise that the device may generate during the image acquisition process, a Gaussian noise with a variance of 0.01 was added to the original image (See Fig. 6j). There were 9 image enhancement methods in this section of the experiment. After the images expansion, the original annotation was still valid. Therefore, this method increases the training sample set, and the diversity of the sample, instead of re-labeling. The addition of the new features enhanced the generalization of the training network.

A total of 1500 original images were collected by the camera. Wherein, 1200 of them were selected as the training sample set, and the remaining 300 images were used as the verification sample set, with the ratio of 4:1. The images did not overlap in the training set and the test set. Table 1 shows the distribution of different growth forms and illumination of the waxberry in the training sample set. The sample set maintained the same ratio in each case to verify network detection in different situations. For the accuracy of the verification effect, the data



(a) standard convolution with a 3×3 kernel

(b) dilated convolution with a 3×3 kernel and dilation rate 2

Fig. 3. Standard convolution and dilated convolution with 3 × 3 kernel.
(a) standard convolution with a 3 × 3 kernel (b) dilated convolution with a 3 × 3 kernel and dilation rate 2.

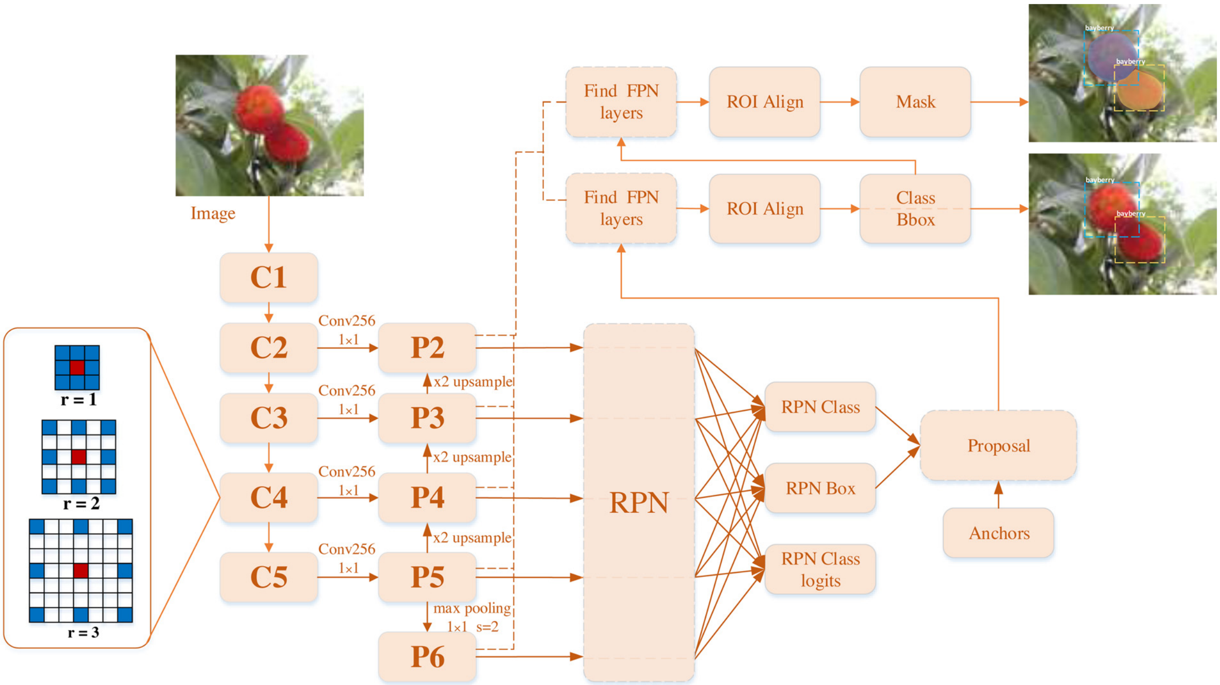


Fig. 4. Structure diagram of D-MRCNN.

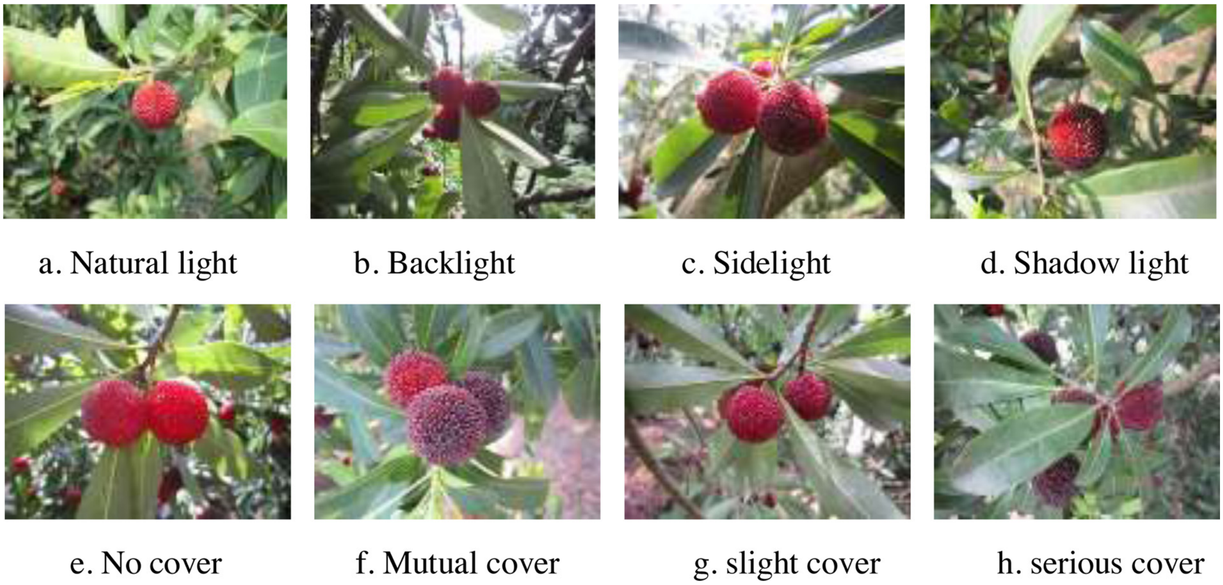

a. Natural light
b. Backlight
c. Sidelight
d. Shadow light
e. No cover
f. Mutual cover
g. slight cover
h. serious cover

Fig. 5. Multiple types of waxberry images.

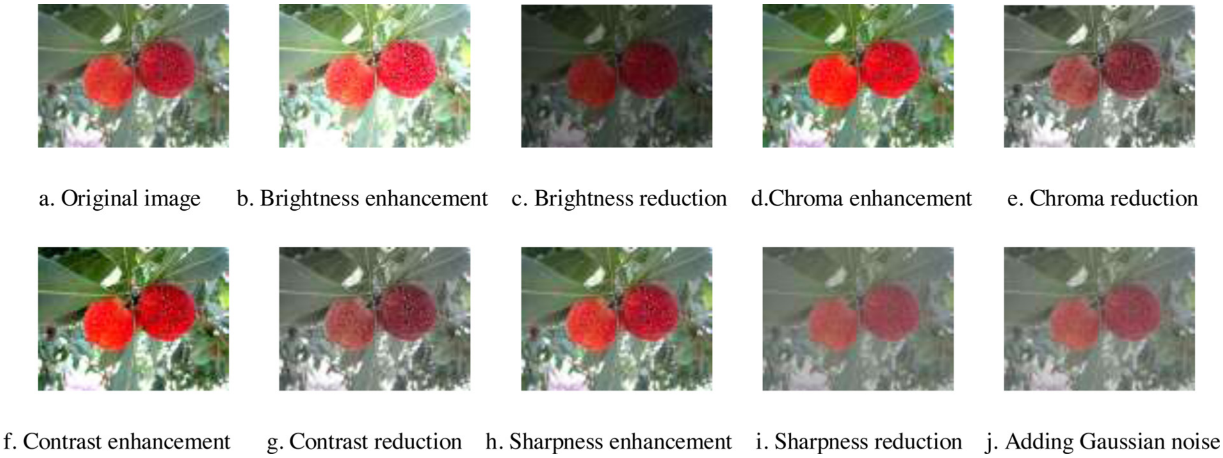
was amplified to 3000 in the same proportion of the verification sample set in each case. In addition to the same data enhancement on the above training sample set, the operations such as mirroring and rotation were carried out to some images. Meanwhile, the re-labeling was performed. By using this validation sample set, we can verify the generalization of the model.

Network training and test experiment were performed on the Ubuntu 16.04 system software platform and the following hardware configurations, with the processor of dual parallel GTX 1080 and the single memory of 8 G. In order to reduce the running time of subsequent experiments, 1,200 original images were scaled to 512 × 384 pixels by linear interpolation algorithm in advance, and then the images were manually labeled. This experiment used the python version of Labelme annotation software. To achieve image segmentation, “polygon” option was applied to the labeling, which needed to fit the waxberry outline and give each individual waxberry a separate label. However, it only labeled one kind of bayberries instead of ones with small pixel proportion in the image or short distance to the image edge. The waxberry dataset in this section was representative.

3.2. *Training of recognition network algorithms*

3.2.1. *Pre-training*

In the training experiment of waxberry identification network model, the COCO dataset (Lin et al., 2014) was used for model pre-training to obtain a file of weighting parameters. What’s more, through the transfer learning method, the network model was retrained with the dataset in the previous section of experiment, and the parameters were fine-tuned to obtain the optimized network. Microsoft has released the COCO dataset widely used in computer vision since 2014. The dataset,


a. Original image b. Brightness enhancement c. Brightness reduction d.Chroma enhancement e. Chroma reduction

f. Contrast enhancement g. Contrast reduction h. Sharpness enhancement i. Sharpness reduction j. Adding Gaussian noise

Fig. 6. Results of data augment.

including 200,000 annotation images, over 1.5 million object instances and a total of 80 categories, was used to initially verify the performance of the Mask RCNN optimized in the work.

Since there were eight Tesla P100 devices used in Ref. (He et al., 2017), and two GTX 1080 in the experiment, in order to ensure the validity of the comparison test, the engineering code was executed on the experimental device. At the same time, the pre-trained number of iterations of the network model in this section was set to 180k, with the learning rate reduced by 10 times in the 120 and 160 K iterations, and the NMS threshold set to 0.5.

To show the difference in the training process between Mask RCNN and D-MRCNN networks, Fig. 7 shows the loss function before and after optimization. The D-MRCNN network follows the loss function of the Mask RCNN network, consisting of three parts (See Eq. (2)).

$$L = L_{class} + L_{box} + L_{mask} \tag{2}$$

Where L_{class} and L_{box} are the same as those in Faster RCNN, using the full connection to predict the category and coordinate figure of rectangular box for each RoI, respectively; L_{mask} is the loss function of Mask.

Fig. 7 shows the convergence rate of D-MRCNN is faster than that of Mask RCNN network. When the iteration cycle reaches 15 times, the D-MRCNN network basically converges. The Mask RCNN network model converges when iterating to about 27 times. The loss value of D-MRCN eventually approaches 0.2, which is lower than the Mask RCNN network.

3.2.2. Transfer learning training of recognition network

After completing the pre-training of the D-MRCNN network, the experiment used the waxberry dataset for the transfer learning training of the network. It aimed to compare the superiority of the network model in the work, and the same dataset was used for the migration learning training of Mask R-CNN network. After the same training of sample set, the performances of Mask R-CNN and D-MR CNN networks were tested with the validation sample set. Fig. 8 shows some of the test images.

Fig. 8a shows four selected sample images including the most common growth states during the picking process. It contains independent fruits with good illumination, severe obscuration by leaves and overlap of fruits. The image on the far right is the most complicated situation, covering several conditions such as uneven light and occlusion of leaves. Fig. 8b and c show that in the boundary and overlap of some fruits, the contour of waxberry can be detected by the algorithm in the work.

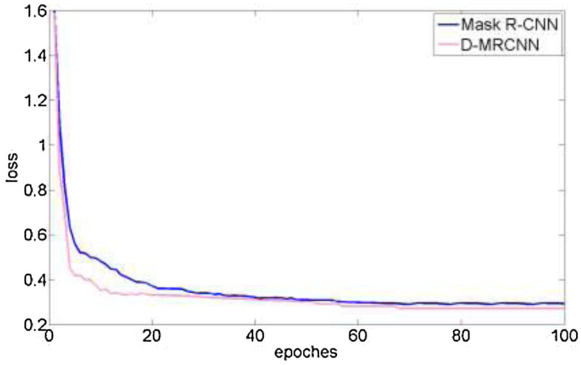


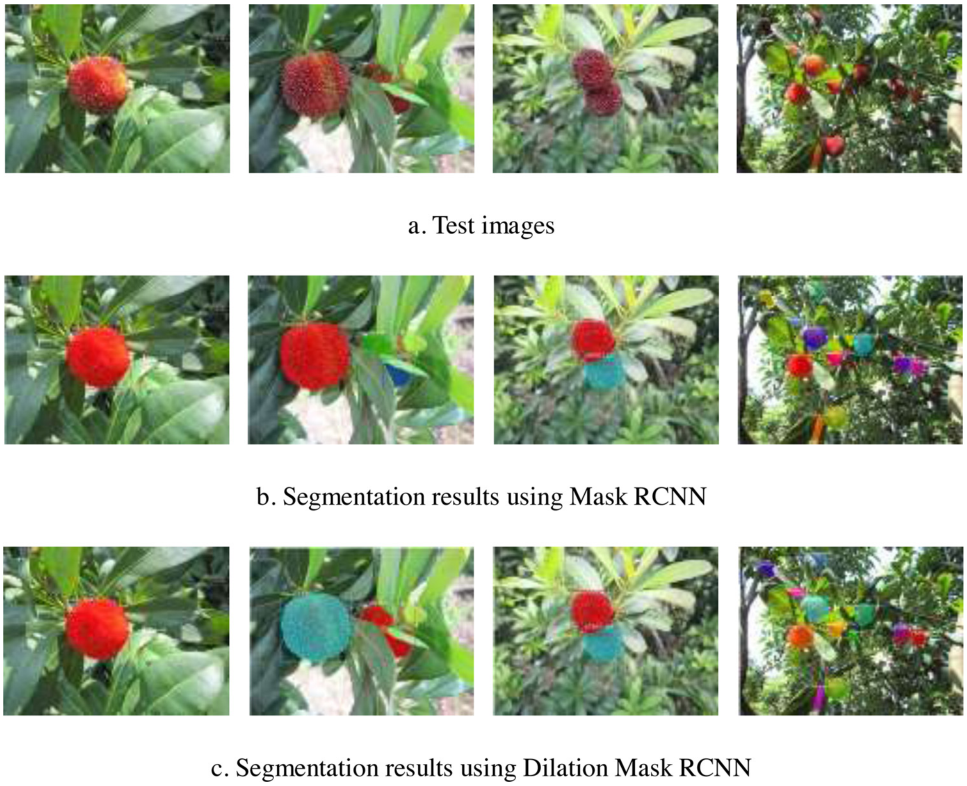
Fig. 7. Loss function curve of two networks.

In order to compare the differences in feature extraction between network models before and after improvement, the thermodynamic chart is used to show the differences in the network feature extraction process (Zhou et al., 2015), and Fig. 9 shows the results.

The "attractiveness" to the network of this region is revealed in thermodynamic char by different colors. Wherein, the red region represents the most impact on the network. As the color changed from red to yellow and eventually to blue, the impact of the weakening range is also shown. Fig. 9a and b show the thermodynamic charts of Fig. 8a in the Mask R-CNN and D-MRCNN networks, respectively. To reflect the effectiveness and superiority of the algorithm in the work, we compare it with other commonly used algorithms of fruit segmentation and recognition.

Table 1
Category and quantity of apple in data set.

Data type		Natural light	Backlight	Sidelight	Leaf shadow	Total
Growth form	Non-occlusion	76	63	58	70	267
	Occlusion	304	187	232	210	933
Total	Original image	380	250	290	280	1200
	Data enhancement	3420	2250	2610	2520	10800



a. Test images

b. Segmentation results using Mask RCNN

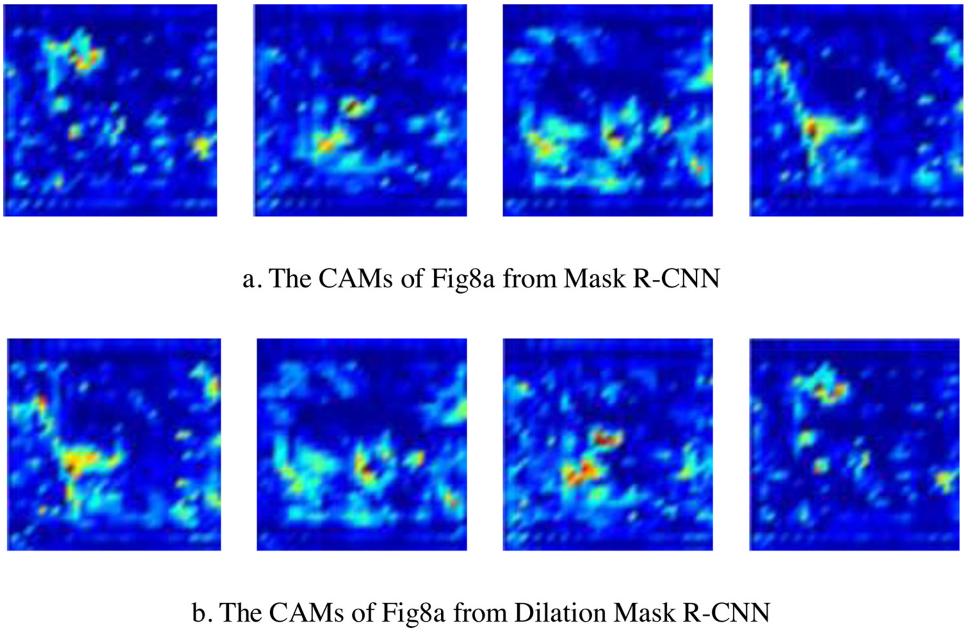
c. Segmentation results using Dilation Mask RCNN

Fig. 8. Examples of the Segmentation results using two network.
a. Test images.
b. Segmentation results using Mask RCNN.
c. Segmentation results using Dilation Mask RCNN.

4. Experimental results and analysis

Before the deep learning of image segmentation method, the commonly used methods include the threshold method, cluster analysis and so on. In the experiment, the Otsu and K-means methods were chosen to compare the segmentation, and Fig. 10 shows the results. By comparison, the Otsu and K-means algorithms (Ray and Turi, 1999) are greatly affected by the light environment and fruit growth morphology, but the algorithm in the work has the high robustness of the algorithm.

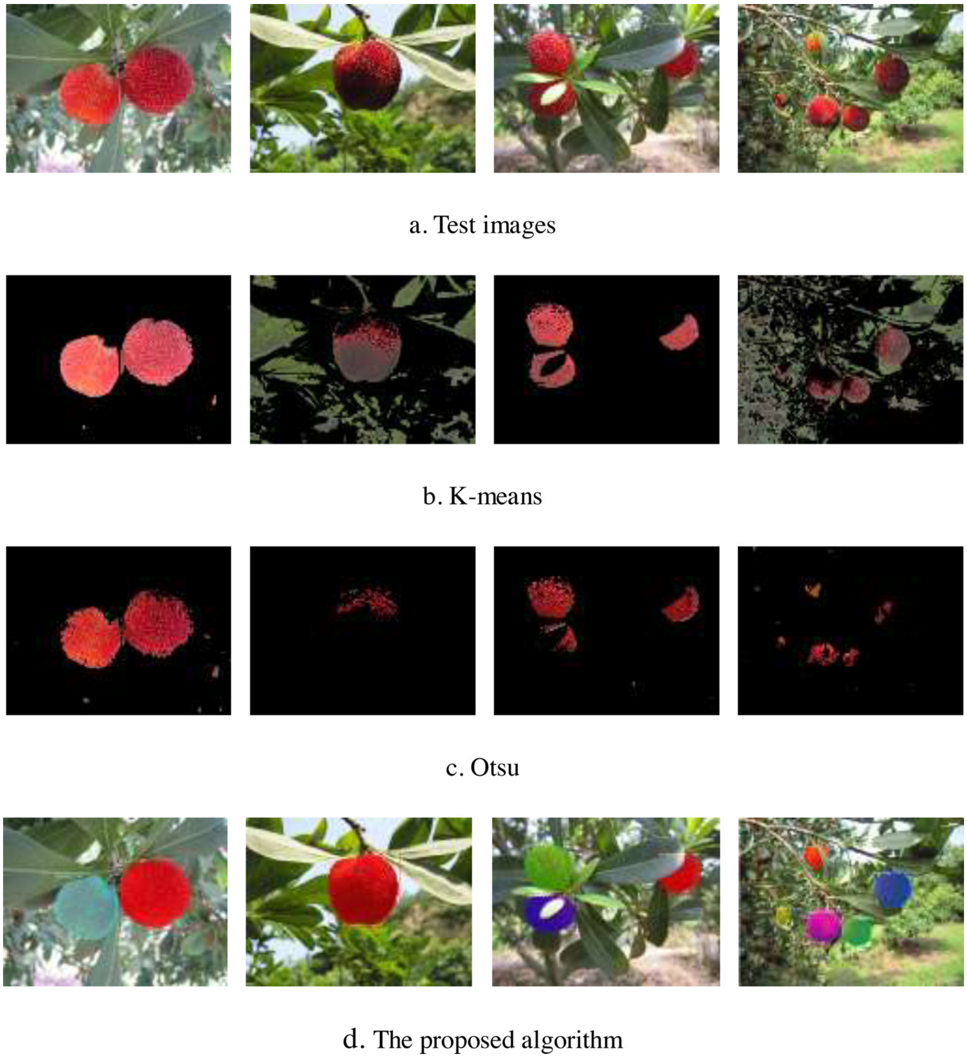
In addition, in order to reflect the superiority of the algorithm in the



a. The CAMs of Fig8a from Mask R-CNN

b. The CAMs of Fig8a from Dilation Mask R-CNN

Fig. 9. The CAMs of Fig. 8a from two networks.
a. The CAMs of Fig. 8a from Mask R-CNN.
b. The CAMs of Fig. 8a from Dilation Mask R-CNN.



a. Test images

b. K-means

c. Otsu

d. The proposed algorithm

Fig. 10. Examples of the segmentation results produced by different algorithm.
a. Test images.
b. K-means.
c. Otsu.
d. The proposed algorithm.

work in the segmentation of bayberries, the work compared the performance of more deep learning segmentation networks, using U-net (Ronneberger et al., 2015) and FCN (Long et al., 2015) networks for segmentation. The results showed that the segmentation algorithm in the work was better through comparison (See Fig. 11). Fig. 11a shows the 4 original images of the selected set of test samples, including backlight with dense fruits, side light, interference objects, and single fruit segmented into pieces by branches and leaves.

Fig. 11 shows that the algorithm in the work can segment and identify the bayberries whether the fruits are dense or the light is uneven; however, it is not ideal for the effect of U-net and FCN network detection. What's more, in the case of unmarked interfering objects, both U-net and FCN networks, except for the algorithm in the work, have false detection. For a waxberry segmented into pieces by occluded leaves, only the algorithm in the work completely identifies each individual and performs pixel-level segmentation on the visible portion. At the same time, under the condition of backlighting, compared with the other two networks, the algorithm network in the work realized the accurate segmentation of waxberry, revealing its learning of the edge characteristics of waxberry.

To quantitatively evaluate the segmentation of different algorithms, 2400 images of bayberries in the validation set were taken for the division experiment. Then, the segmentation results were evaluated using the Precision, Recall, and F1 indicators in Equation 3. Table 2 shows the results.

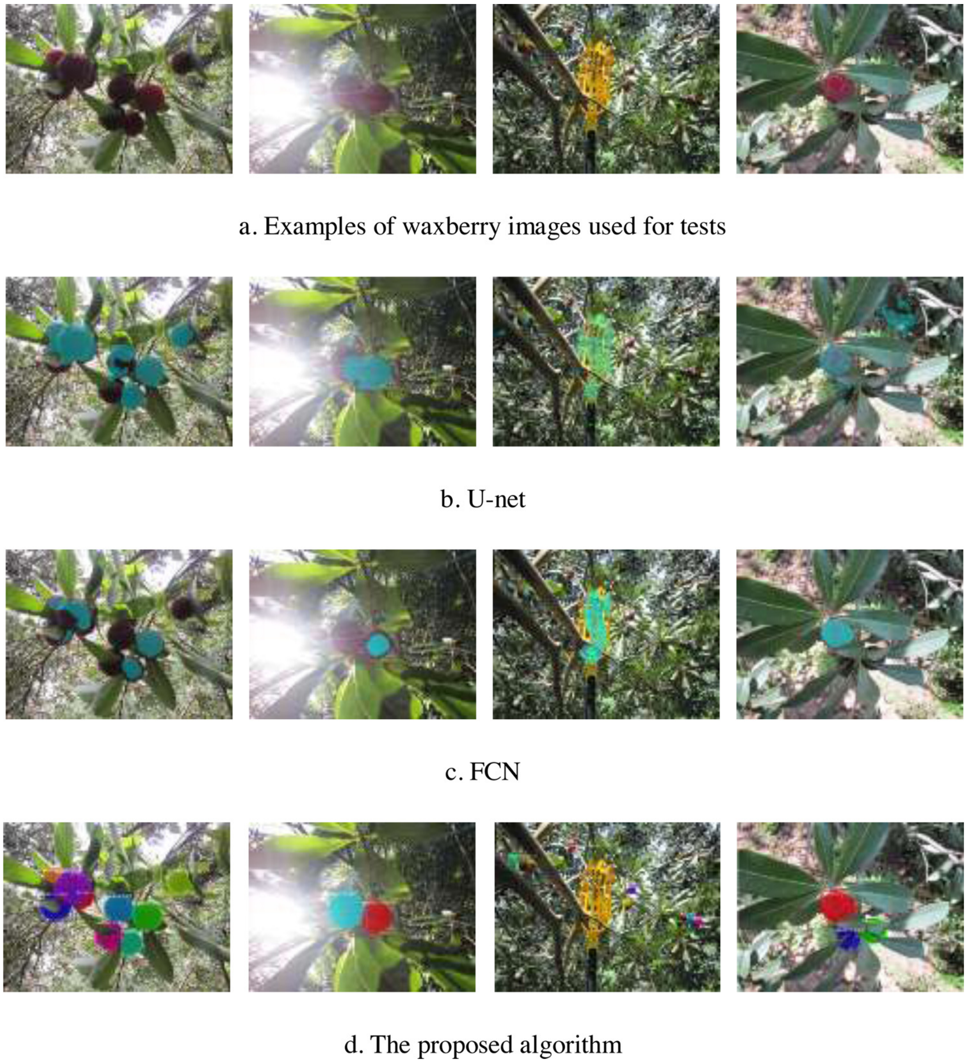
$$\mathrm{P}{=}\frac{TP}{TP+FP}\quad \mathrm{R}{=}\frac{TP}{TP+FN}\quad F_1=\frac{2PR}{P+R} \tag{3}$$

Where P represents the accuracy rate; R the recall rate; TP the number of bayberries identified by the algorithm; FP the number that algorithm mistakes the background as waxberry; FN the number of unidentified bayberries. Under normal circumstances, P increases with the reduced R, and F_1 can be used to determine the detection performance of the algorithm.

Table 2 shows that the Otsu and K-means algorithms are more affected by the light in the natural environment with poor segmentation. In the four deep learning network models, the Dilated Mask R-CNN network has the highest performance.

5. Conclusion

In the work, the waxberry identification network D-MRCNN was obtained by introducing the dilated convolution to the Mask R-CNN


a. Examples of waxberry images used for tests

b. U-net

c. FCN

d. The proposed algorithm

Fig. 11. Examples of the Segmentation results using different Networks.
a. Examples of waxberry images used for tests.
b. U-net.
c. FCN.
d. The proposed algorithm.

Table 2
Evaluation index statistics of different segmentation algorithms.

Segmentation algorithm	Precision	Recall	F_1
K-means	0.76	0.59	0.66
Otsu	0.82	0.72	0.77
U-net	0.93	0.90	0.91
FPN	0.86	0.84	0.85
DMRCNN	0.97	0.91	0.94

network, and established the model of data set training. Results showed that the D-MRCNN network model had high precision and stronger robustness to occlusion, overlap and uneven illumination with accuracy and recall rate of 97 and 91 %, respectively.

1) The dilated convolution of the Mask RCNN backbone network was carried out by a zigzag arrangement of dilated convolution coefficients, which increased the receptive field without causing the loss of feature information. The convergence rate of the training was improved compared with the original network, with better generalization for a single fruit segmented into pieces by branches or leaves. In contrast with the other two deep learning segmentation networks, both the accuracy and recall rate remained high.
2) Compared with other segmentation algorithms, the D-MRCNN in the work can realize the segmentation of waxberry, that is, it has the function of overlapping-fruit separation. In subsequent studies, the network will be trained by increasing the immature waxberry sample dataset to achieve the count of fruits and fruit yield estimation.
3) As a supervised deep learning method, D-MRCNN has high requirements on the quality, quantity and diversity of datasets. What's more, it is important to produce effective datasets, which is time-consuming. In the work, image enhancement processing is adopted without labeling, which greatly improves the detection accuracy of occluded fruits or overlapping fruits. The overlapping fruit has been a difficult problem of picking robot vision system for a long time.

The algorithm adopted in the work can realize the accurate identification of waxberry, yet the research of D-MRCNN will never stop here. More datasets should be collected to train the network for the identification of fruits and vegetables.

Declaration of Competing Interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This work was partly supported by the Natural Science Foundation of Jiangsu Province with Grant NO. BK20140266, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China with Grant NO. 17KJB416002, the Changzhou Sci&Tech Program with Grant NO. CJ20179057, NO. CJ20180021, NO. CE20195025 and the National Natural Science Foundation of China with Grant NO. 61906021.

References

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation.

Finlayson, G.D., Hordley, S.D., Lu, C., Drew, M.S., 2006. On the removal of shadows from images. IEEE Trans. Pattern Anal. Mach. Intell. 28 (1), 59–68.

Fu, L.S., Feng, Y.L., Elkamil, T., Liu, Z.H., Li, R., Cui, Y.J., 2018. Image recognition method of multi-cluster kiwifruit in field based on convolutional neural networks. Trans. Chin. Soc. Agric. Eng. 34 (2), 205–211.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy. pp. 2980–2988.

Henten, E.J.V., Van't Slot, D.A., Hol, C.W.J., Willigenburg, L.G.V., 2009. Optimal manipulator design for a cucumber harvesting robot. Comput. Electron. Agric. 65 (2), 247–257.

Inkyu, S., Zongyuan, G., Feras, D., Ben, U., Tristan, P., Chris, M.C., 2016. Deepfruits: a fruit detection system using deep neural networks. Sensors 16 (1222), 1–23.

Ji, W., Qian, Z.J., Xu, B., Chen, G.G., Zhao, D.D., 2019. Apple viscoelastic complex model for bruise damage analysis in constant velocity grasping by gripper. Comput. Electron. Agric. 162, 907–920.

Kestur, R., Meduri, A., Narasipura, O., 2018. MangoNet: a deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. Eng. Appl. Artif. Intell. 77, 59–69.

Kuwata, K., Shibasaki, R., 2015. Estimating crop yields with deep learning and remotely sensed data. In: IGARSS 2015 - 2015 IEEE International Geoscience and Remote Sensing Symposium. IEEE.

Leu, A., Razavi, M., Langstadtler, L., Ristic-Durrant, D., Raffel, H., Schenck, C., et al., 2017. Robotic green asparagus selective harvesting. IEEE/ASME Trans. Mechatron. 22 (6), 2401–2410.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision. Springer International Publishing, pp. 740–755.

Linker, R., 2017. A procedure for estimating the number of green mature apples in night-time orchard images using light distribution and its application to yield estimation. Precis. Agric. 18 (1), 59–75.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. The IEEE Conference on Computer Vision and Pattern Recognition, arXiv Preprint, Arxiv:1411.4038v2.

Ray, S., Turi, R.H., 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. The 4th International Conference on Advances in Pattern Recognition and Digital Techniques 137–143.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. arXiv preprint, arXiv 1505 (4597).

Tong, Z., Wu, L.M., He, L.G., Wang, Z.J., Sun, Z.H., Jiang, Y.C., 2015. The development status and prospect of waxberry industry in hubei province. Hubei Agric. Sci. 54 (24), 6255–6258.

Xiong, J.T., Liu, Z., Liu, R., Chen, S.M., Chen, W.J., Yang, Z.G., 2018. Unmanned aerial vehicle vision detection technology of green mango on tree in natural environment. Trans. Chin. Soc. Agric. Mach. 49 (11), 23–29.

Xue, Y.J., Huang, N., Tu, S.Q., Mao, L., Yang, A.Q., Zhu, X.M., Yang, X.F., Chen, P.F., 2018. Immature mango detection based on improved YOLOv2. Trans. Chin. Soc. Agric. Eng. 34 (7), 173–179.

Yu, F., Koltun, V., 2016. Multi-scale Context Aggregation by Dilated Convolutions. ICLR.

Zhao, J.Y., Zhang, T.Z., Yang, L., 2006. The vision system of tomato picking robot target extraction. Trans. Chin. Soc. Agric. Mach. 37 (10), 200–203.

Zhao, D.A., Wu, R.D., Liu, X.Y., Zhao, Y.Y., 2019. Apple positioning based on YOLO deep convolutional neural network for picking robot in complex background. Trans. Chin. Soc. Agric. Eng. 35 (3), 164–173.

Zhou, B.L., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2015. Learning Deep Features for Discriminative Localization.