Computers and Electronics in Agriculture 172 (2020) 105380

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot

Weikuan Jia (Ph.D) (Associate Professor)[a,*], Yuyu Tian[a], Rong Luo[b], Zhonghua Zhang[a], Jian Lian[c], Yuanjie Zheng (Ph.D) (Professor)[a]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
[b] School of Light Industry Science and Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250351, China
[c] Department of Electrical Engineering and Information Technology, Shandong University of Science and Technology, Jinan 250031, China

ARTICLE INFO

Keywords:
Overlapped apples
Mask R-CNN
Image segmentation
Target detection

ABSTRACT

In order to better apply the good performance of feature extraction and target detection used in deep learning to fruit detection in orchards, a model of harvesting robot vision detector based on Mask Region Convolutional Neural Network (Mask R-CNN) is proposed. The model was improved to make it more suitable for the recognition and segmentation of overlapped apples. Residual Network (ResNet) combined with Densely Connected Convolutional Networks (DenseNet) can greatly reduce input parameters and is used as a backbone network for feature extraction. Feature maps are input to the Region Proposal Network (RPN) for end-to-end training to generate the region of interest (RoI), and finally the mask is generated by the full convolution network (FCN) to get the region where the apple is located. The method is tested by a random test set with 120 images, and the Precision Rate has reached 97.31%, and the Recall Rate has reached 95.70%. And the recognition speed is faster, which can meet the requirements of the apple harvesting robot's vision system.

1. Introduction

With the gradual maturity of AI technology, the focus of agricultural harvesting is gradually shifting to automation. Vision based fruit detection is a critical component for infield automation in agriculture (Bargoti and Underwood, 2017). The realization of apple's automatic picking robot makes the fruit harvesting work with a small number of people or even no one to work, which can greatly reduce the dependence on labor. The vision system is the most important way for the picking robot to "discover" and "judge" the position and shape of the fruit. Accurate fruit positioning and segmentation provide the possibility for the operation of the visual system (Bac et al., 2014; Zhao et al., 2016; Bac et al., 2017). Therefore, how to achieve accurate identification and localization of target fruits has become the key to vision system research. However, due to the complex background in the actual apple orchard, factors such as fluctuating illumination, dense fruit distribution, occlusion of fruits by branches and leaves, overlapping fruits, camera angle and distance will have a certain impact on target detection, which cause huge difficult and created challenges in accurate identification of fruits.

In recent years, many researchers have proposed different methods for the improvement systems of fruit detection in complex backgrounds. Liu proposed a recognition method of apples in plastic bags based on bock classification, watershed algorithm is adopted to segment original images into irregular block, and then these blocks are classified into fruit blocks and non-fruit blocks by VSM, the new method can restrain the interference of light (Liu et al., 2018). Linker designed a four-step method for green apple recognition and counting based on features such as color, texture, seed area and edge shape, which realized the recognition and counting of green apple under natural illumination. Although it reached 95% recognition rate, when two when there are overlaps in one or more apples, it is difficult for the algorithm to identify them separately (Linker et al., 2012). With the development of machine learning, more and more deep learning methods have been proposed for fruit detection in agriculture, which represent a step advance from algorithms based on hand crafted features such as color, shape and texture (Cheng et al., 2017; Tao et al., 2018; Koirala et al., 2019).

Choi used the three-dimensional information of the depth image to identify the fruit area in the map, and the CNN network was used to identify it accurately, and compared three methods and draw conclusions that NIR images performed best with 96% true positive rate for both the circular object detection and classification (Choi, 2017). Chen proposed a blob detector based on a fully connected CNN is used to

* Corresponding author.
E-mail address: jwk_1982@163.com (W. Jia).

https://doi.org/10.1016/j.compag.2020.105380
Received 19 December 2019; Received in revised form 14 March 2020; Accepted 18 March 2020
Available online 26 March 2020

extract candidate regions in the image, segment the object regions and use a subsequent CNN counting algorithm to calculate the number of fruits (Chen et al., 2017). Tian improved the yolo-v3 network, which enhanced the feature propagation, improved the reusability of features, and realized the detection of apples in different growth stages. Although good results were obtained, the experimental data showed that the algorithm would still Affected by occluded fruits (Tian et al., 2019). Sa applied Faster RCNN on multi-vision sensor to detect peppers, rockmelons and apples through transfer learning (Sa et al., 2016). Bargoti adopted the faster-RCNN model on the detection of apples and mangos in orchards with an F1-score of > 0.9, obtained a high detection accuracy, but the inability to detect all fruit appearing in a cluster (Bargoti and Underwood, 2017). Mask RCNN was proposed and applied to largely solves the problem of difficult identification caused by fruit occlusion, which can be used to segment the fruit (He et al., 2017). Yu proposed a mask-CNN algorithm to detect and quantify strawberries in the wild, fruit detection results of 100 test images showed an average detection precision rate of 95.78%, the recall rate was 95.41% and the mean intersection over union (MIoU) rate for instance segmentation was 89.85% (Yu et al., 2019). Many image segmentation algorithms have been proposed and achieved good results (Deng et al., 2018; Yang et al., 2016; Zhu et al., 2018); inspired by this article, this paper proposed an improved model based on Mask RCNN, making it more suitable for the identification and segmentation of single classification, and applied to Apple's detection.

We will explain in detail the specific steps of the algorithm in the following article. In the second chapter, we describe the acquisition and labeling of the dataset. In the third chapter, clarify how the method is improved on the basis of Mask RCNN to form the basic model of the algorithm through the fact as the feature extraction, RoI acquisition and mask generation. The training methods are described in the fourth chapter, including the extension of datasets and the application of transfer learning. The experiment was finally carried out and the experimental results were given. The overall accuracy can reach 97.31%, and the recall rate reaches 95.70%, which can accurately identify and segment apple fruit.

2. Data acquisition

2.1. Images acquisition

In this study, image was acquired using a camera with 6000 × 4000 pixel resolution with different nature light condition. The image of Apple's image was collected at the Longwangshan apple production base in Fushan District, Yantai City, Shandong Province (agricultural information technology experimental base of Shandong Normal University).

The images data used in this paper were collected in apple orchards during cloudy and sunny weather conditions. The collection periods included 6 a.m. to acquire apple images of the low-light environment, and 11 a.m. to direct glaring and 5: 30 p.m to the sunlight slanted. The illumination conditions included front-lighting, backlighting, side-lighting, and scattered lighting. In particular, we added the nighttime images taken at 8 p.m, which the flash of Canon camera was turned on during shooting. In order to better simulate the visual system during the picking robot operation, we collected the fruit images from multiple directions, and there were many large-area occlusion and coincidence phenomena in the image, including the facts that the fruit was occluded by the branches, the fruit was occluded by the leaves, and the fruit occluded each other. The specific images are listed in Fig. 1 below.

2.2. Images annotation and dataset production

Due to the small number of datasets, we use Imagenet pre-trained weights for model training. To make the model reach lower loss, it is not enough to rely on transfer learning. It depends more on the structural diversity of the model and the richness of the new dataset. As we all known, whether the neural network can process the images collected at different time of the day depends on the integrity of the training dataset. In order to expand the diversity of data, the image is preprocessed from the angle, brightness (Includes simulating light condition as sunny or cloudy days and exposure or weak light and sharpness). The process increases the networks capability to generalize and reduces the probability of overfitting.

In the practical process of picking robots, speed of recognition is a major goal for real-time apple fruit recognition. Pictures taken in real time may be blurred, and low resolution may be an unfavorable factor for recognition. In order to adapt the trained algorithm to target recognition at low resolution more, we set the picture length to 512 pixels and adjust the width to keep the aspect ratio constant. We scrambled and renumbered all images, then manually labeled them. To prevent over-fitting of the neural network, some positive samples that are not clearly defined are discarded. In particular, we have specifically marked the occlusion of the foliage or more than two overlapping fruits by the method of edge complementation, we hope to solve the occlusion problem by this method. The labeled pixel is defined as1, and the others are 0. Totally, we labeled 1020 images.

3. Methodology

Mask R-CNN is a state-of-the-art algorithm that can pricisely detect the target object and accurately segment the target. We proposed an improved Mask R-CNN structure and apply it to the vision system of the apple picking robot. For a long time, the lack of dataset was a huge problem to the research in fruit object recognition, and required a large number of labeled images to train deep networks. This paper proposed that ResNet combined with DenseNet network structure instead of the original backbone network for feature extraction can increase feature transitivity and reusability, which can use less parameters to get an outstanding performance (Huang et al., 2017). The obtained feature maps through the above backbone network will be used as the input of RPN network which used to generate region proposals with the probability of being a object for each feature map respectively. Finally, the mask is generated by the full convolution network to obtain the area where the apple is located. In particular, owing to our goal is to identify and segment the apples in a complex background, the final segmentation results have only one class and do not need to identify different kinds of objects, so we eliminated the classification branch in multi-tasking and defined apple class to achieve smaller loss and faster speeds. The overall model structure is shown in Fig. 2.

The algorithm is applied to the vision system of the apple picking robot, which used to provide precise position information, fruit shape and size information for robot picking. Mask R-CNN is a state-of-the-art recognition and segmentation algorithm, our method is based on the Mask RCNN model, and in order to make it more suitable for real-time segmentation of apple fruit, it has made some adjustments and optimizations.

3.1. Feature extraction (ResNet + DenseNet)

Deep convolutional network with different depth can be established by designing different weight layers, which is widely used in image feature extraction. But as we all known that after the CNN network reaches a certain depth, with the increase of the number of convolutional layers, the training errors will increase, and the classification accuracy of test dataset will become very poor (He et al., 2016; Sui et al., 2017; Lian et al., 2018; Wang et al., 2017). ResNet effectively solved this difficult by using multiple layers with parameter to learn the representation of residuals between inputs and outputs, and significantly improve the training speed and prediction accuracy of the deep network. The core of the ResNet model is to establish a “shortcuts” (skip connection) between the front and back layers, which helps the

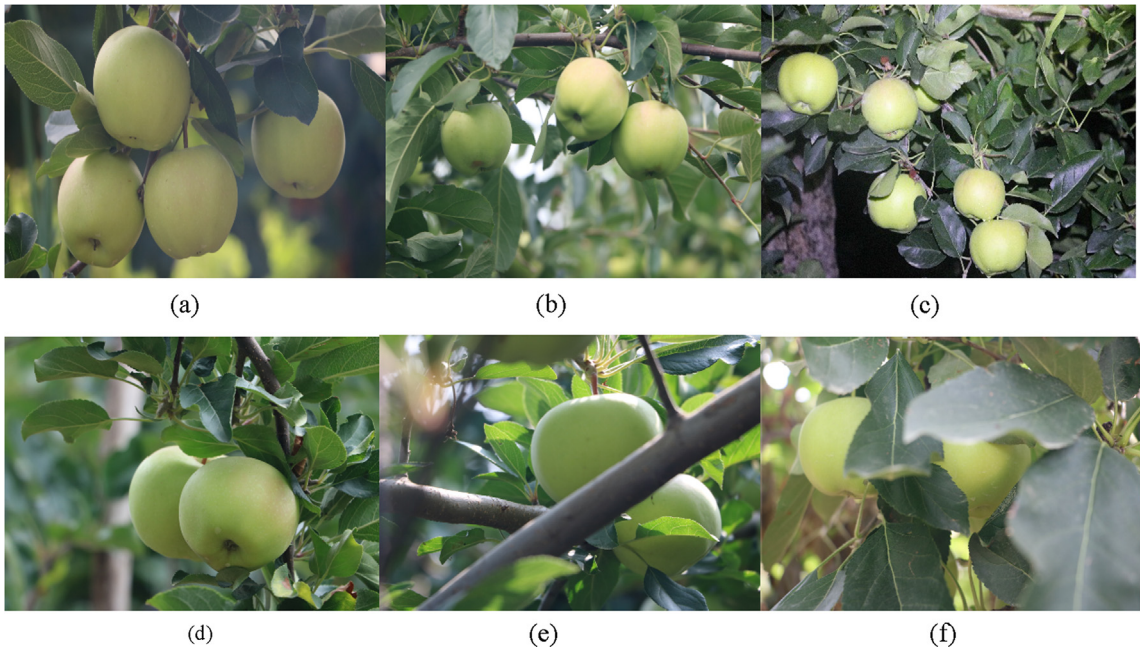

Fig. 1. Apple images collected at different light intensities and at different angles. a–c is low-light, direct light, and nighttime images, respectively. d–f is an image of inter-fruit occlusion, branch occlusion, and leaf occlusion.

back-propagation of the gradient during the training process to training a deeper CNN network. Considering that during the actual training process, many over-small area features will be ignored due to the convolutional layer resolution are getting smaller and smaller, in order to better preserve the small-area features, a dense convolutional network (DenseNet) which can achieve each layer directly get the output of all previous layers is designed to parallel the output to the input inspired by ResNet's addition of output to input. In order to better preserve the feature of fruit images on multiple scales, DenseNet is proposed to improve the backbone network, which connects each layer to every other layer in a feed-forward fashion. Compared with the ResNet combined Feature Pyramid Networks (FPN) which do the feature delivery in the original model, DenseNet improved the reusability of features by one level. For each layer, the feature map of all previous layers is used as input, and its own feature map is used as input of all subsequent layers. It greatly enhances the feature propagation and encourages the feature reuse, which can better identify the fruit with too small area due to occlusion. Therefore, DenseNet combined with ResNet will be used as the backbone network for feature extraction in this paper, which Residuals can help deepen the depth of training, and Densely ensure that low-dimensional features are not completely discarded. The residual dense unit is composed of a plurality of convolution layers and ReLU. The output of each unit establishes a short connection with the output of each convolution layer of the next unit, thereby realizing continuous information transmission. The specific network structure is shown in Fig. 3 below.

For a convolutional network, assume the input image x_0. The network consists of L layers, each of which implements a nonlinear transformation$H_i(\cdot)$, where i represents the i^{th} layer, then the i layer obtains the feature maps x_0, x_1, ……,x_{i-1} of all previous layers as input, as

$$x_i = H_i([x_0, x_1, \cdots\cdots, x_{i-1}]) \tag{1}$$

where $[x_0, x_1, \cdots\cdots, x_{i-1}]$ represents the cascade of feature maps. And defined $H_i(\cdot)$ as a combined function of two consecutive operations: ReLU function and 3 * 3 convolution (Conv).

Defined the input and output of the dth unit be $F_{d-1}$and F_d, respectively, and the number of feature maps is G_0. The output of the cth convolution layer in the unit can be expressed as

$$F_{d,c} = H\{[F_{d-1}, F_{d,1}, \cdots\cdots, F_{d,c-1}]\} \tag{2}$$

Since the connection mode between the input of the unit and the convolutional layer is used, it is necessary to compress the feature map at the end of the unit. Therefore, the number of feature maps controlled

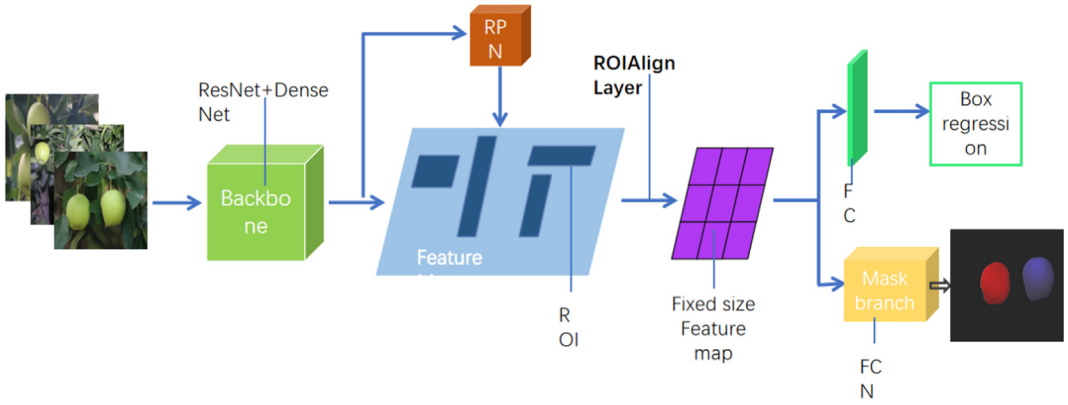

Fig. 2. Overall model structure of improved Mask R-CNN.

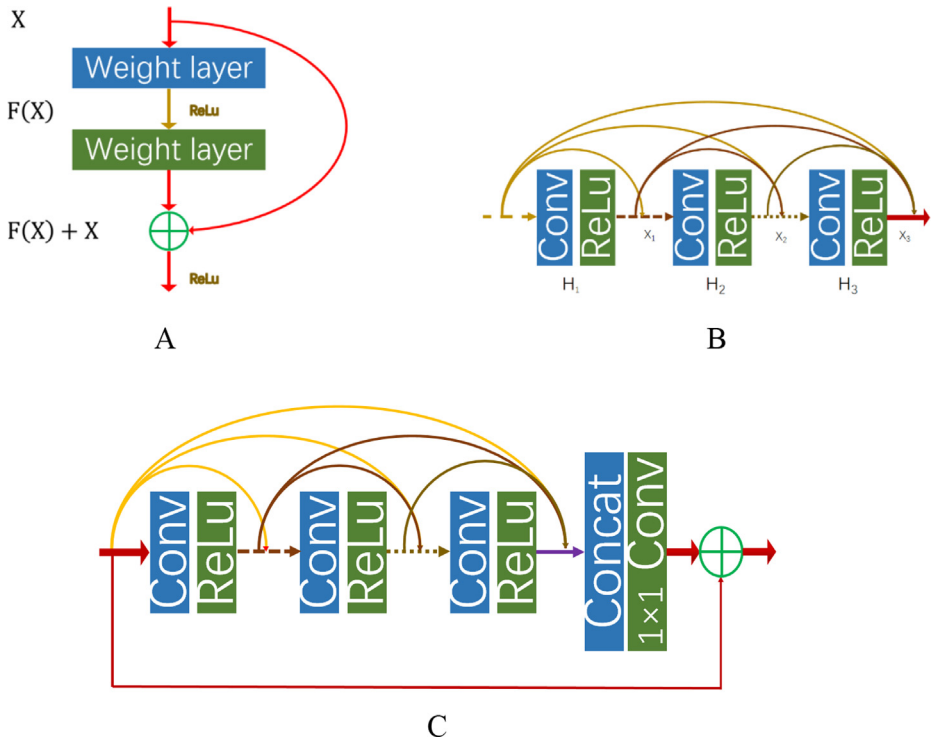

Fig. 3. The specific network structure of backbone. A is structure of residual block. B is structure of dense block (l = 3). C is structure of residual dense block.

by 1 × 1 convolution can be expressed as

$$F_{d,LF} = H^{d}_{LFF}\{[F_{d-1}, F_{d,1}, \cdots\cdots, F_{d,c}]\} \quad (3)$$

where H^{d}_{LFF} represents 1 × 1 convolution. The final output of the unit can be expressed as

$$F_{d} = F_{d-1} + F_{d,LF} \quad (4)$$

In this article, we use 3 layers of dense blocks, that is, three BN + ReLU + Conv (3 * 3) Layer network structures. Input image pixels are 512 * 512. In order to ensure the maximum information flow between each layer, all network layers are directly connected. In order to maintain the feedforward characteristics, the input of each layer is the sum of the mapping output of all the previous layers, and its own feature mapping result is also used as the input of the subsequent layers. The dense layer is followed by a transition layer. The transition layer is a 1 * 1 convolution kernel, the number of channels is 4 k, k is the growth rate, and the growth rate k = 4 means that the feature image output by each dense layer has a dimension of 4.

3.2. Generation of RoIs and RoIAlign

The corresponding feature map will be generated by extracting the features of the apple image through the ResNet combined with DenseNet. The outputs of the above backbone network are used as inputs of the RPN to further generate a series of region proposals with the probability of being a object. Because of the different shooting distance, the size of fruit in different images varies greatly. And in the process of shooting, there may exist the problem of mutual occlusion between fruits, the occlusion can be upper and lower structure or left and right structure, the horizontal and vertical ratio of fruit exposed area varies greatly. According to the actual situation, 3 different area scales including 16 * 16, 64 * 64, and 128 * 128, and 3 aspect ratios as 1:1, 1:2, 2:1 are randomly combined to generate 9 anchor boxes on the original image for each corresponding pixel on the feature map. Inside the RPN, the classification branch (CLS) and the border regression branch (bbox reg) respectively calculate the various anchors. At the end of RPN, the anchor will be initially screened by summarizing the results of the two branches to get "Proposal", that is, 2 * 9 scores represent object probabilities and 4 * 9 coordinates represent four vertex positions of the target box. The application of RPN only makes the extra cost of a two-layer network. The Proposal obtained at this point are mapped to the feature map obtained in the previous step, which is called Regions of Interest(RoI). RoIs and the corresponding feature maps are fed into RoIAlign which is proposed to improve pixel-accurate of predicting masks that removed the harsh quantization of RoIPooling, properly aligning the extracted features with the input (see Fig. 4).

3.3. Target detection and instance segmentation (FCN)

Next to RoIAlign, Fully Convolutional Networks(FCN) is then used for instance segmentation. According to the traditional Mask R-CNN, there would be three branches that have implemented for classification, bounding-box regression and instance segmentation respectively. Considering the goal to be achieved in this paper, we hope to realize the recognition of apple fruits in a complex background with a single target, thus we improved the model of Mask RCNN. In order to improve the running speed of the algorithm, we removed the classification branch without affecting the segmentation effect. FCN algorithm can accurately segment the object in the picture. It is an end-to-end network including convolution and deconvolution, that is, the image is convolved and pooled to reduce the size of its feature map. Then, deconvolution operation is carried out, continuously increase its resolution of feature map by interpolation operation, and finally classify each pixel. So as to realize the accurate segmentation of the input image.

4. Model training and loss function

4.1. Pre-training with ImageNet

Because of the lack of data sets, transfer learning (Zhang et al., 2019; Zhu et al., 2018) is employed to train the deep learning models more efficient and stable. Using ImageNet (Zhu et al., 2018) (containing 1000 object categories and 1.2 million images) pre-trained CNN features, state-of-the-art results have been obtained on a variety of image

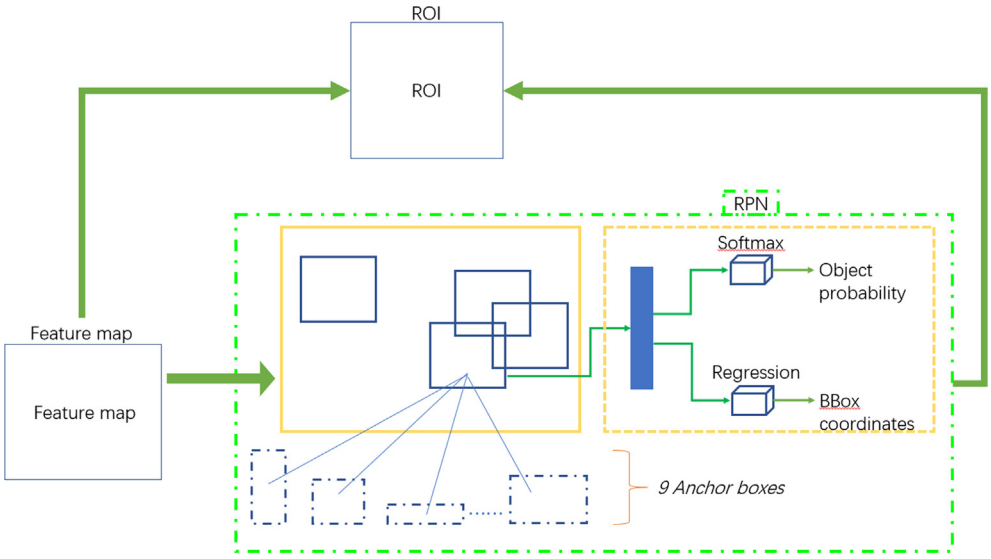

Fig. 4. The structure of RPN.

Table 1
The standard of test.

True	Predicted	Confusion matrix
Positive	Positive	TP
Positive	Negative	FN
Negative	Positive	FP
Negative	Negative	TN

Table 3
Comparison result of Precision.

recognition methods	Unobstructed fruit	Blocking fruit	Overlapping fruit
Literature 29	96.10	87.23	86.89
Literature 30	97.61	91.49	85.25
Literature 31	95.32	89.36	86.89

processing tasks from image classification to image captioning (Krizhevsky et al., 2012; Huh et al., 1608). Considering that there are fruits and even apple-related images in iii, but because of the variety of images and most of them differ greatly from the dataset to be trained, the performance of the original parameters directly using the pre-trained ImageNet to training is unknown. Besides, our labeled dataset is relatively small. Therefore, two alternatives for optimizing the pre-training model are selected here, one is that re-train the layers of the trained model that are close to the output (including the FCN) with our labeled apple image to fine-tune module, and the other is that train the layers that are close to the input with the labeled Apple image training. In the following experiments, we will detail how to compare and select these two methods.

4.2. Loss function

The overall loss of our approach includes two aspects: the loss of classification and regression operations by RPN, and the training loss in the multi-branch predictive network. As shown below:

$$L_{final} = L_{RPN} + L_{multi-branch} \tag{5}$$

where L_{final} represents the total loss, L_{RPN} represents the training loss of the RPN, and $L_{multi-branch}$ represents the training loss due to the branch structure.

Among them, in the training process of RPN, there are two things to do for the generated anchors. The first one is to measure the probability whether the anchor is the foreground or the background, that is, whether the anchor had the target or not. And the second is to perform the preliminary coordinate correction for the anchors belonging to the foreground. Therefore, L_{RPN} consists of RPN classification loss and RPN regression loss. For the previous question, Faster R-CNN's approach is to use Softmax Loss to train directly, eliminated the anchors that transcend the image boundary during training; for the latter problem, use Smooth *L1* Loss for training. So, L_{RPN} is calculated as follows:

$$L_{RPN} = \frac{1}{N_{cls1}} \sum_i L_{cls}(p_i, p_i^*) + \lambda_1 \frac{1}{N_{reg1}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{6}$$

where the formula before the "+" indicates the classification loss, and the latter indicates the regression loss, pi represents the classification probability of anchor i, and pi* represents the ground-truth label probability of anchor *i*; The variable t_i presents the difference between the prediction bounding box and the ground-truth label box in four parameter vectors (the horizontal, vertical coordinate value of the center point in the bounding box; the width and height of the bounding box), and t_i^* indicates the difference between the ground-truth label box and the positive anchor.

With the classification loss of RPN: If an anchor overlaps a box of ground truth with IoU $>=0.7$ then it's positive, and we represented it as 1 here. If an anchor overlaps all box of ground truth with IoU <0.3 then it's negative, represented it as -1. And neutral anchors are those that don't match the conditions above, represented it as 0, and they don't influence the loss function. There are three kinds of real tags: {1, 0, -1}, and the logistic result is distributed from 0 to 1. Since the anchors whose true tags are 0 do not participate in the construction of

Table 2
The specific result with Precision Rate and Recall Rate.

Evaluation parameter	Fruit with an occlusion area $<20\%$	Fruit with an occlusion area $>=20\%$	Overall
Precision Rate/%	98.21%	94.59%	97.31%
Recall Rate/%	97.68%	89.74%	95.70%

Fig. 5. Segmentation result of the algorithm. (a) inter-fruit occlusion, (b) branch occlusion, and (c) leaf occlusion.

the loss function in the RPN classification result, we culled the tag whose labeled 0, and then the −1 label is converted to 0 for cross entropy calculation.

With the regression loss of RPN: Only classes with a label of 1 participate in the regression operation. For target_bbox, although the number of boxes for each picture is the same and equal to the second dimension of rpn_match, only the front N_i boxes are meaningful for picture i (instead of one-to-one correspondence with anchors), it is padded with 0, and the value of N_i is equal to the number of rpn_match of the corresponding picture as 1.

$L_{Mul-Branch}$ is the sum of two branch training losses (SmoothL1 Loss, and Mask Loss (the average binary cross-entropy loss)) in multi-branch predictive networks:

$$L_{Mul-Branch} = L(t_i, t_i^*, s_i, s_i^*) \tag{7}$$

That is

$$L_{Mul-Branch} = \lambda_2 \frac{1}{N_{reg2}} \sum p_i^* L_{reg}(t_i, t_i^*) + \gamma_2 \frac{1}{N_{reg2}} \sum L_{mask}(s_i, s_i^*) \tag{8}$$

In the above formula, the constant N_* represents the number of corresponding anchor points or bounding boxes. The hyperparameters λ_* and γ_* balance the training losses of the regression and mask branches. Here s^* and s respectively represent the mask binary matrices from the prediction and ground-truth label.

When calculating the Loss contribution, the cross entropy will output a value for each box. The class with the largest score of the box does not belong to the Apple class, and In the training process, the class belonging to Apple in the picture is marked as 1, and the other class (which is not belonging to Apple) is marked as 0. When calculating the Loss contribution, the cross entropy will output a value for each box. The class with the largest score of the box does not belong to the Apple class, its Loss will be ignored.

Mask RCNN regression loss function: only calculate the real label non-background class (the number of classifications is greater than 0); Since the prediction has a regression output for each category of each frame, only the regressions of the positive category are calculated.

Mask RCNN mask loss function: In this article, we only specify one class, the Apple class. In the Mask RCNN, the mask branch has an output of Km^2 dimension for each RoI, and K is the number of categories; in this paper, $K = 1$, the mask branch has an output of m^2 dimension for each RoI, and the output is $m * m$. The binary mask. A per-pixel sigmoid is used, and Lmask is defined as the average binary cross entropy loss. Similarly, for this class of RoI, L_{mask} only considers the mask of this class (other mask inputs do not contribute to the loss function).

5. Experiment and results

5.1. Transfer learning method selection

We first tested a total of 368 apple fruits in 120 test sets using the ImageNet pre-training model. There are 34 images in which the target is misidentified, including the fact that the fruit whose occlusion area is too large is missed to recognized, and the leaves are misidentified as fruits. The correct recognition rate of the target fruit reached 86.14%. Obviously, this is an acceptable result. So, it's considered to input our labeled Apple images close to the convolutional neural network's output layer, fine-tuning the training weights and the softmax layer's target to be identified. Convolutional neural networks extract features according to the following rules: shallow convolutional layers extract basic features (such as edges, contours, etc.), and deep convolutional layers extract abstract features (such as shapes, textures, etc.).

The approach taken in this paper is to freeze layers which close to the input of the pre-training module and train the remaining convolutional layers (the convolutional layer near the output and the fully connected layer of the output).

5.2. Precision and recall

According to whether the true sample and the predicted sample match, for the binary classification problem, the sample can be divided into four types: true positive (TP), false positive (FP), true negative (TN) and false negative (FN), as shown in Table1.

Precision(P) and recall(R) are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

120 images of the testset (including 368 fruits) was used to test new module, and the final result was P = 97.16%, R = 96.52%. Table 2 lists the detailed results of each category of apple images.

In order to describe more clearly the effect of recognition and localization method established in this paper, the target fruit recognition and positioning effects of three different growth postures are separately counted here, and the results are listed in Table 3.

It can be seen from Table 1 that several algorithm models can achieve good results for the recognition of unobstructed fruits. However, for the recognition of the target fruit occluded by the foliage and the other fruit, the performance of several methods is different. The new method proposed in this paper has the best recognition and localization effect.

5.3. Evaluation of instance segmentation

We analyzed the generated mask, and achieved a better segmentation effect, especially the recognition of the occluded fruit, and made a great breakthrough. The specific segmentation results are given in Fig. 5.

In order to better observe the segmentation results, we randomly fill each instance area with different colors. It can be seen from the

segmentation results that the algorithm has achieved good results whether the fruits are occluded or the fruits are blocked by the foliage.

6. Conclusions

This paper presents a depth recognition algorithm for the apple picking robot vision system. Successfully extended the deep learning application to the direction of agricultural robot picking. The traditional method has the advantages of low recognition accuracy and poor robustness. Undoubtedly, deep learning has made outstanding progress in feature extraction and target detection, which has made the recognition accuracy a big leap. The vision system of the fruit picking robot is used to process the images captured by the onboard camera. The precision of the recognition and segmentation directly determines the accuracy of the robot picking. In summary, it is necessary to introduce deep learning into the visual system of agricultural picking robots. In the following two directions of identification and segmentation, the algorithm of this paper is summarized.

An algorithm which can automatically detect apples was trained in this paper, and it can output mask of each apples from the model. The fruit detection results of 120 testing images showed that the average detection precision rate and recall rate were 97.31% and 95.70%, respectively. It proves that the algorithm proposed in this paper has high recognition accuracy, and can improve the working efficiency of the picking robot if it is put into practical use. The identification of overlapping fruits has always been a difficult point in the fruit detection. It is difficult to achieve the problem that the small exposed area of the fruit caused by the occlusion of the foliage. The case segmentation proposed from Mask R-CNN solved this problem very well. It can classify the image based on the entity class in the image. There is almost no case where an apple is trapped because of occlusion. Although we have achieved considerable results in this method, there are still many facts that need improvement. The problem of missing samples in deep learning has not been solved. Next, we will work on weakly supervised or unsupervised deep learning model, or find a way to mark the sample instead of manually. And strive to liberate manpower to label samples such as the tedious and monotonous work. Further improve the efficiency of deep learning, so that the method of deep learning is better applied in agricultural image processing.

CRediT authorship contribution statement

Weikuan Jia: Conceptualization, Formal analysis, Resources, Writing - original draft, Supervision, Funding acquisition. **Yuyu Tian:** Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing - original draft, Visualization. **Rong Luo:** Investigation, Resources. **Zhonghua Zhang:** Validation, Investigation. **Jian Lian:** Software, Investigation. **Yuanjie Zheng:** Writing - review & editing, Project administration, Funding acquisition.

Acknowledgments

This work is supported by Focus on Research and Development Plan in Shandong Province (No.: 2019GNC106115); China Postdoctoral Science Foundation (No.: 2018M630797); National Nature Science Foundation of China (No.: 21978139, 61572300); Shandong Province Higher Educational Science and Technology Program (No.: J18KA308); Taishan Scholar Program of Shandong Province of China (No.: TSHW201502038).

References

Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards. IEEE Int. Conf. Robot. Automation (ICRA) 2017, 3626–3633.
Bac, C.W., van Henten, E.J., Hemming, J., et al., 2014. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. J. Field Rob. 31 (6), 888–911.
Zhao, Y., Gong, L., Huang, Y., et al., 2016. A review of key techniques of vision-based control for harvesting robot. Comput. Electron. Agric. 127, 311–323.
Bac, C.W., Hemming, J., Van Tuijl, B.A.J., et al., 2017. Performance evaluation of a harvesting robot for sweet pepper. J. Field Rob. 34 (6), 1123–1139.
Liu, X.Y., Jia, W.K., Ruan, C.Z., et al., 2018. The recognition of apple fruits in plastic bags based on block classification. Precis. Agric. 19 (4), 735–749.
Linker, R., Cohen, O., Naor, A., 2012. Determination of the number of green apples in RGB images recorded in orchards. Comput. Electron. Agric. 81, 45–57.
Cheng, H., Damerow, L., Sun, Y., et al., 2017. Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks. J. Imaging 3 (1), 6.
Tao, Y., Zhou, J., Wang, K., et al. Rapid detection of fruits in orchard scene based on deep neural network. 2018 ASABE Annual International Meeting, 1.
Koirala, A., Walsh, K.B., Wang, Z., et al., 2019. Deep learning–Method overview and review of use for fruit detection and yield estimation. Comput. Electron. Agric. 162, 219–234.
Choi, D., Lee, W.S., Schueller, J.K., et al., 2017. A performance comparison of RGB, NIR, and depth images in immature citrus detection using deep learning algorithms for yield prediction. ASABE Annual International Meeting.
Chen, S.W., Shivakumar, S.S., Dcunha, S., et al., 2017. Counting apples and oranges with deep learning: A data-driven approach. IEEE Rob. Autom. Lett. 2 (2), 781–788.
Tian, Y., Yang, G., Wang, Z., et al., 2019. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput. Electron. Agric. 157, 417–426.
Sa, I., Ge, Z., Dayoub, F., et al., 2016. Deepfruits: A fruit detection system using deep neural networks. Sensors 16 (8), 1222.
He, K., Gkioxari, G., Dollár, P., et al., 2017. Mask R-CNN. Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.
Yu, Y., Zhang, K., Yang, L., et al., 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput. Electron. Agric. 163, 104846.
Deng, X., Zheng, Y., Xu, Y., et al., 2018. Graph cut based automatic aorta segmentation with an adaptive smoothness constraint in 3D abdominal CT images. Neurocomputing 310, 46–58.
Yang, Q., Chen, W.N., Gu, T., et al., 2016. Segment-based predominant learning swarm optimizer for large-scale optimization. IEEE Trans. Cybern. 47 (9), 2896–2910.
Zhu, L., Huang, Z., Li, Z., et al., 2018. Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval. IEEE Trans. Neural Net. Learn. Syst. 29 (11), 5264–5276.
Huang, G., Liu, Z., Laurens, V.D.M., et al., 2017. Densely connected convolutional networks. In: IEEE conference on Computer Vision and Pattern Recognition, pp. 2261–2269.
He, K., Zhang, X., Ren, S., et al., 2016. Identity mappings in deep residual networks. In: European conference on computer vision. Springer, Cham, pp. 630–645.
Sui, X., Zheng, Y., Wei, B., et al., 2017. Choroid segmentation from optical coherence tomography with graph-edge weights learned from deep convolutional neural networks. Neurocomputing 237, 332–341.
Lian, J., Hou, S., Sui, X., et al., 2018. Deblurring retinal optical coherence tomography via a convolutional neural network with anisotropic and double convolution layer. IET Comput. Vision 12 (6), 900–907.
Wang, Q., Zheng, Y., Yang, G., et al., 2017. Multiscale rotation-invariant convolutional neural networks for lung texture classification. IEEE J. Biomed. Health. Inf. 22 (1), 184–195.
Zhang, C., Zhang, H., Qiao, J., et al., 2019. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. IEEE J. Sel. Areas Commun. 37 (6), 1389–1401.
Zhu, L., Huang, Z., Li, Z., et al., 2018. Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval. IEEE Trans. Neural Net. Learn. Syst. 29 (11), 5264–5276.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Adv. Neural Inform. Process. Syst.pp. 1097–1105.
Huh, M., Agrawal, P., Efros, A.A., 2016. What makes ImageNet good for transfer learning? arXiv preprint arXiv:1608.08614.