IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received November 20, 2019, accepted December 3, 2019, date of publication December 11, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958614

Tender Tea Shoots Recognition and Positioning for Picking Robot Using Improved YOLO-V3 Model

HUALIN YANG[1], LONG CHEN[1], MIAOTING CHEN[1], ZHIBIN MA[1], FANG DENG[1], MAOZHEN LI[1,2], AND XIANGRONG LI[1]

[1]College of Mechanical and Electrical Engineering, Qingdao University of Science and Technology, Qingdao 266061, China
[2]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, U.K.

Corresponding author: Fang Deng (dengfhelen@163.com)

This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2019MEE102, in part by the Key Research and Development Program of Shandong Province under Grant 2018GNC112007, and in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J18KA015.

ABSTRACT To recognize the tender shoots for high-quality tea and to determine the picking points accurately and quickly, this paper proposes a method of recognizing the picking points of the tender tea shoots with the improved YOLO-v3 deep convolutional neural network algorithm. This method realizes the end-to-end target detection and the recognition of different postures of high-quality tea shoots, considering both efficiency and accuracy. At first, in order to predict the category and position of tender tea shoots, an image pyramid structure is used to obtain the characteristic map of tea shoots at different scales. The residual network block structure is added to the downsampling part, and the fully connected part is replaced by a 1×1 convolution operation at the end, ensuring accurate identification of the result and simplifying the network structure. The K-means method is used to cluster the dimension of the target box. Finally, the image data set of picking points for high-quality tea shoots is built. The accuracy of the trained model under the verification set is over 90%, which is much higher than the detection accuracy of the research methods.

INDEX TERMS Image recognition, YOLO-v3, convolutional neural network, image pyramid, tea shoot.

I. INTRODUCTION

Currently, the mechanized tea picking still stays at the low-level status, which cuts the tender tea shoots and old leaves at the same time. This cannot meet the picking requirements of high-quality tea. Although many researches about tender tea shoots recognition have been conducted, the effect and speed cannot meet the actual picking requirements. Furthermore, most studies are limited to indoor condition and laboratories, the recognition stability is poor that these operation cannot be applied to the tea garden. Therefore, the recognition of tender tea shoots is an urgent problem needed to be solved. Only by accurately recognizing the tender tea shoots and judging the picking points, can the automatic tea-picking be realized with appropriate mechanical equipment.

Convolutional neural network is also called deep neural network because of its many layers of internal neurons [1]–[3]. Many convolutional neural network algorithms are used to solve the problem of target detection [4]–[11]. These algorithms are represented by R-CNN algorithm [12], [13] and Yolo algorithm proposed by Redmon *et al.* [14] and Redmon and Farhadi [15]. Compared with the traditional target recognition algorithm, R-CNN algorithm greatly improves the accuracy, but the operation speed is slow. Although researchers have proposed Fast R-CNN and other improved algorithms [16]–[18], the processing speed is still slow, until Joseph Redmond's Yolo algorithm makes the operation speed produce a qualitative leap. Then researchers proposed a variety of improved Yolo algorithms [19]–[23].

With the continuous development of computer vision, computer vision has been widely used in the recognition of fruits and plants, etc [24]–[27]. Specific to the tea identification, computer vision has made great progress in the intelligent tea picking. Liu *et al.* [28] used the threshold segmentation method to separate the tender tea shoots from the old leaves and the background, assuming that the shoots tip is at the top, and the shape of the shoots is scanned line

The associate editor coordinating the review of this manuscript and approving it for publication was Yongtao Hao.

by line. Wang [29] extracted the color features and edge features of tea images, and used the region growing method to segment the tender tea shoots in the images. Wu *et al.* [30] first performed median filtering on the captured tea images to reduce noise, and then extracted the G and G-B components in the image, respectively, where the Oust method was used to find the optimal threshold for the tender tea shoots and background. Shao [31] used the rapid watershed algorithm to combine the color information to segment the tender tea shoots of tea images. Finally, the support vector machine was used to divide the segmented tender tea shoots images into one shoots with one leaf, one shoots with two leaves and one shoots with three leaves. Wang and Liu [32] used the segmented samples to train the deep learning model to identify the tender tea shoots. Based on the traditional threshold segmentation method, and combining with K clustering theory, Zhang and Lv [33] extracted the R-B component of the tea image RGB model and the B component of the LAB model for the extraction of the tender tea shoots image.

In the tea garden scene, the picking and identification of tea shoots suffers from problems of the high similarity in color and shape of old and young leaves, and the diversity of lighting conditions caused by the time of day and weather conditions. Although the aforementioned methods can separate the tender tea shoots from the old leaves, they cannot accurately identify the picking points of the shoots, and quickly transfer the coordinates of the picking points to the robot control system. Aiming to solve this problem, this paper proposes an improved Yolo-v3 algorithm for picking points of tender tea shoots.

In this study, We have made a picture data set of high-quality tea with well segmented foreground and background tea shoots. Based on the Yolo-v3 model, the image pyramid structure is first used to fuse different levels of tea shoots, and the tea shoots feature images of different scales are obtained for category and position prediction. Then the target box dimensions are clustered and the prior box is added. The number of anchor boxes enables the network model to learn more information about the edge of the tea shoots. Finally, during the training process, multi-size images are used for training, so that the model can adapt to different resolution pictures. Experiments show that the improved Yolo-v3 algorithm can improve the detection rate while ensuring the detection accuracy.

II. REQUIREMENTS FOR RECOGNITION

According to the filed photography of the tea garden, it is found that there are three situations for the camera installed on the picking manipulator, namely, face-up, side-view and look-down on the tea shoots. Only facing up to the shoots, the Yolo-v3 can quickly and accurately identify the position of the picking point. If the other two situations are identified, the manipulator should be controlled to deflect the corresponding angle.

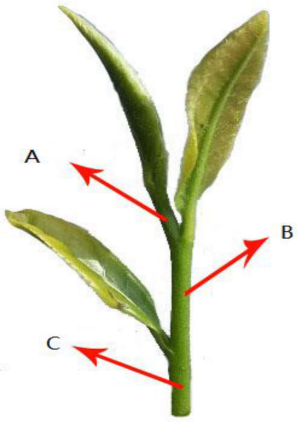


FIGURE 1. Schematic diagram of standard picking points.

A. ANALYSIS ON THE IDENTIFICATION OF PICKING POINTS

According to the different product grade of tea and the process of stir-frying tea, the picking points of shoots can be divided into single shoot picking (A), one shoot with one leaf (B) and one shoot with two leaves (C), among which one shoot with two leaves picking is most widely used. The schematic diagram of the corresponding standard picking point is shown in Fig. 1, which is about 3mm below the junction of the tender leaf and the stem. The picture is an ideal picking situation. In the actual situation, if the shooting angle is not suitable, the shoots will be covered and overlapped, and the picking point cannot be determined. In addition, the posture of tea shoots is various, which is not easy to identify. Focus on the above problems, this paper utilizes the deep convolution neural network to identify the shoot picking points, and obtains the specific picking position coordinates. For the case of leaf occlusion overlap, the network will learn to distinguish in the training process, and classify the occlusion situation, and in addition, control the manipulator to carry out the subsequent corresponding operations.

B. SELECTION OF IDENTIFICATION SCHEME FOR PICKING SHOOTS

When the picking robot works, the control system needs to get the position of the picking points, which requires the identification system to transfer the coordinates of the picking points. The idea is to identify the more characteristic position in the shoots and improve the recognition accuracy. In addition, according to the actual situation, the system is mainly to identify two kinds of shoots, one shoot with one leaf and one shoot with two leaves. There are two ways of identification.

1) IDENTIFY THE JUNCTION BETWEEN TENDER LEAF AND STEM

From Fig. 1, it can be seen that most of the positions where the leaf and stem are connected belong to the area where the shoot and the background coexist, and the green of shoot has a gradual change, which can be used as the basis of recognition. After the recognition, the picking point is calculated

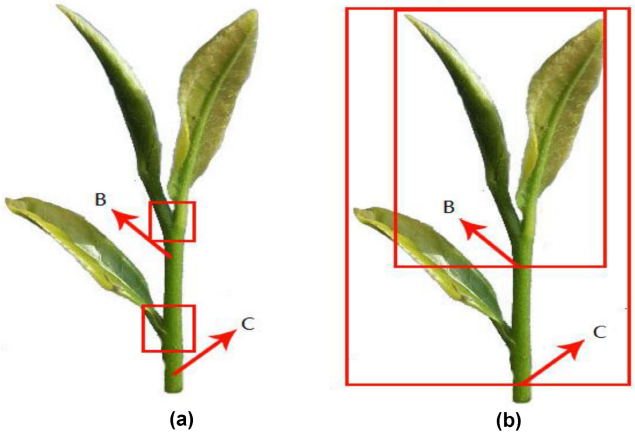


FIGURE 2. Comparison of two identification schemes.

according to the recognition position. Since the tea stem is long and thin, it is easy to detect. Therefore, the tea stem is detected in the tea shoot image. The position extending a fixed length from the center of the recognition frame to the direction of the stem is defined as the picking point. As shown in Fig. 2(a), the box is the connection of the leaf and the stem recognized by the network, and the root of the arrow is the picking point position.

2) IDENTIFY THE WHOLE PICKING SHOOT

The identification of the whole shoot is mainly based on the number of leaves in the image. It is easy to identify the number of leaves in the image based on neural network, and it does not need to consider the change of the subtle color of buds, so the recognition accuracy will be higher. The contact position between the tea stem and the identification box is found by detecting the tea stem, and it is set as the picking point position, see Fig. 2(b). The box represents the whole shoot recognized by the neural network. The root of the arrow is the intersection of the tea stem and the box, which is defined as the shoot picking point.

Comparing the two methods, we can see that the first method has obvious recognition features, but this feature is the same for the recognition of one shoot with one leaf, one shoot with two leaves, so the convolution network can identify the junction but cannot determine which leaf it belongs to. While in the second method, the convolution network only needs to calculate the number of leaves in the box, so the recognition is easier. The determination of subsequent picking points is also simpler than the former, so the latter is finally selected.

3) THE PRINCIPLE OF THE IMPROVED YOLO-V3

The method of multi-scale fusion is used in the prediction of targets by Yolo-v3 deep convolution neural network. A fusion method similar to feature pyramid networks (FPN) is used to predict the location and category of features on multiple scales to improve the accuracy. Use dimension clusters as a prior box to predict the bounding box. K-means [18] method is used to cluster the target frames in the data set, and 9 priori boxes of different sizes are divided into multiple scale feature maps. The larger scale feature maps use smaller priori boxes. This method makes the number of priori boxes finally obtained is more than that of Yolo-v2, and the feature extraction effect is better. In the category prediction, Yolo-v3 does not use Softmax to classify each box, but uses multiple independent logical classifiers. In the training process, binary cross entropy loss is used to predict the category.

C. FEATURE EXTRACTION NETWORK

Compared with the function of the Darknet-19 network in Yolo-v2, Yolo-v3 uses the residual block structure of the residual network to extract features. Residual learning means that each subsequent layer in a deep neural network is only responsible for fine tuning the output from a previous layer by just adding a learned ''residual'' to the input. This differs from a more traditional approach that each layer had to generate the whole desired output. It uses the continuous 3×3 and 1×1 volume accumulation layers and expands to 53 layers, including 53 volume accumulation layers and 5 maximum pooling layers. In order to prevent the occurrence of over fitting, the batch normalization operation is added after each volume accumulation layer under the condition of meeting the speed requirements of real-time detection, the detection accuracy of this network is much higher than that of Darknet-19 network, and the detection effect is better than Resnet-101 and Resnet-152. Yolo-v3 has a good effect on the recognition of small objects, especially for the recognition of the characteristic position of tea shoots. Resnet-101 and Resnet-152 can't make a good recognition of the picking position of small shoots. Specifically, Resnet-101 and Resnet-152 can't give the bounding box including the picking point correctly.

D. IMPROVED METHOD

Firstly, based on the image pyramid structure, the improved YOLO-v3 model is used to fuse the feature maps of different levels. Three groups of predicted feature maps are obtained, where the position and category are predicted on. Secondly, K-means algorithm is used to determine the prior box dimension in the self-made data set of tea shoots picking point. The prior box dimension parameters are distributed to three different scales. For the specific task of identifying the tender shoots picking points, it is necessary to cluster the specific data set to get the corresponding cluster center. Finally, since the full connection layer is removed from the Yolo-v3 model, the input size can be changed during the training process, so that the trained model can adapt to the input images of different scales, in which the input image is the self-made data set of tender shoots picking point.

E. MULTISCALE FEATURE DETECTION BASED ON IMAGE PYRAMID

When using convolution neural network to identify the target, it is found that for the features extracted from the convolution layers, the features of low level have rich location

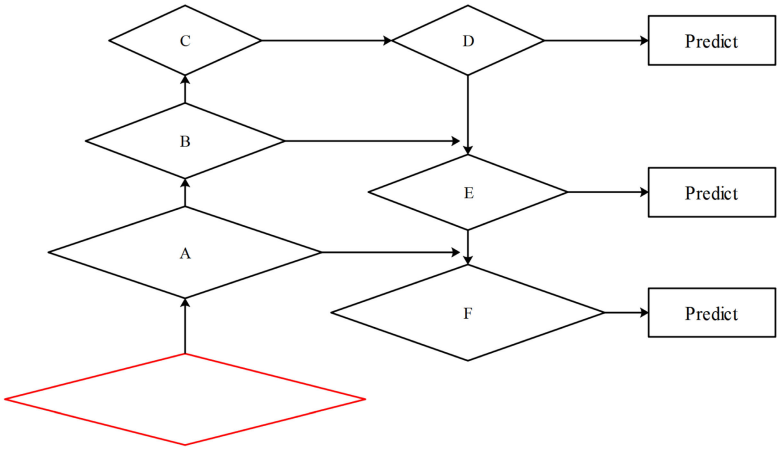


FIGURE 3. Feature fusion pyramid.

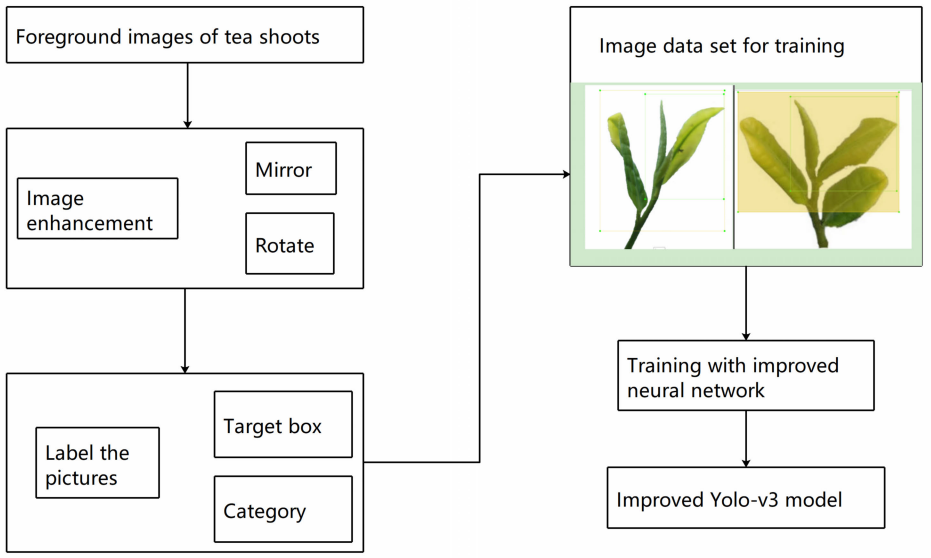


FIGURE 4. Training process of identification network.

and detail information, and the features of high level have rich semantic information. From low level to high level, the location and detail information is reduced, and the semantic information is increased. For prediction location, more low-level features are needed, while for prediction category, more high-level information is needed. Therefore, considering the principle of image pyramid, this paper adds up sampling to the network to fuse the features of low level and high level, so as to achieve the purpose of predicting the location and category of tea shoots picking points at the same time.

The feature pyramid on the right side in Fig. 3 is generated by the feature pyramid on the left side. The whole process is as follows. First, the input image is deeply convoluted, then the features above B are convoluted, and the features in D are upsampled to have the same size. Then the convolution sum is performed on the processed B and D, and the results are input into the E layer. In the same way, feature fusion is carried out among multiple layers to get multiple sets of feature maps for prediction. Based on this method, the processed low-level features and high-level features are accumulated. The purpose of doing this is that the low-level features can provide more accurate location information, and many times of down sampling and up sampling operations make the location information of the deep-seated network produce errors, so it is used together to build a deeper feature pyramid and integrate the multi-level feature information. The prediction is made on different characteristic maps.

Based on the aforementioned idea about image pyramid feature fusion, the improved YOLO-v3 network method is proposed. The upper layer features and the lower layer

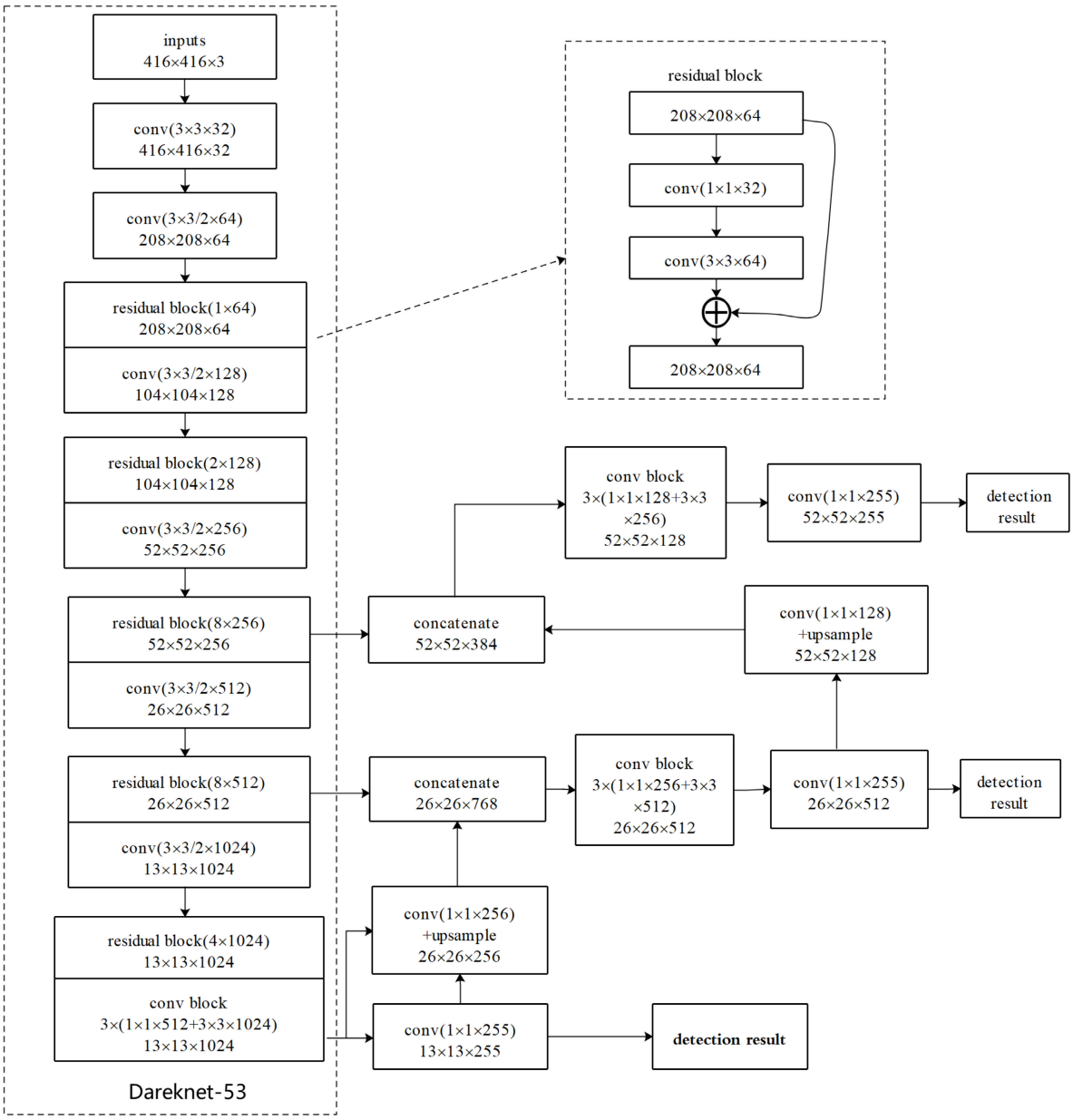


FIGURE 5. The network model for identifying picking points of tea shoot.

features are merged by upsampling, and three sets of feature maps with different scales are obtained. The three sets of feature maps with different scales are used for prediction.

III. MODEL BUILDING AND TRAINING

The training process of the shoot identification network is shown in Fig. 4. To increase the number of images, the images that have been segmented out of the foreground need to be mirrored and rotated. The training samples are labeled manually, using the boundary box labeling and the category labeling method. The improved Yolo-v3 network is used as the training model, multi-scale training strategy is adopted, and the boundary box and foreground area information of training data set are used for training.

A. CONSTRUCTION OF NETWORK STRCTURE

1) THE PRINCIPLE OF NETWORK

The specific network structure of the shoot recognition model is shown in Fig. 5. The left side is the downsampling part of the model, which is called Darknet-53. The whole part includes 53 convolution operations. There are 5 downsampling operations in total, each sampling step is 2. The downsampling operation reduces the image to one quarter of the

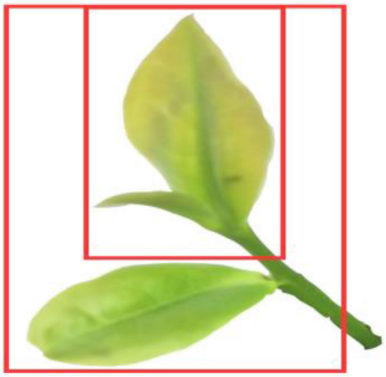


FIGURE 6. The target boxes of shoot.

original image. After 5 downsampling, the resolution changes from 416 × 416 to 13 × 13.

In order to better retain the image information, the residual block structure in the residual network is added in the downsampling process. The structure is shown in the dotted box above, which can avoid the network structure to be too deep under the premise of ensuring accurate results. Darknet-53 uses convolution operation instead of the original maximum pooling operation to ensure that the image details are not lost. At the same time, the last full connection part is replaced by 1 × 1 convolution operation, which can achieve the same effect but greatly reduce the running time.

The right side of Fig. 5 is the feature map of three scales, 13 × 13, 26 × 26 and 52 × 52, upsampled from the bottom. Taking the feature map of 13 × 13 as an example, the input image is divided into 13 × 13 or 169 copies. Each point in the feature map is the prediction result of the corresponding region of the original image.

Another two scale feature map can be obtained by upsampling the downsampled results and merging with the original corresponding layer. The more intensive the scale division, the smaller objects can be detected by the network, which can ensure more accurate target positioning. Finally, three scale feature map outputs the prediction box with the highest Intersection over Union (IOU) value after non maximum suppression of the detected object. IOU evaluates the performance of the model by calculating the overlap ratio between the predicted bounding box and the true bounding box as follows.

$$IOU = \frac{S_1}{S_2}, \tag{1}$$

where S_1 is the area of intersection of the predicted bounding box and the true bounding box, S_2 is the area of the union of the two bounding boxes.

The details of the specific improvements are as follows. Firstly, the feature pyramid of image is obtained by Darknet-53, and a set of Yolo layers to be processed are obtained by continuous convolution operation of Conv53 layer. Then, a group of convolution operations are carried out to get the small scale Yolo layer. At the same time, the layer is upsampled, making convolution sum with Conv45 layer in Darknet-53. In the same way, continuous convolution operation is used to get the second set of the successive Yolo layer, and then a set of convolution operation is carried out to get the mesoscale Yolo layer. Meanwhile, the layer is upsampled, making convolution sum with Conv29 layer in Darknet-53, and the same continuous convolution operation is used to get the third set of the successive Yolo layer, and then a set of convolution operation is carried out to get the large-scale Yolo layer. Finally, after the above operations, three groups of Yolo feature layers with different scales are obtained, and these three groups of feature layers are used to predict the location and category.

2) ELEMENTS OF NETWORK OUTPUT

The recognition of tea shoots picking points requires not only the prediction of categories but also the location of targets. The location can be determined by the coordinates of the upper left corner and the lower right corner of the rectangular box. There are four elements in the coordinates of the two points and four categories in the identification of tea shoots. One shoot with one leaf is marked with ''One'', one shoot with two leaves is marked with ''Two'', obstructed shoot shot by the camera on the side of tea is marked with ''Side'' and the tea-stem free buds shot by the camera on the top of tea is marked with ''Top''. In order to ensure the accuracy of the prediction box, it is necessary to calculate the confidence scores, which is used to measure the probability that the prediction box containing the tea shoots and the IOU value between the prediction box and the manual annotation. A group of output results contains nine elements, including four location elements, four category elements and one confidence element. The *Confidence* is defined as follows:

$$Confidence = p_r(Object) \times IOU, p_r(Object) \in \{0, 1\}. \tag{2}$$

Noting that when the target is in the grid, $p_r(Object) = 1$ and 0 otherwise.

3) THE DESIGN OF THE ANCHOR BOXES

In the process of recognition, the ''Side'' and the ''Top'' do not appear together, but the ''One'' and the ''Two'' often appear together. As shown in Fig. 6, the target box of ''One'' overlaps the target frame of ''Two''. In order to recognize both of them in the detection process, the target box style of each category can be set, that is, the size of anchor boxes can be defined. Most of the target boxes of ''One'' and ''Two'' are rectangles with a larger aspect ratio, while the target box of ''One'' and ''Two'' tend to be square. In the network, the K-mean clustering algorithm is used to cluster the target boxes in the annotation file, and nine anchor boxes are obtained, which are (161,246), (271,345), (311,534), (457,400), (461,664), (607,899), (660,524), (897,802) and (1209,1306). Nine anchor boxes are randomly and evenly assigned to three feature maps of different scales. Each anchor box corresponds to a group of outputs when the results are output, so the final output dimension is 3 × 9 = 27.

FIGURE 7. Three kinds of pictures of High-quality tea.

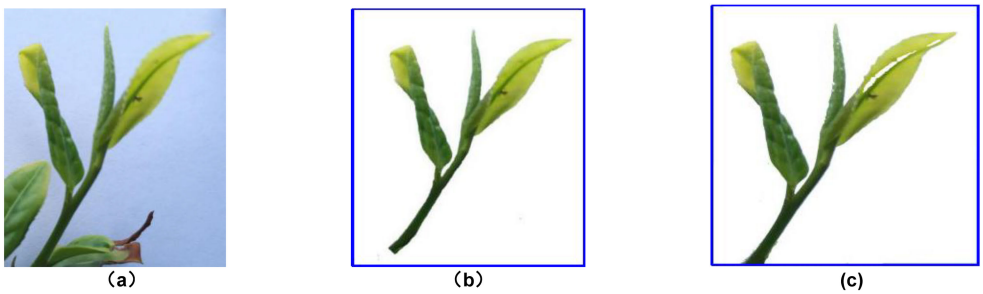


FIGURE 8. Sample images of tender shoots.

B. TRAINING OF NETWORK MODEL

1) EXPERIMENT PLATFORM

In this paper, all the training and testing of the recognition model are written by Python, the Tensorflow-GPU version is installed by Anaconda 3 platform, the Keras library is called, the program is written by Pycharm software and run under Windows 10 system.

2) COLLECTION OF HIGH-QUALITY TEA PICTURES

More than 5000 images were collected in several tea gardens in Laoshan District, Qingdao City, Shandong Province, China. As shown in Fig. 7, in case (a) where the camera is facing the shoot, the junction of stem and shoot is clearly visible, and the picking position is easy to determine. In case (b), the camera is on the side of the shoot, and the shoot overlap with the tea stem. The junction of stem and shoot is invisible, and the picking point cannot be recognized. In case (c), the camera shoots the shoot in a bird' eye view, the shoot tip and the leaves are mostly visible, the tea stem are completely obscured, and the picking points cannot be recognized. All of the above three cases of tea images will appear. For the shoot in side view and top view state, the camera can be controlled to rotate to a certain angle and then take a positive view of the shoots to complete the identification of the picking points.

3) PREPROCESSING OF IMAGE DATA SET

Two methods are used to segment all the images, one is manual segmentation, the other one is the improved PSO-SVM algorithm. There are two reasons for this. One is to compare with the effect of complete artificial classification of tea shoots in the recognition experiment, and the other is to increase the number of different types of image samples. In SVM algorithm, the determination and optimization of penalty factor c and kernel function parameter g has a great influence on the final image segmentation results. Therefore, PSO algorithm is used to optimize the parameters of c and g in SVM to get the best image segmentation results. As shown in Fig. 8, Fig. 8(a) is the original image of the shoot, Fig. 8(b) is the manually segmented image of shoot, and Fig. 8(c) is the image of shoot segmented by the improved PSO-SVM algorithm. Comparing the two images, we can see that there are some flaws in the algorithm segmented shoot, which is shown in the figure as the elongated cavities caused by over segmentation in the leaves. The defects in other segmented images and the impurities caused by insufficient segmentation are fused with the shoots. However, the final identification of the tea shoots picking points is still to be carried out in the images segmented by the algorithm, so the tea shoot sample segmented by the algorithm is adopted to build the tea shoots picking points data set.

In the case of small training set and lack of samples, the overfitting problem with high accuracy in training set and low accuracy in test set often occurs. Because of the complexity of the determination of the shoot picking point, a large number of samples are needed. In order to avoid the occurrence of overfitting, it is necessary to enhance the original shoot sample data set. As shown in Fig. 11, the number of samples is mainly increased by mirroring, rotating, scaling and changing the background of all shoot image samples.

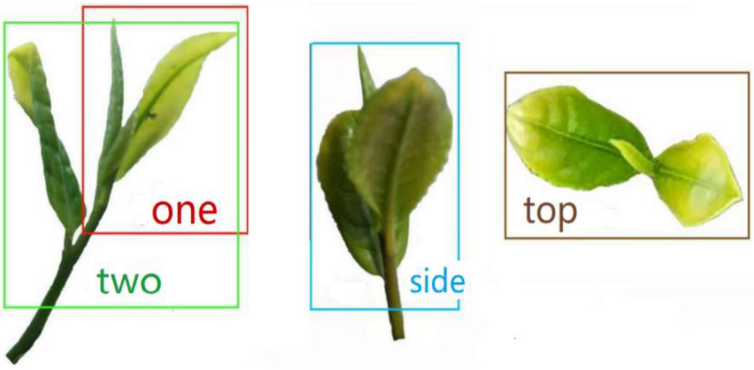


FIGURE 9. Labeling of shoots images.

After manual and algorithmic segmentation of the tea image, data enhancement operation is carried out, and the final number of samples of shoot image is 40,000, which are trained, cross verified and tested respectively according to the ratio of 9:0.5:0.5. At the same time, 70% of the samples in the cross validation set and the test set are segmented by the algorithm, which can achieve a more real and effective training effect.

After making the samples of tea image, we need to use a software named Labeling to mark the tender shoot in the image. As shown in Fig. 9, there are four types of shoot targets, namely, ''One'', ''Two'', ''Side'' and ''Top''. It is of great importance to identify the picking position on the tea stem. Although the overall IOU of shoots is not the most important factor, the frame selection of shoots in the labeling process is still to try to ensure the integrity of the tender shoots and improve the accuracy of neural network training. The annotation information of tea image will be saved in the form of XML file, including image name, storage path, resolution, category of annotation box and coordinates of annotation box.

4) TESTING AND EVALUATION OF NETWORK MODEL

Loss function is used to estimate the inconsistency between the predicted value and the real value. It is a non negative real-valued function. The smaller the loss function, the better the robustness of the model. The shoot image samples of different scale are read into the picking point recognition model. In the training process, the batch size is set to 2, which can ensure the normal operation of the computer and reduce the training time. After 50 hours of training, the recognition model that meets the recognition requirements is obtained. Fig. 10 shows loss changes with each complete iteration. As the number of iterations increases, the loss function of the model decreases. In the early stage of model training, the loss function is as high as 4,000. After 10,000 times of training, the loss function quickly drops to 20. The decline curve in the early stage has little effect on the training completion of the analysis model, which is not shown in the figure. It can be seen that it is difficult to continue to reduce the loss value after it reaches 12. When the loss value drops below 9, the curve

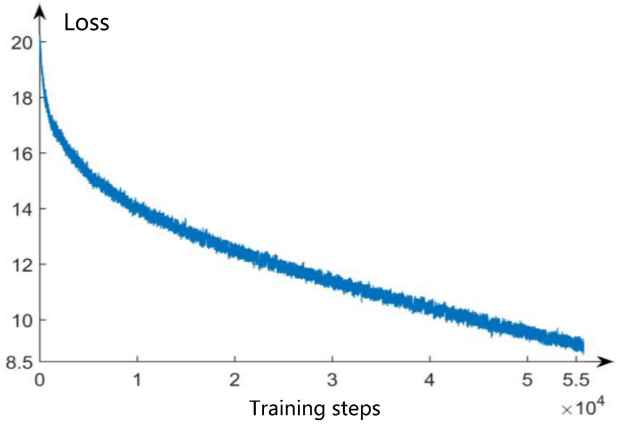


FIGURE 10. Loss decline curve of the model.

is almost flat, and the model has been trained well. It means that we can use this model for identification.

IV. EXPERIMENTS AND DISCUSSION

A. IMAGE OF RECOGNITION RESULTS

The trained model is used to identify the image of tea shoot, and the recognition time for each picture is kept at 4*s*, The recognition results are shown in Fig. 12. The manually segmented image of tea shoots are shown in Fig. 12(a)-(c). Results show that the determination of the types of shoot and the box selection of the target position are basically correct, but the box selection of ''One'' and ''Two'' is wrong, and the tea stem is framed at the bottom of the third leaf in Fig. 12(c). In this case, the shoots extracted are one shoot with three leaves. In Fig. 12(d)-(f), the final recognition effect of the tea shoot image segmented by the algorithm meets the requirements of picking. In Fig. 12(d), there is over-segmentation at the tea stem, so the final position of the identified tea stem is not very accurate, but it is caused by incomplete image preprocessing. In Fig. 12(g)-(h), there are multiple tea shoots in an image. From the test results, it can be seen that when the number of tea shoots in an image is not unique, the model can also identify and locate the shape of each shoot well.

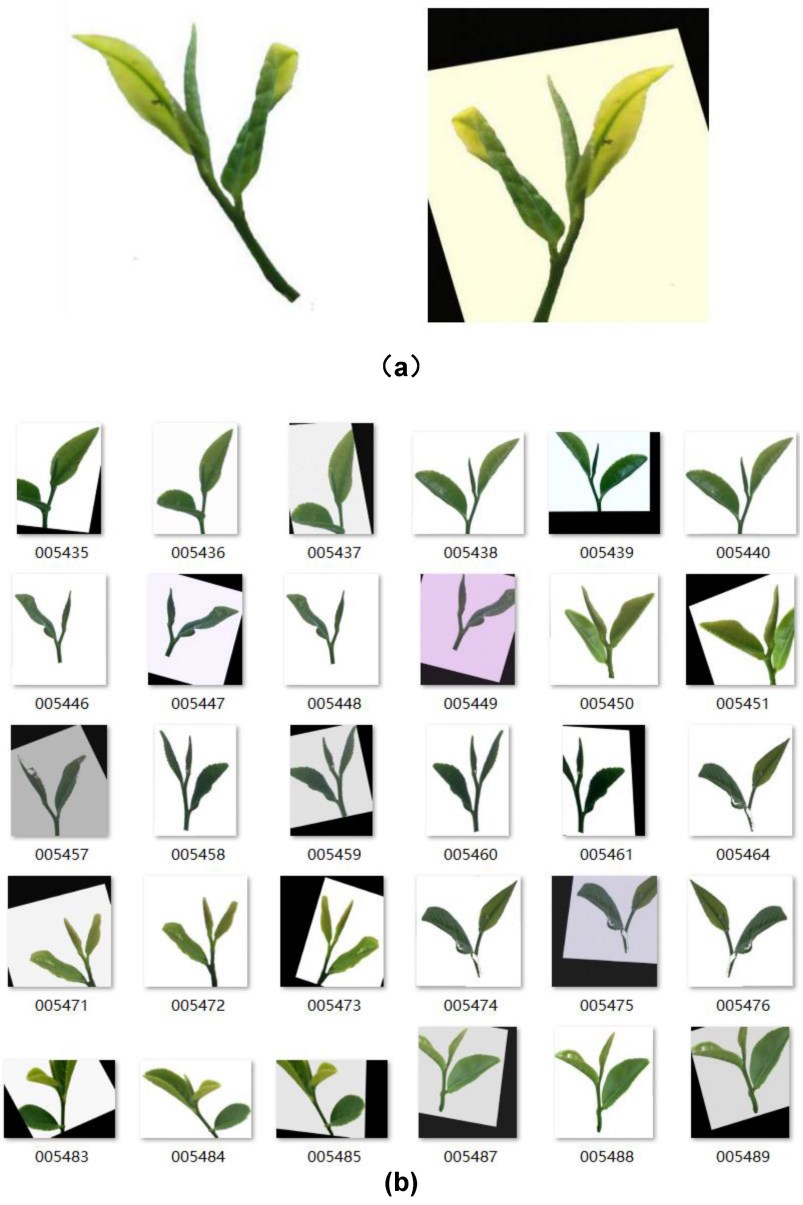


FIGURE 11. Data enhancement operation.

B. ASSESSMENT FOR THE EFFECTIVENESS OF THE NETWORK MODEL

Fig. 12 shows the recognition results, it can be seen that the model has the recognition ability for the shoot image with better shooting angle, and the position of the shoot prediction box is also relatively accurate. The prediction boxes of the eight maps all appear the problem that the range of the box is too large or the leaves are cut, which is also the reason that the prediction confidence is not high. However, the ultimate purpose of the experiment is to find the shoot picking point, and to determine the correct position of the stem in the prediction box. The above-mentioned bounding boxes all contain the position of the picking point accurately and the size of prediction box will not affect the final picking results, so the confidence is not the strict criterion of this experiment.

In order to better evaluate the performance of the model, the model is used to recognize the tea pictures in the test set. The related indicators for evaluating the effectiveness of the neural network models include three aspects: precision, recall and F-measure.

For the problem of tea bud recognition, the output samples can be divided into the three types: true positive (TP) which refers to the number of correctly identified samples, false positive (FP) which refers to the number of samples with wrong identification, true negative (TN) which refers to the number of unrecognized samples. Precision (P) and recall (R)

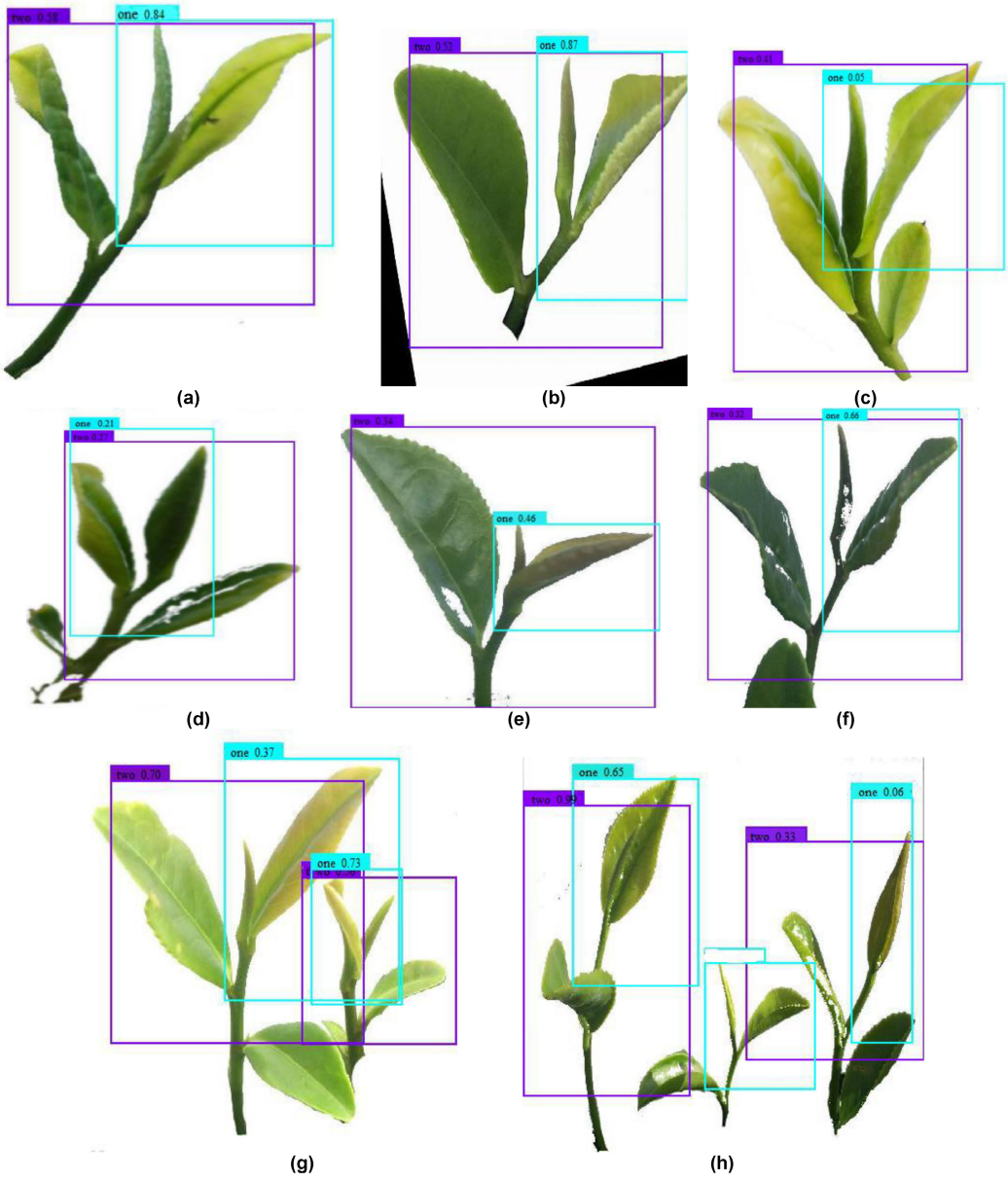


FIGURE 12. Recognition results.

are defined as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + TN}. \tag{3}$$

Noting that when there is a large gap between Precision(P) and recall(R), F-measure is required to comprehensively consider these two parameters:

$$F - Measure = \frac{2 \times P \times R}{P + R}. \tag{4}$$

The loss function in YOLO is defined as follows:

$$Loss = Error_{coord} + Error_{iou} + Error_{cls}, \tag{5}$$

$$Error_{coord} = \lambda_{coord} \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] + \lambda_{coord} \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} [(w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2], \tag{6}$$

$$Error_{iou} = (1 + \lambda_{noobj}) \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} [(C_i - \hat{C}_i)^2 + (h_i - \hat{h}_i)^2], \tag{7}$$

$$Error_{cls} = \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2, \tag{8}$$

where λ_{coord} is the weight of the coordinate error, S^2 is the number of grids in the input image, and B is the

TABLE 1. Test results of four types of shoots.

Type	TP	FP	TN	P	R	F-Measure
One	1762	153	189	92.01%	90.31%	91.15%
Two	1802	144	176	92.60%	91.10%	91.84%
Side	130	8	115	94.20%	53.06%	67.88%
Top	183	11	26	94.33%	87.56%	90.82%
Total	3877	316	506	92.46%	88.46%	90.42%

TABLE 2. Test results of four types of shoots by changing image samples.

Type	TP	FP	TN	P	R	F-Measure
One	1725	164	172	91.32%	90.93%	91.12%
Two	1788	151	164	92.21%	91.60%	91.90%
Side	194	20	21	90.65%	90.23%	90.44%
Top	190	18	16	91.35%	92.23%	91.79%
Total	3897	353	373	91.69%	91.26%	91.47%

TABLE 3. Test results of two kinds of shoots.

Type of sample	TP	FP	TN	P	P	F-Measure
A	1625	128	80	92.70%	95.31%	93.99%
B	1555	168	100	90.25%	93.96%	93.07%

TABLE 4. Experimental results of four algorithms under different lighting conditions.

Light conditions	Networks	F-Measure/%			
		The first	The second	The third	Average
Day time	Faster RCNN	78.36	75.99	79.21	76.23
	Yolo-V2	84.21	84.66	83.65	84.03
	Yolo-V3	90.36	88.98	89.01	89.56
	Improved Yolo-V3	92.33	91.65	91.68	92.19
Night fall	Faster RCNN	70.23	70.65	71.23	71.56
	Yolo-V2	80.33	81.56	81.37	81.58
	Yolo-V3	86.36	86.65	85.91	86.02
	Improved Yolo-V3	90.15	89.96	88.65	90.23
Cloudy	Faster RCNN	72.36	71.89	71.48	72.63
	Yolo-V2	81.26	81.48	81.96	81.45
	Yolo-V3	87.28	87.49	87.34	87.66
	Improved Yolo-V3	90.12	90.66	90.64	90.88
Sunny	Faster RCNN	79.88	79.30	80.02	80.11
	Yolo-V2	84.22	84.51	84.76	85.05
	Yolo-V3	88.23	88.71	89.23	89.14
	Improved Yolo-V3	92.36	92.87	91.56	93.06

number of bounding boxes generated by each grid. Referring to the original parameters in the Yolo-v3 model, $\lambda_{coord} = 5$, $S = 7$, and $B = 9$ were selected in this study. $1_{ij}^{obj} = 1$ denotes that the object falls into the bounding box in grid i, otherwise $1_{ij}^{obj} = 0$. $(\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ are values of the center coordinate, height, and width of the predicted bounding box. (x_i, y_i, w_i, h_i) are true values. The parameter λ_{noobj} is the weight of the *IOU* error. Referring to the original parameters of the Yolo-v3 model, $\lambda_{noobj} = 0.5$ was selected in this study. $\hat{C}_i$ is the predicted confidence, C_i is the true confidence. The parameter c refers to the class to which the detected target belongs. $p_i(c)$ refers to the true probability that the object belonging to class c is in grid i. $\hat{p}_i(c)$ is the predicted value. The $Error_{cls}$ for grid i is the sum of classification errors for all the objects in the grid.

C. COMPARISON OF DIFFERENT EXPERIMENTS

TABLE 1 shows the test results of four types of shoots excepting the "Side" shoots, we can see that the F-measure of the other three types of shoots is more than 90%, and the difference is not big. The false positive number of the "side" shoot is low, but the number of unrecognized samples is almost the same as the number of correctly recognized samples, which results in a high precision (P) of the "Side" shoots, but a very low recall rate. This situation is mainly due to the fact that a small part of the exposed bud tip is often included in the side shooting of tea, which leads to the misidentification of "Side" shoots as "One" and "Two". After the remake of the image dataset, F-measure is greatly improved by adding a part of image samples with exposed tip to the "Side" shoots as shown in TABLE 2. TABLE 3 shows the results of using the manual segmented samples (A) and

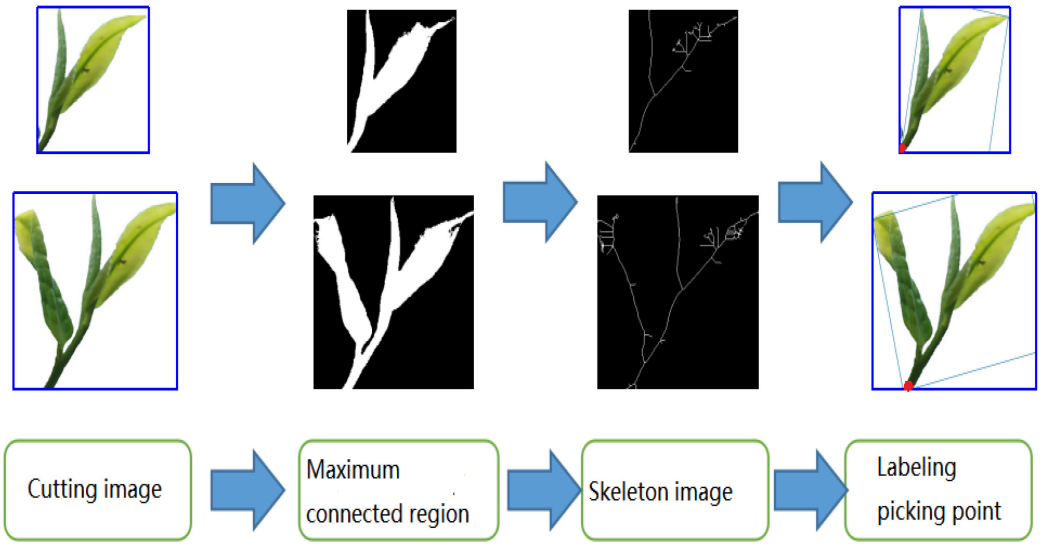


FIGURE 13. Calculation process of picking points.

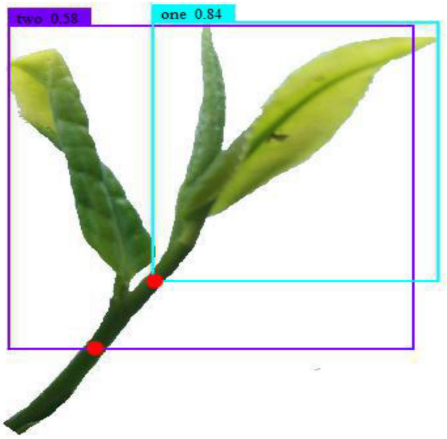


FIGURE 14. The result of picking points recognition.

the algorithm-segmented samples (B) to identify under equal conditions. It can be seen that there is not much difference between the accuracy of the two identifications. In order to show that the neural network model has a good performance in tea recognition, four commonly used algorithms are used for tea image recognition experiments with light conditions as control variables, and the results are compared, as shown in TABLE 4. Through the check experiment, we can see that the F-measure of the picking point recognition model is more than 90% under normal picking conditions, which meets the requirements of High-quality tea picking. It is proved that this algorithm can be used to identify the picking points of High-quality tea.

V. COORDINATE OF THE TEA PICKING POINT

The neural network model not only outputs the shoot type of the image, but also outputs the coordinate of the upper left corner and the lower right corner of the prediction box, as shown in Fig. 12. The output coordinate values are: One: (174,27) and (366,253), Two: (17,24) and (394,327), respectively. Using these coordinate values, it is very convenient to cut the image of the target area, so that the next step is to get the accurate shoot picking point. The cutting image is used to get the picking points. The idea is to extract the skeleton information of tea shoots and find the location of tea stems. According to the growth characteristics of tea shoots and shooting angle of camera, it can be seen that tea shoot and tea stem are located at the bottom of the image, which may be left or right. However, as long as finding the smallest circumscribed rectangle of the shoots and scaning the skeleton point at the bottom of the rectangle, we can determine it as the shoot picking point. Since some other leaf pixels may be included in the cutting process, it is necessary to find the largest connected area in the image and remove other leaves before extracting the shoot skeleton, which is also an important step to ensure the accuracy of the calculation of shoot picking point.

The specific process of calculating picking point is shown in Fig. 13. The binary processing is carried out for the clipped image, and the pixel points in the discrete area of the binary image are counted, only the area with the largest number of pixel points is reserved, so as to get the maximum connected area of the shoot. The central skeleton of the shoot image is obtained by cyclic corrosion of the largest connected region. Calculating the minimum circumscribed rectangle of the central skeleton, finding the lowest point of the skeleton according to the boundary information of the rectangle, and relying on the minimum circumscribed rectangle can effectively prevent the interference of oblique growth of shoots on the judgment of picking points.

After a series of processing, the coordinates of the picking points are: the coordinates of the picking points of ‘‘One’’ is

(185,249) and "Two" is (132,324). The determination of the picking points in the original drawing is shown in Fig. 14.

VI. CONCLUSION

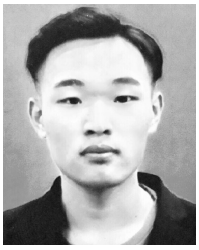
According to the requirements of High-quality tea shoots picking, the appropriate recognition method is determined and the improved Yolo-v3 is proposed. Images dataset with large number of tea images is built for model training. Experimental results show that the model has high detection precision and stronger robustness to shoots with different posture and occlusion. The results are concluded as follows:

- According to the principle of image pyramid structure, Yolo-v3 model is improved to obtain the characteristic map of different scales of tea shoots, so that the model can recognize the types and positions of shoots better.
- Four kinds of network models are used to do comparative experiments, the recognition precision is more than 90%. which verifies the superior performance of the improved Yolo-v3 model for the High-quality tea picking.
- The picking points of the identified tea shoots are calculated. The skeleton is extracted from the image of the buds in the prediction frame, and the lowest point of the skeleton is found to determine the position of the picking points. The experiment shows that the method is feasible, which can accurately locate the picking points and solve the two-dimensional coordinates.

REFERENCES

[1] S.-J. Lee, T. Chen, L. Yu, and C.-H. Lai, "Image classification based on the boost convolutional neural network," *IEEE Access*, vol. 6, pp. 12755–12768, 2018.
[2] A. Esmaeilzehi, M. O. Ahmad, and M. N. S. Swamy, "Compnet: A new scheme for single image super resolution based on deep convolutional neural network," *IEEE Access*, vol. 6, pp. 59963–59974, 2018.
[3] Y. Shan and S. Li, "Descriptor matching for a discrete spherical image with a convolutional neural network," *IEEE Access*, vol. 6, pp. 20748–20755, 2018.
[4] P. A. Dias, A. Tabb, and H. Medeiros, "Apple flower detection using deep convolutional networks," *Comput. Ind.*, vol. 99, pp. 17–28, Aug. 2018.
[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2012, pp. 1097–1105.
[7] W. Rawat and Z. Wang, "Deep convolution neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017.
[8] J. Tang, D. Wang, Z. Zhang, L. He, J. Xin, and Y. Xu, "Weed identification based on K-means feature learning combined with convolutional neural network," *Comput. Electron. Agricult.*, vol. 135, pp. 63–70, Apr. 2017.
[9] H. Cheng, L. Damerow, Y. Sun, and M. Blanke, "Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks," *J. Imag.*, vol. 3, no. 1, pp. 6–18, Jan. 2017.
[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[11] B. Kang, S. Tripathi, and T. Q. Nguyen, "Real-time sign language fingerspelling recognition using convolutional neural networks from depth map," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, 2015, pp. 136–140.
[12] X. Deng, Y. Zhang, S. Yang, P. Tan, L. Chang, Y. Yuan, and H. Wang, "Joint hand detection and rotation estimation using CNN," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 1888–1900, Apr. 2018.
[13] L. Yang, Z. Qi, and Z. H. Liu, "An embedded implementation of CNN-based hand detection and orientation estimation algorithm," *Mach. Vis. Appl.*, vol. 30, no. 6, pp. 1071–1082, Jun. 2019.
[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.
[15] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6517–6525.
[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.
[17] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[19] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved YOLO-V3 model," *Comput. Electron. Agricult.*, vol. 157, pp. 417–426, Feb. 2019.
[20] J. Moran, L. Haibo, Z. Wang, H. Bin, and C. Zheng, "The application of improved YOLO V3 in multi-scale target detection," *Appl. Sci.*, vol. 9, no. 18, pp. 3775–3788, Sep. 2019.
[21] W. He, Z. Huang, Z. Wei, C. Li, and B. Guo, "TF-YOLO: An improved incremental network for real-time object detection," *Appl. Sci.*, vol. 9, no. 16, pp. 3225–3240, Aug. 2019.
[22] X. Wang, T. Xu, J. Zhang, S. Chen, and Y. Zhang, "SO-YOLO based WBC detection with Fourier ptychographic microscopy," *IEEE Access*, vol. 7, pp. 51566–51576, Aug. 2018.
[23] H. Chen, Z. He, B. Shi, and T. Zhong, "Research on recognition method of electrical components based on YOLO V3," *IEEE Access*, vol. 7, pp. 157818–157829, Oct. 2019.
[24] Y. Xue, N. Huang, S. Tu, L. Mao, A. Yang, X. Zhu, X. Yang, and P. Chen, "Immature mange detection based on improved YOLOv2," *Trans. Chin. Soc. Agricult. Eng.*, vol. 34, no. 7, pp. 173–179, Apr. 2018.
[25] D. Zhao, R. Wu, X. Liu, and Y. Zhao, "Apple positioning based on YOLO deep convolutional neural network for picking robot in complex background," *Trans. Chin. Soc. Agricult. Eng.*, vol. 35, no. 3, pp. 164–173, Feb. 2018.
[26] X. Wu, L. Gao, M. Yan, and F. Zhao, "Flower species recognition based on fusion of multiple features," *J. Beijing Forestry Univ.*, vol. 39, no. 4, pp. 86–93, Apr. 2017.
[27] L. Fu, Y. Feng, T. Elkamil, Z. Liu, R. Liu, and Y. Cui, "Image recognition method of multi-cluster kiwifruit in field based on convolutional neural networks," *Trans. Chin. Soc. Agricult. Eng.*, vol. 34, no. 2, pp. 205–211, Jan. 2018.
[28] Z. Liu, Y. Tian, L. Yang, F. Yang, and Q. Yang, "Automatic detection of tea shoots under overlapping conditions," *Chin. J. Stereol. Image Anal.*, vol. 14, no. 2, pp. 129–132, 2009.
[29] J. Wang, "Segmentation algorithm of tea combined withthe color and region growing," *J. Tea Sci.*, vol. 31, no. 1, pp. 72–77, 2011.
[30] X. Wu, F. Zhang, and J. Lv, "Study on the method of tea tender leaf recognition based on image color information," *J. Tea Sci.*, vol. 33, no. 6, pp. 584–589, 2013.
[31] M. Shao, "Research on computer vision based recognition methods of Longjing tea sprouts," China Jiliang Univ., Hangzhou, China, Tech. Rep., 2013.
[32] K. Wang and D. Liu, "Intelligent identification for tea statebased on deep learning," *J. Chongqing Inst. Technol.*, vol. 29, no. 12, pp. 120–126, 2015.
[33] K. Zhang and J. Lv, "Studyon automatic segmentation of tea sprouts under natural conditions," *J. Heilongjiang August First Land Reclamation Univ.*, vol. 28, no. 2, pp. 100–104, Apr. 2016.

HUALIN YANG received the M.S. degree in mechanical manufacturing and automation from the Qingdao University of Science and Technology, Qingdao, China, in 2003, and the Ph.D. degree in chemical machinery from Zhejiang University, Hangzhou, China, in 2006. His current research interests include intelligent manufacturing, computer vision, mechanical design, control, and automation.



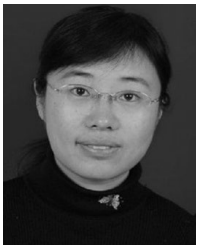
LONG CHEN received the B.S. degree in mechanical engineering from the Qingdao University of Science and Technology, Qingdao, China, in 2018, where he is currently pursuing the M.S. degree. His current research interests include intelligent manufacturing, mechanical design, and computer vision.

MIAOTING CHEN received the M.S. degree from the Qingdao University of Science and Technology, Qingdao, China, in 2019. She is currently working as a Teaching Assistant with the Qingdao University of Science and Technology. Her current research interests include deep learning and computer vision.

ZHIBIN MA is currently pursuing the M.S. degree with the Qingdao University of Science and Technology, Qingdao, China. His current research interests include intelligent manufacturing, mechanical design, and computer vision.

FANG DENG received the M.S. degree in chemical machinery from Zhejiang University, Hangzhou, China, in 2006, and the Ph.D. degree in mechanical engineering from the Qingdao University of Science and Technology, Qingdao, China, in 2019. Her current research interests include nonlinear control and estimation, adaptive control, and marine crafts control.

MAOZHEN LI received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, in 1997. He is currently a Professor with the Department of Electronic and Computer Engineering, Brunel University London, U.K. His main research interests include high performance computing, big data analytics, and intelligent systems with applications to smart grid, smart manufacturing, and smart cities. He has over 180 research publications in these areas, including four books. He is also a Fellow of the British Computer Society and the IET. He has served over 30 IEEE conferences. He is on the Editorial Board of a number of journals.



XIANGRONG LI received the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2006. Her research interests include signal acquisition and processing, and mechatronics design.