Identification of Growing Points of Cotton Main Stem Based on Convolutional Neural Network

CHUNSHAN WANG1,2,3,4, SIQI HE1, HUARUI WU2,3, GUIFA TENG1,4, CHUNJIANG ZHAO2,3, AND JIUXI LI5

1School of Information Science and Technology, Hebei Agricultural University, Baoding 071001, China
2National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
3Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China
4Hebei Province Key Laboratory of Agricultural Big Data, Baoding 071001, China
5School of Mechanical and Electrical Engineering, Hebei Agricultural University, Baoding 071001, China

Corresponding author: Chunjiang Zhao (zhaocj@nercita.org.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871041 and Grant 61771058, in part by the Innovation Funding Project for Postgraduates in Hebei Province under Grant CXZZBS2020103, and in part by the Scientific Research Projects of Universities in Hebei Province under Grant KY202004. They respectively provided equipment development, experimental location, experimental conditions and technical support for the research work of this article.

ABSTRACT Identification of growing points of cotton main stem is the key to realize intelligent and precise machine topping and chemical topping. The identification of growing points by applying the traditional target detection model is subject to a series of shortcomings such as low accuracy, slow identification speed, large number of model parameters, huge storage cost and high calculation workload. On the basis of the advantages and disadvantages of YOLOv3, this paper proposed a modified YOLOv3 lightweight model for identifying the growing points of cotton main stem, which realized the multiplexing integration of low-level semantic features and high-level semantics features by adding dense connection modules and modifying the nonlinear transformation within the dense modules. This model significantly reduced the number of model parameters by utilizing deep separable convolution and improved the learning ability of multi-scale features by applying the hierarchical multi-scale method. Our model achieved an accuracy rate of 90.93% based on a self-prepared dataset, which is higher than the accuracy of the original YOLOv3 model by 1.64%, while the number of training parameters was significantly reduced by 48.90%. Compared with other target detection models under different illumination conditions and actual complex environments, the modified YOLOv3 model proposed in this paper showed better robustness, higher accuracy and higher speed in identifying the growing points of cotton main stem.

INDEX TERMS Cotton topping, convolutional neural network, YOLOv3, target detection, multi-scale features, lightweight.

I. INTRODUCTION

Cotton is one of the most important economic crops in China, with an annual output of about 600104t and a sown area of over 300106hm2 [1]. Cotton topping (also known as removing the growing points of cotton main stem) is an important process of cotton cultivation and management. Topping can weaken the top growing advantage, change the direction of nutrient delivery, promote cotton plants to produced more cotton bolls, and increase the cotton yield and quality. Chemical topping and machine topping are currently the mainstream topping methods of cotton, which are both characterized by high speed and high efficiency. However, due to the inability to accurately identify and locate the growing points of cotton main stem in real time, it is difficult to achieve precise chemical topping targeting at cotton tops. Large-scale and widespread chemical spraying not only inhibits the growth of other parts of cotton, but also easily causes wastage and environmental pollution. Similarly, the cutting tool used in machine topping can only calculate the cutting position according to the plant height. Unfortunately in most cases, there is not a definite quantitative relationship between the position of cotton top and the plant height. The calculation of cutting position merely based on the plant height can lead to serious problems of over-topping or under-topping. Therefore, comprehensive research on the identification
and positioning of main stem growing points based on machine vision plays a key role in realizing precise chemical spraying and accurate control of cutting tool for machine topping [2], [3].

As an important branch of artificial intelligence, the image recognition technology based on deep learning has been widely used in various detection fields and has achieved great success due to its non-contact, real-time and low-cost characteristics [4]–[8]. Liu Junqi [9] proposed an automatic recognition method of cotton top images on the basis of neural network. Zhai Ruiyang [10] extracted the mean, variance and standard deviation of the RGB three-channel values of cotton plant tops, leaves and other parts as the features, so as to apply BP neural network to identify cotton tops. Shen Xiaochen [11] realized the identification of cotton plant tops in complex fields by investigating the in-depth architecture, and then proposed a modified VGG deep learning identification framework. Zhu Rongjie et al [12] designed a cotton plant positioning device based on binocular vision ranging to reconstruct the three-dimensional coordinates of cotton points cloud and then realized the identification of the relative position between cottons. Liu Qingfei [13] established a convolutional neural network based on the SSD method for locating the cotton top center, and then calculated the cotton top center spatial coordinates by combining with the depth camera data and the camera internal parameters. This method achieved a testing accuracy of 73.1%.

Although existing studies have applied different technologies to identify and locate the growing points of cotton main stem, they are often subject to various shortcomings such as low accuracy, slow identification speed, large number of model parameters, huge storage cost and high calculation workload [14]–[16]. On the basis of YOLOv3 [17], this paper proposed an improved YOLOv3 lightweight identification model of cotton main stem growing points, which realized the multiplexing integration of low-level semantic features and high-level semantics features by adding dense connection modules and modifying the nonlinear transformation within the dense modules. This model significantly reduced the number of model parameters by utilizing the deep separable convolution and improved the learning ability of multi-scale features by applying the hierarchical multi-scale method. Under different illumination conditions and actual complex environments, our model achieved an accuracy rate of 90.93% with a detection time on each image of 21ms. It provides a valuable technical means and realization ideas for the research of machine topping and chemical topping based on image recognition.

II. MATERIALS AND METHODS
The present study focuses on the growth morphology of cotton during the topping period as the main research object. All the images were collected at the experimental base of the State Key Laboratory of Crop Improvement and Regulation in North China during June to July. Specifically, the cotton plant images were captured at a position between 200mm and 400mm above the top of the cotton plant in the morning (9:00-11:00), early afternoon (13:00-14:00) and late afternoon (16:00-18:00) respectively, in order to ensure that the images could reflect different photographing distances, illumination conditions, and shooting angles.

In total, 12,000 images were collected and 11,001 images were qualified after manual screening and labeling. All the labelled images were saved in PASCAL VOC format.

III. MODEL MODIFICATION
A. THE BASIC YOLOv3 NETWORK
The YOLO series of algorithms are known for their fast identification speed and good effectiveness. This method directly predicts the target object boundary box and combines the two stages of candidate area and object recognition into one. The YOLO algorithm is no longer window sliding, but directly divides the original image into small squares that do not overlap with each other. YOLOv3 inherits the advantages of the YOLO algorithm while making some extra improvements. First, it uses darknet53 as the feature extraction network, which draws on the idea of ResNet [18] by adding residual modules into the network. Second, it also adds a batch normalization (BN) layer and linear units into the network after the convolution of each layer, so as to alleviate the problem of gradient disappearance caused by excessive depth of layers. Thus, while maintaining operation speed, YOLOv3 can effectively improve the accuracy. However, for identifying the growing points of cotton main stem, YOLOv3 is still subject to the disadvantages of large number of model parameters, huge storage cost and high calculation workload.

B. YOLOv3 MODIFICATION METHOD
The original YOLOv3 network was designed for multi-type, multi-object problems, which involves too many layers in the network. However, in our study, we only targeted at a single-type problem, which is the growing points of cotton main stem. Therefore, our modification strategy was to reduce the number of network layers and modify the feature extraction network so as to further improve the object identification accuracy and decrease network parameters.

1) DENSE CONNECTION NETWORK
The dense connection network [19] directly connects all layers together on the premise of retaining the feature information, i.e., the input of each layer is the output of all previous layers. In the dense connection network, different layers are not directly added; instead, the input of each layer is spliced together and then passed to the subsequent layer to maximize the information flow between layer and layer throughout the network.

This connection mode can realize multi-layer feature integration, so as to reduce the number of parameters and alleviate the problem of gradient disappearance to a certain extent by fully making use of the features of each layer. In order to prevent the feature dimension from growing too fast due to...
the increase of network depth, the dense module of different feature dimensions are connected through a conversion layer, which further includes a convolution layer and a pooling layer. The convolution layer converts the input feature dimension into half of the current dimension, and then passes it through the pooling layer. The feature extraction network of YOLOv3 is composed of 5 residual modules. In this paper, the first two residual modules were replaced by two dense connection modules in order to increase the information flow between layer and layer.

The dense connection network consists of dense modules and the conversion layer. The characteristic relationship between layer and layer in the dense module can be expressed as:

\[ x_m = H_m([x_0, x_1, \cdots, x_{m-1}]) \] (1)

where \( x_0 \) is the input feature map, \( x_m \) is the output feature map of the \( m \)-th layer, \( H_m \) is the nonlinear transformation to the \( m \)-th layer. Matrix \([x_0, x_1, \cdots, x_{m-1}]\) is the splicing of the output feature map from the 0-th layer to the \((m-1)\)-th layer.

The nonlinear transformation of \( H_m \) is BN, Relu, 11 convolution and BN, Relu, 33 convolution.

2) LIGHTWEIGHT NETWORK STRUCTURE

The MobileNet network [20] is a highly-efficient model proposed for mobile terminals and embedded devices, which is characterized by light weight. The concept of depth separable convolution proposed in this technique is an important part of the entire network. By drawing on the model structure of the Mobilenet network, we applied the concept of depth separable convolution to the nonlinear transformation in the dense module, so that the original nonlinear transformation was replaced by the depth separable convolution. The network
The diagram of depth separable convolution is illustrated by Figure 3 below.

**FIGURE 3. Depth separable convolution.**

Depth separable convolution aims to split a standard convolution into a deep convolution and a point-by-point convolution (i.e., the ordinary $1 \times 1$ convolution). If the input feature map is $(P, P, M)$, after passing through the standard convolution ($D_k \times D_k \times M \times N$), the output will be $(P, P, N)$. Its calculation amount can be expressed by the following formula:

$$D_k \times D_k \times M \times N \times P \times P$$  

(2)

The conversion of standard convolution into depth separable convolution refers to the process to pass through the depth convolution ($D_k \times D_k \times 1 \times M$) and the point-by-point convolution ($1 \times 1 \times M \times N$) successively. Its calculation amount can be expressed by the following formula:

$$D_k \times D_k \times M \times P \times P + M \times N \times P \times P$$  

(3)

The calculation amount of the second method is far less than that of the first method. The ratio of the calculation amount between the two methods is as follows:

$$\frac{D_k \times D_k \times M \times P \times P + M \times N \times P \times P}{D_k \times D_k \times M \times N \times P \times P} = \frac{1}{N} + \frac{1}{D_k^2}$$  

(4)

Assuming that the size of the convolution kernel is $3 \times 3$ and the output channel is 32, the ratio of the calculation amount between the two methods will be $41/288$ (approximately $1/7$). The larger the number of channels, the greater its proportion in Equation (4) is. When the inverse ratio of the number of channels gradually approaches 0, the calculation amount of the two methods will approximate to the inverse ratio of the square of the convolution kernel, that is, the calculation amount of depth separable convolution can be reduced to approximately $1/9$ of that of the standard convolution.

3) **MUTLI-SCALE CONVOLUTIONAL NETWORK**

As a new convolution module, Res2Net [21] constructs hierarchical residual connections in a single residual block.

In the residual module structure, the input is processed through $1 \times 1$, $3 \times 3$, and $1 \times 1$ convolution operations successively, and then the result is added to the original input to derive the output. In this process, there lacks multi-scale feature extraction. In the actual image acquisition scenario, the heights vary among different cotton plants, but the camera is always fixed at a certain height. Therefore, the cotton top buds in the captured images are of varying sizes, where a multi-scale method is required for learning the features. Res2Net adopts a hierarchical multi-scale method. The features extracted by the same convolution are different in different channels. Thus, the channels are divided into four layers, and features are extracted layer by layer to increase the size of the sensory field. Then, the obtained features are integrated and passed through a $1 \times 1$ convolution again to empower it with stronger feature extraction ability. In addition, this layer-by-layer feature extraction method also plays an important role in reducing the number of parameters.

4) **NETWORK CONSTRUCTION FOR IDENTIFYING THE GROWING POINTS OF COTTON MAIN STEM**

In this paper, we modified the YOLOv3 network for the purpose to use the dense connection network and Res2Net to reconstruct the feature extraction network. The original feature extraction network is darknet-53, which contains 5 residual modules. Each residual module is composed of the convolution and a varying number of residual units. The internal structure of the residual unit is illustrated by the figure below, where DBL contains convolution, BN and other operations.

**FIGURE 5. Residual unit.**

In this experiment, the number of residual modules was reduced to 3, each containing a DBL and 2 residual units. The residual units were further replaced by Res2Net modules, and were placed into the deep part of the feature extraction network. Then, 2 dense connection modules were added into the shallow part of the network, each containing 5 bottleneck layers, and there is a depth separable convolution in each
bottleneck layer. The 2 dense connection modules are connected by a conversion layer, which consists of the operations of BN, Relu, 1 x 1 convolution and average pooling.

Dense connection can better transfer shallow features to the deep network, and realize the multiplexing and integration of features by extracting features from the 2 dense modules. This method takes into account both the shallow and the deep semantic features of the image. Then, by further deepening the network structure through the residual modules, it is possible to reduce the number of model parameters and ensure the accuracy and timeliness of model detection. When the network extraction feature map size reaches 52 x 52 and 26 x 26, parallel two way network convolution operation respectively and concatenated with the feature map of the previous branch. Finally, the whole network outputs three feature maps of different sizes. The diagram of the modified YOLOv3 network structure is shown below, where sep_conv is a depth separable convolution.

IV. MODEL EXPERIMENT AND ANALYSIS
A. EXPERIMENTAL DATASET
The dataset of our experiment consists of 11001 images, which were divided into the training set (8800 images) and testing set (2201 images) respectively according to the ratio of 8:2. The dataset was acquired by using a mobile phone (iPhone7). All the images were standardized to the size of 416 x 416 first, and then were labelled manually using the LabellImg tool. The outcome file is in xml format, which contains the width and height of the image, the target category information, as well as the x_min, x_max, y_min and y_max values of the target rectangle relative to the upper left corner of the image. After sorting and matching the images and the xml file, the dataset was saved in the format of PASCAL VOC for the purpose to train and test of the YOLOv3 network model.

B. MODEL TRAINING
The hardware configuration for model training and testing is as follows: Intel® Core i9-9820X CPU @ 3.30GHz 20, 64G memory, GeForce RTX 2080 Ti graphics card, 64-bit operating system Ubuntu 18.04.1 LTS, CUDA Version 10.0, TensorFlow Version is 1.13.1.

With adjustment of the YOLOv3 model parameters, the accuracy of the final results can be improved by modifying the learning rate. Different learning rates will lead to different accuracy rates. After performing repeated tests on the model, we found that the model exhibited a relatively high accuracy rate when the learning rate was equal to 4e-5. Thus, the initial learning rate was set to be 4e-5, and the learning rate would decrease gradually with the increase of the number of training iterations. Eventually, the final learning rate was set to be 1e-6, the IOU threshold was set to be 0.5, the batch size was set to be 3, the confidence score was set to be 0.5, and the number of iterations was set to be 40. In general, there are two methods for model training. The first one is to train the model with randomly-initialized weights, and the second one is to train the model with pre-trained weights. In this paper, the first method was applied to train the model and to compare between different modification methods.

C. RESULT ANALYSIS
In order to better evaluate our model, the precision P, the recall rate R, the harmonic mean F1 and the average accuracy mAP were chosen as the indicators for comparative analysis. Specifically, the calculation equations for these indicators are as follows:

$$P = \frac{TP}{TP + FP} \quad (5)$$

$$R = \frac{TP}{TP + FN} \quad (6)$$
TABLE 1. The scheme design.

| Scheme | Dense block | Depth separable convolution | Modification of residual structure |
|--------|-------------|-----------------------------|-----------------------------------|
| 1      | √           |                             |                                   |
| 2      | √           | √                           |                                   |
| 3      | √           | √                           |                                   |

TABLE 2. YOLOv3 model comparison.

| Model   | P/%  | R/%  | F1/%  | mAP/% | Total parameters | Weight/MB | Detection time/ms |
|---------|------|------|-------|-------|------------------|-----------|------------------|
| YOLOv3  | 92   | 90   | 90.98 | 89.29 | 61,523,734       | 994.9     | 21               |
| Scheme1 | 90   | 86   | 87.95 | 84.43 | 41,765,318       | 680.9     | 22.5             |
| Scheme2 | 92   | 91   | 91.49 | 89.88 | 41,283,398       | 671.2     | 22.5             |
| Scheme3 | 92   | 92   | 92    | 90.93 | 31,437,254       | 515.9     | 21               |

FIGURE 7. Loss curve of schemes.

\[
F1 = \frac{2PR}{P + R} \tag{7}
\]

\[
mAP = \frac{1}{S} \sum_{k=1}^{N} P(k)R(k) \tag{8}
\]

where TP is the number of times that cotton top is detected correctly, FP is the number of times that the cotton top is detected mistakenly, FN is the number of images without cotton top, S is detection type. In this study, we had only one detection type, so \(S = 1\). \(N\) refers to the number of times that the threshold value is quoted, and \(k\) refers to the threshold value.

The network training saves ten weight files by default. After testing the weight files with relatively a low test loss, the one with the greatest mAP value was chosen to test the images of cotton plants that were not in the dataset, and the results were saved.

In this paper, we compared the three modification schemes based on YOLOv3, and comprehensively analyzed the influence of different modification methods on the overall detection effect in terms of a series of indicators including P, R, F1, mAP, weight file size, total number of model parameters, and the average time of detecting a single image. Specifically, for Scheme 1, the darknet-53 feature extraction network was modified into 2 dense connection modules and 3 residual modules. On the basis of Scheme 1, the non-linear transformation in the dense connection module was changed to a depth separable convolution, which is referred to as Scheme 2. Then on the basis of Scheme 2, the residual structure was further changed to the Res2Net residual structure, which is referred to as Scheme 3. Scheme 3 is the model that was eventually used to detect the growing points of cotton main stem.

The TABLE 1 below visually compare the characteristics of different modification schemes:

It can be seen from TABLE 2 that the original YOLOv3 network model has both a large number of parameters and a large weight size, while its mAP is 89.29%. Relative to the original YOLOv3 network, the number of parameters and weight size of Scheme 1 are significantly reduced, but the accuracy drops by about 5%. The main reason for the decrease of accuracy lies in the reduction of residual units. The number of residual units in the original network is 23, which is reduced to 6 in Scheme 1. Although dense blocks are added, the accuracy still drops. Scheme 2 is added with the depth separable convolution on the basis of Scheme 1, which slightly reduces the number of parameters and weight size and improves the accuracy rate of the original model by 0.59%. Scheme 3 utilizes the hierarchical multi-scale method to modify the residual structure on the basis of Scheme 2, which improves the accuracy rate of the original model by 1.64% and reduces the number of parameters and weight size by nearly half. It is noteworthy that the different network models take roughly the same time for detecting a single image (the difference is only about 2ms).

By comparing the loss value change curve between the original network and the modified networks, it can be found...
that the loss value of all the models can eventually stabilize with the increase of the number of iterations. Generally speaking, the loss value is higher in the first iteration and will be greatly reduced in the subsequent iterations. The difference will become very small at last. Although the loss value of Scheme 1 shows some fluctuations in the training process, it can still get flattened after 20 epochs. The loss values of other network models drop rapidly in the first 5 epochs and then decrease slowly after that, exhibiting an overall process of steady decline.

V. COMPARISON WITH OTHER DETECTION MODELS

A. COMPREHENSIVE COMPARISON

In order to further verify the effectiveness of the network model established in this paper for identifying the growing points of cotton main stem, a total of 11,001 images in the same dataset were selected to compare our model with other existing detection networks in terms of detection accuracy, detection time of a single image, the number of model parameters, and the weight file size.

It can be seen from TABLE 3 that the mAP of all the network models are above 85%, but the differences in detection time and weight file size are apparent. Specifically, Faster RCNN uses VGG16 as the backbone network, so it has a slightly lower accuracy, a longer detection time and a larger number of model parameters. RetinaNet uses ResNet50 as the backbone network. Although its accuracy is good, it takes a longer time to detect a single image. In view of that the actual cotton topping process requires real-time visual detection and lightweight embedded hardware, CenterNet, YOLOv4 and our algorithm can better meet the practical requirements in terms of comprehensive evaluation. Out of these three models, our algorithm has the highest accuracy with the total number of parameters reduced by 50.83% compared to YOLOv4. Therefore, our algorithm has a great advantage in volume.

B. ROBUSTNESS COMPARISON

In order to analyze the robustness of different detection networks, the testing set reflecting different shooting distances, different occlusions and different illuminations were
FIGURE 9. Detection effect when the target site involves occlusion phenomenon.

TABLE 3. Comparison of different algorithms for target detection.

| Model          | mAP/% | Total parameters | Weight/MB | Detection time/ms |
|----------------|-------|------------------|-----------|------------------|
| Faster RCNN (VGG16) | 86.51 | 136,428,864      | 1126.4    | 50               |
| RetinaNet (ResNet50) | 87.26 | 36,382,957       | 436.4     | 45               |
| CenterNet       | 85.6  | 20,168,589       | 226       | 15               |
| YOLOv4          | 86.7  | 63,937,686       | 767.8     | 14.7             |
| Our algorithm   | 90.93 | 31,437,254       | 515.9     | 21               |

used to further compare CenterNet, YOLOv4 and the model proposed in this paper. The detection results for different shooting distances are shown in Figure 8, 9, and 10.

By comparing the effects of the three algorithms in identifying multiple small targets, it can be seen that CenterNet and YOLOv4 could identify all the targets, while our algorithm missed one target. The main reason of missing detection is that the target was too small and its characteristics were not obvious enough. More specifically, the main stem growing points of this target were blocked by other leaves, which led to inaccurate detection. As the positions of the leaves around the main stem growing point are inconsistent, for the same cotton plant, the main stem growing point sometimes cannot be identified by photographing from an oblique top angle, but can be identified by photographing directly above the growing point.
By comparing the detection results of the three algorithms above in the presence of partial occlusion over the target, it can be seen that all the network models have basically learned the feature of the cotton main stem growing points. When the top bud and new leaf were blocked by the surroundings, all the three algorithms could achieve satisfactory identification effect. More specifically, CenterNet exhibited the problem of repeated frames during detection, while YOLOv4 and our algorithm had better results. In general, there are newly unfolded leaves nearby the cotton main stem growing points, and different cotton plants are in different growing states with different sizes of leaves. In addition, some new leaves are fully expanded whereas others are partially expanded, and some cotton plants not only have new leaves on the main stem but also on the side branches. Therefore, all the identification models are prone to the problem of single target with multiple identifications.

CenterNet and our algorithm could roughly identify the target site under different illumination conditions. From Figure 10 above, it can be observed that CenterNet had a slight deviation from the target site, as it identified the leaf nearby the main stem growing point but failed to identify the site of growing point. This may cause positioning error in the machine topping process. Comparatively, YOLOv4 had the problem of missing detection.

VI. DISCUSSIONS

Most of the images in the dataset of our experiment can clearly show the sites of the cotton main stem growing points. There are only a few images with large-area occlusion or with holes on the newly expanded leaves around the top buds. In view of the differences in growing states due to different planting methods, the growing points of cotton main stem may be completely blocked by the surrounding large leaves. Furthermore, in the actual photographing process, the acquired images may be blurred due to camera shaking. Similarly, out of the impact of weather conditions such as wind, the sites of main stem growing points may change constantly, which can also lead to unclear images. Therefore, there are still a lot of difficulties to be overcome in practical applications. Our experimental model needs to be
further optimized in terms of the application universality and practicality.

VII. CONCLUSION
In this paper, we proposed an improved convolutional neural network based on YOLOv3 model for identifying the growing points of cotton main stem and tested this model in the actual environment. By adding dense connection modules, modifying the nonlinear transformation in the dense modules and enhancing the modes multi-scale feature extraction ability, our model achieved an accuracy rate of 90.93% with the detection time of 21ms for a single image. Compared with the original YOLOv3 model, our model showed a higher accuracy, a reduced weight file size and less parameters (reduced by nearly 50%) while ensuring that the detection time did not change too much. Meanwhile, we compared our model with some other deep learning algorithms in identifying the growing points of cotton main stem, and analyzed the detection effects of different algorithms under different conditions based on the same images. The results suggest that our improved YOLOv3 model is featured with the advantages of fast speed, high accuracy and good robustness, and it provides a strong basis for the research of machine vision focusing on the application of cotton precise machine topping and chemical topping.

REFERENCES

[1] China Statistical Yearbook 2019. Beijing: China: China Statistics Press, 2019.
[2] L. Zhaoqi, Z. Weiran, L. Haitao, and B. J. Hanxin, “Application status and development trend of cotton topping technology,” Modern Rural Sci. Technol., no. 7, p. 16, 2020.
[3] L. Xuru, J. Xiaoyue, and N. Jiahui, “The present situation and prospects of cotton industry development in China,” Scientia Agricultura Sinica, vol. 51, no. 1, pp. 26–36, 2018.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.
[6] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2961–2969.
[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 318–327, Feb. 2020.
[9] L. Junqi, “Research on cotton plant top identification system,” Ph.D. dissertation, Shihzei Univ., Shihzei, China, 2009.
[10] Z. Ruiyang, “Cotton plant recognition system based on the machine vision technology,” Ph.D. dissertation, Shihzei Univ., Shihzei, China, 2013.
[11] S. Xiaochen, “The height of the cotton plant identification technology for Cotton top-cutting machine,” Ph.D. dissertation, Shihzei Univ., Shihzei, China, 2018.
[12] R. Zhu, Y. Zhu, L. Wang, W. Lu, H. Luo, and Z. Zhang, “Cotton positioning technique based on binocular vision with implementation of scale-invariant feature transform algorithm,” Trans. CSAE, vol. 32, no. 6, pp. 182–188, 2016.
[13] L. Qingfei, “Application of convolutional neural network in agricultural scene,” Ph.D. dissertation, Xinjiang Univ., Ürümqi, China, 2019.
[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556
[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 21–37.
[16] L. Yuan, W. Boyu, and C. Xu, “Research progresses of target detection technology based on deep learning,” Semicond. Optoelectron., vol. 41, no. 1, pp. 1–10, 2020.
[17] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767
[18] K. He et al., “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[19] G. Huang et al., “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2017, pp. 4700–4708.
[20] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861. [Online]. Available: https://arxiv.org/abs/1704.04861
[21] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, “Res2Net: A new multi-scale backbone architecture,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Aug. 30, 2019, doi: 10.1109/TPAMI.2019.2938758.

CHUNSHAN WANG received the B.S. and M.S. degrees in electrical engineering from Hebei Agricultural University, China, in 2001 and 2004, respectively. He is currently pursuing the Ph.D. degree in agricultural engineering with the National Engineering Research Center for Information Technology in Agriculture, Beijing, China. In 2004, he joined the faculty of the School of Information Science and Technology, Hebei Agricultural University, where he has been an Associate Professor, since 2012. From 2016 to 2017, he was a Visiting Professor with the Department of Electrical Engineering and Computer Science, Washington State University, Pullman, USA. His research interests include smart agriculture, deep learning, and image processing.

SIQI HE was born in Hebei, China, in 1996. She received the bachelor’s degree in computer science and technology from Hebei Agricultural University, China, in 2018, where she is currently pursuing the master’s degree in computer technology. Her current research interests include computer vision, deep learning, and agricultural image recognition and detection.

HUARUI WU received the Ph.D. degree from China Agriculture University, Beijing, China. He was a Master Tutor. He is currently a Research Fellow, the Deputy Director of the Beijing Agricultural Information Technology Research Center and the Chief Scientist of the National Agricultural Science and Education Cloud Platform of Ministry of Agriculture and Rural Affairs of the People’s Republic of China. He has authored more than 280 articles in related journals. He has participated in the projects 863 Program. His current research interests include agricultural intelligent systems and wireless sensor network technology.
GUIFA TENG received the B.S. and M.S. degrees in agricultural mechanization engineering from the Hebei Agricultural University, China, in 1983 and 1988, respectively, and the Ph.D. degree in information science from Peking University, Beijing, China, in 2005. From 1996 to 2001, he was a Visiting Scholar with the Institute of Engineering, Hiroshima University, Japan. He is currently a Doctoral Supervisor Tutor with Hebei Agricultural University and the Dean of the School of Information Science and Technology, Hebei Agricultural University. He has authored more than 100 articles in related journals. His current research interests include agricultural intelligent equipment, agricultural big data technology, artificial intelligence, and big data.

CHUNJIANG ZHAO received the Ph.D. degree from China Agriculture University, Beijing, China. He is currently a Doctoral Supervisor Tutor Academician with the Chinese Academy of Engineering, the Chief Scientist of the Beijing Agricultural Information Technology Research Center, the Director of the National Agricultural Information Engineering Technology Research Center, expert enjoying the special government allowance, and China’s expert with great contributions. He has more than 30 years of industry and university experience, as a Fellow, a Chief Scientist, and a Distinguished Member of Technical Staff. He has authored more than 400 articles in related journals. His current research interests include image analysis, intelligent agriculture, deep learning, and agricultural information technology.

JIUXI LI received the B.S. degree in industrial design from Yanshan University, China, in 2001, and the M.S. degree in industrial design from the Beijing Institute of Technology, China, in 2007. In 2001, she joined the faculty of the School of Mechanical and Electrical Engineering, Hebei Agricultural University, China, where she has been an Associate Professor, since 2013. From 2016 to 2017, she was a Visiting Professor with the Department of Mechanical Engineering, Washington State University, Pullman, USA. Her research interests include industrial design, smart agriculture, and image processing.