

RESEARCH ARTICLE

A Deep Learning-Based Object Detection Scheme by Improving YOLOv5 for Sprouted Potatoes Datasets

GUOWEI DAI¹, LIN HU^{1,2}, JINGCHAO FAN^{1,2}, SHEN YAN¹, AND RUIJING LI¹¹National Agriculture Science Data Center, Agricultural Information Institute of CAAS, Beijing 100081, China²National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572024, China

Corresponding author: Jingchao Fan (fanjingchao@caas.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF0704200, in part the Agricultural Science and Technology Innovation Program under Grant CAAS-ASTIP-2016-AII, and in part by the Major Project of Science and Technology of Inner Mongolia Autonomous Region under Grant 2021ZD0043.

ABSTRACT Detecting and eliminating sprouted potatoes is a basic measure before potato storage, which can effectively improve the quality of potatoes before storage and reduce economic losses due to potato spoilage and decay. In this paper, we propose an improved YOLOv5-based sprouted potato detection model for detecting and grading sprouted potatoes in complex scenarios. By replacing Conv with CrossConv in the C3 module, the feature similarity loss problem of the fusion process is improved, and the feature representation is enhanced. SPP is improved using fast spatial pyramid pooling to reduce feature fusion parameters and speed up feature fusion. The 9-Mosaic data augmentation algorithm improves the model generalization ability; the anchor points are reconstructed using the genetic algorithm k -means to enhance small target features, and then multi-scale training and hyperparameter evolution mechanisms are used to improve the accuracy. The experimental results show that the improved model has 90.14% recognition accuracy and 88.1% mAP, and the mAP is 4.6%, 7.5%, and 12.4% higher compared with SSD, YOLOv5, and YOLOv4, respectively. In summary, the improved YOLOv5 model, with good detection accuracy and effectiveness, can meet the requirements of rapid grading in automatic potato sorting lines.

INDEX TERMS Object detection, convolutional neural network, sprouting potato recognition, mosaic, hyperparametric optimization, spatial pyramid pooling.

I. INTRODUCTION

Potatoes are one of the most consumed food crop worldwide, and is cultivated in more than 100 countries. As the fourth largest crop after maize, wheat, and rice, potato production is of significant concern to the food industry, supporting many research projects and, in particular, potato storage which is the basis of the potato industry [1]. Potatoes are taken as a kind of strategic staple food in China, which indicates a need to improve potato processing. The bottleneck in this strategy is the selection of suitable raw materials, technologies, and equipment for processing potatoes. Potato sprouting during storage is lethal for the whole industry, with sprouted potatoes

containing as little as 0.2 mg/g of solanine leading to asphyxiation and even death [2], [3], with resulting net industrial losses and increased food waste.

Since the beginning of the twentieth century, many researches has been conducted on identifying and grading of sprouted potatoes, some of which are based on the external quality detection of traditional computer vision systems, mainly: Jing *et al.* [4] proposed the Otsu method to remove potato image background. He used the perception learning algorithm (PLA) to classify two kinds of potatoes, and finally used the K-nearest neighbor classification algorithm (KNN) to identify surface sprouted potatoes. Weidong and Zhong [5] proposed a method to extract the feature vectors of multidimensional representations of shapes within a single potato region using principal component analysis (PCA) to reduce

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar¹.

the dimensionality and a ten-fold cross-validation technique to bring a single potato image into a parameter-optimized SVM model for automatic potato classification. There are also studies on how hyperspectral and multispectral imaging techniques can be used to acquire full spatial multidimensional features. Hai-Long *et al.* [6] presented a comparative study on high penetration emission and reflection imaging techniques using reflection and transmission spectroscopy to select randomly placed potato damage identification through subwindow alignment analysis (SPA). Ye *et al.* [7] proposed a non-destructive detection method based on the moisture method to obtain the average spectrum of each potato by masking, using the correlation coefficient algorithm SA optimization, for SNV processing data for dimensionality reduction, and applied grid search algorithm to optimize modeling parameters to achieve the classification of potato damage degree.

With the continuous development of deep learning technology in recent years, the target detection technology has wide application in various industries due to its good in-depth feature perception capability [8], [10], [11], [12]. Among the applications in agriculture for example, apple-related recognition [13], [14], [15], tomato recognition [16], [17], [18] and crop disease recognition [20], [21], etc. In the research of potato recognition, Marino *et al.* [22] mainly focused on a weakly supervised deep learning method to classify and localize segment defects on potatoes, by improving convolutional neural network(CNN) using defect activation map (DAM), motivating CNN to predict key regions of a certain class, and using a coarse-to-fine segmentation method to obtain more accurate defect sizes thus enables a method for classifying defects in multiple potato images. Rui *et al.* [23] Improved the Faster R- CNN method for potato bud-eye recognition by also acquiring image data under a fixed image acquisition system and optimizing the NMS algorithm in R-CNN using a Gaussian weight reduction function, thus in order to improve the recognition of a potato bud-eye. The recognition methods mentioned above are proposed in a specific scenario that is predicted, and the data model constructed needs to satisfy the image acquisition system set up. These techniques proposed by previous studies mainly satisfy the quality grading of potatoes in an actual state or a positive scene using different detection means. In contrast, the detection and grading of potatoes in various complex scenarios using computer vision techniques are less involved.

This paper presents a deep learning-based image detection algorithm for sprouted potatoes. The algorithm aims to achieve efficient and accurate intelligent grading and screening of potatoes in the storage process and daily quality inspection, since the use of human power to detect sprouted potatoes is a common practice throughout the potato grading process while relying on human power to be able to detect sprouted potatoes in complex scenarios. Therefore, the identification of sprouted potatoes in various scenarios must be satisfied. Suppose real-time, online, and high-throughput detection of sprouted potatoes can be achieved. In that case, it will save

labour costs and avoid errors caused by distracted staff, erroneous detection by human eyes, and missed detection, which will make a significant contribution to the whole production line. In summary, this paper proposes a germinating potato image recognition method based on an improved YOLOv5 model based on the structure of the YOLOv5 model. Moreover, the following two contributions are made.

-Firstly, we built the sprouted potato dataset, covering three types of potatoes: healthy potatoes, sprouted potatoes and rotten potatoes. We also propose improved data augmentation methods for image augmentation and optimization to strengthen small target features and improve model generalization.

-Secondly, we optimize the YOLOv5 algorithm from several perspectives. A new feature extraction network is proposed, which improves the original convolution module using CrossConv based on intra-graph convolution to enhance feature similarity. SPPF (Fast Spatial pyramid pooling) replaces SPP in the original network structure, which reduces the number of fusion parameters and accelerates the fusion speed. The k -means of the genetic algorithm is used to reconstruct the anchor size and improve the detection accuracy, followed by fine-tuning the accuracy of the model using the hyperparameter evolution mechanism and training with a multi-scale strategy to improve the generalization ability of the model.

II. RELATED WORK

A. YOLOv5 FRAMEWORK

Ultralytics launched YOLOv5 [24] in June 2020. So far, seven iterations have been introduced since the introduction of YOLOv5-v1.0, which incorporates better experimental networks as a parameter structure into the backbone. Cross Stage Partial Network (CSPNet) is used to improve backbone, mixed precision (FP16) is used to accelerate small target inference, and PANet [25] is used to reduce parameters. Its detection speed and accuracy on COCO datasets are better than previous YOLOv4 and YOLOv3 algorithms. Four models with different weights, v5s, v5m, v5l and v5x, are developed, and the depth and width of the models are controlled by the `depth_multiple` and `width_multiple` parameters, corresponding to the four different model levels of YOLOv5. The mAP for each model is enhanced sequentially, and the speed is decreased sequentially while the parameters are increased. The YOLOv5 network structure consists of an input layer of 640×640 image, Backbone Network, Neck Network and Detector (shown in Fig. 1).

The Backbone Network is the core structure of YOLOv5, consisting of Focus, Conv, C3 (CSPNet Bottleneck with three convolutions), and Spatial Pyramid Pooling (SPP) modules. The Focus module slices the Mosaic image input vertically and horizontally and then stitches it together. Compared with convolutional downsampling, the Focus output depth is increased four times, and more image information is retained. Conv is the basic convolution unit of YOLOv5, which sequentially performs two-dimensional convolution,

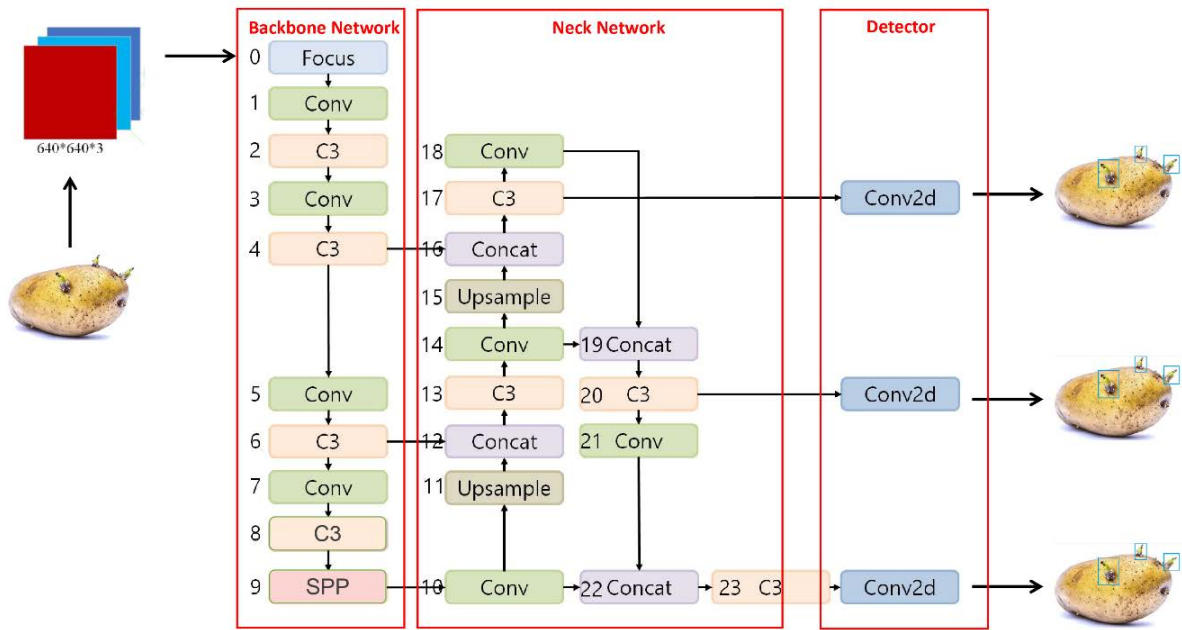


FIGURE 1. YOLOv5 model structure.

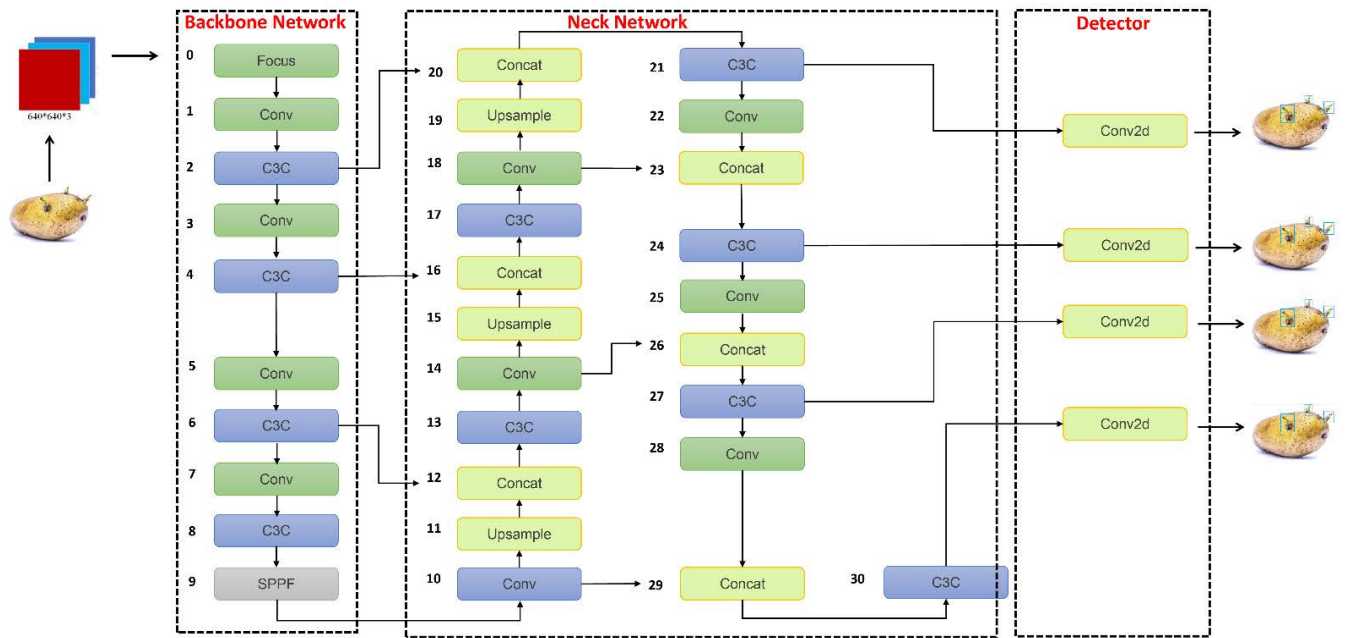


FIGURE 2. Improved YOLOv5 model structure.

regularization, and activation operations on the input. C3 is composed of 3 Convs and several bottlenecks. The Concat operation is carried out with the original value after the input is subjected to a 1×1 three-layer convolution with the convolution kernel. Bottleneck completes the residual feature transfer without reducing the features. SPP is a spatial pyramid pooling layer. SPP performs three different sizes of maximum pooling operations on the input, and the output

result is spliced into Concat, and the output depth is the same as the input depth. The 11 and 15 layers of the Neck Network use the upsample module to expand the features. The number 4 and 15 layers after upsampling are further fused using Concat to form a more extensive feature map and output to the Detector for prediction. YOLOv5 has three feature detection scales suitable for feature detection of different sizes.

B. IMPROVED NETWORK MODEL DETECTION SCALE

The advanced feature mapping of the YOLOv5 network model has a broader acceptance domain and focuses on representing abstract semantic information. However, because it focuses on extracting more comprehensive features, it is insensitive to small target features, resulting in poor performance on tiny target datasets. YOLOv5 includes three feature detection scales. The largest feature map is 80×80 . When the input image size is 640×640 , means that feature maps smaller than 8×8 are not detectable. In this paper, a detection scale is added. The maximum feature map is 160×160 , which satisfies the detection target of more than 4×4 pixels and can meet the detection requirements of more acceptable targets. The structure of the improved YOLOv5 model is shown in Fig. 2, and the parameters are shown in Table 1.

TABLE 1. Improved YOLOv5 network structure.

Layers	From	Params	Module	Arguments
0	-1	7040	Focus	[3,64,3]
1	-1	74240	Conv	[64,128,3,2]
2	-1	156928	C3C	[128,128,3]
3	-1	295936	Conv	[128,256,3,2]
4	-1	1611264	C3C	[256,256,9]
5	-1	1181696	Conv	[256,512,3,2]
6	-1	6433792	C3C	[512,512,9]
7	-1	4722688	Conv	[512,1024,3,2]
8	-1	9971712	SPPF	[1024,1024,[5]]
9	-1	2624512	C3C	[1024,1024,3,False]
10	-1	525312	Conv	[1024,512,1,1]
11	-1	0	Upsample	[None,2,'nearest']
12	[-1,6]	0	Concat	[1]
13	-1	2757632	C3C	[1024,512,3,False]
14	-1	263168	Conv	[512,512,1,1]
15	-1	0	Upsample	[None,2,'nearest']
16	[-1,4]	0	Concat	[1]
17	-1	2626560	C3C	[768,512,3,False]
18	-1	131584	Conv	[512,256,1,1]
19	-1	0	Upsample	[None,2,'nearest']
20	[-1,2]	0	Concat	[1]
21	-1	657920	C3C	[384,256,3,False]
22	-1	590336	Conv	[256,256,3,2]
23	[-1,18]	0	Concat	[1]
24	-1	690688	C3C	[512,256,3,False]
25	-1	590336	Conv	[256,256,3,2]
26	[-1,14]	0	Concat	[1]
27	-1	2626560	C3C	[768,512,3,False]
28	-1	2360320	Conv	[512,512,3,2]
29	[-1,10]	0	Concat	[1]
30	-1	9971712	C3C	[1024,1024,3,False]

The parameter ‘‘From’’ indicates that the input from the previous layer of the network is accepted; for the Concat layer, ‘‘From’’ indicates that the existing layer is integrated

with the features of the specified layer. ‘‘Params’’ are the network layer parameters of the current layer; modules indicate the name of the model used in the current layer. In this paper, a detection head is added to the original detection scale, an up-sampling layer is added, the layer with serial number 20 is merged with the second layer, and the output feature map is regarded as a 4-fold sampling detection. The values in the ‘‘Arguments’’ column indicate the number of input channels, the number of output channels, the size of the convolution kernel, and information about the step size of the module, respectively.

C. CROSSCONV IMPROVED C3

CrossConv [26] transfers features between two feature map structures to be matched. A similarity matrix is constructed by calculating the similarity of any two vectors between the two graphs. The Sinkhorn algorithm is used to solve the similarity matrix to obtain a matching relationship, and the predicted matching relationship is used as the span between the two graph structures. The weight of the graph is updated. Therefore, the cross-graph convolutional layer simultaneously considers the information of the two graphs to be matched during the matching process. Through CrossConv update the initially relatively similar features between the two graphs will be more similar. Therefore, the C3C module in Table 1 represents the C3 module before the improvement, while Fig. 3 represents the use of CrossConv to improve the Conv in the YOLOv5 model C3.

$$\{h_{1i}^{(k)}\} \leftarrow \text{CrossConv} \left(\hat{S}, \{h_{1i}^{(k-1)}\}_{i \in v_1}, \{h_{2j}^{(k-1)}\}_{j \in v_2} \right) \quad (1)$$

$$\{h_{2j}^{(k)}\} \leftarrow \text{CrossConv} \left(\hat{S}^T, \{h_{2j}^{(k-1)}\}_{j \in v_2}, \{h_{1i}^{(k-1)}\}_{i \in v_1} \right) \quad (2)$$

$$k\text{-th layer features } \{h_{1i}^{(k)}, h_{2j}^{(k)}\}_{i \in v_1, j \in v_2} \quad (3)$$

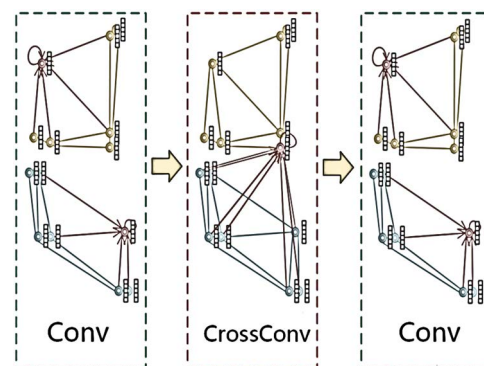


FIGURE 3. CrossConv improved C3 structure of YOLOv5.

In formula (1), (2), (3), \hat{S} replaces the adjacency matrix, \hat{S}^T is the transposed matrix of \hat{S} , and the transposed matrix \hat{S}^T represents the transformation of primitive information; the image augmentation operation is regarded as a

fine-grained matrix transposition; Select v_1 of the $k-1$ layer from \hat{S} or \hat{S}^T , v_2 is the feature vector $\{h_{2j}^{(k)}\}$ obtained by CrossConv, $\{h_{1i}^{(k)}\}$ is the updated weight.

D. IMPROVE MOSAIC DATA AUGMENTATION

The proposal of Cutmix [27] changed the method of data augmentation that relied on experience in the past. The included Mixup, Cutout, and Cutmix combinations enhance the data, and many verifications show the correctness of this method. The 9-Mosaic data augmentation method proposed in this paper is improved on Cutmix, and a new data processing method is proposed: load an original picture, randomly select eight pictures, combine them, and use the hyperparameters translate, scale, and scale in Table 3 shear to deal with it. The effect diagram is shown in Fig. 4. The advantage of this method is to enrich the background of the detected object; especially the random scaling increases the tiny target so that the robustness of the network is improved to a certain extent. On the other hand, with 9-Mosaic data augmentation, the model input will calculate nine images, which implicitly increases the batch size and allow the model to converge quickly and reduces the requirement for GPU performance.

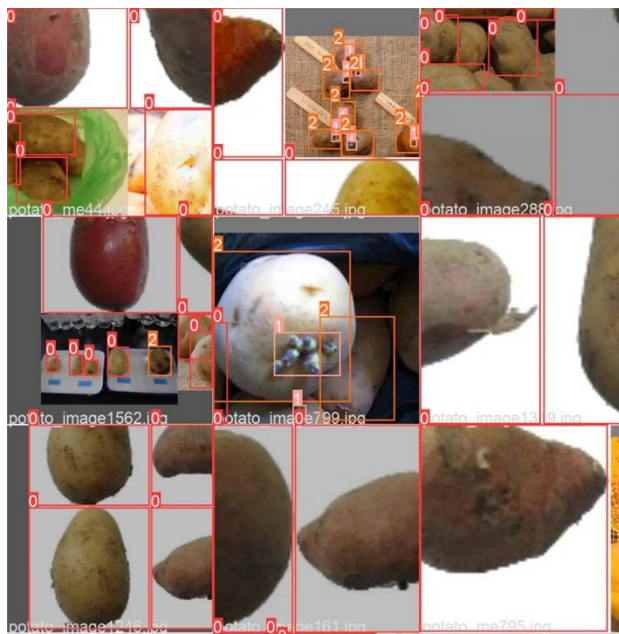


FIGURE 4. 9-Mosaic data augmentation.

E. K-MEANS CLUSTERING ALGORITHM TO REGAIN THE INITIAL ANCHOR

The Detector structure prepares initial anchors of different widths and heights for the three Detect modules. The initial anchors contain prior knowledge of the target data, and their selection will positively affect the learning of the network and the target detection effect. This paper uses k -means of genetic algorithm to match the best anchor points, as shown in Fig. 5. The iterations are 1500 times with $img_size=640$

or 940 and $thr=0.4$, where img_size is the size of the input image. Increasing img_size helps to improve the recognition accuracy but increases the computational burden. The thr determines the aspect ratio of the target box. Under the same thr parameter, the number of iterations, img_size and the number of anchor points K together determine the optimal anchor point. Improving these parameters can help improve the accuracy of recognition. Fig. 6 shows the relationship between the change in gain due to different parameter combinations as the number of iterations increases. When $K=9$ transmitted iterations 500 times, the fitness of $img_size=960$ is greater than $img_size=640$. After 800 iterations of genetic iterations, the fitness remains steady; by comparison, when $K=12$, the fitness of the first three cycles fitness improves quickly and remains steady after the 800th genetic cycle. Compared with $K=9$, there is a certain improvement effect. In this paper, $img_size=640$, four layers of anchor points, namely $K=12$, 1500 genetic iterations to reacquire the initial anchor as shown in Table 2.

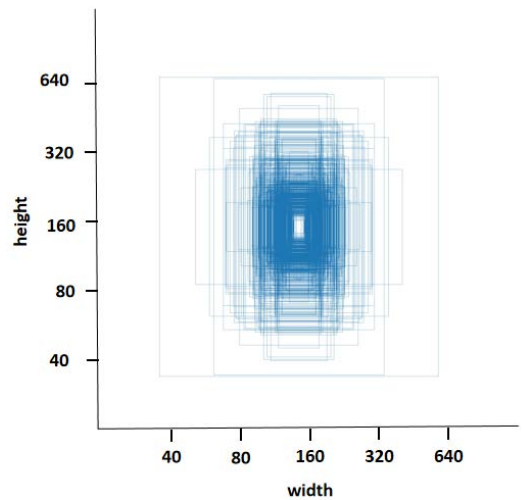


FIGURE 5. Genetic algorithm to calculate the anchor process.

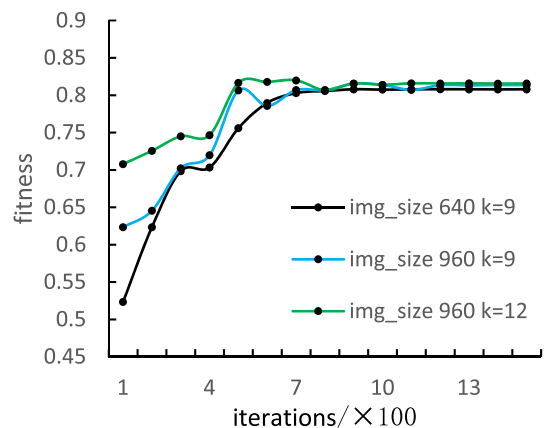


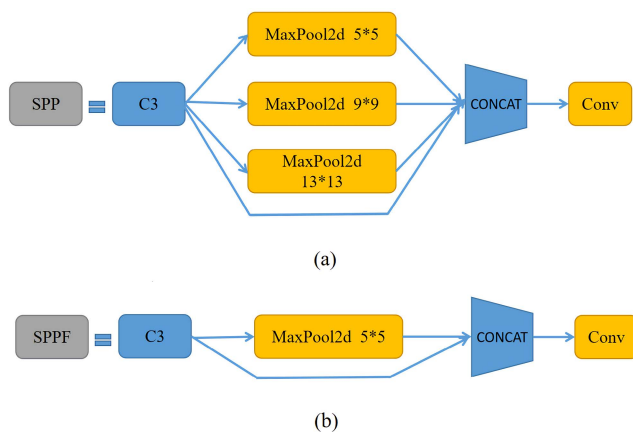
FIGURE 6. Calculation results of genetic algorithm using different combinations of parameters.

TABLE 2. Genetic iteration to obtain anchor results.

Anchor	Object type	Sampling multiples
[30,35,36,59,53,41]	tiny	4 times
[58,62,54,107,85,69]	small	8 times
[83,101,123,73,72,185]	medium	16 times
[129,113,129,198,218,183]	large	32 times

F. SPPF IMPROVES SPP

In the improved YOLOv5 network structure, Table 1 serial number 8 can generally be understood as the eighth layer of the model network; the eighth layer is the Backbone layer, SPP [28] is generally at the end of the Backbone. It is generally believed that the ability to extract features is enhanced with the deepening of the network. SPP in the original network structure shown in Fig. 7(a) uses three scales 5, 9, 13 for multi-scale fusion; if the shape of the fused potato feature map is more uniform, small target location information with overlap will become inaccurate during the fusion process, or even lost in severe cases. Due to many potato shoots in the dataset that overlap and cross with rotten potatoes, there will be missed detections and reduced accuracy. Here SPPF is used instead of SPP in the original network structure. the improved SPPF shown in Fig. 7(b) uses a single parameter5 for pooling, which reduces the number of fusion parameters, accelerates fusion pooling, and improves the accuracy of potato detection after fusion.

**FIGURE 7. SPPF improvement SSP comparative analysis.**

G. HYPERPARAMETER EVOLUTION MODEL MECHANISM

Hyperparametric optimization [29] occupies an essential place in machine learning research. We constantly want to tune hyperparametric optimization models as part of the scientific process. Hyperparameters are different from model parameters and need to be set before the model is trained. Hyperparameters are usually not intuitive, but they have a significant impact on the model's performance. Unlike model parameters, hyperparameters are usually unpredictable. Therefore, optimization of hyperparameters usually requires an empirical search to evaluate

the model's performance on training and validation samples. The efficiency of hyperparameter tuning has been further improved after automatic hyperparameter optimization has been proposed, but not all datasets are suitable for automatic hyperparameter tuning. This paper uses automatic stochastic search to explore the more necessary hyperparameters. The stochastic search sets a random value for the hyperparameters, evaluates the model after each training, and then selects the optimal value for the parameter settings; since the dataset of this paper has only 2316 images, all data are included in the hyperparameter evolution process. The hyperparameters were evolved 300 times, and finally, the hyperparameters were finalized during the evolution of 256 epochs (as shown in Table 3), and training the model with these hyperparameters usually results in better performance.

TABLE 3. Hyperparameters for optimizing YOLOv5.

No.	Params	Value
0	momentum	0.980
1	weight_decay	0.0005
2	warmup_epochs	2.640
3	warmup_momentum	0.854
4	warmup_bias_lr	0.0997
5	box	0.0462
6	cls	0.415
7	cls_pw	1.110
8	obj	1.130
9	obj_pw	1.300
10	iou_t	0.200
11	anchor_t	4.310
12	hsv_h	0.0152
13	hsv_v	0.185
14	hsv_s	0.725
15	translate	0.0862
16	scale	0.516
17	mosaic	0.987
18	shear	0.000

Note: The table is only important hyperparameters, not all parameters.

III. MODEL VALIDATION

The potato dataset consists of a web crawler and data extracted from Fruits360 [30]. (The dataset is available at <https://www.agridata.cn/data.html#/datadetail?id=289632>). The data were filtered according to the rule of excluding potatoes with differentiated potato buds, sprouted potatoes with a single scene and non-potato images, and finally obtained a total of 513 images that met the rule, plus 1803 potato images from Fruits-360, totaling 2316 images; used to create training and test datasets. The dataset consisted of healthy potatoes with sprouted potatoes and rotten potatoes, As shown in Fig. 9. The data acquired through the network is not well standardized, and there exists a large number of images of various types of different sizes, in order

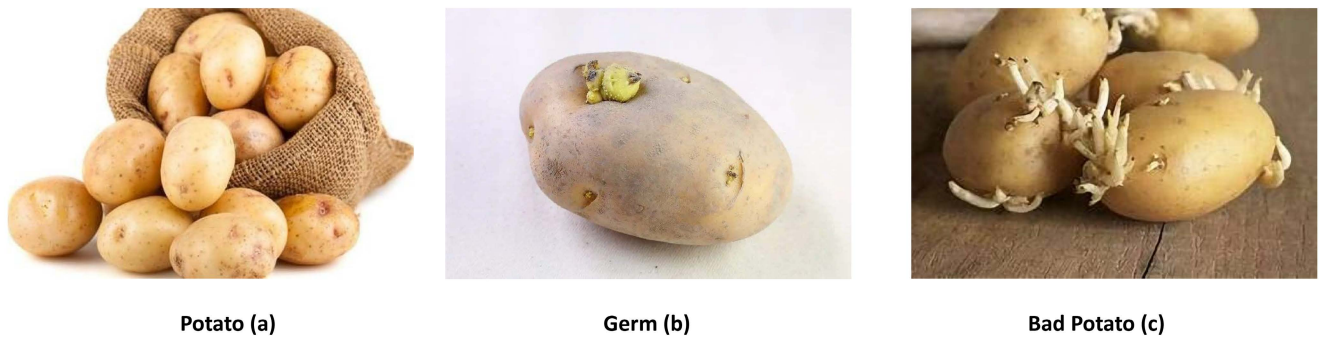


FIGURE 8. Data set samples healthy potatoes (a), sprouted potatoes (b), rotten potatoes (c).

to unify the data; the potato image data was processed using waifu2x-caffe [31] with parameters set in RGB image mode with noise reduction Level1 and conversion mode two and saved in JPG format. The potato samples were divided into the training set and test set in the ratio of 8:2. The number of images in the training set was 1843, and the number of images in the test set was 461, of which the number of label categories in the training set is shown in the table below.

TABLE 4. Number of label categories in the training data set.

Label type	Samples
Potato	3620
Germ	1225
Bad Potato	536

After pre-processing, the potato dataset was analyzed using the Table 2 anchor points, and the data visualization results were obtained by taking a normalized approach. In Fig. 8(a), x and y refer to the location of the center point, and the blue square represents the location of the point the more concentrated the center point of the target box. The width and height in Fig. 8 (b) represent the width and height of the objects in the dataset labels, respectively. From Fig. 8 (a)(b), it can be seen that the dataset objects are more evenly distributed, and the proportion of small and medium-sized objects is more prominent, which also indicates that the dataset is suitable for fusion features with small sampling values. In addition, the presence of darker colored square blocks for the coordinate values in Fig. 8 (a) indicates that there is much overlap in the potato germination in the dataset. In this paper, the problem of identifying overlapping germinating potatoes can be improved by using CrossConv to modify the C3 model.

A. TRAIN THE NETWORK MODEL

The experimental model in this paper was trained using ultralytics YOLOv5. The model provides powerful model configuration capabilities for each module on a fine-grained basis, without the worry of tedious process control. Since the potato dataset is not correlated with ImageNet [32], we do not use the preset weights here. We performed training with 300 epochs using the hyperparameters in Table 3, with batch

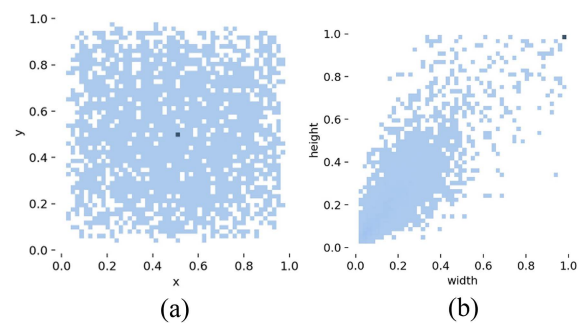


FIGURE 9. Dataset analysis graph.

size adjusted to 64 and image size adjusted to 640×640 , using approximately 5 hours. Table 5 shows our experimental environment configuration, and Fig. 10 shows the different performances of the improved model in the training and validation sets.

TABLE 5. Experimental environment configuration.

Parameter	Configuration
CPU	Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz
GPU	NVIDIA GeForce RTX 3090
System Environment	Ubuntu 20.04
Language	Python 3.9
Acceleration Environment	CUDA 11.1

Three different types of losses are shown in Fig. 10: box loss (box_loss), target loss (obj_loss), and classification loss (cls_loss). Box loss indicates the extent to which the algorithm can locate the center of an object and the extent to which the predicted bounding box covers the object. Target loss is essentially a measure of the probability that an object exists in the proposed region of interest. If the objectivity is high, the image window may contain objects. Classification loss gives the ability of the algorithm to predict the correct class of a given object. The model improves rapidly in precision, recall, and average accuracy, leveling off after about 200 epochs. The validation data's box, target, and classification loss also

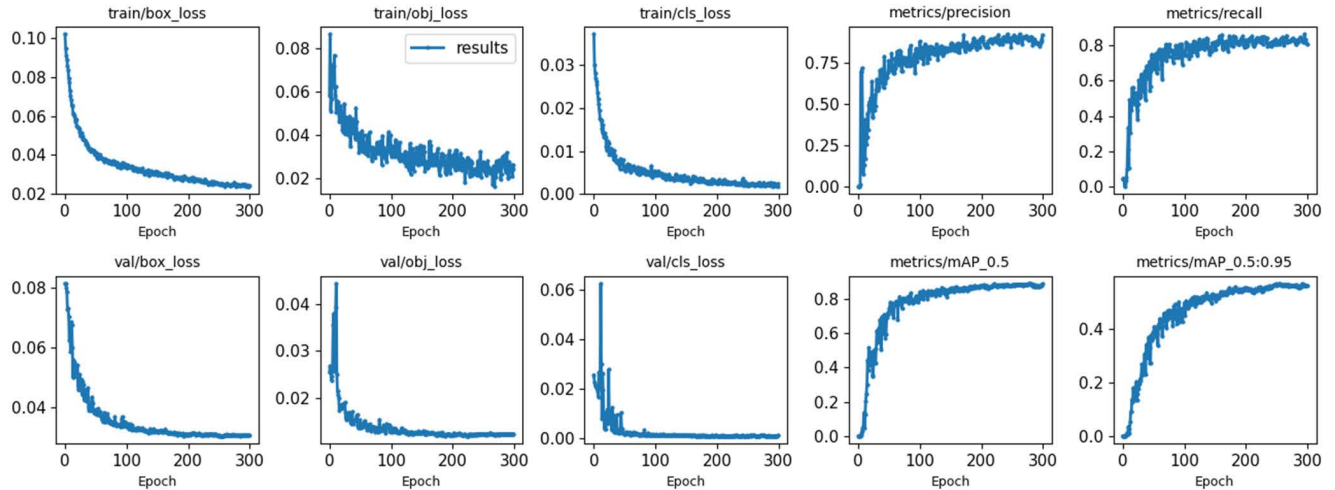


FIGURE 10. Plots of box loss, objectness loss, classification loss, precision, recall and mean average precision (mAP) over the training epochs for the training and validation set.

show a rapid decrease. We use an early stopping strategy to select the best weights and stop training at 295 epochs.

B. MODEL EVALUATION METRICS

In order to achieve the evaluation of potato detection models, the models were evaluated using the most commonly used metrics in the field of target detection, Precision, Recall, F1, and Mean Average Precision (mAP). This paper uses test data consisting of 461 potato images for performance testing. Except for the three categories of labels in the potato dataset in Table 4, other categories of images will not be detected.

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{\text{all ground truths}} \quad (4)$$

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

$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6)$$

$$\text{AP} = \frac{1}{n} \sum_{i=1}^n \text{Precision}_i \quad (7)$$

$$\text{mAP} = \frac{\sum_{i=1}^Q \text{AP}_i}{Q} \quad (8)$$

In equation (4), TP includes the proportion of positive examples of rotten potatoes that can be found correctly; the proportion of positive examples of healthy potatoes that can be found correctly; and the proportion of buds of positive examples of potatoes that can be found correctly. FP includes misclassification of negative cases of rotten potatoes, the misclassification of negative cases of healthy potatoes, misclassification of the proportion of negative cases of potatoes with shoots. FN includes the proportion of misclassified positive examples of rotten potatoes, the proportion of misclassified positive examples of healthy potatoes, and the proportion of misclassified positive examples of potatoes with shoots. The three categories of potato classification described above are represented by the label categories in Table 4. In Equation (8),

mAP was calculated as the average AP value for multiple individuals in the germinating potato test set, mAP is defined as the mean value of the AP under all categories at an Intersection over Union (IOU) threshold of 0.5, as shown in (4-8).

C. ANALYSIS OF TEST RESULTS

According to the experimental results in Table 6 and Table 7, it is clear that the improved YOLOv5 model in this paper can effectively improve the detection accuracy of healthy potatoes, potato sprouts, and rotten potatoes. The test dataset consisting of 461 potato images was tested under the improved model and the original YOLOv5s model separately. The improved model had an average detection accuracy of 90.1% for all classes (all), 81.2% for sprouted potato buds (germ), 81.2% for healthy potatoes (potato), and 97.1% for rotten potatoes (badpotato). The mAP of the three classifications reached 88.1%, which was 7.4% higher than the mAP of the original YOLOv5s model. Under the potato category, the mAP of the original model is 2.6% higher than that of the improved model; compared to the original model, there is a bizarre situation of a decrease in mAP. By comparing the F1 of the model before and after the improvement are 91%, it is found that the improved model has negligible effect on the negative gain of accuracy under this category, and the problem of the negative optimization model of the original model can be ignored. Under the germ and badpotato, the improved model is 15.1% and 9.9% higher than the original model mAP, respectively, indicating that the improved model has a more significant effect on the accuracy rate improvement of the latter two. This shows that the model is excellent in potato detection and grading accuracy and can meet the requirements of potato grading detection with high accuracy in multiple scenarios.

In order to compare the specific performance of the improved methods proposed in this paper with YOLOv5,

TABLE 6. Ablation test of YOLOv5s with different improvement methods.

No.	Models	P/%	R/%	F1/%	mAP/%
1	YOLOv5s	78.7	78.5	78.0	80.6
2	YOLOv5s+9-Mosaic	82.1	79.8	80.9	82.6
3	YOLOv5s+SPPF	81.3	80.7	80.9	80.9
4	YOLOv5s+k-means	82.3	78.4	80.3	83.1
5	YOLOv5s+Hyperparameter evolution	80.7	79.5	80.0	81.1
6	YOLOv5s+CrossConv	86.6	82.3	84.3	85.3
7	YOLOv5s+9-Mosaic+SPPF	83.9	79.8	81.7	84.3
8	YOLOv5s+9-Mosaic+SPPF+k-means	84.7	80.2	82.3	86.1
9	YOLOv5s+9-Mosaic+SPPF+k-means+Hyperparameter evolution	85.1	81.2	83.1	86.9
10	YOLOv5s+9-Mosaic+SPPF+k-means+Hyperparameter evolution+ CrossConv	90.1	80.4	84.0	88.1

TABLE 7. Performance of the YOLOv5s model.

Class	Images	Labels	P/%	R/%	F1/%	mAP/%
all	461	1043	78.7	78.5	78	80.6
potato	461	716	86.3	96.5	91	97.8
germ	461	231	72.5	56.1	63	62.1
badpotato	461	96	77.4	82.8	77	81.8

TABLE 8. Improving the performance of the YOLOv5s model.

Class	Images	Labels	P/%	R/%	F1/%	mAP/%
all	461	1043	90.1	80.4	84	88.1
potato	461	716	92	89.6	91	95.2
germ	461	231	81.2	76.3	78	77.2
badpotato	461	96	97.1	75.3	84	91.7

we tested each of the five improved methods with YOLOv5, followed by a combination of methods in a cumulative manner, and the full test results are shown in Table 8. The proposed five improvement method (No.10) improves 7.5% compared to the YOLOv5s model mAP. CrossConv improvement C3 (No.6) has a large improvement in model accuracy with 7.9% improvement in accuracy and 4.7% improvement in mAP. In addition, the hyperparameter improvement for YOLOv5s (No.5) has the least improvement in performance among the five improvement methods. Overall, the proposed improvement methods all improve the YOLOv5s model performance.

The performance index comparison results of the model proposed in this paper with SSD, Faster R-CNN, YOLOv3, YOLOv4, YOLOv5, and YOLOX target detection models are shown in Table 9. The bolded fonts are the optimal values of the models, and it is clear from the data in the table that the model proposed in this paper can obtain higher detection accuracy. Regarding time, the average detection time of 100 images is higher than YOLOv5 and SSD. However, it has a good detection speed compared with the other four models to obtain higher accuracy at the expense of less time. Therefore, combining accuracy and speed, the model in this paper is more suitable for the detection task of sprouting potatoes.

According to the analysis of the above results, the model in this paper has a higher performance advantage than the other six models. The improved model makes fuller use of low-level feature information, which improves the detection rate in small target detection; it improves the effect of occlusion target detection and improves the performance of the model more substantially.

D. PRACTICAL SCENARIO TEST ANALYSIS

We selected three representative potato images from the validation set, representing the recognition of potatoes in complex scenes, and the recognition results can be generalized to a certain extent to guarantee the effectiveness of potato recognition in natural scenes. As shown in Fig. 11, Fig. 11 (a) represents the potato recognition results under the original YOLOv5s model, and Fig. 11 (b) represents the potato recognition results under the improved YOLOv5 model in this paper. Where the red arrows in Fig. 11 (a) indicate the difference from the identification in Fig. 11 (b), the difference being precisely in the blue boxes missing to varying degrees; the black arrows in Fig. 11 (a) indicate the negation of the identification result in Fig. 11 (b), specifically in the blue boxes incorrectly identifying the non-defined semantic label categories, as present feature targets. By analysis, Fig. 11(a) from top to bottom relative to Fig. 11(b), the unidentified three types of feature boxes are 5, 2, and 5, respectively, with a total of 12 feature labels lost, and an average of 4 labels lost in one image. The images in the middle of Fig. 11 (a) and Fig. 11 (b), compared with the other two types of images, exhibit the characteristics of larger potato targets and less masking, and the actual results of detection by the original YOLOv5 model perform slightly better compared with the other two types of results in Fig. 11(a); however, for the intermediate images, Fig. 11(a) shows the problem of repeatedly identifying the same potatoes and repeatedly identifying the same targets compared to Fig. 11(b), which indicates that the original YOLOv5 model is shallow for feature extraction. Meanwhile, the germ feature pointed by the black arrow in the bottom image of Fig. 11 (a) does not appear in Fig. 11 (b), indicating that the original YOLOv5 model has

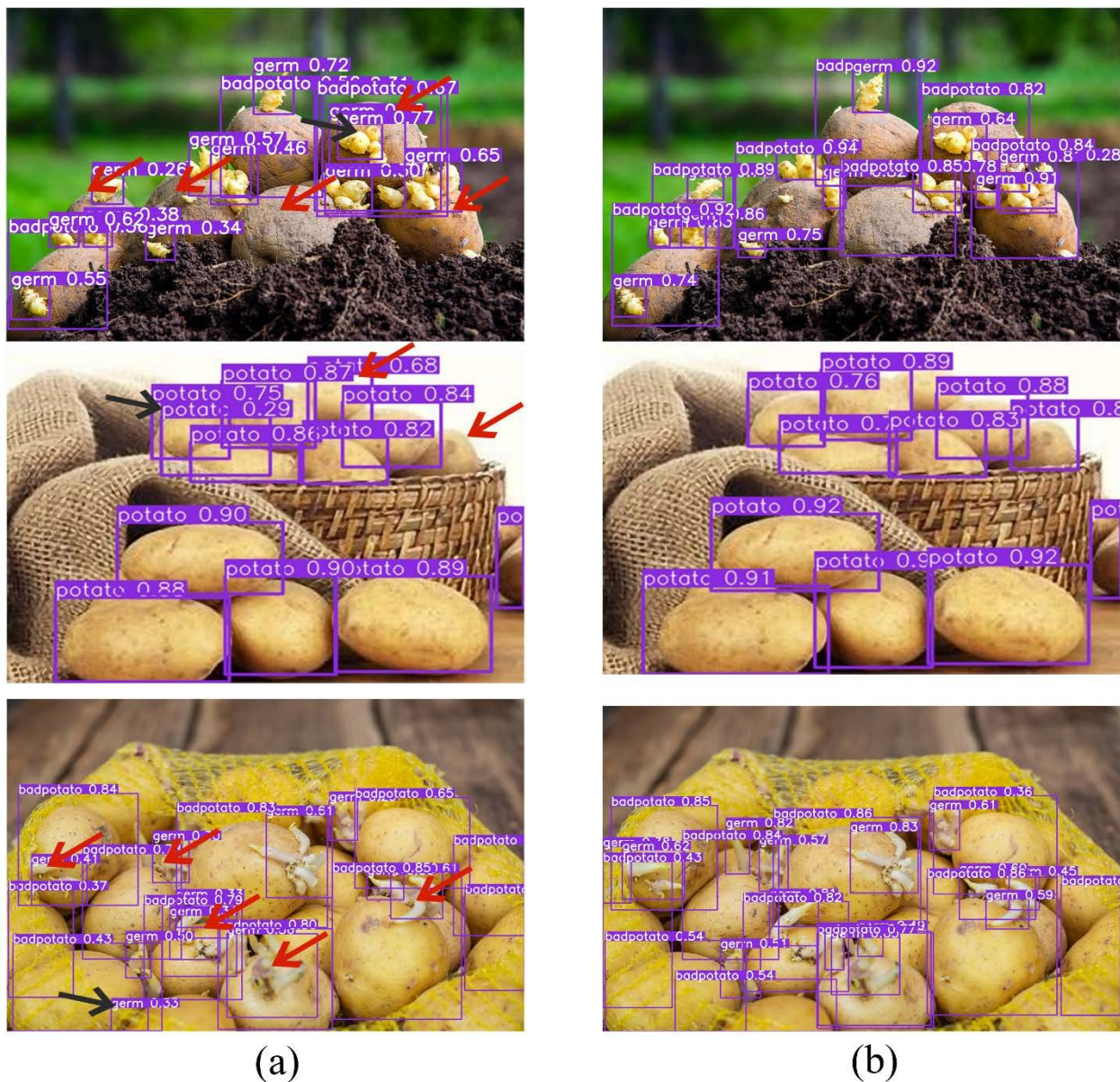


FIGURE 11. Images form the test dataset showing the performance for detecting the three classes potato; germ; badpotato.

TABLE 9. Comparison of performance indicators for different detection algorithms.

Models	AP/%			mAP%	Average detection time/s
	potato	germ	badpotato		
Faster R-CNN	62.4	72.8	68.3	67.8	0.026
SSD	87.4	86.5	76.6	83.5	0.017
YOLOv3	69.7	59.8	70.9	66.8	0.027
YOLOX	88.3	87.1	88.9	88.1	0.170
YOLOv4	77.5	68.4	81.2	75.7	0.048
YOLOv5	78.7	78.5	78	80.6	0.016
Proposed approach	90.1	80.4	84	88.1	0.020

poor generalization for generalization feature recognition. It also shows that the improved YOLOv5 model detects small

targets and multiple occlusions better than the original model. In addition, we tested the performance of the improved

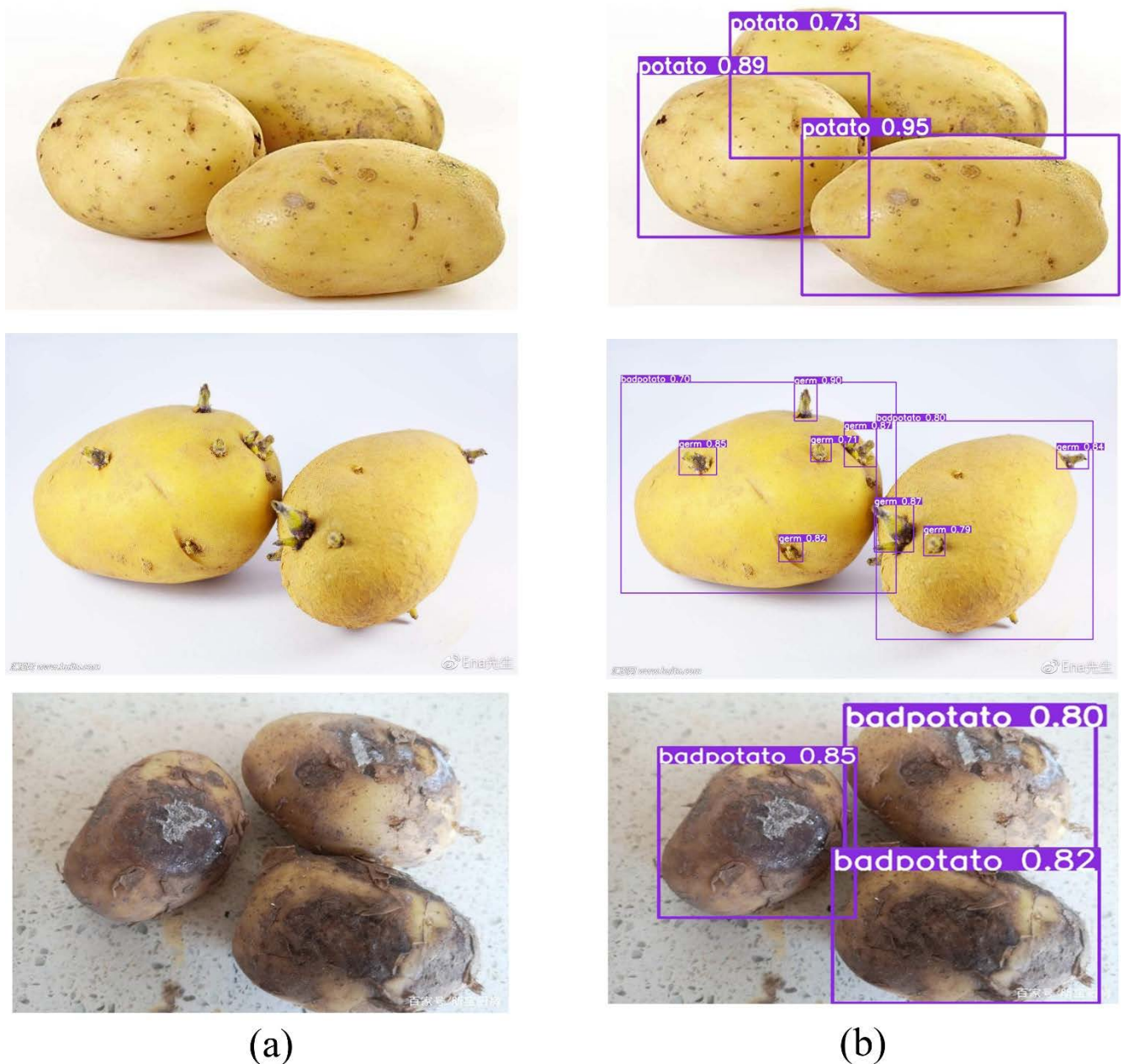


FIGURE 12. The test data set images show the performance of detecting three classes of normal potatoes, sprouted potatoes and rotten potatoes without sprouting.

YOLOv5 model on images of regular potatoes, sprouted potatoes, and rotten potatoes without sprouting. Fig. 12 shows that the model proposed in this paper can correctly identify the three types of labels in the potato dataset. It is also able to identify the smaller potato shoots for localization correctly. In the problem of complex sprouted unspoiled and rotten unspoiled potatoes, the model can effectively distinguish between the two types of cases, and potatoes with a high number of sprouts can also be treated as rotten potatoes promptly, basically realizing the critical need for potato quality grading. In summary, the improved YOLOv5 model effectively

solves the potato high-throughput grading problem with good robustness in complex environments.

IV. CONCLUSION

The purpose of this paper is to use the improved YOLOv5 detection algorithm to detect sprouting potatoes in various complex environments. First, we built a sprouted potato dataset containing 2316 images of sprouted potatoes in complex scenes. Next, we optimized the YOLOv5 model structure and adjusted the transfer rules of model features by CrossConv instead of Conv in the C3 module to connect

two intra-graph convolutions and enhance feature similarity. The introduced 9-mosaic data augmentation, genetic algorithm-based k -means with fast spatial pyramid SPPF to improve recognition accuracy, enhance feature similarity and strengthen small target features. In addition, hyperparametric evolution with a multiscale training mechanism is employed to improve the accuracy further. It solves the problems of YOLOv5, such as insensitivity to medium and large targets, minor defect misses and false detections, and overcomes the existing research on detecting sprouting potatoes based on specific scenarios. Finally, for practical application needs, the improved YOLOv5 model was shown to detect the three states of potatoes with higher accuracy, detecting mAP consistently above 88.1%. In the same test set, the accuracy was improved by 7.4% compared to the original model. All tests and results demonstrate that the model has good performance and sufficient speed to generalize strongly in multiple scenarios of sprouting potato recognition and maintain stability in recognition accuracy.

Nevertheless, since the data set contains sprouted, healthy and rotten potatoes, the model can only detect three states of potatoes. In actual production, mechanical damage to potato tubers and skin greening is also crucial for grading potatoes for quality inspection. Future research will include collecting images of potato mechanical damage and skin greening for detection. In addition, consideration should be given to simplifying and deploying the model on a mobile platform to build a more practical potato quality grading system to meet practical agricultural needs.

ACKNOWLEDGMENT

Author Contributions: conceptualization: Guowei Dai and Shen Yan; data creation: Guowei Dai and Shen Yan; methodology: Guowei Dai; software: Guowei Dai; validation: Guowei Dai; visualization: Guowei Dai; writing (original draft): Guowei Dai; writing (review and editing), Guowei Dai, Lin Hu, Jingchao Fan, and Ruijing Li; project administration: Guowei Dai and Shen Yan; funding acquisition: Jingchao Fan and Lin Hu; and resources: Jingchao Fan and Lin Hu. All authors have read and agreed to the published version of the manuscript.

REFERENCES

- [1] M. Boivin, N. Bourdeau, S. Barnabe, and I. Desgagne-Penix, "Sprout suppressive molecules effective on potato (*Solanum tuberosum*) tubers during storage: A review," *Amer. J. Potato Res.*, vol. 97, no. 5, pp. 451–463, Oct. 2020, doi: [10.1007/s12230-020-09794-0](https://doi.org/10.1007/s12230-020-09794-0).
- [2] S. Langkilde, M. Schröder, D. Stewart, O. Meyer, S. Conner, H. Davies, and M. Poulsen, "Acute toxicity of high doses of the glycoalkaloids, α -solanine and α -chaconine, in the Syrian Golden hamster," *J. Agricult. Food Chem.*, vol. 56, no. 18, pp. 8753–8760, Sep. 2008, doi: [10.1021/jf8012794](https://doi.org/10.1021/jf8012794).
- [3] J. R. Velasco-Chong, O. Herrera-Calderón, J. P. Rojas-Armas, R. D. Hañari-Quispe, L. Figueroa-Salvador, G. Peña-Rojas, V. Andía-Ayme, R. Á. Yuli-Posadas, A. F. Yepes-Perez, and C. Aguilar, "TOCOSH FLOUR (*Solanum tuberosum* L.): A toxicological assessment of traditional Peruvian fermented potatoes," *Foods*, vol. 9, no. 6, p. 719, Jun. 2020, doi: [10.3390/foods9060719](https://doi.org/10.3390/foods9060719).
- [4] X. Jing, H. Zhiliang, T. Linyue, and X. Juntao, "Research of potato quality detection technology based on computer vision," *Comput. Eng. Appl.*, vol. 54, no. 5, pp. 165–169, 2018.
- [5] X. Weidong and Z. Zhonggai, "Potato shape sorting based on PCA-SVM algorithm," *Control Eng. China*, vol. 27, no. 2, pp. 246–253, 2020.
- [6] G. Hai-Long, L. Xiao-Yu, X. Sen-Miao, T. Hai-Long, L. Xiao-Jin, and S. Jin-Feng, "Comparative study of transmission and reflection hyperspectral imaging technology for potato damage detection," *Spectrosc. Spectral Anal.*, vol. 33, no. 12, pp. 3366–3371, Dec. 2013, doi: [10.3964/j.issn.1000-0593\(2013\)12-3366-06](https://doi.org/10.3964/j.issn.1000-0593(2013)12-3366-06).
- [7] D. Ye, L. Sun, W. Tan, W. Che, and M. Yang, "Detecting and classifying minor bruised potato based on hyperspectral imaging," *Chemometrics Intell. Lab. Syst.*, vol. 177, pp. 129–139, Jun. 2018, doi: [10.1016/j.chemolab.2018.04.002](https://doi.org/10.1016/j.chemolab.2018.04.002).
- [8] F. Jubayer, J. A. Soeb, A. N. Mojmunder, M. K. Paul, P. Barua, S. Kayshar, S. S. Akter, M. Rahman, and A. Islam, "Detection of mold on the food surface using YOLOv5," *Current Res. Food Sci.*, vol. 4, pp. 724–728, Jan. 2021, doi: [10.1016/j.crf.2021.10.003](https://doi.org/10.1016/j.crf.2021.10.003).
- [9] I. S. Walia, D. Kumar, K. Sharma, J. D. Hemanth, and D. E. Popescu, "An integrated approach for monitoring social distancing and face mask detection using stacked ResNet-50 and YOLOv5," *Electronics*, vol. 10, no. 23, p. 2996, Dec. 2021, doi: [10.3390/electronics10232996](https://doi.org/10.3390/electronics10232996).
- [10] M.-E. Otgonbold, M. Gochoo, F. Alnajjar, L. Ali, T.-H. Tan, J.-W. Hsieh, and P.-Y. Chen, "SHEL5K: An extended dataset and benchmarking for safety helmet detection," *Sensors*, vol. 22, no. 6, p. 2315, Mar. 2022, doi: [10.3390/s22062315](https://doi.org/10.3390/s22062315).
- [11] Z. Zheng, J. Zhao, and Y. Li, "Research on detecting bearing-cover defects based on improved YOLOv3," *IEEE Access*, vol. 9, pp. 10304–10315, 2021, doi: [10.1109/ACCESS.2021.3050484](https://doi.org/10.1109/ACCESS.2021.3050484).
- [12] B. Wei, Y. Zhang, Y. Pu, Y. Sun, S. Zhang, H. Lin, C. Zeng, Y. Zhao, K. Wang, and Z. Chen, "Recursive-YOLOv5 network for edible mushroom detection in scenes with vertical stick placement," *IEEE Access*, vol. 10, pp. 40093–40108, 2022, doi: [10.1109/ACCESS.2022.3165160](https://doi.org/10.1109/ACCESS.2022.3165160).
- [13] B. Darwin, P. Dharmaraj, S. Prince, D. E. Popescu, and D. J. Hemanth, "Recognition of bloom/yeild in crop images using deep learning models for smart agriculture: A review," *Agronomy*, vol. 11, no. 4, p. 646, Mar. 2021, doi: [10.3390/agronomy11040646](https://doi.org/10.3390/agronomy11040646).
- [14] L. J. Biffi, E. Mitishita, V. Liesenberg, A. A. D. Santos, D. N. Gonçalves, N. V. Estrabis, J. D. A. Silva, L. P. Osco, A. P. M. Ramos, J. A. S. Centeno, M. B. Schimalski, L. Rufato, S. L. R. Neto, J. Marcato, Jr., and W. N. Gonçalves, "ATSS deep learning-based approach to detect apple fruits," *Remote Sens.*, vol. 13, no. 1, p. 54, Dec. 2020, doi: [10.3390/rs13010054](https://doi.org/10.3390/rs13010054).
- [15] Y. Bin, F. Pan, L. Xiaoyan, L. Zhijie, and Y. Fuzeng, "A real-time apple targets detection method for picking robot based on improved YOLOv5," *Remote Sens.*, vol. 13, no. 9, p. 1619, May 2021, doi: [10/gmmbpc](https://doi.org/10/gmmbpc).
- [16] K. Ko, I. Jang, J. H. Choi, J. H. Lim, and D. U. Lee, "Stochastic decision fusion of convolutional neural networks for tomato ripeness detection in agricultural sorting systems," *Sensors*, vol. 21, no. 3, p. 917, Jan. 2021, doi: [10.3390/s21030917](https://doi.org/10.3390/s21030917).
- [17] S. A. Magalhães, L. Castro, G. Moreira, F. N. dos Santos, M. Cunha, J. Dias, and A. P. Moreira, "Evaluating the single-shot MultiBox detector and YOLO deep learning models for the detection of tomatoes in a greenhouse," *Sensors*, vol. 21, no. 10, p. 3569, May 2021, doi: [10.3390/s21103569](https://doi.org/10.3390/s21103569).
- [18] O. M. Lawal, "Development of tomato detection model for robotic platform using deep learning," *Multimedia Tools Appl.*, vol. 80, no. 17, pp. 26751–26772, May 2021, doi: [10.1007/s11042-021-10933-w](https://doi.org/10.1007/s11042-021-10933-w).
- [19] G. Gao, S. Wang, C. Shuai, Z. Zhang, S. Zhang, and Y. Feng, "Recognition and detection of greenhouse tomatoes in complex environment," *Traitement du Signal*, vol. 39, no. 1, pp. 291–298, Feb. 2022, doi: [10.18280/ts.390130](https://doi.org/10.18280/ts.390130).
- [20] F. Saeed, M. A. Khan, M. Sharif, M. Mittal, L. M. Goyal, and S. Roy, "Deep neural network features fusion and selection based on PLS regression with an application for crops diseases classification," *Appl. Soft Comput.*, vol. 103, May 2021, Art. no. 107164, doi: [10.1016/j.asoc.2021.107164](https://doi.org/10.1016/j.asoc.2021.107164).
- [21] M. E. H. Chowdhury, T. Rahman, A. Khandakar, M. A. Ayari, A. U. Khan, M. S. Khan, N. Al-Emadi, M. B. I. Reaz, M. T. Islam, and S. H. M. Ali, "Automatic and reliable leaf disease detection using deep learning techniques," *AgriEngineering*, vol. 3, no. 2, pp. 294–312, May 2021, doi: [10.3390/agriengineering3020020](https://doi.org/10.3390/agriengineering3020020).
- [22] S. Marino, P. Beausery, and A. Smolarz, "Weakly-supervised learning approach for potato defects segmentation," *Eng. Appl. Artif. Intell.*, vol. 85, pp. 337–346, Oct. 2019, doi: [10.1016/j.engappai.2019.06.024](https://doi.org/10.1016/j.engappai.2019.06.024).

- [23] X. Rui, J. Kai, Z. Wanzhi, L. Zhaoqin, and H. Jialin, "Recognition method for potato buds based on improved faster R-CNN," *Trans. Chin. Soc. Agricult. Machinery*, vol. 51, no. 4, pp. 216–223, 2020.
- [24] J. Solawetz. *YOLOv5 New Version—Improvements and Evaluation*. Accessed: Apr. 2021. [Online]. Available: <https://blog.roboflow.com/yolov5-improvements-and-evaluation>
- [25] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9196–9205, doi: [10.1109/ICCV.2019.00929](https://doi.org/10.1109/ICCV.2019.00929).
- [26] R. Wang, J. Yan, and X. Yang, "Learning combinatorial embedding networks for deep graph matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3056–3065, doi: [10.1109/ICCV.2019.00315](https://doi.org/10.1109/ICCV.2019.00315).
- [27] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031, doi: [10.1109/ICCV.2019.00612](https://doi.org/10.1109/ICCV.2019.00612).
- [28] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jun. 2015, doi: [10.1109/TPAMI.2015.2389824](https://doi.org/10.1109/TPAMI.2015.2389824).
- [29] L. Hertel, P. Baldi, and D. L. Gillen, "Reproducible hyperparameter optimization," *J. Comput. Graph. Statist.*, vol. 31, no. 1, pp. 84–99, Jan. 2022, doi: [10.1080/10618600.2021.1950004](https://doi.org/10.1080/10618600.2021.1950004).
- [30] M.-D. Minut and A. Iftene, "Creating a dataset and models based on convolutional neural networks to improve fruit classification," in *Proc. 23rd Int. Symp. Symbolic Numeric Algorithms Sci. Comput. (SYNASC)*, Dec. 2021, pp. 155–162, doi: [10.1109/SYNASC54541.2021.00035](https://doi.org/10.1109/SYNASC54541.2021.00035).
- [31] lltcggie, "Rewrite image conversion software using caffe," Zenodo, Sep. 5, 2020. [Online]. Available: <https://zenodo.org/record/6605147#.Yve3ki5By4Q> and <https://github.com/lltcggie/waifu2x-caffe>, doi: [10.5281/zenodo.6605147](https://doi.org/10.5281/zenodo.6605147).
- [32] K.-E. Margrit, H. Nico, B. Stian, S. Tom, M. Oystein, and K. P. Egil, "Short communication: Detecting heavy goods vehicles in rest areas in winter conditions using YOLOv5," *Algorithms*, vol. 14, no. 4, p. 114, Apr. 2021, doi: [10/gmxrvh](https://doi.org/10/gmxrvh).



JINGCHAO FAN received the B.S. degree in computer science and technology from Liaoning University, in 2003, the M.S. degree in management science and engineering from the Chinese Academy of Agricultural Sciences, in 2006, and the Ph.D. degree in crop information science from the Chinese Academy of Agricultural Sciences, in 2011. He is currently the Deputy Director of the Scientific Data Research Department, Agricultural Information Institute of CAAS (AIICAAS).

He is mainly engaged in research on agricultural information technology, Chinese web information retrieval, and fruit tree model construction. He has presided over and participated in seven projects, such as 863, transformation of scientific and technological achievements, and basic scientific research business fund. He has participated in the construction of six national science and technology infrastructure platforms as a Technical Backbone. He has received three scientific and technological achievements, published 12 papers; authorized 15 utility model patents, applied for four invention patents, and 16 software copyrights.



SHEN YAN received the Ph.D. degree in crop genetics and breeding from Sichuan Agricultural University, in 2020. He is currently working as a Postdoctoral Researcher at the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. His research interest includes agricultural information technology.



GUOWEI DAI received the B.S. degree in computer science and technology from Wuhan Polytechnic University, in 2021. He is currently pursuing the M.S. degree with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. His research interests include agricultural artificial intelligence and agricultural information technology. He serves as a Reviewer for *Journal of Agricultural Engineering*.



LIN HU received the B.S. degree in forest conservation and the Ph.D. degree in forest management from Beijing Forestry University, in 1988 and 2006, respectively. He is currently the Director of the Science Data Research Department, Institute of Agricultural Information of the Chinese Academy of Agricultural Sciences (AIICAAS). He has been engaged in data analysis and visualization research for a long time. He has presided over and participated in more than 20 national and provincial projects, received three scientific and technological achievements awards at provincial and ministerial levels, published 50 papers, received more than ten patents, and published three books. He has supervised three M.Sc. and Ph.D. students. His research interests include agricultural information technology, data analysis, and visualization.



RUIJING LI received the B.A. degree in sociology from Northern State University, USA, in 2010, the M.A. degree in research method (science, technology and society) and the Ph.D. degree in science and technology from the University of Nottingham, in 2012 and 2017, respectively, and the Postdoctoral degree from the Beijing Academy of Science and Technology, in 2021. She is currently working with the Scientific Data Center, Institute of Agricultural Information, Chinese Academy of Agricultural Sciences. She has published more than 15 articles in relevant fields. Her research interests include technology and knowledge transfer.