IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 November 2023, accepted 23 January 2024, date of publication 13 February 2024, date of current version 6 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365559

RESEARCH ARTICLE

Research on a Lightweight Method for Maize Seed Quality Detection Based on Improved YOLOv8

SIQI NIU[1], XIAOLIN XU[1], AO LIANG[1], YULIANG YUN[2], LI LI[3], FENGQI HAO[4], JINQIANG BAI[4], AND DEXIN MA[1,5]

[1]College of Animation and Communication, Qingdao Agricultural University, Qingdao 266109, China
[2]College of Mechanical and Electrical Engineering, Qingdao Agricultural University, Qingdao 266109, China
[3]Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China
[4]Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China
[5]Intelligent Agriculture Institute, Qingdao Agricultural University, Qingdao 266109, China

Corresponding author: Dexin Ma (madexin@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 42201458, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2022MC152, in part by the Central Government Guiding Local Science and Technology Development Special Plan under Grant 23-1-3-6-zyyd-nsh, and in part by the Shandong Province Key Research and Development Plan under Grant 2023TZXD023.

ABSTRACT Seeds are the most basic and important means of production for agriculture. During the production and processing of seeds, they may undergo potential mechanical damages and mildew alterations, which might jeopardize their germination viability. Hence, checking the quality of seeds before sowing is of paramount importance for the benefit of the sower and the safety of agricultural production. In order to achieve an efficient detection of maize seed quality, our experiment assembled a dataset composed of 2,128 seeds with four different health statuses of maize: healthy, broken, moth-eaten, and mildewed. In this paper, we proposed a lightweight maize seed quality detection model for small objects based on improved YOLOv8: I-YOLOv8. Firstly, we introduced a multi-scale attention mechanism called EMA to efficiently retain information across channels and reduce computational load. Next, we chosen the SPD-Conv module for low-resolution images and small objects, and applied it to the backbone, which addressed the loss of fine-grained information and the less efficient learning of feature representations present in YOLOv8. Lastly, we reduced the large detection layer, which directed the network to pay more attention to the location, channel, and dimensional information of smaller objects, and we also replaced the loss function with WIoUv3. We validated our model using ablation studies and compared it with YOLOv5, YOLOv6, and YOLOv8. The mAP (Mean Average Precision) of the improved model I_YOLOv8 reaches 98.5%, which is 6.7% higher than YOLOv8. The average recognition time per image was 163.9fps, a boost of 5.2fps compared to YOLOv8. This study lays a theoretical foundation for the efficient, convenient, and rapid detection of maize quality, while also offering a technical basis for advancing automated maize quality detection means.

INDEX TERMS YOLOv8, object detection, lightweighting, maize seed.

I. INTRODUCTION

Maize (Zea mays L.) is one of the most widely distributed crops in the world [1]. About one-third of the world's population depends on maize as a staple food [2]. It has high nutritional and economic value [3]. During the production and processing of maize seeds, they may undergo potential mechanical damages and mildew alterations, which might jeopardize their germination viability. Sowing maize seeds with damaged quality will reduce the germination rate and waste of labor, thus affecting economic benefits. Hence, checking the quality of seeds before sowing is of paramount

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

importance for the benefit of the sower and the safety of agricultural production.

Traditional methods for detecting maize seed quality are categorized into empirical, physical, and chemical methods. These methods are often cumbersome, time-consuming, and limited by experimental locale [4]. The application of machine vision technology can achieve rapid and accurate detection and identification of maize seeds, which has very important application value [5].

In the early stages of conventional machine learning, some scholars used image processing techniques to simply process maize seed images for recognition. Chen et al. [6] proposed a method based on image In HSV and Otsu method based on genetic algorithm optimization, which achieved more accurate segmentation and recognition of the disease of color and shape features, and enhanced the real-time and accuracy of the image of maize disease detection and recognition. Subsequently, many researchers began to use neural network approaches for maize seed image analysis. Kiratiratanapruk and Sinthupinyo [7] extracted color histograms from RGB and HSV color spaces, along with textures based on Gray-Level Co-Occurrence Matrix (GLCM) and Local Binary Patterns (LBP), and then applied Support Vector Machine (SVM) to classify maize seed defects.

Traditional machine learning approaches have achieved some applications in the recognition of maize seeds. However, these traditional methods are subject to limitations such as their reliance on manually selected features, and the problem such as high computational demands and costs, which can impact recognition accuracy. In contrast, deep learning can autonomously learn complex features from raw data without the need for manual feature extraction. Presently, the most pervasive deep learning technique in machine learning is Convolutional Neural Network (CNN) [8]. For the detection of maize seeds, there are two main directions: image classification and target detection. Unlike image classification, which only categorizes objects within an image, object detection techniques perform image segmentation based on geometric and statistical features of targets, combining the tasks of object segmentation and recognition, so it has excellent accuracy and real-time processing capabilities.

Object detection algorithms based on deep learning are mainly divided into two categories: two-stage and single-stage methods. The two-stage object detection algorithm initially generates RP (Region Proposals), followed by sample classification using convolutional neural networks. Representative two-stage object detection algorithms include R-CNN [9], SPP-Net [10], Faster R-CNN [11], Mask R-CNN [12]. Zhao et al. [13] designed four distinct network models based on Faster R-CNN, and achieved superior recognition results for the selection of maize kernels by directly inputting color images. Velesaca et al. [14] used the Mask R-CNN algorithm for segmenting and extracting maize group images, and designed a lightweight network CK-CNN to classify good kernels, defective kernels, and impurities. Although two-stage detection algorithms exhibit high accuracy, their slower detection speed renders them unsuitable for real-time detection. The single-stage object detection algorithm can directly generate the class probability and location coordinates of the target in one stage without generating RP. Representative two-stage object detection algorithms include SSD [15], YOLO [16], RetinaNet [17]. Although these algorithms may exhibit moderate accuracy, their fast detection speed makes them suitable for real-time detection tasks. Liu and Wang [18]proposed a method for detecting damaged maize kernels based on YOLOv3-tiny. The proposed maize detection method is implemented on NVIDIA TX2 and can achieves the speed up to 10fps speed, which can perform almost real-time detection. Li et al. [19] investigated a maize seed breakage detection device based on YOLOv4-tiny, which is applicable to combine harvesters, addressing the issue of low accuracy in existing methods for maize seed integrity assessment. Thangaraj Sundaramurthy et al. [20] proposed an object detection method based on YOLOv5 to accurately detect maize seeds infected with Fusarium Head Blight (FHB), enabling real-time detection of FHB-infected maize seeds on the processing line. Wang et al. [21] aimed to rapidly and accurately identify broken maize kernels, proposing a model BCK-YOLOv7 based on an improved YOLOv7, which fine-tuned the model's positive sample matching strategy and incorporated Transformer encoding modules and CA attention mechanisms, enhancing the model's accuracy to 96.9%, recall to 97.5%, and mAP to 99.1%.

The study proposed a more streamlined convolutional neural network model, I_YOLOv8, based on YOLOv8 benchmark, addressing the loss of fine-grained information and low-efficiency feature representation learning inherent to YOLOv8, it also reducing the model's complexity for ease deployment on mobile devices. This experiment established a dataset for maize seed variety recognition, encompassing 2128 maize seeds of four types: healthy, broken, moth-eaten, and mildewed. Extensive comparative experiments were conducted with the I_YOLOv8, YOLOv5, YOLOv6, and YOLOv8 using this dataset. The results indicated that the proposed I_YOLOv8 significantly outperformed other methods, providing technical support for the automated recognition and non-destructive testing of maize seed quality.

II. MATERIALS AND METHODS

A. DATASET CONSTRUCTION

This experiment collected four types of maize seeds with different health conditions to establish a dataset, as shown in Figure 1, which are healthy, broken, moth-eaten, and mildewed. In order to enhance the robustness of the model, we chose five maize seeds varieties, namely JINYU118, KENUO58, LIYUAN296, HUIYU18, and TIEYAN630. Under natural lighting conditions, these seeds were randomly arranged in a ratio of healthy, broken, moth-eaten, and mildewed at 5:1:1:1. A total of 133 photographs were taken, each featuring 16 seeds, culminating in a dataset of

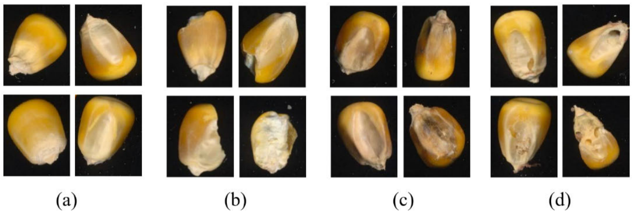

FIGURE 1. Maize seeds with four different health conditions. (a) Healthy maize. (b) Broken maize. (c) Mildewed maize. (d)Moth-eaten maize.

2128 seeds, the images of each maize seed category were randomly divided into training, validation, and test sets in a ratio of 7:2:1.

To further enhance the robustness and generalization capabilities of the model, this study augmented the training image number of the dataset through data augmentation techniques, these enhancements primarily included rotation, exposure adjustment, and mosaic techniques. The augmented maize seed dataset is shown in Figure 2. Figure 2(a) showed the quantity of maize seeds across different categories. Figure 2(b) depicted the spatial distribution of the maize seeds bounding boxes, indicating a relatively uniform spread of maize seeds without excessive clustering. Figure 2(c) presented the dimensions of the maize seeds bounding boxes, it can be seen that the height and width of the bounding box are relatively uniform.

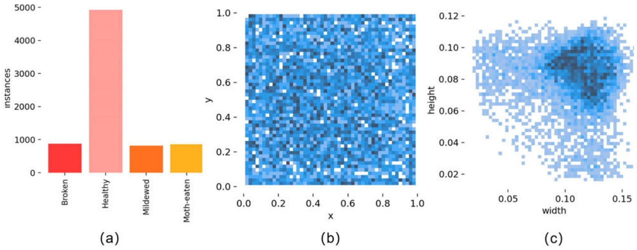

FIGURE 2. Visualization of the dataset. (a) Number of annotations per class. (b) The statistical distribution of the bounding box position. (c) The statistical distribution of the bounding box sizes.

B. MAIZE SEED DETECTION MODEL: I_YOLOv8

Since its initial release in 2015, YOLO (You Only Look Once) series of computer vision models has consistently been one of the most popular in the field of deep learning. YOLOv8 is the latest version of the YOLO series of algorithms, which can be used for object detection, segmentation, classification tasks, and learning of large-scale datasets. Compared to previous outstanding models in the YOLO series, such as YOLOv5 and YOLOv7, YOLOv8 offers higher detection accuracy and speed. YOLOv8 is a detection algorithm known for its fast detection speed and high accuracy. It performs well on some open-source datasets but requires improvement for seed detection tasks.

In order to address the challenges such as small seed size and low resolution in maize seed quality detection, this paper improved and optimized on the basis of YOLOv8, and proposed an algorithm I-YOLOv8 for maize seed quality detection.

The proposed model architecture was illustrated in Figure 3, and the specific improvements were summarized as follows:

(1) By incorporating attention mechanisms to enhance the object detection capability of the network and extract regions of interest. As illustrated in Figure 3, an efficient multi-scale attention mechanism EMA was introduced in the Backbone and added in front of the SPPF structure to efficiently retain the information on each channel and reduce the computational load.

(2) By integrating SPD-Conv module to boost detection capabilities of low-resolution images and small objects. As demonstrated in Figure 3, SPD-Conv was applied within the Backbone at the following stages: the 1st, 3rd, 5th, and 7th convolutional layers. This addressed the loss of fine-grained details and learning of less effective feature representations in YOLOv8.

(3) By removing one of the large object detection layers (P5), the network was directed to focus more on the location, channel, and dimensional information of smaller objects, thereby enhancing the detection of small-scale targets. As indicated in Figure 3, the detection layers of YOLOv8n were simplified from large, medium, and small to medium and small. The 24th layer (Convolution layer), the 25th layer (Concat module), and the 26st layer (C2f module) of the feature fusion layers were removed.

(4) By altering the loss function, the model's accuracy and overall performance were further improved. The loss function CIoU was replaced with WIoUv3, which balanced the ratio between low and high-quality samples, addressing issues of detecting small, blurry objects and those with overlapping occlusions.

1) EFFICIENT MULTI-SCALE ATTENTION MODULE

The attention mechanism is a mechanism that simulates human vision, focusing on important features while suppressing unnecessary ones. Remarkable effectiveness of the channel or spatial attention mechanisms for producing more discernible feature representation are illustrated in various computer vision tasks. However, modeling the cross-channel relationships with channel dimensionality reduction may bring side effect in extracting deep visual representations. Ouyang et al. [22] proposed a novel efficient multi-scale attention (EMA) module. The design prioritizes the retention of information from each channel while minimizing computational overhead. By reshaping certain channels into a batch dimension and segmenting the channel dimension into multiple sub-features, it ensures a homogeneous distribution of spatial semantic attributes within each feature subset. Specifically, apart from encoding the global information to re-calibrate the channel-wise weight in each parallel branch, the output features of the two parallel branches are further aggregated by a cross-dimension interaction for capturing pixel-level pairwise relationship.

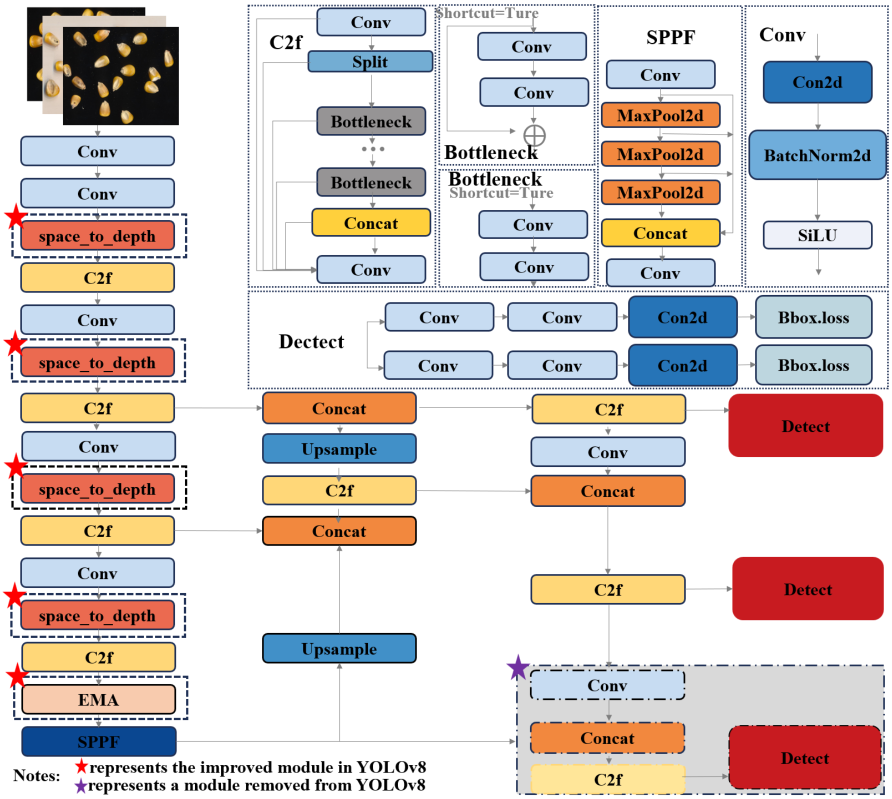

FIGURE 3. Improved YOLOv8 network structure.

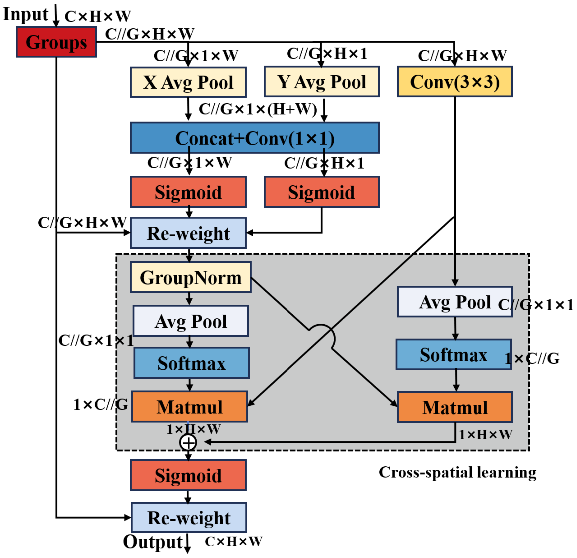

FIGURE 4. Efficient multi-scale attention structure.

The parallel substructures help the networks avoid more sequential processing and large depth. As shown in Figure 4, the EMA module employs a parallel processing strategy.

(1) Feature Grouping. For any given input feature map $X \in \mathbb{R}^{C\times H\times W}$, EMA will divide X into G sub-features across the channel dimensions direction for learning different semantics, where the groups-style can be donated by $X = [X_0, X_i, \ldots, X_{G-1}]$, $X \in \mathbb{R}^{C\times H\times W}$.

(2) Parallel Subnetworks. The large local receptive fields of neurons enable the neurons to collect multi-scale spatial information. Accordingly, EMA conducts that three parallel routes are exploited to extract attention weight descriptors of the grouped feature maps. Two of parallel routes is in 1x1 branch and the third one route is that the 3x3 branch. For capturing dependencies across all channels and relieving the computation budgets, they model the cross-channel information interaction at channel direction. To be more specific, there are two 1D global average pooling operations employed to encode the channel along two spatial directions respectively in 1x1 branch and only a single 3x3 kernel is stacked in 3x3 branch for capturing multi-scale feature representation.

Given the truth that there is no batch coefficient in the dimension of the convolution function for the normal convolution, the number of convolution kernels are independent of the batch coefficients of the forward operational inputs. Accordingly, the group G is reshaped and replaced into the batch dimension, and the shape of the input tensor is redefined as $C//G \times H \times W$. On one hand, the two encoded features are concatenated against the images height direction and share the same 1x1 convolution, without dimensionality reduction in the 1x1 branch. After factorize the outputs of 1x1 convolution into two vectors, two non-linear Sigmoid functions are employed to fit the 2D Binormial distribution upon linear convolutions. In order to implement different cross-channel interaction features between two parallel routes of 1x1 branching, the attention maps of the two channels are aggregated within each group by a simple multiplication. On the other hand, the 3x3 branch captures the local cross-channel interaction via a 3x3 convolution to enlarge the feature space.

(3) Cross-spatial learning. EMA provides a cross-spatial information aggregation method at different spatial dimension direction for richer feature aggregation. Please note that here, two tensors are still introduced: one from the output of the 1x1 branch and the other from the output of the 3x3 branch. Then, global spatial information is encoded into the output of the 1x1 branch using 2D global average pooling. The output of the 3x3 branch is directly transformed to the corresponding dimension shape before the joint activation mechanism of channel features, i.e., ${\mathbb{R}_1}^{1\times\mathbb{C}//G} \times {\mathbb{R}_3}^{C//G\times HW}$. The 2D global pooling operation formula is shown in Equation (1):

$$z_c = \frac{1}{H \times W}\sum_{j}^{H}\sum_{i}^{W} x_c(i,j) \tag{1}$$

which is designed for encoding the global information and modeling the long-range dependencies. For efficient computation, the natural non-linear functions Softmax for 2D Gaussian maps is employed at the outputs of 2D global average pooling to fit the upon linear transformations. By multiplying the outputs of above parallel processing with matrix dot-product operations, the first spatial attention map can be derived. To observe this, it collects different scale spatial information in the same processing stage. Moreover, global spatial information in the 3x3 branch is encoded using 2D global average pooling and the 1x1 branch will be transformed to the correspond dimension shape directly before the joint activation mechanism of channel features, i.e., ${\mathbb{R}_1}^{1\times\mathbb{C}//G} \times {\mathbb{R}_3}^{C//G\times HW}$. After that, the second spatial attention map, which preserves the entire precise spatial positional information is derived. Finally, the output feature map within each group is calculated as the aggregation of the two generated spatial attention weight values followed by a Sigmoid function. It captures pixel-level pairwise relationship and highlights global context for all pixels. The final output of EMA is the same size of X, which is efficient yet effective to stack into modern architectures.

2) SPD-CONV MODULE

Convolutional Neural Networks (CNNs) have achieved significant success in various computer vision tasks such as image classification and object detection. However, their performance rapidly deteriorates in more challenging tasks characterized by lower image resolutions or smaller objects. This limitation arises from a common yet flawed design in existing CNN architectures, which involves the use of strided convolutions and/or pooling layers. These components lead to the loss of fine-grained information and inefficient learning of features. SPD-Conv [23] serves as a novel CNN module, capable of replacing each strided convolution and pooling layer. SPD-Conv is composed of an SPD layer and a non-strided convolution layer.

a: SPACE-TO-DEPTH (SPD)

The SPD component extends the (original) image transformation technique to downsample feature maps inside CNNs and throughout the entire CNN. As illustrated below, consider any intermediate feature map X of size $S \times S \times C_1$, where a series of sub-feature maps is extracted as

$$\begin{aligned}
&f_{0,0} = X[0:S:scale, 0:S:scale],\\
&f_{1,0} = X[1:S:scale, 0:S:scale], \ldots,\\
&f_{scale-1.0} = X[scale-1:S;scale, 0:S:scale];\\
&f_{0,1} = X[0:S:scale, 1:S:scale], f_{1,1}, \ldots,\\
&f_{scale-1,1} = X[scale-1:S:scale, 1:S:scale];\\
&\qquad\qquad\vdots\\
&f_{0,scale-1} = X[0:S:scale, scale-1:S:scale],\\
&f_{1,scale-1}, \ldots,\\
&f_{scale-1,scale-1} = X[scale-1:S:scale, scale-1:S:scale].
\end{aligned}$$

In general, for any given (original) feature map X, a sub-map $f_{x,y}$ is formed by all stripes $X(i+y)$ where $i+x$ and $i+y$ are divisible by a scaling factor. Thus, each sub-map undergoes downsampling on feature map X according to the scaling factor.

Next, these sub-maps are concatenated along the channel dimension, resulting in a feature map X', where its spatial dimensions are reduced by a scaling factor, and the channel dimension is increased by a scaling factor. In other words, SPD transforms the feature map $X(S, S, C_1)$ into an intermediate feature map $X'(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1)$.

b: NON-STRIDED CONVOLUTION

After the SPD feature transformation layer, a non-strided (i.e., stride=1) convolutional layer with C_2 filters, where $C_2 < scale^2 C_1$, is added, and further transforms $X'(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1) \rightarrow X''(\frac{S}{scale}, \frac{S}{scale}, C_2)$. The use of non-strided convolution aims to preserve all discriminative feature information as much as possible.

3) DETECTION LAYER MODULE

The input size of YOLOv8 is set to 640 × 640, and the output layer modules P3, P4, and P5 are designed for detecting large, medium, and small objects, respectively. The maize seeds are characterized by their small volume and minimal variation within the image, so the detection belongs to small object detection. Based on this, the detection layers of YOLOv8n were simplified from large, medium, and small to medium and small. The 24th layer (Convolution layer), the 25th layer (Concat module), and the 26st layer (C2f module) of the feature fusion layers were removed. Reducing the number of detection layers effectively decreased model parameters and computational complexity, thus leading to improved detection speed.

4) DETECTION LAYER MODULE

The YOLOv8 uses CIoU [24] bounding box loss function, which incorporates the overlap area, center point distance, and aspect ratio during bounding box regression, which enhanced the precision of regression localization. However, CIoU still suffers from the following problem: during the regression process of the prediction frame, if the height and width aspect ratios between the predicted box and the ground truth box are linearly proportional, the penalty for the relative proportions degenerates zero, thereby affecting the optimization of the network.

Due to the inherent imbalance in the dataset, there were inevitably low-quality samples in the training data. In order to address the issue of sample quality imbalance, based on WIoUv1, WIoUv3 was proposed to balance the ratio of low-quality to high-quality samples, thereby addressing challenges related to the unclear delineation of small targets and difficulties in detecting overlapping and occluded objects [25]. Consequently, it further enhanced the model's accuracy and overall detection performance. We replaced the loss function with WIoUv3, the spatial relationship between the ground truth box and the predictive box is shown in Figure 5.

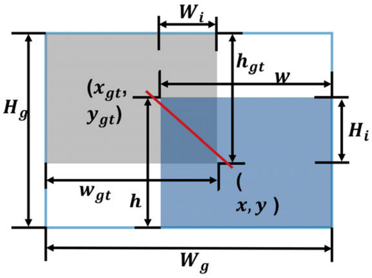

FIGURE 5. The spatial relationship between the ground truth box(red) and the predictive box(blue). *w* and *h* represent the width and height of the predicted box respectively; w_{gt} and h_{gt} represent the width and height of the ground truth box; W_i and H_i respectively indicate the width and height of the overlapping rectangle between predictive box and the ground true box; W_g and H_g are the width and height of the minimum enclosing rectangle of the predictive box and the ground truth box.

The calculation formula for WIoUv1 is shown in Equations (2) - (4):

$$L_{IoU} = 1 - IoU = 1 - \frac{W_iH_i}{wh + w_{gt}h_{gt} - W_iH_i} \tag{2}$$

$$L_{WIoUv1} = R_{WIoU}L_{IoU} \tag{3}$$

$$R_{WIoU} = \exp(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}) \tag{4}$$

The abnormality degree of the anchor box β is represented by the ratio of $L^*_{IoU}$and $\overline{L_{IoU}}$, as shown in Formula (5):

$$\beta = \frac{L^*_{IoU}}{\overline{L_{IoU}}} \in [0, +\infty) \tag{5}$$

Applying β to WIoUv1 to construct nonmonotonic focusing coefficients yields WIoUv3 as shown in Eq. (6):

$$L_{WIoUv3} = rL_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{6}$$

where the mapping of outlier degree is β and gradient gain is r, which is controlled by the hyper-parameters α and δ; The value of a is 1.9 and the value of δ is 3.

C. MODEL TRAINING ENVIRONMENT CONFIGURATION

All model training and testing procedures were conducted on the same workstation (Windows 11 with 64-bit operating system, Intel (R) Core (TM) i9-13900HX, NVIDIA GeForce RTX 4060), PyTorch2.0.0, Python 3.8, and CUDA 11.7 was utilized for training acceleration. Detailed hyperparameters of the experiment are shown in Table1.

TABLE 1. Detailed hyperparameters of experiment.

Parameters	Value
image size	640×640
batch size	16
Classes	4
epochs	200
optimizer	Adam
iou	0.7
lr0	0.01
lrf	0.01
momentum	0.937
weight_decay	0.0005

D. MODEL PERFORMANCE EVALUATION METRICS

We used evaluation metrics in object detection models are confusion matrix, precision (P), recall (R), average precision (AP), mean average precision (mAP), and model size (MB). The confusion matrix usually has four indexes including True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). P and R are typically used to evaluate the model's ability to predict a specific category. mAP evaluates the model's detection performance over the entire dataset and is calculated as the average of the mean precision over all the categories. FPS is the number of images processed per second by the model and is used to measure the speed of detection. Table 2 provides a summary and brief description of the formulas.

III. RESULT AND ANALYSIS

A. IMPACT OF DIFFERENT ATTENTION MECHANISMS

We added an attention mechanism, which helped the network focus on the key information related to effective targets, extracted regions of interest, and further improved the network's detection capabilities. Various attention mechanisms, including CBAM, SE, EMA, SimAM, and CoTAttention, were individually added.

TABLE 2. Model performance evaluation index.

Evaluation Metrics	Formulas	Content
Precision(P)	$P = \frac{TP}{TP + FP}$	TP represents the number of samples that are actually positive but are predicted to be positive. FP represents the number of samples that are actually negative but are predicted to be positive.
Recall(R)	$R = \frac{TP}{TP + FN}$	FN represents the number of samples that are actually positive but predicted to be negative.
Average Precision (AP)	$AP_i = \int_0^1 P_i R_i dR_i$	AP_i represents the average accuracy of the model's detection for each category.
Mean Average Precision(mAP)	$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$	mAP represents the average of multiple categories of AP.

TABLE 3. Impact of different attention mechanisms.

Model	P (%)	R (%)	mAP (%)	Parameters	FPS
YOLOv8	82.2	87.2	91.8	3006428	166.7
YOLOv8-CBAM	90.4	85.7	93.5	3027318	158.7
YOLOv8-SE	83.2	92.4	92.8	3014620	144.9
YOLOv8-SimAM	82.4	87.2	92.6	3006428	147.1
YOLOv8-CoTAttention	86.5	85.8	92.2	3583452	138.9
YOLOv8-EMA	89.1	88	94.4	3016796	153.9

Table 2 indicated that the addition of EMA led to a noticeable improvement in the network detection performance, with mAP increasing by 2.6%, exceeding the performance enhancements yielded by the SE, CBAM, SimAM, and CoTAttention attention mechanisms. EMA demonstrated superior efficiency in filtering effective feature information, so we adopted EMA to further improve the network's performance.

B. IMPACT OF ATTENTIONAL MECHANISM ADDED IN DIFFERENT LOCATIONS

The adoption of attention mechanisms brought multiple advantages, but the specific advantages depended on the location you add the attention module. We conducted a comparative experiment to evaluate various positions of EMA integration, investigating the impact of adding EMA at different positions on detection performance.

The experiment results are presented in Table 4. In Experiment A, the mAP increased, but the number of parameters also increased. This is because although the C2f_EMA module could better fuse shallow feature maps and deep feature maps, the introduction of the attention mechanism increased the depth of the model.

Experiment B achieved higher accuracy, but recall and mAP were lower. This was due to the addition of EMA to the small object detection layer aiding the model in accurately locating targets but not accurately obtaining the characteristic information of individual maize seeds.

In experiment C, the number of parameters increased, but the network detection performance was greatly improved. This is because adding EMA to the front of the Backbone's SPPF helped the Backbone selectively focus on different parts of the input feature map. This made it easier for the model to learn complex image patterns and improved the accuracy of maize seed quality detection.

TABLE 4. The impact of EMA adding in different locations.

Experiment	P (%)	R (%)	mAP (%)	Parameters	FPS
A	87.2	86.4	92.6	3016796	169.5
B	91.1	83.2	92.4	3007100	158.7
C	89.1	88	94.4	3016796	153.9

The C2f module added to the Backbone, forming the new C2f_EMA module. B. Added to the small object detection layer (P3). C. Added to the front of the Backbone's SPPF.

C. THE IMPACT OF DIFFERENT LOSS FUNCTIONS

In order to verify the superiority of WIoUv3, we conducted comparative experiments on YOLOv8 and I-YOLOv8 using WIoUv3 and some mainstream loss functions.

The experiment results are presented in Table 5. Both models exhibited the highest mAP when using WIoUv3 as the bounding box loss function, this indicated that using WIoUv3 as bounding box regression results in the best detection performance. Furthermore, the improved YOLOv8 achieved a 2.1% higher mAP when using WIoUv3 compared to CIoU, which demonstrated the effectiveness of introducing WIoUv3.

IV. DISCUSSION

A. MODEL PERFORMANCE ANALYSIS

We analyzed the performance of I-YOLOv8 from the training phase to the testing phase and evaluated its ability to detect the quality status of various types of maize seeds.

As depicted in Figure 6, as the iteration proceeded, the loss of I-YOLOv8n decreased and mAP improved correspondingly. After approximately 100 epochs of training, higher mAP and lower loss were achieved. Between 100 and 150 epochs, fluctuations occurred in the training process, this was due to the absence of a corresponding pre-trained model, which led to randomness in the weights during gradient

TABLE 5. The impact of different loss functions.

Model	loss function	P (%)	R (%)	mAP (%)
YOLOv8	DIoU	89.7	88.6	94.2
	GIoU	88.1	88.2	94
	Focal_ CIoU	90.7	88.2	94.5
	Focal_ EIoU	91.1	87.5	94.9
	CIoU	92.2	87.2	91.8
	WIoUv3	92.8	94.9	97.4
I-YOLOv8	DIoU	89.6	91.1	95.9
	GIoU	92.3	89.1	95.2
	Focal_ CIoU	92.3	90.2	97.3
	Focal_ EIoU	88.2	93	95.6
	CIoU	93.7	88.5	96.4
	WIoUv3	96.4	95.4	98.5

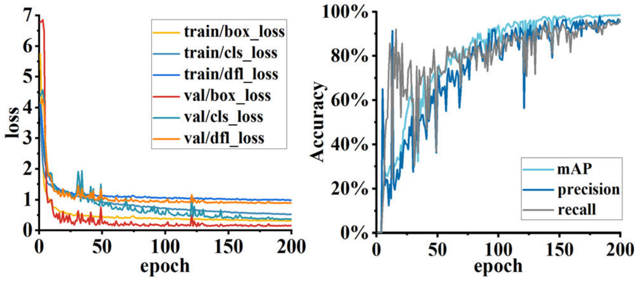

FIGURE 6. Loss and Accuracy Values during Training of I-YOLOv8.

TABLE 6. Performance Comparison between YOLOv8 and I-YOLOv8.

Model	P (%)	R (%)	mAP (%)	Parameters	Model Size (MB)	FPS
YOLOv8	82.2	87.2	91.8	3006428	5.94	166.7
I-YOLOv8	96.4	95.4	98.5	2264728	4.5	169.5

descent and made it challenging to achieve optimal training results. After 150 epochs, the model tended to stabilize with both mAP and loss maintaining a relatively stable state.

It was evident from Table 6 that the improved model exhibited enhanced performance compared to the original YOLOv8. The precision had increased by 14.2%, the recall had improved by 8.2%, and the mAP value had risen by 6.4%.

Figure 7 illustrated the detection results of YOLOv8 and I-YOLOv8 under different backgrounds. The left two images displayed the detection results of YOLOv8, while the right two images showed the results of I-YOLOv8. The red bounding box represented detected broken maize seeds, the pink bounding box represented healthy maize seeds, the orange bounding box represented mildewed maize seeds, and the yellow bounding box represented moth-eaten maize seeds. The confidence scores for each detected maize seed in the images were displayed to the right of the corresponding bounding boxes. In Figure 7(a), the predictions were generally accurate, but it misidentified the infested maize seed in the upper right maizeer as a damaged one. In contrast, Figure7(b) correctly detected all maize seeds without any misidentifications or omissions. In Figure7(c), it misidentified the healthy and damaged maize seeds in the upper left maizeer as infested ones, while Figure 7(d) accurately detected all maize seeds without any misidentifications or omissions.

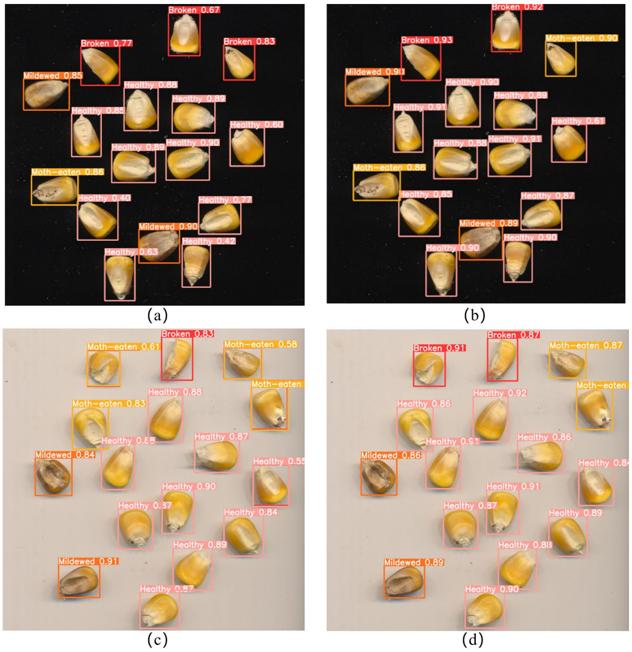

FIGURE 7. Comparison of Detection Results between YOLOv8 and I-YOLOv8. (a) Represents the detection result of YOLOv8 on a black background. (b) represents the detection result of I-YOLOv8 on a black background. (c) represents the detection result of YOLOv8 on a white background. (d) represents the detection result of I-YOLOv8 on a white background.

In summary, YOLOv8 could detect all seeds, which was why it was chosen as the baseline model. However, compared to the unimproved YOLOv8, I-YOLOv8 could better distinguish between damaged and moth-eaten maize seeds, making it more suitable for maize seed quality detection.This was because broken and moth-eaten corn seeds exhibited fundamental similarities in texture and contour edges, with only slight variations in color at the locations of insect holes or broken edges, making it challenging for the model to extract features. I-YOLOv8 introduces an efficient multi-scale attention mechanism (EMA) in the Backbone section, allowing it to focus more on the deep features of the seeds, thereby improving accuracy in identifying damaged and infested corn seeds. Therefore, through targeted improvements to YOLOv8, there was a successful enhancement in its performance in the detection of specific seeds.

B. ANALYSIS OF EXPERIMENT RESULTS UNDER DIFFERENT MODELS

In order to validate the performance of the improved YOLOv8, the enhanced I-YOLOv8 was compared with the original YOLOv8, YOLOv5, and YOLOv5. Precision rate, recall rate, average precision mean, and detection speed were used as performance evaluation metrics.

As can be seen from Figure 8, the accuracy, recall, and mAP values of the improved YOLOv8 were 96.4%, 95.4%, and 98.5% respectively, which surpassed the other three networks in performance. In terms of model detection speed, I-YOLOv8 processes each image in 0.059 seconds (169.5 fps), which was the fastest among the four networks. Additionally, the model size of I-YOLOv8 was only 4.5 MB and the number of parameters was 2,264,728, both of them were far smaller than the other three network models. I-YOLOv8 could still achieve high detection accuracy and speed even in small size. This indicated that the improved network model exhibits superior recognition capabilities for maize seed quality detection compared to other detection models.

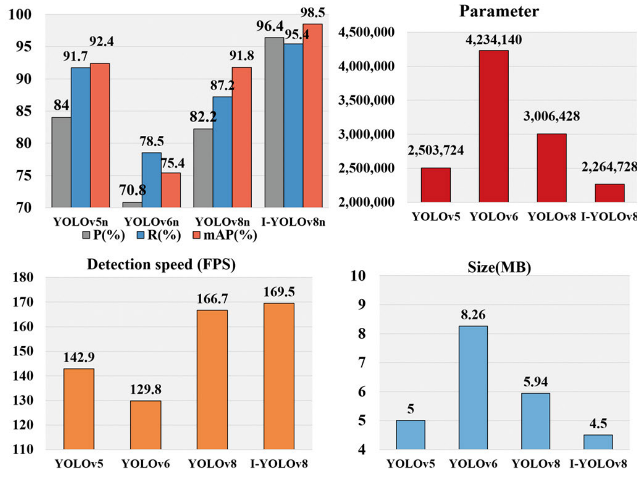

FIGURE 8. Performance comparison of four object detection networks.

In general, increasing the depth and width of a neural network model can improve its performance to a certain extent. However, when performance reaches a certain level, further increasing the depth and width of the network may no longer lead to performance improvement, it may lead to problems such as gradient instability, network degradation, a significant increase in computation complexity and the number of parameters. This study introduced attention mechanisms and SPD-Conv modules into YOLOv8, enhancing the network's performance to some extent but also increasing the number of parameters. Therefore, the study chose to prune the model on this basis, reducing redundancy by eliminating a significant amount of irrelevant semantic information in the model. This is why I-YOLOv8 achieves superior detection speed and network size compared to the other three models while maintaining the highest accuracy.

As depicted in Figure 9, we used a confusion matrix for visual performance evaluation of the four different models. The color of the matrix represented the effectiveness of predictions, the darker the color of the matrix block, the higher the probability of occurrence. The deeper the colors of the blocks along the diagonal of the matrix, the higher the predictive accuracy in this category.

It can be seen from Figure 9 that the color of the diagonal matrix blocks of I-YOLOv8 was darker than that of the other models, and its overall correct recognition rate was higher than other models in the experiment. The correct recognition rates of I-YOLOv8 for healthy and mildewed maize seeds were almost the same as that of other models, but the correct recognition rate in broken and moth-eaten maize seeds was much higher than that of other models. This is because healthy and mildewed maize seeds exhibit noticeable differences in color and texture, while broken and moth-eaten maize seeds are very similar in the contours of the missing parts, with only slight differences in color. In summary, the overall performance of I-YOLOv8 significantly outperforms the other models in the study.

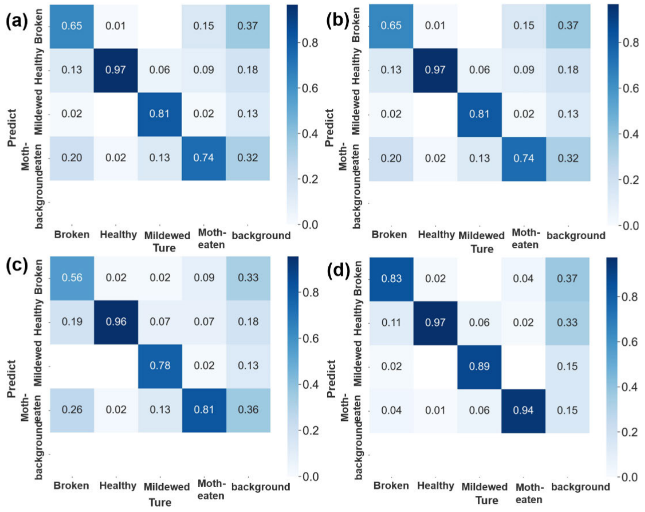

FIGURE 9. Confusion matrix for different models (a)YOLOv5 (b)YOLOv6 (c)YOLOv8 (d)I-YOLOv8.

C. ABLATION STUDY

To validate the effectiveness of each proposed improvement strategy in this study, we designed ablation studies based on the baseline model YOLOv8n to evaluate its effectiveness. The results of the ablation experiments were presented in Table 7, where "$\surd$" indicated the usage of the corresponding module, while its absence indicated the non-usage of the module.

Table 7 demonstrated that each improvement strategy had effectively enhanced the detection performance. Experiment 2 introduced the efficient multi-scale attention mechanism EMA, resulting in a 2.6% increase in mAP, while the number of parameters only increased by 0.34%. This indicated that EMA could efficiently retain information on each channel, and the number of parameters was not significantly increased while ensuring accuracy. Experiment 3 added the SPD-Conv module on the basis of Experiment 2, and the mAP increased by 1.2%, which solved the problem of a loss of

TABLE 7. Detection results after introducing different improvement strategies.

Experiment	EMA	SPD-Conv	P3,p4	WIOU	P(%)	R(%)	mAP(%)	Parameters	FPS
1					82.2	87.2	91.8	3006428	158.7
2	√				89.1	88	94.4	3016796	153.9
3	√	√			91.8	91.1	95.6	3277916	142.9
4	√	√	√		93.7	88.5	96.4	2264728	163.9
5	√	√	√	√	96.4	95.4	98.5	2264728	163.9

fine-grained information caused by the strided convolution or pooling layer in the existing CNN architecture, and could effectively reduce the missed detection rate for small targets.

To mitigate the increased parameters resulting from the addition of attention mechanisms and the SPD-Conv module, a large object detection layer was reduced to decrease a large amount of irrelevant semantic information in the model, which enhanced network speed and reduced network size. Experiment 4 reduced a large target detection layer based on Experiment 3, the number of parameters was reduced by 30.9%, the FPS was increased by 20f/s, and the mAP value was also improved.

Experiment 5 was the improved model proposed in this article, it was based on Experiment 4 and the loss function was replaced with WIoUv3. WIoUv3 was used to weigh the ratio of low-quality samples to high-quality samples, which increased mAP by 1.2% and solved the problem of small target are blurry and difficult to detect. Compared to the baseline model, the number of parameters decreased by 24.7%, mAP increased by 6.7%, recall improved by 8.2%, precision increased by 14.2%, and FPS decreased by 5.2f/s, this indicated that the improved network had excellent performance.

V. CONCLUSION

In response to the characteristics of dense image distribution and small targets in maize seed object detection, we proposed a lightweight maize seed quality detection model based on the improved YOLOv8: I-YOLOv8.

I-YOLOv8 achieved mAP of 98.5%, a 6.7% increase compared to YOLOv8, and an average recognition speed of 163.9 frames per second, a 5.2 frames per second improvement over YOLOv8. Furthermore, when compared to YOLOv5, YOLOv6, and YOLOv8 network models, I-YOLOv8 outperformed these three models in various aspects, and showed a significant improvement in detection performance. The improved model can provide more efficient computing performance and rapid real-time decision-making in agricultural deployment, helping to realize intelligent agricultural management in farm equipment, airborne equipment and edge computing, and improve production efficiency and resource utilization efficiency.

The combination of deep learning and machine vision can achieve non-destructive and efficient identification of corn seeds. The application of deep learning and machine vision in quality detection for maize seeds is expected to bring about significant transformation in agricultural production, promoting the development of a more intelligent, efficient, and sustainable direction for agriculture. In the next steps, we will further optimize the model and increase the variety and quantity of samples to enhance the model's applicability.

ACKNOWLEDGMENT

The authors thank the Shandong Academy of Agricultural Sciences for providing experimental support.

REFERENCES

[1] O. Erenstein, J. Chamberlin, and K. Sonder, "Estimating the global number and distribution of maize and wheat farms," *Global Food Secur.*, vol. 30, Sep. 2021, Art. no. 100558.

[2] F. Guzzon, L. W. A. Rios, G. M. C. Cepeda, M. C. Polo, A. C. Cabrera, J. M. Figueroa, A. E. M. Hoyos, T. W. J. Calvo, T. L. Molnar, L. A. N. León, T. P. N. León, S. L. M. Kerguelén, J. G. O. Rojas, G. Vázquez, R. E. Preciado-Ortiz, J. L. Zambrano, N. P. Rojas, and K. V. Pixley, "Conservation and use of Latin American maize diversity: Pillar of nutrition security and cultural heritage of humanity," *Agronomy*, vol. 11, no. 1, p. 172, Jan. 2021.

[3] K. Tu, S. Wen, Y. Cheng, T. Zhang, T. Pan, J. Wang, J. Wang, and Q. Sun, "A non-destructive and highly efficient model for detecting the genuineness of maize variety 'JINGKE 968' using machine vision combined with deep learning," *Comput. Electron. Agricult.*, vol. 182, Mar. 2021, Art. no. 106002.

[4] M. Kharbach, M. Alaoui Mansouri, M. Taabouz, and H. Yu, "Current application of advancing spectroscopy techniques in food analysis: Data handling with chemometric approaches," *Foods*, vol. 12, no. 14, p. 2753, Jul. 2023.

[5] T. U. Rehman, M. S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, "Current and future applications of statistical machine learning algorithms for agricultural machine vision systems," *Comput. Electron. Agricult.*, vol. 156, pp. 585–605, Jan. 2019.

[6] G. Chen, Y. Meng, J. Lu, and D. Wang, "Research on color and shape recognition of maize diseases based on HSV and OTSU method," in *Proc. Int. Conf. Comput. Comput. Technol. Agricult.*, vol. 509. Cham, Switzerland: Springer, 2016, pp. 298–309.

[7] K. Kiratiratanapruk and W. Sinthupinyo, "Color and texture for corn seed classification by machine vision," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Dec. 2011, pp. 1–5.

[8] B. B. Traore, B. Kamsu-Foguem, and F. Tangara, "Deep convolution neural network for image recognition," *Ecol. Informat.*, vol. 48, pp. 257–268, Nov. 2018.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[10] K. He, X. Zhang, S. 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, Sep. 2015.

[11] 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.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[13] C. Zhao, L. Quan, H. Li, R. Liu, J. Wang, H. Feng, Q. Wang, and K. Sin, "Precise selection and visualization of maize kernels based on electromagnetic vibration and deep learning," *Trans. ASABE*, vol. 63, no. 3, pp. 629–643, 2020.

[14] H. O. Velesaca, R. Mira, P. L. Suárez, C. X. Larrea, and A. D. Sappa, "Deep learning based corn kernel classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 294–302.

[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.*, Oct. 2016, pp. 21–37.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[18] Z. Liu and S. Wang, "Broken corn detection based on an adjusted YOLO with focal loss," *IEEE Access*, vol. 7, pp. 68281–68289, 2019.

[19] X. Li, Y. Du, L. Yao, J. Wu, and L. Liu, "Design and experiment of a broken corn kernel detection device based on the YOLOv4-tiny algorithm," *Agriculture*, vol. 11, no. 12, p. 1238, Dec. 2021.

[20] R. P. Thangaraj Sundaramurthy, Y. Balasubramanian, and M. Annamalai, "Real-time detection of fusarium infection in moving corn grains using YOLOv5 object detection algorithm," *J. Food Process Eng.*, vol. 46, no. 9, Jun. 2023, Art. no. e14401.

[21] Q. Wang, H. Yang, Q. He, D. Yue, C. Zhang, and D. Geng, "Real-time detection system of broken corn kernels based on BCK-YOLOv7," *Agronomy*, vol. 13, no. 7, p. 1750, Jun. 2023.

[22] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[23] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, 2023, pp. 443–459.

[24] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.

[25] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000.

SIQI NIU was born in 2000. She is currently pursuing the master's degree in agricultural engineering and information technology with Qingdao Agricultural University.

XIAOLIN XU was born in 1999. He is currently pursuing the master's degree in agricultural engineering and information technology with Qingdao Agricultural University.

AO LIANG was born in 1999. He is currently pursuing the master's degree in agricultural engineering and information technology with Qingdao Agricultural University.

YULIANG YUN received the Ph.D. degree in agricultural engineering from China Agricultural University, China. He is currently an Associate Professor with the School of Mechanical and Electrical Engineering, Qingdao Agricultural University, China. His current research interests include agricultural artificial intelligence and decision support systems.

LI LI received the Ph.D. degree in agricultural electrification and automation from China Agricultural University, Beijing, China, in 2007. She is currently an Associate Professor with the Information and Electrical Engineering College, China Agricultural University. Her current research interests include agricultural artificial intelligence and greenhouse environmental regulation.

FENGQI HAO received the master's degree in computer system architecture from Shandong University, China, in 2006. He is currently an Associate Researcher with the Qilu University of Technology, Jinan, China. His current research interests include intelligent control, computer vision, and deep learning.

JINQIANG BAI received the Ph.D. degree from Beihang University, Beijing, China, in 2020. He has been a Research Assistant with the Qilu University of Technology, Jinan, China, since 2020. His current research interests include computer vision, deep learning, and intelligent agriculture.

DEXIN MA received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014. He is currently a Professor with the Intelligent Agriculture Institute, Qingdao Agricultural University, China. His current research interests include agricultural artificial intelligence and agricultural informatization.