Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 3, 2025

Improvement and Implementation of a Potato Recognition Algorithm Based on YOLOv8

Dongxuan Huang, Mingge Sun *, Sen Ye, Jiaxuan Chai

School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, Jilin 132022, China

*** Corresponding author:** Mingge Sun

Abstract: To address issues such as poor recognition accuracy and low efficiency caused by the large number of potato targets that need to be processed at once during potato identification, this paper proposes an improved YOLOv8-based potato recognition algorithm. The improvements include: introducing DynamicConv in the backbone network to enhance model representation capability; replacing the C2f module in the backbone network with C2f_SAConv to strengthen feature extraction in object detection and segmentation tasks; incorporating Slim-neck in the Neck layer to reduce model parameters and floating-point operations; and adding an EMA attention mechanism to improve the model's focus on different potatoes. The improved YOLOv8 algorithm achieves an average recognition accuracy of 89.7%, reduces computational load by 27.2%, and facilitates deployment on resource-constrained embedded devices.

Keywords: Deep Learning; YOLOv8; Attention Mechanism; Potato Recognition.

1. Introduction

Potatoes have now become China's fourth largest staple crop after corn, rice, and wheat, serving as an important food resource. China boasts vast potato planting areas, particularly in Northeast China, North China, and Southwest China [1-2]. Potatoes serve multiple purposes: they can be cooked as vegetables, consumed as staple food, or used as industrial raw materials for manufacturing products, making them a versatile grain resource [3]. Currently, potato classification primarily relies on manual sorting, which suffers from low efficiency, high labor intensity, and significant variability between different workers. Traditional manual sorting can no longer meet the processing requirements for subsequent potato production. Therefore, research on automatic potato sorting robots holds significant practical importance.

In potato classification, operations are achieved by extracting feature information such as size, color, and texture of different potato varieties [4-5]. With advancements in science and technology, numerous machine vision-based classification methods for various items have emerged. Al-Kateb G et al. enhanced potato disease management using generative models and convolutional neural networks (CNNs). Traditional disease identification methods achieved over 95% success in fully accurate detection and 85% success in early-stage detection [6]. Liu Yijun et al. developed a potato sprouting and surface damage detection model based on the Faster-RCNN neural network. They improved the model using a feature extraction network combining ResNet50 residual networks and multi-scale fusion of pyramid pooling networks, enabling effective detection and evaluation of sprout size and surface damage severity [7]. Akther J et al. created a CNN model for real-time detection of early and late blight in potatoes. The method was evaluated using classification optimizers, metrics, and loss functions, with further optimization through layer-wise TensorBoard analysis. The final model achieved 96.09% accuracy on the investigated dataset [8]. Zhang W et al. proposed a seed potato bud recognition method based on an improved YOLOv3-tiny. By introducing the CIoU bounding box regression loss function to enhance regression performance, the model achieved precision, recall, average precision, and F1 scores of 88.33%, 85.97%, 91.18%, and 87.13%, respectively, for bud recognition [9].

This study proposes an improved YOLOv8 algorithm to further enhance the recognition accuracy for different potato varieties. The algorithm primarily optimizes the Backbone and Neck modules: DynamicConv is introduced into the backbone network, replacing traditional static convolution with dynamic convolution to strengthen the model's expressive capabilities; Slim-neck is incorporated into the Neck layer to reduce model parameters and computational complexity, broadening its applicability; and the EMA attention mechanism is integrated to improve feature extraction for diverse characteristics. Through these enhancements, the algorithm achieves reduced computational demands while significantly improving potato recognition accuracy.

2. Introduction to YOLOv8

YOLOv8 was introduced by Ultralytics in early 2023 and can perform object detection, segmentation, and classification tasks by selecting appropriate weight files for different scenarios. The network architecture of YOLOv8, as shown in Figure 1, consists of five main components: Input, Backbone, Neck (feature fusion network), Head (detection head), and Output. The Backbone of YOLOv8 adopts the ELAN concept from YOLOv7 [10], serving to provide computational power and feature extraction. It comprises CBS, C2f, and SPPF modules. The C2f module enhances gradient flow diversity through residual connections, optimizing network training. For the loss function, YOLOv8 replaces traditional IOU matching or unilateral ratio allocation with the TaskAlignedAssigner positive sample allocation strategy—a dynamic approach that selects positive samples based on weighted scores combining classification and regression metrics.

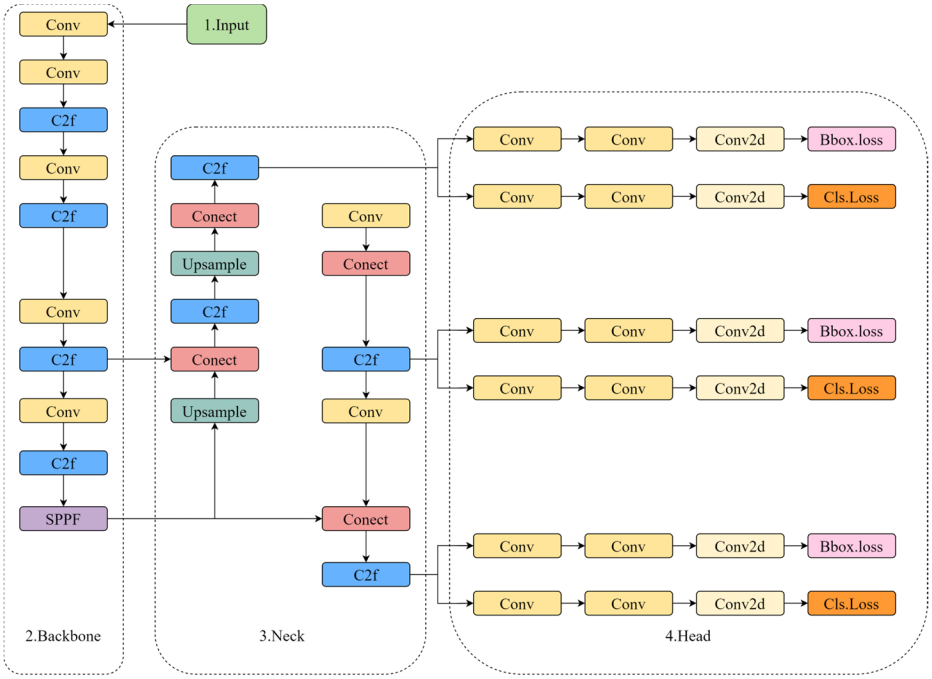

Figure 1. YOLOv8 Network Structure

3. Improved YOLOv8

3.1. DynamicConv

Traditional 3×3 convolution kernels have fixed shapes and sizes, which suffer from drawbacks such as large parameter counts, poor scale invariance, and limited local receptive fields during convolution operations [11]. To address these limitations, the DynamicConv module is constructed by integrating dynamic convolution (DynamicConv) into the Backbone network. DynamicConv dynamically selects or combines different convolution kernels based on input samples, enabling the network structure to adaptively adjust its parameters according to varying input data. By employing multiple convolution kernels within a single convolutional layer and integrating attention mechanisms to fuse information from different kernels, DynamicConv enriches feature extraction without significantly increasing computational complexity.

As shown in Figure 2, DynamicConv sets K convolution kernels with the same scale and input data channel dimensions at specific layers. These kernels are fused through their respective attention weights π_k to obtain the convolution kernel parameters for that layer. The π_k is obtained through the attention mechanism based on the input data x. First, global average pooling (avg_pool) is performed to obtain global segmentation features. Then, two FC layers map these features to K-dimensional space, followed by softmax normalization to allocate the obtained K attention weights to the K convolution kernels of this layer. By multiplying the obtained weight coefficients with their corresponding biases, the data information at that convolution position can be obtained. The summation of all convolution weights is then computed, and the matrix multiplication of this summation result with x completes the DynamicConv operation. The dynamic perception model of DynamicConv is shown in Equation (1).

$$y = g\left(\widetilde{W}(x)^T x + \tilde{b}(x)\right), \tag{1}$$

The weights and biases required by the dynamic perception model can be derived from Equations (2) and (3).

$$\widetilde{W}(x) = \sum_{k=1}^{K} \pi_k(x)\widetilde{W}_k, \tag{2}$$

$$\tilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\tilde{b}_k, \tag{3}$$

$$s.t. 0 \le \pi_k(x) \le 1, \sum_{k=1}^{K} \pi_k(x) = 1, \tag{4}$$

Equation (4) represents the constraint on the weight coefficients. The weight coefficients are not fixed and can be adaptively selected based on the input data, thereby enabling the model to possess stronger feature representation capabilities.

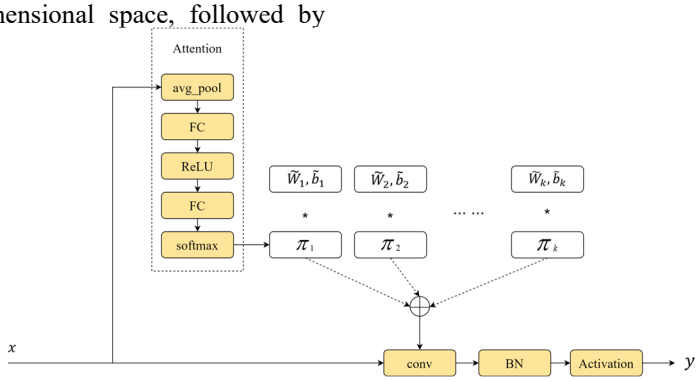

Figure 2. DynamicConv Structure

3.2. SAC

In current object detection research, it has been demonstrated that mechanisms employing dual observation and reasoning can exhibit superior performance. SAC (Switchable Atrous Convolution) integrates this dual observation and reasoning mechanism into the backbone network of object detection systems, deploying this mechanism at both macro and micro levels. This approach significantly enhances detector performance without substantially increasing model parameters or weight size, thereby maintaining inference speed.

The core concept of SAC lies in applying convolution operations with different dilation rates to the same input features. Dilated convolution expands the receptive field by introducing additional spaces (dilation) in the convolution kernel without increasing parameters or computational load. SAC leverages this characteristic to capture multi-scale features. By utilizing identical weights for convolutions with different dilation rates, SAC can convert traditional convolution layers into SAC layers, enabling soft-switching between different dilation rates during convolution operations. Although SAC switches between different dilation rates, all operations share the same weights with only one trainable variation. This design reduces model complexity while preserving flexibility. SAC also incorporates two global context modules that augment input features with image-level information. This operation helps the network better understand and process the holistic content of images, thereby improving the quality and accuracy of feature extraction. Figure 3 illustrates the specific implementation of SAC.

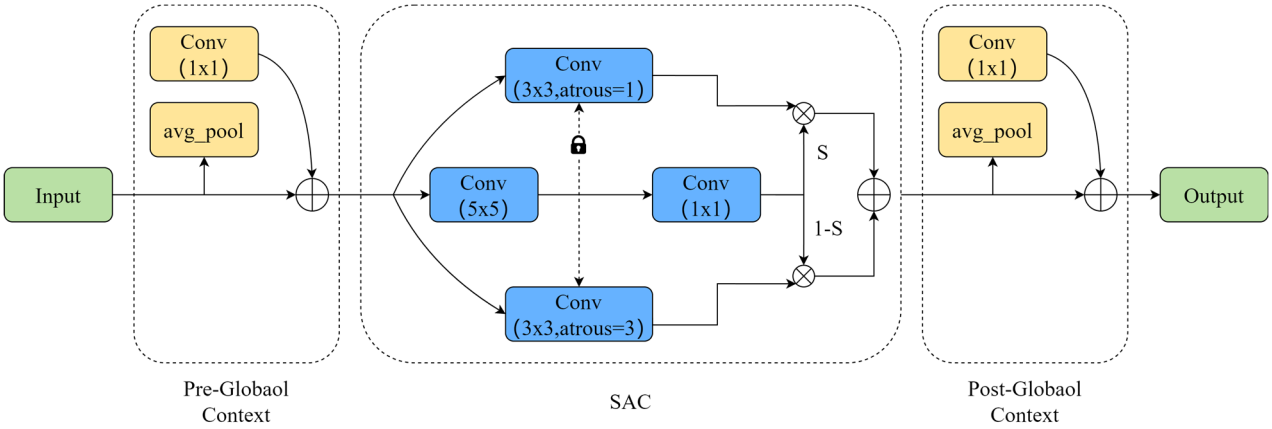

Figure 3. Detailed Architecture of SAC Transformation

3.3. Feature Fusion Network

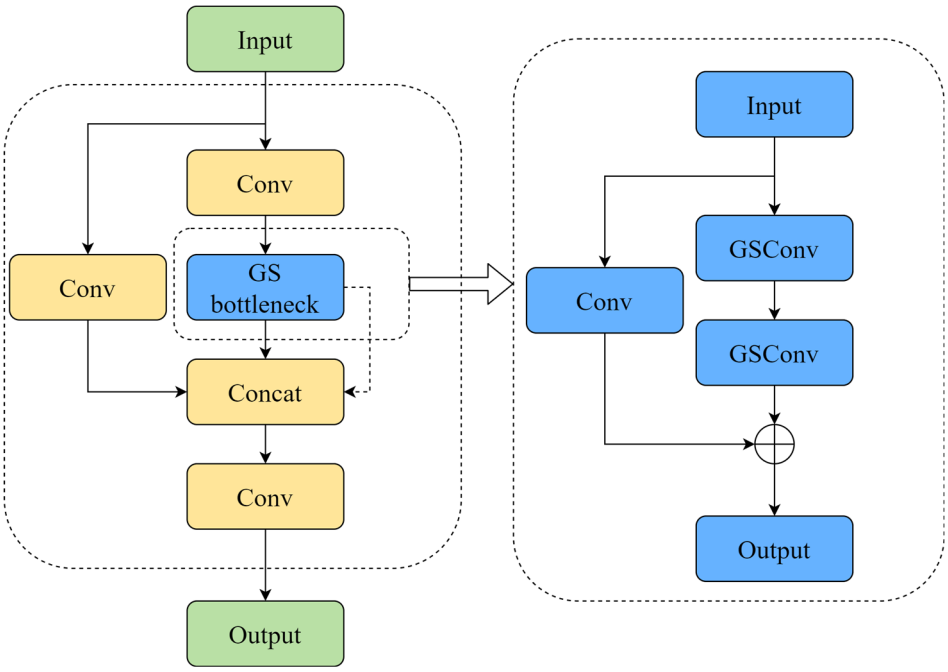

Figure 4. VOV-GSCSP module

The Neck is the component that connects the Backbone and Head, responsible for processing and fusing feature maps of different scales. In potato recognition scenarios with dense occlusions, the YOLOv8 algorithm may lead to missed detections. Therefore, the Slim-neck structure is introduced to improve the Neck network of YOLOv8. Slim-neck is constructed using modern lightweight depthwise separable convolution (GSConv), GS bottlenecks, and a cross-stage partial network module (VOV-GSCSP). Traditional convolutional neural networks convert spatial information into channel information, but this process inevitably leads to partial loss of semantic information during each spatial

compression and channel expansion of feature maps. GSConv independently performs convolution on each channel of the input feature map, concatenates the convolution results, and then applies a shuffle operation to rearrange feature channels, thereby enhancing information flow between features. The GS bottleneck built upon GSConv improves nonlinear feature expression and information reuse. The VOV-GSCSP module, designed with a one-time aggregation method, enables effective information fusion between feature maps at different stages. This approach reduces computational complexity while maintaining sufficient accuracy, which is particularly beneficial for resource-constrained environments. These modular designs embody the Slim-neck philosophy, aiming to reduce computational complexity and inference time without compromising accuracy. Through such modular design, flexible network architectures can be constructed to suit specific task requirements. The structure of the VOV-GSCSP framework is illustrated in Figure 4.

3.4. Attention Module

The Efficient Multi-scale Attention (EMA) module is a novel attention mechanism that reshapes partial channel and batch dimensions while grouping channel dimensions into multiple sub-features. This approach preserves critical channel information, reduces computational overhead, and enhances the model's feature processing capability. By encoding global information, the EMA module recalibrates channel weights in each parallel branch and captures pixel-level relationships through cross-dimensional interactions, thereby strengthening feature representation. As shown in Figure 5, G represents the number of groups into which input channels are divided, while X_avg_pool and Y_avg_pool denote 1D average pooling along horizontal and vertical directions, respectively. The EMA partitions multi-scale spatial information into three parallel pathways: two pathways in the 1×1 branch utilize 1D global average pooling operations along horizontal and vertical spatial directions for channel encoding, while the third pathway in the 3×3 branch employs stacked 3×3 kernels to capture multi-scale feature representations. Output features from both branches undergo sigmoid activation and normalization, then merge through a cross-dimensional interaction module to establish pixel-level pairwise relationships. After final sigmoid modulation, the enhanced/attenuated feature maps are combined with original inputs to produce the ultimate output.

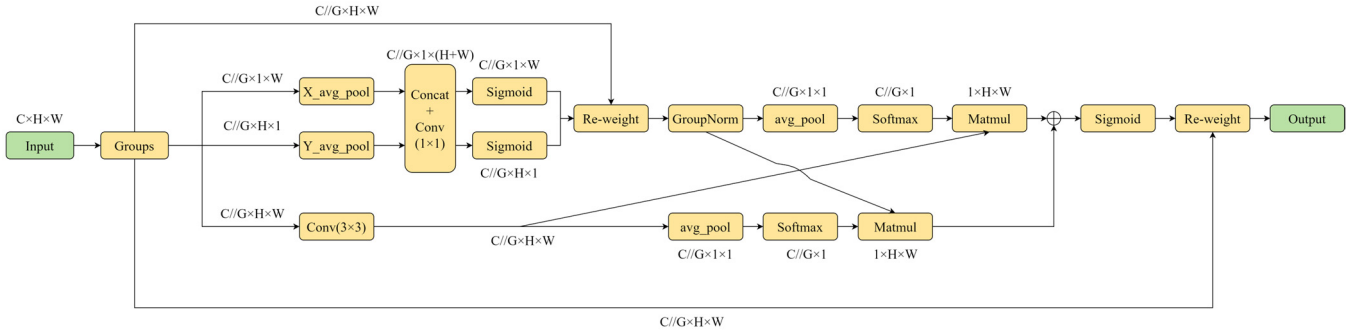

Figure 5. EMA Architecture Diagram

4. Experiments and Results Analysis

4.1. Experimental Environment

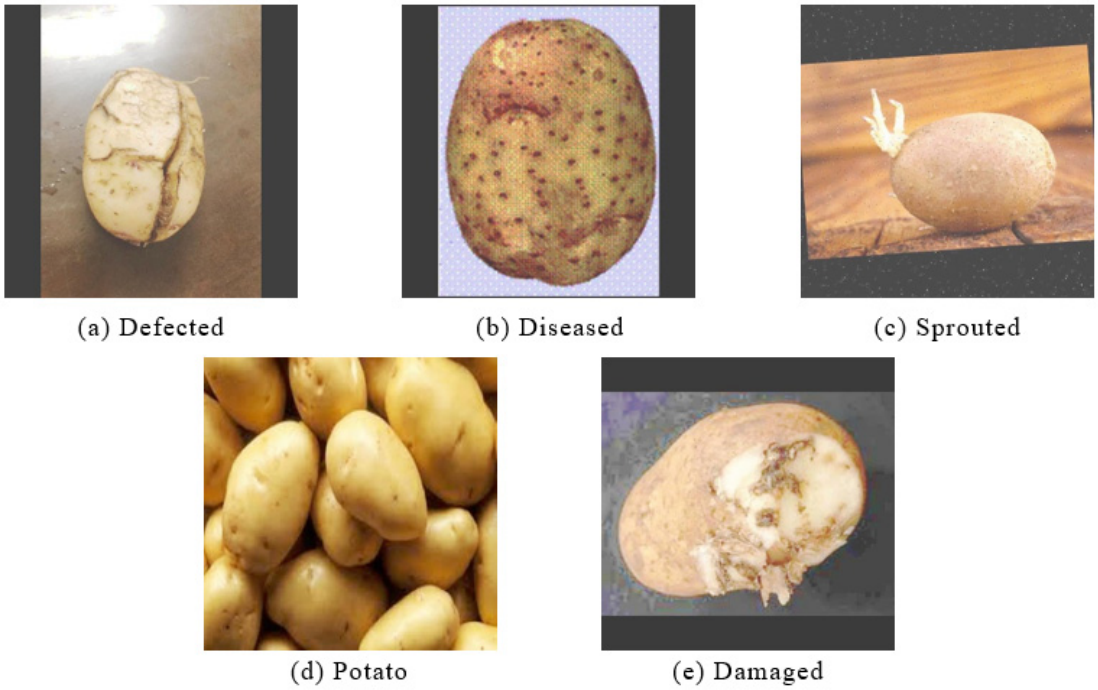

Figure 6. Image examples of five categories

The equipment operating system used in this experiment was Windows 11. The model framework utilized the PyTorch 2.2.1 deep learning framework. The central processing unit (CPU) was a 12th Gen Intel(R) Core (TM) i7-12700H, with an NVIDIA GeForce RTX 3060 graphics card providing 6GB of video memory. The device was equipped with 16GB of RAM, running on program compiler version 2023 with Python 3.11 and CUDA version 12.1.

4.2. Dataset

The dataset used in this experiment is potato-eruvaka, with example images shown in Figure 6. This dataset contains 8,003 images of five potato categories: Normal Potato, Damaged Potato, Defected Potato, Diseased Potato, and Sprouted Potato. Each image has a resolution of 640×640 pixels. The dataset is divided into training, validation, and test sets with an 8:1:1 ratio of image quantities.

4.3. Evaluation Metrics

This experiment employs precision (P), recall (R), mean average precision (mAP), computational complexity (GFLOPs), and frames per second (FPS) as evaluation metrics. Precision represents the proportion of correctly predicted positive samples among all predicted positive samples; recall denotes the proportion of correctly predicted positive samples among all actual positive samples; mean average precision (mAP) is obtained by averaging the results of integrating the precision-recall (P-R) curves for each category. The equations for precision, recall, and mean average precision are shown in formulas (5) to (7).

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$mAP = \frac{\sum_{i=1}^{c} \int_0^1 P(R)\mathrm{d}\,R}{c} \tag{7}$$

TP represents the number of correctly classified positive samples, FP represents the number of correctly classified negative samples, and FN represents the number of incorrectly classified positive samples.

4.4. Ablation Experiment

As can be seen from Table 1, the mAP value of the unimproved YOLOv8 algorithm is 86.4%, while the improved YOLOv8 algorithm achieves a mAP of 89.7%, representing an increase of 3.3 percentage points. The average precision after implementing different module improvements reaches 89.1%, 88.8%, and 89.5% respectively, which are 2.7%, 2.4%, and 3.1% higher than the original YOLOv8 algorithm. Improvement 4 demonstrates the best performance in detecting Potato, Damaged, Diseased, and Sprouted categories, proving that our algorithm enhancements can effectively recognize different potato types. After implementing DynamicConv and SAC module improvements, the model's average precision shows consistent improvement. Although Improvement 2 exhibits a slight decrease in average precision compared to Improvement 1, it successfully reduces model size, decreases computational requirements, and enhances detection speed. Improvement 3 further elevates average precision while continuing to reduce computational costs and model size. In Table 1, Potato, Damaged, Defected, Diseased, and Sprouted represent the precision values corresponding to each respective category.

Table 1. Ablation Experiment Results

Project	DynamicConv	SAC	VOV-GSCSP	EMA	AP/%	
					Potato	Damaged
YOLOv8	×	×	×	×	90.0	80.0
Project 1	√	×	×	×	94.1	88.0
Project 2	√	√	×	×	88.8	88.3
Project 3	√	√	√	×	94.3	88.4
Project 4	√	√	√	√	94.8	89.8

AP/%			*P*/%	*R*/%	*mAP*/%	GFLOPs/10^9
Defected	Diseased	Sprouted				
85.0	87.6	89.5	80.3	77.5	86.4	8.1
82.7	89.1	91.8	82.2	82.9	89.1	7.1
82.0	88.3	91.1	82.2	83.4	88.8	6.4
84.6	89.0	91.2	82.5	89.5	89.5	5.8
82.5	89.5	91.9	82.2	89.7	89.7	5.9

4.5. Controlled Experiment

To further validate the effectiveness of the improved algorithm in enhancing performance, we compared it with SSD [12], Faster-RCNN [13], YOLOv5[14], and YOLOv10 [15], with the comparison results shown in Table 2.

The comparative experiments demonstrate that the improved algorithm outperforms others in terms of recognition accuracy, mean average precision (mAP), and computational complexity across all categories. The complexity of the Faster-RCNN algorithm significantly exceeds that of other methods, while its mAP and FPS show no clear advantages, indicating poor performance on our dataset. Compared to the SSD algorithm, the improved algorithm achieves a 16% increase in mAP and a 90.6% reduction in computational load. Although YOLOv5 and YOLOv10, as members of the YOLOv8 series, exhibit comparable performance to the improved algorithm in most aspects, the proposed method still shows improvements: mAP increases by 1.9% and 2.3%, respectively, while computational load decreases by 16.9% and 28.1%. These results confirm that our improved algorithm demonstrates strong potato recognition capabilities. Despite a slight reduction in FPS, its detection accuracy and computational efficiency still meet practical requirements, making it more suitable for deployment on resource-constrained embedded devices.

Table 2. Comparative Experiment Results

Project	*AP*/%					*mAP*/%	GFLOPs/10^9	FPS/ms
	Potato	Damaged	Defected	Diseased	Sprouted			
SSD	79.9	74.3	72.6	63.1	78.4	73.7	62.7	134
Faster-RCNN	82.6	75.3	71.5	61.2	82.3	74.6	369.8	72
YOLOv5	94.2	88.0	78.4	88.8	89.6	87.8	7.1	158
YOLOv10	91.5	88.1	81.2	88.7	87.4	87.4	8.2	119
Improved YOLOv8	94.8	89.8	82.5	89.5	91.9	89.7	5.9	61

5. Conclusion

The improved YOLOv8 algorithm was subjected to various comparative experiments and ablation experiments. The enhanced algorithm achieved a detection accuracy of 89.7%, representing a 3.3% improvement over the original algorithm while reducing instances of missed and false detections. To validate the robustness and generalizability of the improved algorithm, multiple experiments were conducted on a potato dataset. The modified algorithm demonstrated enhanced detection sensitivity and accuracy in potato classification tasks, with a 27.2% reduction in computational load compared to the original algorithm, making it more suitable for deployment on resource-constrained embedded devices.

However, the improved algorithm still exhibits limitations in detection speed. Future research will focus on employing more flexible backbone networks and image preprocessing techniques to refine the algorithm. The goal is to optimize it into a more lightweight and faster detection model while maintaining the current level of detection accuracy.

References

[1] Tang Jianzhao, Wang Jing, Xiao Dengpan, et al. Research Progress and Development Prospects of Potato Growth Models [J]. Scientia Agricultura Sinica,2021,54(05):921-932.

[2] Hao Ruoshi, Lyu Jianfei, Wang Aoxue, et al. New Characteristics and Development Trends of China's Potato Industry Under the Background of High-Quality Development [J]. Agricultural Outlook,2024,20(01):7-12.

[3] Li Zhonghui, Zhang Haocheng, Wang Xiuli. Current Status and Development Prospects of Potato Supply and Demand in China [J]. Agricultural Outlook,2022, 18(06): 79-85.

[4] Liang Jurong, Huang Biao, Tian Haijiang, et al. Research Status of Automatic Potato Grading Machines [J]. Hans Journal of Agricultural Sciences, 2021, 11: 138.

[5] Li Qi, Wang Jun. Potato Grading Method Based on Fusion of Color and Texture Features[J]. Science Technology and Engineering, 2019, 19(25): 273-279.

[6] Al-Kateb G, Mijwil M M, Aljanabi M, et al. AI-PotatoGuard: Leveraging Generative Models for Early Detection of Potato Diseases [J]. Potato Research, 2024: 1-15.

[7] Liu Yijun. Research on Potato External Quality Grading Method Based on Machine Vision [D]. Chinese Academy of Agricultural Mechanization Sciences, DOI: 10.27629/d. cnki. gznjk. 2023.000009.

[8] Akther J, Nayan A A, Harun-Or-Roshid M. Potato leaves blight disease recognition and categorization using deep learning[J]. Engineering Journal, 2023, 27(9): 27-38.

[9] ZHANG W, Yuelin H A N, Huang C, et al. Recognition method for seed potato buds based on improved YOLOv3-tiny[J]. INMATEH-Agricultural Engineering, 2022, 67(2).

[10] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[11] C G S, Upadhyay A, Zhang Y, et al.Field-based multispecies weed and crop detection using ground robots and advanced YOLO models: A data and model-centric approach[J].Smart Agricultural Technology, 2024, 9100538-100538.

[12] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[13] Shih K H, Chiu C T, Lin J A, et al. Real-time object detection with reduced region proposal network via multi-feature concatenation[J]. IEEE transactions on neural networks and learning systems, 2019, 31(6): 2164-2173.

[14] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.

[15] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. arXiv preprint arXiv:2405. 14458, 2024.