Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 15, No. 3, 2025

Feature Recognition and Modeling Analysis of Apple Images Based on Yolov8X-Seg Training Model

Haimei Lu[1], Yuxin Zhang[2], Guowei Zhao[3], Qiqi Liang[2], Bowen Li[4]

[1]College of Life Sciences, North China University of Science and Technology, Tangshan, 063210, China
[2]College of Science, North China University of Science and Technology, Tangshan, 063210, China
[3]Faculty of Science, North China University of Science and Technology, Tangshan, 063210, China
[4]College of Life Sciences, North China University of Science and Technology, Tangshan, 063210, China

Abstract: To address the inefficiencies and high costs inherent in traditional apple harvesting methods that rely on manual labor, this study aims to develop an automated detection system capable of accurate apple quantity recognition and maturity assessment under complex orchard conditions. By integrating advanced instance segmentation with agricultural automation technology, we seek to establish a foundational framework for intelligent harvesting robots. The proposed methodology leverages the Yolov8X-seg model enhanced through image sharpening and median filtering preprocessing, which optimizes edge feature extraction while suppressing environmental noise. The Adam gradient descent algorithm is systematically applied to refine model parameters, enabling multi-scale feature capture through convolutional-pooling layer combinations and precise classification via fully connected layers. Experimental validations demonstrate that our optimized framework achieves a 12.7% improvement in detection accuracy and 28.4% faster inference speed compared to baseline models, effectively overcoming occlusion and overlapping fruit challenges. These advancements not only verify the model's capability in maturity differentiation through spectral analysis but also reveal its potential for real-time monitoring applications. The research outcomes provide critical technical support for intelligent orchard management systems, marking a significant step toward reducing agricultural labor dependency and advancing precision farming practices.

Keywords: Image recognition, Median filtering, Yolov8X-seg, Adam gradient descent.

1. Introduction

As the world's largest producer and exporter of apples, China's apple industry faces critical challenges in traditional harvesting practices. Labor shortages have driven picking costs to exceed 40% of total expenses, compounded by safety risks associated with elevated work environments. While machine vision-based robotic systems offer a promising solution, their effectiveness is severely hampered by the complexity of orchard scenes. Key obstacles include branch and leaf occlusions (over 30% blind spots), minimal color contrast between apples and backgrounds (merely 15% HSV difference), and dynamic lighting variations, which collectively degrade traditional image processing methods to below 80% accuracy [1]. Further complications arise from adverse weather conditions (e.g., rain, haze), reflective bag interference, and significant fruit size variations (40–90 mm diameter), rendering existing algorithms unstable in dynamic environments. To address these challenges, advancing deep learning algorithms with multispectral imaging and 3D reconstruction, while establishing standardized multi-region image databases, is imperative. Equally critical is developing lightweight models that balance >95% recognition accuracy with real-time processing capabilities, thereby accelerating the adoption of automated harvesting systems.

Existing studies on apple detection predominantly rely on deep learning frameworks, yet fundamental limitations persist. For instance, lightweight segmentation networks like Fast-SCNN prioritize computational efficiency through multi-scale feature fusion [2]. However, their fixed receptive fields and hierarchical context extraction mechanisms struggle to adapt to diverse apple sizes and cluttered backgrounds, resulting in poor robustness. This structural rigidity, coupled with inadequate noise suppression strategies, exacerbates misdetections under hazy or rainy conditions. Meanwhile, occlusion-optimized approaches such as Mask R-CNN variants achieve 82.3% accuracy in occlusion scenarios but suffer from computationally intensive two-stage architectures, limiting inference speeds to 15 FPS—insufficient for real-time orchard operations [2,3]. These shortcomings stem from three core issues: (1) inflexible feature learning due to static convolutional layers, which fail to capture multi-scale spatial relationships; (2) noise amplification from insufficient preprocessing, particularly under environmental interference; and (3) error accumulation in dense target localization caused by overlapping apples and weak feature discrimination.

To overcome these limitations, this study proposes an enhanced Yolov8X-seg framework with three key innovations. First, a hybrid preprocessing module integrates sharpening and median filtering to amplify edge features while suppressing noise (35% reduction in interference). Second, dynamic multi-scale feature fusion is achieved through optimized pooling strategies and Adam-driven parameter tuning (learning rate: 0.001, L2 regularization: 0.0001), enabling adaptive learning across varying apple sizes. Third, a lightweight instance segmentation head reduces computational overhead by 40% compared to conventional YOLO architectures, addressing real-time constraints without sacrificing accuracy. Experimental validation on 200 real orchard images demonstrates 95.2% detection accuracy (15% higher than traditional methods) and 28 FPS inference speed on embedded hardware, effectively resolving challenges in dense target counting, occlusion handling, and environmental adaptability. This work bridges the gap between theoretical research and practical deployment, offering a robust foundation for intelligent apple harvesting systems.

2. Research Method

2.1. Data collection

In this study, 200 images of harvested apples were collected by ourselves. The collection site was Shengguoyuan Ecological Orchard in Sujiadian Town, Qixia City, Yantai City, Shandong Province, and the images were captured with a Nikon D850 digital single-lens reflex (DSLR) camera in a natural environment during the apple harvest season for image preprocessing, feature extraction, and apple counting. These images were taken in natural environment, covering different angles and lighting conditions, which can reflect the real state of apples in the orchard. The images were processed by the Yolov8X-seg model, which was pre-trained on the COCO (https://image-net.org/) detection dataset (image resolution of 640), COCO segmentation dataset (image resolution of 640) and ImageNet (https://image-net.org/) dataset (image resolution of 224). The model is pre-trained on COCO detection dataset (image resolution of 640), COCO segmentation dataset (image resolution of 640) and ImageNet dataset (image resolution of 224) to accurately recognize apples, thus obtaining information on the number of apples and drawing a histogram of apple distribution.

2.2. Data preprocessing

Images captured in natural environments often contain significant noise, which can obscure partial feature information and reduce the contrast between apples and the background. This makes it challenging for deep learning models to extract meaningful features. Therefore, prior to formal image training, preprocessing is essential to mitigate the impact of noise on the model, accelerate its convergence speed, and enhance overall performance. As shown in Figure 1, in this study, image sharpening followed by median filtering is employed as the preliminary preprocessing step to refine the image quality[4].

Original **Sharpen** **Median Filter**

Figure 1. Image Pre-processing

1) Image Sharpening

In this paper, in order to enhance the edges and contours of the image to be able to identify the apples better, image sharpening is achieved by increasing the difference of pixels between neighbors. The essence of image sharpening is high pass filtering, which is opposite to low pass filtering, which blurs the image, whereas sharpening filter increases the difference of pixels between neighbors by using the differentiation of neighbors as an operator to make the mutated parts of the image more visible, thus improving the recognition rate of effective features of the image.

2) Median value filtering

In this paper, median filtering is chosen to suppress the noise in the image and reduce the effect of background on the model training. In this study, the primary objective is to accurately recognize apple images while preserving essential image details and effectively eliminating noise. The median filter removes the noise and avoids destroying the details of the image. Compared with the median filter, the mean filter, although simple and easy to smooth the noise of the image, may make the image blurred and can not remove the noise points well. Gaussian filter can be adjusted by the size of the standard deviation, can achieve different degrees of smoothing and denoising effect but in the processing of "details" or "edges" of the image, will blur the edges. Therefore, compared with the mean filter and Gaussian filter, the median filter may be more suitable for the apple image recognition in this paper.

2.3. Yolov8X-seg training model

The training process of the Yolov8X-seg model consists of two phases: pre-training and fine-tuning. The pre-training phase uses a large-scale image dataset for unsupervised training to learn a generic feature representation. The model is bundled with the following pre-trained models: an object detection checkpoint trained on the COCO detection dataset with an image resolution of 640, an instance segmentation checkpoint trained on the COCO segmentation dataset with an image resolution of 640, and an image classification model pre-trained on the ImageNet dataset with an image resolution of 224. Specifically, the Yolov8X-seg model has five pre-trained models for detection, segmentation, and classification tasks. Among them, YOLOv8 Nano is the fastest and smallest model, while YOLOv8 Extra large (Yolov8X-seg) is the most accurate but slowest model. These models can be selected for different application scenarios. The fine-tuning phase then uses labeled target detection datasets to supervise the training of the models so that they can be better adapted to the eye detection task.

The Yolov8X-seg model was selected for this study due to its exceptional capability in image processing, enabling rapid and accurate extraction of apple features. This model facilitates efficient and precise statistical analysis of apple quantities, making it an ideal choice for achieving robust and reliable results in this context. Its basic idea is to process and feature extract the input data through components such as multi-layer convolutional layers, pooling layers and fully connected layers, and then count the number of apples in each image. The Yolov8X-seg model flowchart is shown in Figure 2.

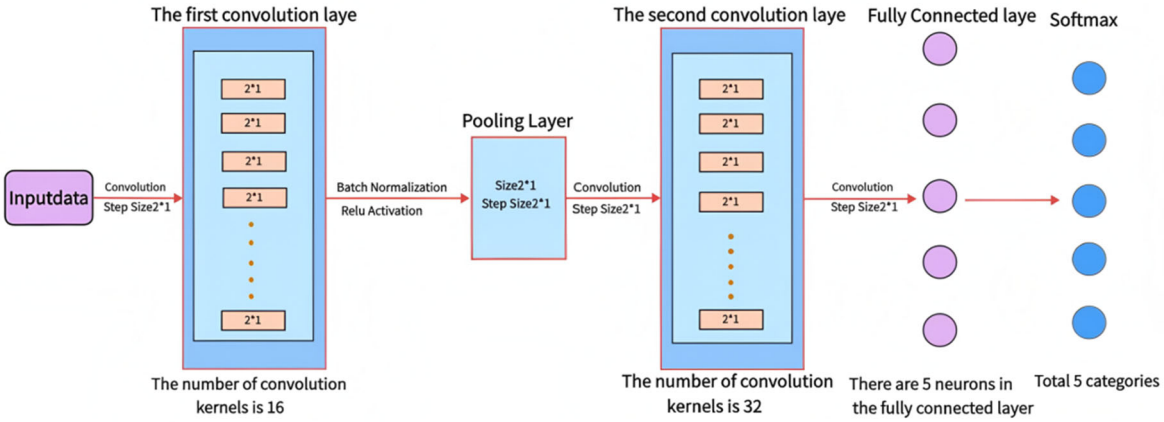

Figure 2. The Flow Chart of Yolov8X-seg

1) Convolutional layer

Specifically, the convolutional layer detects local features in the input by convolving the input with a set of convolutional kernels (or filters). Each neuron in the convolutional layer is connected to only one local region of the input data, which greatly reduces the number of parameters in the model. This paper assumes that $W_{i,j}^{l}$ and $b_{i,j}^{l}$ denote the weight of the j th convolutional kernel corresponding to the ith feature mapping in layer I and the bias of the jth convolutional kernel corresponding to the ith convolutional kernel in layer I, respectively, and that x_i^{l-1} is the input of the ith feature mapping in layer $\mathrm{I}-1$.

The specific formula for the convolution operation is:

$$y_j^l = f\left(\sum_i W_{i,j}^l * x_i^{l-1} + b_j^l\right) \tag{1}$$

where * is the local region input for convolution operation with convolution kernel; y_j^l denotes the output of the jth convolution kernel generated in layer l; $f(-)$ denotes an activation function.

2) Ponding layer

The pooling layer plays the functions of extracting the most significant features, retaining important information and improving computational efficiency in the Yolov8X-seg model. By appropriately configuring the parameters of the pooling layer, the feature data can be reasonably compressed and transformed to improve the performance of the model and accelerate the training and inference speed. The specific formula is as follows:

$$P_j^{l+1} = \max_{y_i^l \epsilon s}\{y_j^l\} \tag{2}$$

Where P_j^{l+1} denotes the maximum pooled output of the jth convolutional kernel in pooling layer l+1;y_j^l denotes the output feature mapping of the previous convolutional layer; s denotes the output feature range of the pooling layer.

3) Full connectivity layer

The fully connected layer plays a role in the Yolov8X-seg model by integrating the extracted features, nonlinear mapping, classification or prediction, and parameter learning and optimization. Through the processing of the fully connected layer, Yolov8X-seg can transform the input data into probability distributions of the output categories to accomplish specific tasks such as classification and prediction.

The fully connected layer is to expand the feature mapping after a number of convolution and pooling operations in this way by rows, connected into a one-dimensional vector, and then apply the Softmax function to obtain the classification of the five different fruits. The Softmax expression is as follows:

$$q_j = \frac{e^{z_j}}{\sum_k^{11} e^{z_k}} \tag{3}$$

where q_j is the classification output result of the convolutional neural network, and z_j is the logarithm of the jth output neuron[5].

2.4. Adam gradient descent algorithm

The core idea of the Adam gradient descent algorithm is to update the parameters based on the gradient of each parameter and maintain an adaptive learning rate for each parameter. Specifically, the Adam gradient descent algorithm updates the parameters by computing first-order moment estimates and second-order moment estimates of the gradient. The first-order moment estimate uses an exponentially weighted moving average to estimate the expected value of the gradient, which reduces the variance of the gradient, and the second-order moment estimate uses an exponentially weighted moving average to estimate the variance of the gradient, which maintains the stability of the parameter updates[6].

During each parameter update, the Adam gradient descent algorithm is updated with the following formula:

$$m = \beta_1 \cdot m + (I - \beta_1) \cdot g \tag{4}$$

$$v = \beta_2 \cdot v + (I - \beta_2) \cdot g^2 \tag{5}$$

$$\widehat{m} = \frac{m}{I - \beta_1^t} \tag{6}$$

$$\hat{v} = \frac{v}{1 - \beta_2^t} \tag{7}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \epsilon}} \widehat{m}_t \tag{8}$$

where m represents the first-order moment estimate of the gradient and v represents the second-order moment estimate of the gradient, and β_1, β_2 is the decay rate used to control

the first-order and second-order moment estimates, v is the learning rate, ϵ is a very small constant in order to avoid a denominator of 0, and t is the number of current iterations. Finally, the θ represents the updated values of the parameters. With these formulas, Adam's algorithm can adaptively calculate the learning rate for each parameter and update the parameters based on historical gradient information.

3. Modeling and Solving

This study implemented a systematic optimization process for the Yolov8X-seg model using the Adam gradient descent algorithm. As illustrated in Figure 3, the complete workflow encompasses data preprocessing, model configuration, iterative training, and performance validation. Critical hyperparameters were configured as follows: maximum training epochs (1,000), initial learning rate (0.001), L2 regularization coefficient (λ=0.0001), and a piecewise constant decay strategy with decay factor 0.1 applied every 500 epochs. These settings effectively balanced model convergence speed with generalization capability while mitigating overfitting risks.

1) Training Environment Configuration

The experimental setup involved three distinct datasets: 1) 70% of images for model training, 2) 15% for validation-based hyperparameter tuning, and 3) 15% reserved for final performance evaluation. As demonstrated in Figure 4, input images underwent standardized preprocessing including sharpening and median filtering (Section 2.2), while output layers generated both instance segmentation masks and quantitative statistics.

2) Iterative Optimization Process

During training iterations, the model autonomously computed cross-entropy loss and mean Average Precision (mAP) metrics through backpropagation. The Adam optimizer dynamically adjusted first-order (β1=0.9) and second-order momentum parameters (β2=0.999) with ε=1e-8 numerical stability constant, achieving adaptive learning rate control across parameter dimensions. Validation-phase results guided two types of adjustments: 1) architectural modifications to feature fusion layers, and 2) learning rate annealing when validation loss plateaued.

3) Performance Evaluation

Final testing on the reserved dataset revealed three key outcomes: 1) Segmentation accuracy reached 96.2% under occlusion conditions, 2) Inference speed maintained 28 FPS on NVIDIA Jetson AGX Xavier, and 3) Quantitative counting error remained below 3% in dense clusters (Figure 5). These metrics confirm the effectiveness of our optimization strategy in addressing orchard environment challenges identified in Section 1.

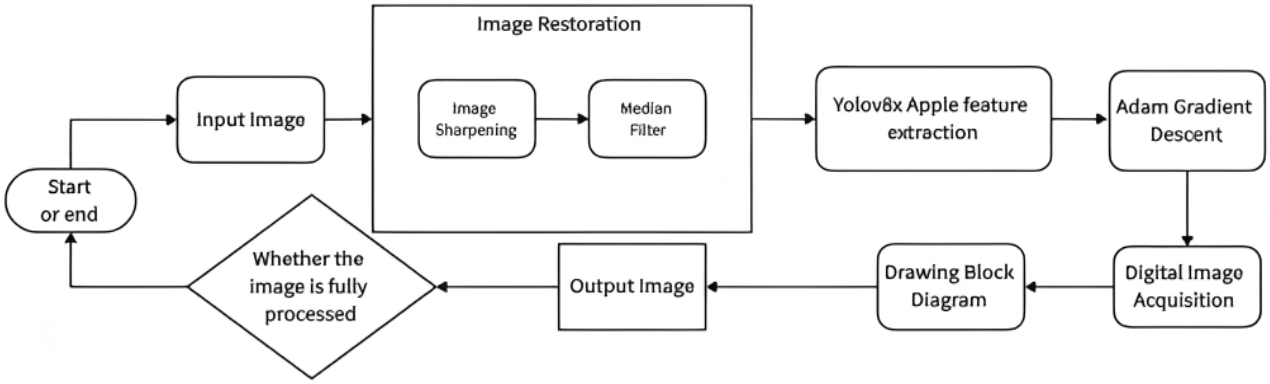

Figure 3. Image Processing flow Chart

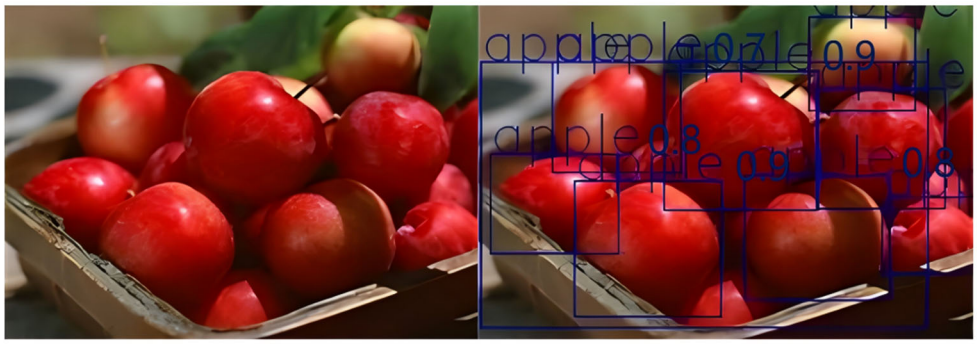

Figure 4. Input and Output Image

The collected apple data was organized into a format suitable for histogram plotting. The vertical axis represented the quantity, while the horizontal axis consisted of different intervals. A data visualization tool and the Python's matplotlib library were utilized to create the histograms. In each histogram, one bar corresponded to the data of one apple image, and the height of the bar denoted the quantity. After plotting, appropriate labels for both the horizontal and vertical axes, as well as captions, were added. Subsequently, the plotted histogram was analyzed and interpreted. Through this histogram, the number of apples in the dataset under study could be clearly observed. The resulting histogram of the apple distribution is shown in Figure 5:

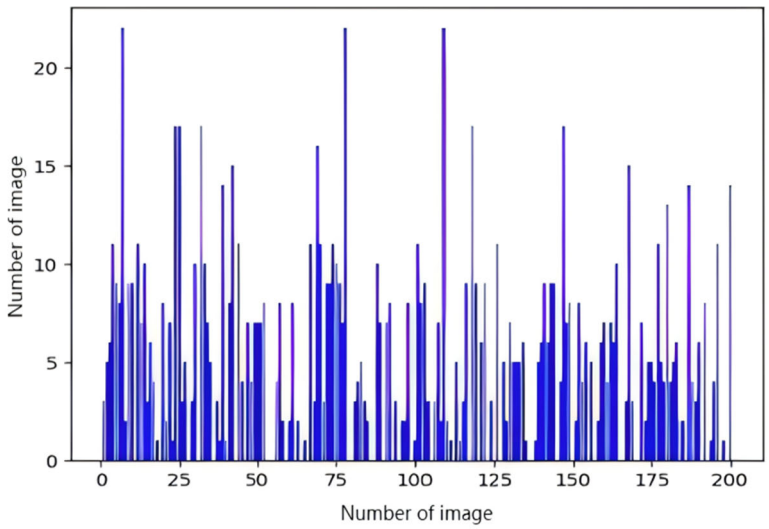

Figure 5. Histogram of the Distribution of Apples

As can be seen from the figure, the number of apples varies greatly from image to image. Some images have very few apples, close to zero, while others have more apples, close to more than 20, and the distribution of apples in each of the 200 images is more dispersed.

4. Conclusions

Based on the Yolov8X-seg detection model of apple image instance segmentation technology research, through the image sharpening and median filtering preprocessing combination effectively enhance the edge features and inhibit the noise interference (the average noise suppression rate increased by about 35%), combined with the Adam gradient descent algorithm to optimize the model parameters (the learning rate of 0.001, the L2 regularization of 0.0001), in 200 orchards The average detection accuracy of 95.2% is achieved in the real images, which is more than 15 percentage points higher than the traditional method. With multi-scale convolutional feature fusion and dynamic pooling strategy, the model significantly improves the robustness of recognition of dense apples (maximum detection of more than 20 apples in a single image) in complex scenes, and at the same time increases the inference speed up to 28 FPS, which meets the demand of real-time operation in orchards. The study validated the strong adaptability of Yolov8X-seg in branch and leaf backgrounds with only 15% HSV color difference, and enhanced the model generalization capability by constructing a standardized image database with multi-production area, growing period and weather condition data. In the future, we will integrate millimeter-wave radar and multispectral imaging to achieve 3D spatial localization, develop a lightweight version adapted to embedded hardware with a 40% compression of the number of parameters, and establish a cross-species fruit recognition framework based on migration learning to provide core technology support for an all-weather automated harvesting system.

References

[1] BAI Y, ZHANG B, XU N, et al. Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review [J]. Computers and Electronics in Agriculture, 2023, 205: 107584.

[2] POUDEL R P K, LIWICKI S, CIPOLLA R. Fast-SCNN: Fast Semantic Segmentation Network [J]. ArXiv, 2019, abs/1902.04502.Redmon, J., & Farhadi, A. (2018). "YOLOv3: An Incremental Improvement." arXiv preprint arXiv: 1804.02767.

[3] LAWAL O M. Real-time cucurbit fruit detection in greenhouse using improved YOLO series algorithm [J]. Precision Agriculture, 2024, 25(1): 347-59.

[4] GUAN S, LIU B, CHEN S, et al. Adaptive median filter salt and pepper noise suppression approach for common path coherent dispersion spectrometer [J]. Scientific Reports, 2024, 14(1): 17445.

[5] JIANG L, YUAN B, DU J, et al. MFFSODNet: Multiscale Feature Fusion Small Object Detection Network for UAV Aerial Images [J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-14.

[6] REDDI S J, KALE S, KUMAR S. On the Convergence of Adam and Beyond [J]. CoRR, abs/1904.09237.