RESEARCH ARTICLE

A Robotic System for Fine-Grained Non-Destructive Grading of Visually Similar Fruits Based on Improved YOLOv11 and Multi-modal Perception

Zhenhao Ma[1,2], **Bin Zhang**[1,2,*] **and Tianzhen Yin**[1,2]

[1] College of Engineering, China Agricultural University, Beijing 100083, China
[2] National Sub-Center for R&D of Agro-Products Processing Technology and Equipment, Beijing 100083, China

Abstract

To address key challenges in post-harvest fruit grading—namely the difficulty of distinguishing visually similar varieties, the invisibility of internal quality, and mechanical damage during grasping—this study develops an intelligent robotic grading system that integrates advanced computer vision, Near-Infrared (NIR) spectroscopy, and flexible force-controlled grasping. First, an improved object detection algorithm, YOLOv11-TFE, is proposed to mitigate visual confusion between Qixia Fuji apples and Beijing Pinggu peaches and to handle the irregular geometry of Nanshui pears. By embedding the parameter-free SimAM attention mechanism into the backbone to explicitly enhance and decouple surface texture features (gloss versus tomentum) and incorporating the DySample upsampling operator to better preserve complex edge information, the discriminative capability of the detector is significantly improved. Second, a non-destructive internal quality detection module based on NIR spectroscopy (650–1100 nm) is constructed. Through the optimization of preprocessing strategies—specifically Improved Derivative Correction (IDC) and Standard Normal Variate (SNV)—robust PLSR models for soluble solids content (SSC) prediction are established for fruits with differing skin textures. Finally, a dynamic actuation unit for a biomimetic flexible gripper has been designed to achieve non-destructive sorting under continuous flow conditions. Experimental results show that the improved visual algorithm achieves an average mAP@0.5 of 94.6%, with detection accuracies for apples and peaches reaching 96.0% and 97.8%, respectively, markedly reducing inter-class confusion. The root mean square error of prediction (RMSEP) for SSC across all three fruit varieties is kept within 0.65%. System-level validation further demonstrates an overall dynamic grasping success rate of 91.7% without causing visible damage. Overall, the proposed system achieves precise and comprehensive grading for multiple high-value fruit varieties.

Keywords: fruit grading robot, YOLOv11, fine-grained classification, NIR spectroscopy, flexible grasping.

Academic Editor:
Jianlei Kong

Submitted: 04 December 2025
Accepted: 29 January 2026
Published: 17 April 2026

Vol. 3, **No.** 2, 2026.
10.62762/TIS.2025.566749

***Corresponding author:**
✉ Bin Zhang
zhangbin64@cau.edu.cn

Citation
Ma, Z., Zhang, B., & Yin, T. (2026). A Robotic System for Fine-Grained Non-Destructive Grading of Visually Similar Fruits Based on Improved YOLOv11 and Multi-modal Perception. *ICCK Transactions on Intelligent Systematics*, 3(2), 94–107.

1 Introduction

Post-harvest sorting and grading are critical for enhancing the commercial value of fruits, ensuring quality stability, and establishing brand identity [1, 2]. For high-value varieties such as Qixia Fuji apples, Beijing Pinggu peaches, and Nanshui pears, traditional manual sorting and conventional machine vision methods struggle to balance grading accuracy with processing efficiency because of their highly similar visual appearances and subtle differences in internal quality, which can lead to frequent downgrading of premium fruits and misallocation of low-quality fruits into high-end channels [3].

With the rapid development of deep learning, computer vision has become a core enabling technology for intelligent fruit grading. Object detection algorithms are generally categorized into two-stage methods, such as Faster R-CNN, and one-stage methods represented by the YOLO (You Only Look Once) series; the latter have been widely adopted in agricultural scenarios due to their favorable trade-off between detection accuracy and real-time performance [4–7]. Existing studies have demonstrated promising results for apples, citrus, strawberries, and other fruits in complex orchard or packing environments, and have begun to explore fine-grained tasks such as cultivar- or ripeness-level detection using lightweight backbones and attention mechanisms. Nevertheless, most current datasets and models still focus on single-species or color-dominant differences and are less effective in scenarios with "small inter-class variance and large intra-class variance," such as distinguishing smooth-cuticle apples from fuzzy peaches or detecting irregularly shaped pears, where subtle texture cues and complex edge geometries are easily lost in standard feature extraction pipelines [8, 9].

In addition to external appearance, internal quality indicators such as soluble solids content (SSC) are increasingly recognized as essential for high-end market positioning, yet most industrial grading systems still lack synchronous online assessment of internal attributes [10]. Near-Infrared (NIR) and hyperspectral techniques have shown strong potential for non-destructive SSC prediction and internal quality evaluation across multiple fruit species, but the recorded spectra are highly sensitive to epidermal microstructure and surface texture—such as smooth waxy skins, surface fuzz, or stone-cell-rich tissues—which induce pronounced scattering and absorption variability and challenge the robustness and transferability of calibration models across varieties [10]. Although various spectral preprocessing and modeling strategies have been proposed to mitigate these effects, systematic integration of texture-aware vision, scatter-robust NIR modeling, and execution-level design remains limited [11, 12].

At the execution level, traditional rigid grasping mechanisms provide high positional stability but tend to cause surface indentations or abrasions on delicate fruits under high-throughput conditions, whereas recent flexible bionic grippers and multimode sensing approaches improve compliance and reduce mechanical damage while enabling richer interaction with the fruit surface [13]. However, challenges persist in ensuring dynamic stability, response speed, and closed-loop force control under continuous-flow operation, particularly when handling fruits with varying sizes, shapes, and surface textures, and truly integrated systems that combine fine-grained visual recognition, internal quality detection, and non-destructive flexible grasping on a single robotic platform are still scarce.

To address these gaps, this study targets three key challenges—difficulty in distinguishing visually similar fruits, invisibility of internal quality, and susceptibility to mechanical grasping damage—and develops a multi-modal synergistic robotic system for intelligent fruit grading. Specifically, an improved YOLOv11-TFE algorithm is designed to enhance fine-grained texture and edge representation for visually similar and geometrically irregular fruits; an NIR spectroscopy module adaptable to different skin textures is constructed to establish a high-precision SSC prediction model; and a biomimetic flexible dynamic gripping unit was implemented to achieve high-success-rate, low-damage grasping under continuous-flow operation conditions. Through deep fusion of visual, spectral, and tactile information, the proposed system provides an engineering-feasible solution for fine-grained, non-destructive grading of multi-variety high-value fruits.

2 Related Work

2.1 Material Preparation

An experimental dataset comprising three categories of typical commercial fruits was constructed for this study, consisting of 108 Qixia Fuji apples, 120 Beijing Pinggu peaches, and 120 Nanshui pears. To ensure the statistical representativeness and significance of

the experimental data, and to maximize the reduction of uncontrolled interference caused by individual developmental variations, all samples were harvested from standardized core production areas specific to each variety and screened according to strict physical morphological and physiological indicators.

The specific selection criteria were as follows: Qixia Fuji apple samples were selected with a transverse diameter of 65–85 mm and a maturity level of 8–12; Beijing Pinggu peach samples were screened for a transverse diameter of 60–80 mm, a maturity level of 8–11, and intact surface fuzz; Nanshui pear samples were selected with a transverse diameter of 70–90 mm, a maturity level of 7–10, and a regular shape. Upon arrival at the laboratory, all samples were subjected to manual re-inspection to exclude individuals with mechanical damage or pest and disease infestation. Subsequently, the samples were placed under constant temperature and humidity conditions for a specific period to equilibrate their physiological states, thereby ensuring the consistency and comparability of conditions for subsequent visual and spectral data acquisition.

2.2 System Architecture

The overall system employs a closed-loop "Perception-Processing-Execution" architecture to achieve online detection, non-destructive grasping, quality assessment, and grading functions. Specifically, the perception layer comprises an industrial camera and a Near-Infrared (NIR) spectroscopy module coaxially mounted on the end-effector, while pressure sensors are integrated within the gripper to provide real-time force control feedback. The processing layer utilizes an NVIDIA RTX 4080 platform to deploy an improved YOLOv11n model for object detection and trajectory tracking, triggering spectral acquisition and multi-modal fusion decision-making only upon confirmation of "grasp stability." The execution layer consists of a lightweight 6-DOF collaborative robot, a flexible gripper, and a conveyor belt, featuring bio-inspired fin-ray soft fingers that effectively reduce contact surface pressure through an enveloping grasp mechanism. To ensure operational precision, the system maintains spatiotemporal consistency during motion via hand-eye calibration and timestamp synchronization; furthermore, a safety fallback mechanism is automatically triggered upon

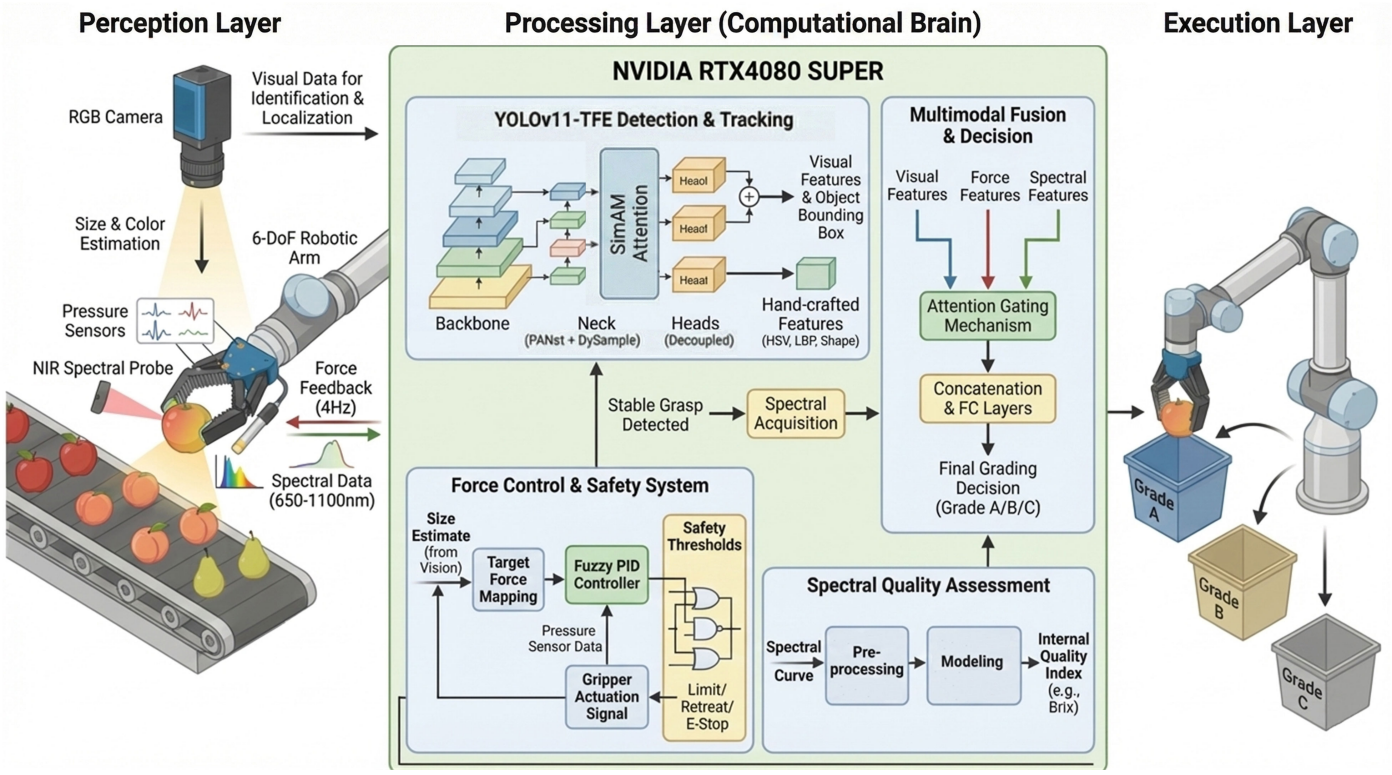


Figure 1. Schematic diagram of the intelligent fruit grading system architecture. The system integrates three core modules: (Left) System Input & Perception, featuring an RGB camera for visual localization and a NIR spectral probe for internal quality detection; (Middle) Intelligent Processing & Decision, deployed on an NVIDIA RTX4080, comprising the improved YOLOv11-TFE algorithm, spectral analysis modeling, and a multimodal fusion decision network; (Right) Automated Sorting & Grading, executing non-destructive sorting via a 6-DoF robotic arm with flexible force control.

Table 1. Statistical breakdown by category for three fruit datasets.

Category	Number of real images	Expand image count	Total	Training/ Validation/Testing
Qixia Red Fuji	1500	2,500	4,000	2800/ 800 / 400
Beijing Pinggu peaches	1500	2,500	4,000	2800 / 800 / 400
Nanshui pear	1500	2,500	4,000	2800 / 800 / 400

detection of pressure threshold violations, target loss, or communication anomalies. The overall system workflow is illustrated in Figure 1.

An industrial camera (resolution 2048×1536, frame rate 30 fps) and a Near-Infrared (NIR) spectrometer (spectral range 650–1100 nm, spectral resolution 2.2 nm, Ocean Optics) are employed to acquire external appearance images and internal quality spectral information, respectively, while an NVIDIA RTX 4080 GPU undertakes computational tasks including object detection, multi-modal fusion, and grading decision-making. The execution end utilizes a lightweight 6-DOF collaborative robot with a repeatability of $\pm$0.02 mm, equipped with a stepper motor-driven rack-and-pinion gripper that delivers a maximum clamping force of 5 N with a force control precision of $\pm$0.1 N. Furthermore, two pressure sensors (range 0–100 N, accuracy $\pm$0.5%) are integrated into the gripper fingertips to enable adaptive threshold adjustment during grasping at a sampling frequency of 4 Hz. Through the synchronous triggering of the vision and spectral modules with a constant-speed closed-loop conveyor belt, the system stably executes an integrated workflow of "visual localization—force-controlled grasping—spectral detection—intelligent sorting" under continuous flow conditions.

2.3 Object Detection Algorithms in Fruit Grading

With the rapid development of deep learning, computer vision has become the core technology for intelligent fruit sorting [2]. Object detection algorithms are generally categorized into two-stage methods (e.g., Faster R-CNN [14]) and one-stage methods represented by the YOLO (You Only Look Once) series [15]. While two-stage detectors offer high accuracy, their computational cost often creates a bottleneck for real-time online grading lines. Consequently, the YOLO series has become the mainstream choice in agricultural robotics due to its superior balance between inference speed and detection precision [16]. Early versions, such as YOLOv3 and YOLOv5, have been extensively applied to the detection of apples, citrus, and pears in complex orchard environments [17]. Recently, YOLOv11 [18], the latest iteration, achieved State-of-the-Art (SOTA) performance by optimizing the backbone with C3k2 blocks. However, standard YOLO models primarily rely on prominent visual features and often struggle with fine-grained classification tasks involving visually similar fruits—such as distinguishing between the smooth cuticle of Qixia Fuji apples and the fuzzy texture of Beijing Pinggu peaches. Therefore, incorporating attention mechanisms to enhance texture-feature awareness within the YOLOv11 architecture is essential for precise multi-variety grading.

3 Methodology

3.1 Dataset Construction

To construct a robust benchmark for fine-grained fruit classification, a dedicated detection dataset was established containing three target categories: Qixia Red Fuji apple, Beijing Pinggu peach, and Nanshui pear. Image acquisition covered both natural orchard scenes and standardized sorting/supermarket environments, incorporating challenging real-world conditions such as strong specular highlights, backlighting, limb and fruit occlusions, and scale variations caused by different shooting distances. After strict data cleaning and augmentation, all images were manually annotated in YOLO format with class labels and bounding boxes. As summarized in Table 1, the final dataset contains 12,000 images, uniformly distributed across the three categories (4,000 images per class, including 1,500 real images and 2,500 augmented or hard-mined images). For each class, the data were split using stratified sampling into 2,800 training, 800 validation, and 400 test images, yielding an overall partition of 8,400 / 2,400 / 1,200 for training, validation, and testing, respectively.

3.2 Improved YOLOv11-TFE Algorithm Architecture

To enhance the inter-class separability among visually similar fruit varieties, we reconstructed the YOLOv11 architecture by integrating a texture-sensitive attention mechanism, geometric feature preservation, and an

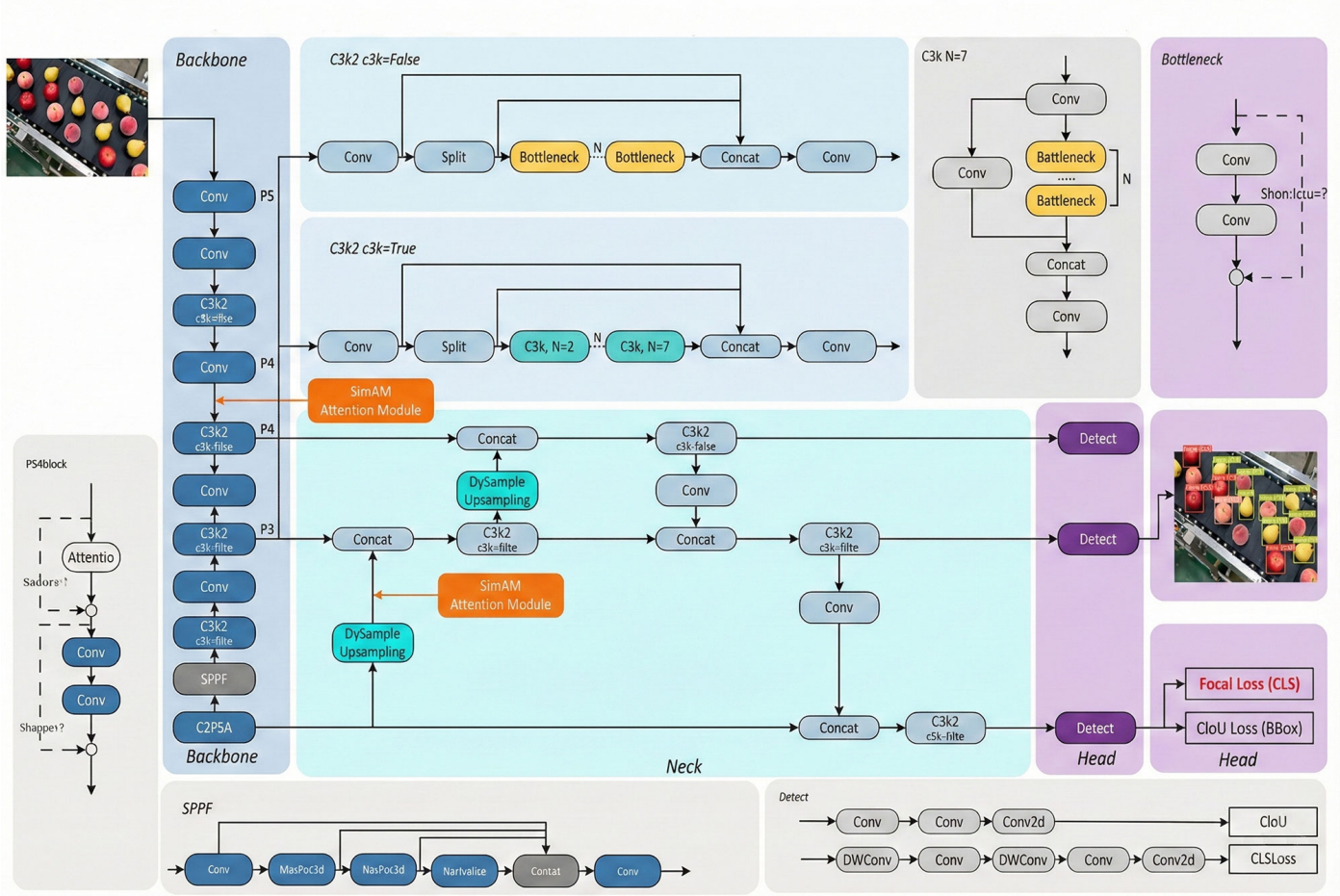


Figure 2. The network topology of the proposed YOLOv11-TFE architecture. Key improvements include the embedding of the SimAM parameter-free attention module in the backbone to enhance texture feature extraction, the integration of DySample content-aware upsampling in the neck to preserve geometric details, and the adoption of Focal Loss in the detection head to address hard sample mining.

adaptive loss re-weighting strategy. The overall network topology is illustrated in Figure 2.

First, we embedded the SimAM parameter-free attention mechanism into the P3 and P4 stages of the CSPDarknet backbone. By evaluating a spatial energy function, this module explicitly amplifies discriminative surface features—specifically the contrast between the diffuse reflection of peach fuzz and the specular reflection of apple cuticles—thereby reducing feature entanglement.

Secondly, to mitigate the degradation of shape information during feature aggregation, we introduced the DySample content-aware upsampling operator to optimize the PANet. By reconstructing feature maps through dynamic point sampling, this operator faithfully preserves the unique obovate boundaries of Nanshui pears, avoiding the edge blurring typically associated with traditional interpolation methods.

Finally, addressing the ambiguity at decision boundaries for morphologically similar varieties, we replaced the standard Binary Cross Entropy (BCE) classification objective function with Focal Loss. This strategy biases training gradients towards hard samples by adjusting a focusing parameter, with its mathematical definition expressed in Equation 1.

$$L_{cls} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

where p_t represents the model's predicted probability for the category, α_t denotes the balancing factor, and γ is the focusing parameter (set to $\gamma = 2.5$ in this study), thereby significantly improving classification accuracy.

3.3 Training and Optimization Strategy

To further exploit the model's potential and strike a balance between inference speed and accuracy, this study implemented a combinatorial optimization strategy.

In terms of data augmentation, we combined geometric transformations (rotation, scaling, flipping) with photometric distortions (brightness, contrast, saturation adjustments). Additionally, Mosaic augmentation and Gaussian blur/noise perturbations were introduced to simulate complex lighting environments and enhance the model's robustness to texture details.

In terms of training dynamics, we adopted a "progressive unfreezing" strategy initialized with COCO pre-trained weights. In the first stage, the backbone parameters were frozen to adapt the detection head; in the second stage, the entire network was unfrozen to fine-tune deep features. Simultaneously, multi-scale training was enabled, while an image pyramid strategy was employed during the inference phase to adapt to inputs of varying resolutions.

3.4 Spectral Data Acquisition Configuration

To acquire high-fidelity spectral fingerprints of the internal fruit quality, we constructed a spectral acquisition system based on diffuse reflection geometry. A 12 V / 2 W broadband tungsten-halogen lamp was used as the light source, illuminating the fruit surface via optical fibers at a 45°incident angle. The detection probe was mounted perpendicular to the sample surface, with the working distance strictly controlled at 5 mm to maximize the Signal-to-Noise Ratio (SNR) and minimize specular reflection from the surface.

To eliminate the influence of environmental stray light and device dark current, a standard black-and-white calibration strategy was employed. Before acquisition, the dark current intensity (with the light source off) and the reflection intensity of a Polytetrafluoroethylene (PTFE) standard white reference were recorded. The raw spectral intensity was then converted into relative reflectance and subsequently into absorbance for downstream processing.

3.5 Spectral Preprocessing

Raw diffuse reflectance spectral data inevitably suffer from non-linear interferences, including high-frequency electronic noise from instruments, light scattering from the sample surface (e.g., fuzz or cuticle reflections), and baseline drift. To improve the spectral SNR and enhance effective chemical fingerprint features, this study evaluated and implemented a comprehensive preprocessing strategy encompassing smoothing/denoising, scatter correction, and feature sharpening.

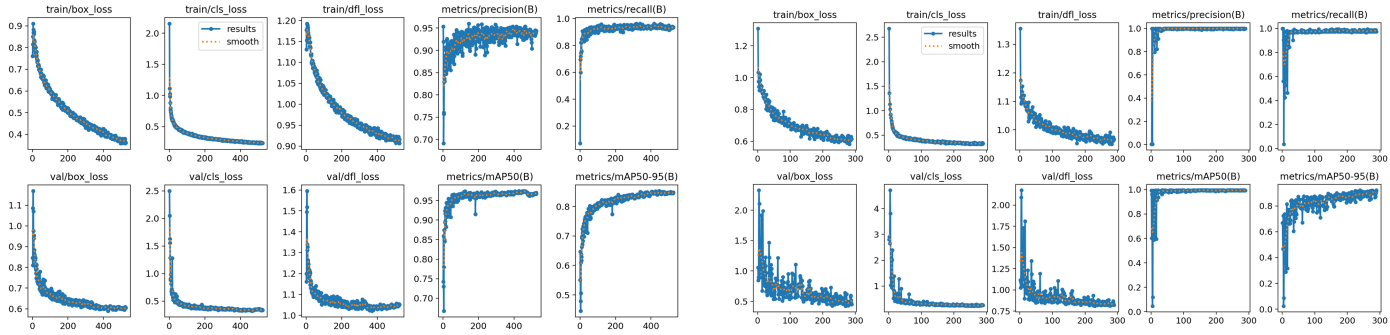
First, the Savitzky-Golay (S-G) convolution smoothing algorithm was applied to filter full-band data, suppressing high-frequency random noise while preserving waveform features. Second, to address optical path differences caused by the physical morphology (e.g., size, curvature) and skin texture of different individual fruits, Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) algorithms were introduced to eliminate additive and multiplicative effects through normalization. Furthermore, to further remove baseline tilt caused by light source fluctuations and to separate overlapping peaks, polynomial baseline correction and derivative transformations (including first derivative and Improved Difference Correction, IDC) were investigated. Finally, based on the distinct skin optical properties of Fuji apples, Beijing Pinggu peaches, and Nanshui pears, the optimal combination of preprocessing algorithms was selected via grid search and 5-fold Cross-Validation to ensure the robustness of subsequent modeling.

3.6 Determination of Soluble Solids Content

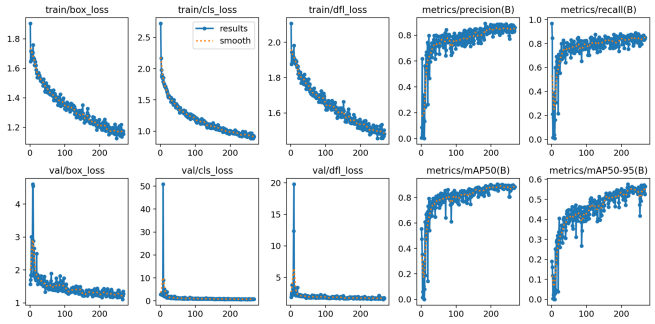
The determination of SSC was conducted in accordance with the Agricultural Industry Standard NY/T 2637-2014. A digital refractometer (Model PAL-BX/AC5, ATAGO, Japan) was used to measure the SSC of apples via a destructive method. The refractometer features a measurement range of 0–60.0%, a resolution of 0.1%, and an accuracy of $\pm$0.2%. following spectral acquisition, approximately 30 g of pulp was extracted from the spectral scanning region. The pulp was juiced using a hand-held juicer and homogenized in a beaker. Subsequently, the juice was applied to the refractometer's prism using a rubber-tipped dropper for measurement. The mean value of three readings per sample was recorded as the final SSC.

3.7 Feature Selection and Quantitative Modeling

To address the issues of multicollinearity and redundant information in full-band spectral data, the Competitive Adaptive Reweighted Sampling (CARS) algorithm was employed for feature wavelength selection [19]. Simulating Darwin's principle of "survival of the fittest," the CARS algorithm iteratively eliminates wavelength variables with low contribution to the model, thereby constructing a feature subset containing key biochemical information.



(a) Apple Model Training Results **(b) Peach Model Training Results**



(b) Training Results for the Nan Shui Li Model

Figure 3. Training convergence curves of the YOLOv11-TFE model. The graphs illustrate the trends of loss functions (box loss, classification loss, DFL loss) and evaluation metrics (Precision, Recall, mAP@0.5, mAP@0.5:0.95) over epochs. The rapid decline in loss and the steady rise in mAP indicate the model's fast convergence and high stability.

On this basis, a Partial Least Squares Regression (PLSR) model was established to achieve quantitative prediction of internal quality indicators such as fruit sugar content (SSC) and firmness [20]. The optimal number of Latent Variables (LVs) was determined using Leave-One-Out Cross-Validation (LOOCV) to avoid model overfitting or underfitting. The performance of the final model was comprehensively evaluated using the coefficient of determination (R_c^2, R_p^2) and the Root Mean Square Error ($RMSEc$, $RMSEp$) for the calibration and prediction sets, respectively. Where an independent test set was available, external validation was further conducted to quantify the model's generalization ability and robustness.

3.8 Evaluation Metrics

This study established a multi-dimensional evaluation framework to comprehensively assess system performance.

1) Detection and grading accuracy were evaluated using mAP@0.5, Recall, Macro F1-score, and normalized confusion matrices.

2) Metrology and physicochemical inversion were assessed by quantifying fruit diameter errors with MAE, RMSE, and MAPE, while the predictive ability and generalization of the SSC model were evaluated using the coefficient of determination (R^2) and RMSE.

3) System efficiency was characterized by end-to-end FPS and the 95th-percentile (P95) latency.

4) Non-destructive grasping performance was verified through double-blind visual inspection and quantified using Cohen's kappa consistency coefficient.

Statistical analysis was conducted using a bootstrapping procedure to construct 95% confidence intervals, and the significance of performance differences between models was determined by paired t-tests or Wilcoxon signed-rank tests, with $p < 0.05$ indicating statistical significance.

4 Experiments

4.1 YOLOv11-TFE Model Training Results and Analysis

To verify the robustness of the YOLOv11-TFE architecture in learning the features of different fruit categories, this study conducted independent training and testing on Qixia Fuji apples, Beijing Pinggu peaches, and Nanshui pears.

Figure 3 illustrates the decline curves of the loss functions and the evolution of accuracy metrics for

Table 2. Comparative experimental results of detection performance between the baseline and the proposed method.

Model	FPS	Fruit Variety	Precision (%)	Recall (%)	mAP@0.5 (%)
YOLOv11n (Baseline)	120	Qixia Red Fuji	90.5	89.7	93.2
		Beijing Pinggu Peach	92.6	90.4	91.3
		Nanshui Pear	86.0	85.1	88.2
		Mean	89.7	88.4	90.9
YOLOv11-TFE (Ours)	95	Qixia Red Fuji	**92.7**	**92.4**	**96.0**
		Beijing Pinggu Peach	**98.1**	**96.4**	**97.8**
		Nanshui Pear	**89.3**	**86.5**	**90.1**
		Mean	**93.3**	**92.1**	**94.6**

the three varieties during the training process. From the overall training trends, the improved model demonstrated excellent convergence characteristics across all three datasets without exhibiting obvious overfitting. This validates the effectiveness of Focal Loss in balancing the gradients between hard and easy samples and in suppressing background noise.

For the Qixia Fuji apple dataset, the training process was extended to 400 epochs to ensure convergence due to the strong interference caused by specular reflections from the cuticular layer on the fruit surface. As indicated by the result curves, the Localization Loss (Box Loss) descended rapidly from an initial value of approximately 2.0 and stabilized near 1.0 after 200 epochs. Although the Classification Loss exhibited certain fluctuations during the early training stages, the mean Average Precision (mAP@0.5) showed a steady upward trend—facilitated by the SimAM module's enhanced adaptability to lighting variations—and ultimately stabilized at around 0.9.

In contrast, the training convergence speed for Beijing Pinggu peaches was extremely fast, reaching performance saturation in approximately 100 epochs. Benefiting from the high frequency-domain discriminability of the unique diffuse reflection features of the peach fuzz, the model achieved a rapid surge in accuracy within the first 20 epochs. The mAP@0.5 ultimately approached 0.99, and the Bounding Box Regression Loss quickly dropped below 0.7, indicating that the model was able to capture edge contour features with extreme precision.

Regarding the Nanshui pear dataset, the training curves revealed a significant reduction in Bounding Box Regression Loss, attributed to the DySample upsampling operator. This demonstrates that the operator effectively preserved the obovate geometric features, overcoming the shape distortion issues typically associated with traditional sampling methods. The training results are presented in Figure 3.

To quantitatively evaluate the final detection performance of YOLOv11-TFE and its breakthrough in inter-class discriminability, detailed evaluation metrics for each category were calculated on an independent test set and compared against the benchmark YOLOv11n model. Experimental results demonstrate that the improved algorithm achieves significant advantages in comprehensive performance, particularly in resolving the visual confusion between "Qixia Red Fuji" apples and "Beijing Pinggu" peaches.

As shown in Table 2, although the inference speed of YOLOv11-TFE decreased slightly from the baseline 120 FPS to 95 FPS, it remains well above industrial real-time detection requirements while achieving a substantial qualitative improvement in accuracy metrics. In the highly confusing categories of Qixia Red Fuji and Beijing Pinggu Peach, the mAP@0.5 of the improved model climbed to 96.0% and 97.8%, respectively. Compared to the baseline model, the former increased by 2.8 percentage points, while the latter achieved a remarkable growth of 6.5 percentage points. This significant increase in precision is primarily attributed to the explicit decoupling of surface texture features by the SimAM attention mechanism, which enables the network to acutely distinguish the subtle differences between the smooth cuticle and the fuzzy texture. Concurrently, for the irregularly shaped Nanshui Pear, detection accuracy (mAP@0.5) increased to 90.1% with a Recall stabilizing at 86.5%, effectively demonstrating the efficacy of the DySample operator in capturing complex geometric edges and mitigating the degradation of shape information. Overall, the Mean mAP@0.5 of YOLOv11-TFE reached 94.6%, exhibiting superior robustness and accuracy in fine-grained multi-variety fruit grading tasks.

To further evaluate the category-level discriminative

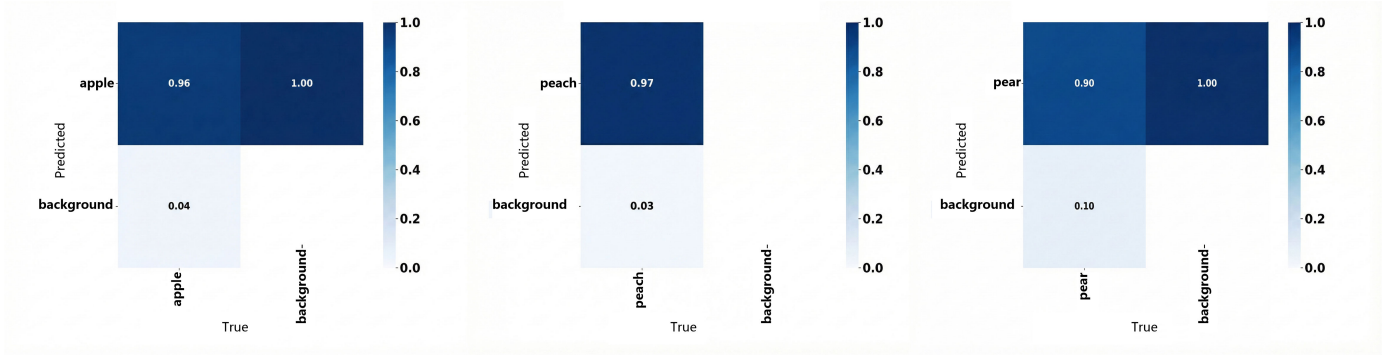


Figure 4. Normalized confusion matrices for the target fruit categories against the background. The matrices display the classification performance for (a) Apple, (b) Peach, and (c) Pear. The high diagonal values represent the recall rates, demonstrating the model's low missed detection rate.

capability of the proposed algorithm, normalized confusion matrices of the YOLOv11 TFE model on the test set were computed for Qixia Red Fuji apple, Beijing Pinggu peach, and Nanshui pear, as shown in Figure 4. The diagonal values indicate the correct classification probability for each class, whereas the off diagonal terms represent the proportion of samples misclassified as background. As illustrated in Figure 4(a–c), the true positive rates for Qixia Red Fuji apple and Beijing Pinggu peach reach 0.96 and 0.97, respectively, while that of Nanshui pear is 0.90. All background samples are correctly identified, and the only errors arise from fruit targets being misjudged as background, with rates of 0.04, 0.03, and 0.10 for apple, peach, and pear, respectively. Notably, no cross category confusion is observed in the test set. This indicates that the incorporation of the SimAM attention mechanism and the DySample operator not only enhances the extraction of surface texture and contour features, but also effectively increases the separability between fruit categories and the background in the feature space. Overall, the confusion matrix results demonstrate that YOLOv11 TFE achieves high detection accuracy and strong class decoupling ability. The remaining errors are mainly due to severely occluded or partially visible fruits that are classified as background, which is acceptable for practical multi variety mixed stream grading applications.

4.2 Spectral Internal Quality Prediction Model

To investigate the impact of different signal processing methods on the accuracy of internal quality detection and to identify the optimal processing pipeline, this study conducted a comparative analysis of four preprocessing algorithms: Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), Difference Correction (DC), and Improved Difference Correction (IDC). The processed spectral response profiles are illustrated in Figure 5. The quantitative modeling results based on Partial Least Squares Regression (PLSR) are presented in Table 3, and the experimental data indicate that the choice of preprocessing strategy significantly influences the prediction performance of the model.

Table 3. Apple modeling results.

Preprocessing	R_c^2	R_v^2	RMSEC	RMSEV
MSC	0.9090	0.9011	0.5570	0.6132
SNV	0.8871	0.8802	0.6625	0.6956
DC	0.8957	0.8905	0.6506	0.6918
IDC	0.9249	0.9143	0.5196	0.5797
IDC+CARS	0.9385	0.9255	0.4920	0.5310

While single SNV or DC processing corrected inter-sample optical path variations and baseline drift to a certain extent, their resulting model performance was relatively limited. The coefficients of determination for the validation set (R_v^2) were only 0.8802 and 0.8905, respectively, accompanied by relatively high prediction errors. In contrast, MSC preprocessing outperformed SNV and DC by achieving more effective scatter correction, improving the R_v^2 to 0.9011. However, the IDC strategy adopted in this study demonstrated superior comprehensive processing capability. By synergistically optimizing baseline correction and feature sharpening, IDC achieved the highest coefficients of determination for both the calibration (R_c^2) and validation (R_v^2) sets in the apple soluble solids content (SSC) model, reaching 0.9249 and 0.9143, respectively. Simultaneously, it significantly reduced the Root Mean Square Error of Validation ($RMSEV$) to 0.5797, surpassing all other individual preprocessing methods.

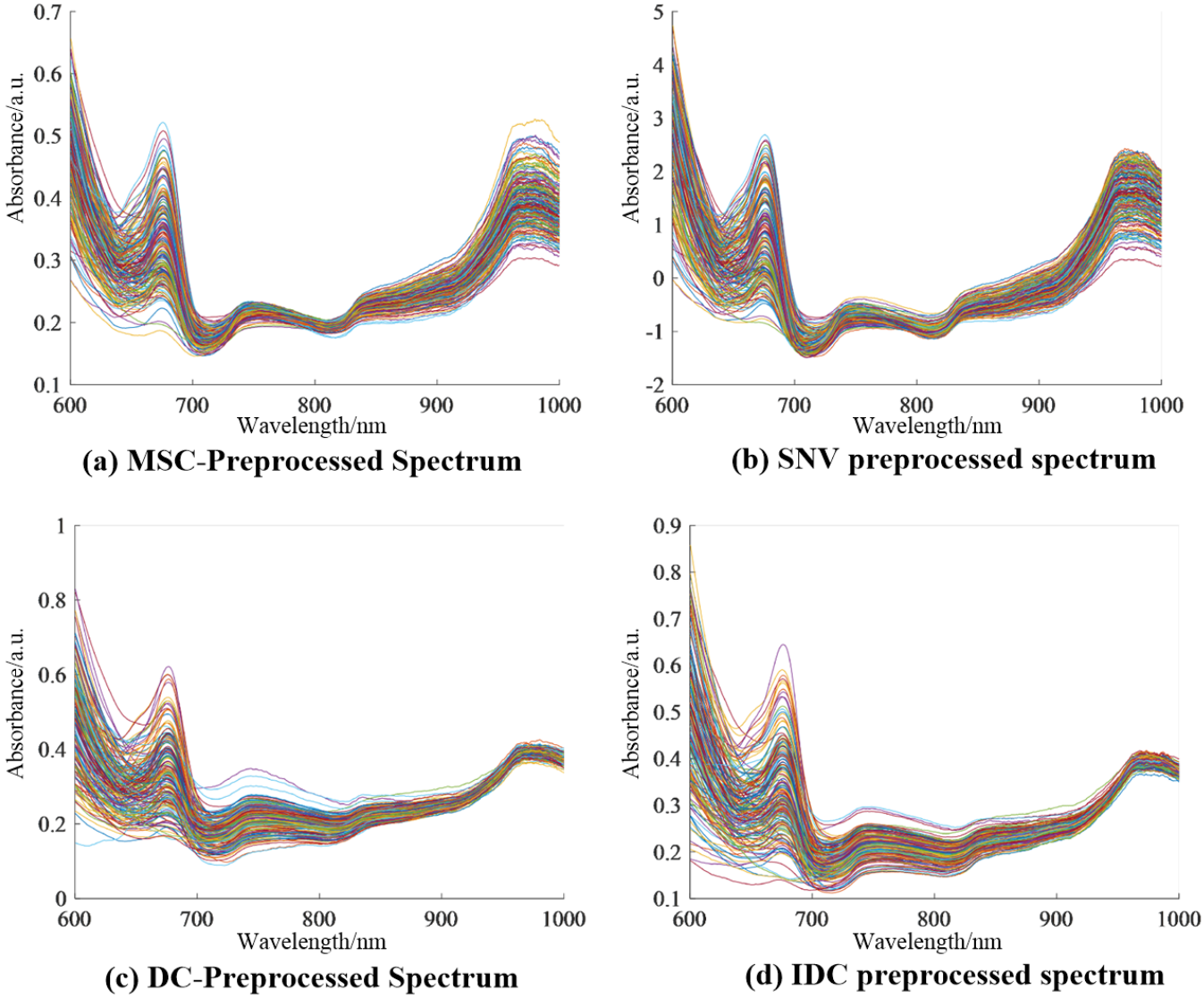


Figure 5. Comparison of diffuse reflectance spectral curves under different preprocessing methods. Subplots illustrate the effects of (a) MSC, (b) SNV, (c) DC, and (d) IDC algorithms on signal baseline correction and feature enhancement for the fruit samples.

Building upon this, to further enhance model robustness, the CARS (Competitive Adaptive Reweighted Sampling) algorithm was introduced for feature variable selection across the full spectrum. The results demonstrate that after the elimination of redundant wavebands, the prediction accuracy of the model achieved a further substantial improvement: R_v^2 ultimately reached 0.9255, and $RMSEV$ decreased to 0.5310. These findings strongly validate the superiority of the combined strategy of "IDC feature enhancement + CARS wavelength selection" in extracting effective spectral information and mitigating noise interference.

Building upon the spectral modeling methodology validated on Qixia Red Fuji apples, this study further extended the application of Near-Infrared Spectroscopy to the internal quality detection of Beijing Pinggu Peaches and Nanshui Pears. The objective was to evaluate the generalization capability of this technology across fruits with distinct epidermal textures.

Accounting for the strong light scattering induced by the surface pubescence of peaches and the unique structural presence of stone cells in pear skins, this experiment investigated the impact of three preprocessing algorithms—Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and Normalization (NOR)—on the performance of PLSR models relative to the raw spectra. The spectral response profiles of the two fruit varieties under different preprocessing conditions are illustrated in

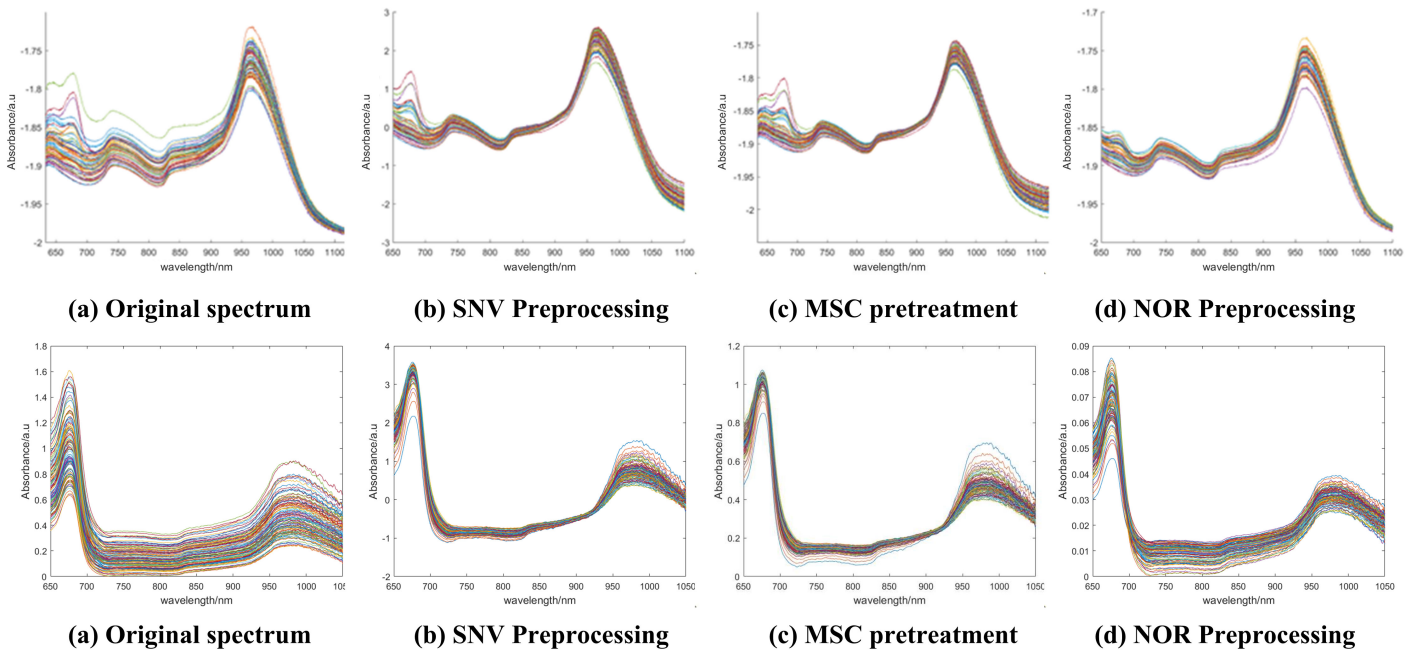


Figure 6. Comparison of spectral response characteristics under different signal preprocessing strategies. The subplots illustrate the effects of (a) Raw original spectrum, (b) Standard Normal Variate (SNV), (c) Multiplicative Scatter Correction (MSC), and (d) Normalization (NOR) algorithms on baseline correction and feature convergence for the fruit samples.

Figure 6.

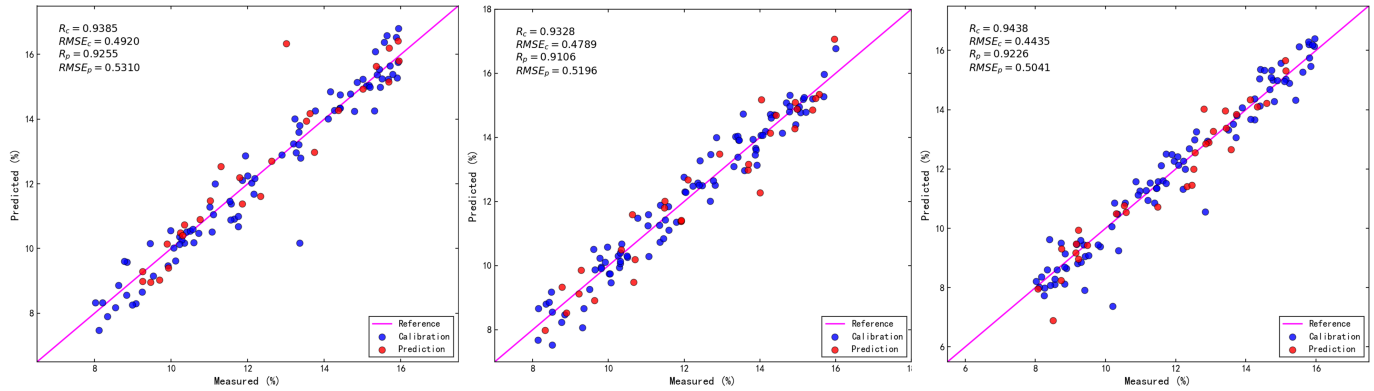
Table 4. Modeling results for peaches and Nanshui pears.

	Preprocessing	R_c^2	R_v^2	*RMSEC*	*RMSEV*
Pinggu Peaches	NONE	0.9055	0.8705	0.6398	0.7009
	NOR	0.9216	0.9102	0.5797	0.6257
	SNV	0.8957	0.8905	0.6318	0.7506
	MSC	0.9299	0.9028	0.5037	0.5228
	MSC+CARS	0.9328	0.9106	0.4789	0.5196
NanShui Pear	NONE	0.8816	0.8702	0.6197	0.8657
	NOR	0.9090	0.9011	0.6132	0.7073
	SNV	0.9264	0.9202	0.5507	0.5982
	MSC	0.9181	0.9111	0.5684	0.6275
	SNV+CARS	0.9438	0.9226	0.4435	0.5041

Spectral morphology analysis demonstrated that preprocessing effectively mitigated baseline drift and significantly enhanced the resolution of characteristic peaks. Furthermore, the quantitative modeling results presented in Table 4 underscore that the optimal signal processing strategy is dictated by the distinct epidermal microstructures of each fruit variety.

For Beijing Pinggu Peaches, the validation set coefficient of determination (R_v^2) of the PLSR model constructed from raw spectra was only 0.8705, accompanied by a high Root Mean Square Error of Validation. This indicates that the dense surface pubescence generated strong non-linear diffuse scattering interference. Comparative analysis demonstrates that the Multiplicative Scatter Correction (MSC) strategy performed best in suppressing particle scattering noise, significantly reducing the RMSEV to 0.5228, which was superior to both SNV and NOR methods. Building on this, the CARS algorithm was introduced for feature wavelength selection. This effectively eliminated redundant variables, optimizing model performance: R_v^2 increased to 0.9106, and RMSEV further decreased to 0.5196.

Regarding Nanshui Pears, although the raw spectra model performed adequately on the calibration set, the validation set error was substantial. This reflects a weak generalization capability, making the model susceptible to the influence of skin thickness and the uneven distribution of internal stone cells. Following preprocessing with Standard Normal Variate (SNV), the additive effects caused by optical path variations were effectively corrected. This significantly elevated the R_v^2 to 0.9202 while drastically reducing the RMSEV to 0.5982. When further combined with the CARS algorithm, the model accuracy was propelled to the highest level across all experimental groups. The regression performance of the final optimized models, comparing predicted versus true values, is illustrated in Figure 7. The scatter points are clustered closely around the ideal diagonal, confirming that this combined strategy achieves high-precision prediction of internal quality across multi-variety fruits.



(a) Apple modeling results **(b) Beijing Pinggu peaches Modeling Results** **(c) Nan Shui Li modeling results**

Figure 7. Scatter plots of the optimal PLSR models for the three fruit varieties. (a) Qixia Red Fuji Apple using IDC+CARS; (b) Pinggu Peach using MSC+CARS; (c) Nanshui Pear using SNV+CARS. The linear regression results (R_p^2 and $RMSEP$) demonstrate the effectiveness of the differentiated modeling strategies.

Table 5. Measurement error and dynamic grasping accuracy statistics for three fruit types.

Fruit Variety	Experimental Setup		System Performance Metrics		
	Sample Size (N)	Ref. Diameter (mm)	Visual Error (mm)	Positioning Error (mm)	Grasping Success Rate (%)
Qixia Red Fuji	30	87.5	1.2	2.8	96
Beijing Pinggu Peach	30	73.6	2.2	3.7	93
Nanshui Pear	30	67.7	1.8	3.7	86

Note: Ref. Diameter = Mean Reference Diameter

4.3 System Integration Performance and Independent Verification

To evaluate the integrated robustness of the multimodal grading system in a continuous operation environment, this study constructed an independent validation set. A total of N=90 samples (30 samples each from Qixia Red Fuji apples, Beijing Pinggu peaches, and Nanshui Pears) were selected for end-to-end testing under continuous conveyor belt operation. The evaluation primarily focused on three core dimensions: external dimension measurement accuracy, dynamic grasping stability, and internal quality SSC detection accuracy.

Utilizing the fruit diameter and localization coordinates provided by the improved YOLOv11-TFE visual system, the robotic manipulator was driven for dynamic grasping. Two key error metrics were recorded: the deviation between the machine-measured diameter and the ground truth obtained via a vernier caliper, and the Euclidean distance error between the end-effector's actual placement point and the centroid computed by the algorithm. The detailed measurement errors and dynamic grasping success rates for each fruit variety are summarized in Table 5.

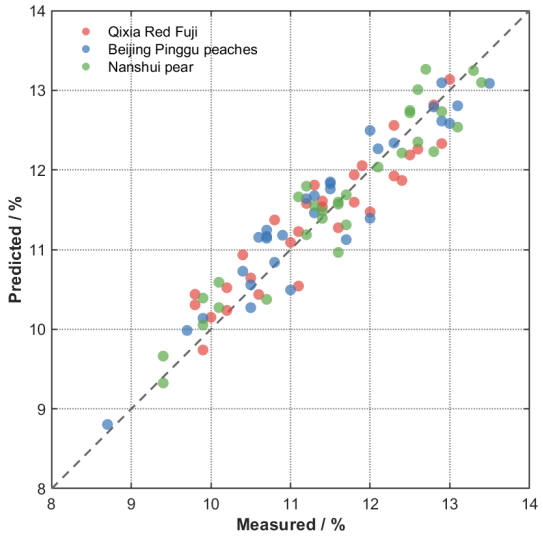


Figure 8. Independent validation results of online SSC detection for three fruit varieties. The scatter plots illustrate the correlation between reference measured values and online predicted values for (a) Qixia Red Fuji Apple, (b) Pinggu Peach, and (c) Nanshui Pear under dynamic operating conditions.

As illustrated in Figure 8, even under dynamic, online operation conditions, the system's Prediction Coefficient of Determination (R_p^2) for Qixia Red Fuji

apples remained above 0.93. Furthermore, for fruits with complex surface textures, such as the Beijing Pinggu Peach and Nanshui Pear, the Root Mean Square Error of Prediction ($RMSEP$) was consistently maintained below 0.55. This signifies that the system is capable of differentiating fruit soluble solids content (SSC) with a precision of $\pm$0.5%, a performance level that fully satisfies the operational requirements for distinguishing between "Premium Grade" and "Grade I" products in commercial grading.

5 Conclusion

This study successfully developed an intelligent robotic system integrating visual recognition, spectral detection, and flexible grasping, effectively addressing critical challenges in post-harvest grading, such as visual feature similarity, invisible internal quality, and mechanical damage risks. Through targeted algorithmic optimization and deep multimodal integration, the system achieved precise, non-destructive grading for Qixia Red Fuji, Beijing Pinggu Peach, and Nanshui Pear. The key conclusions are summarized as follows:

1) Improved YOLOv11-TFE Effectively Resolves Fine-Grained Visual Confusion
To address the specific challenge of distinguishing visually similar fruits, the proposed YOLOv11-TFE algorithm strategically integrates the SimAM attention mechanism and the DySample upsampling operator. This combination explicitly decouples texture features (gloss vs. tomentum) and enhances the capture of irregular contours. The algorithm elevated the average detection accuracy (mAP@0.5) to 94.6% while maintaining an industrial-grade speed of 95 FPS. Notably, it achieved significant accuracy improvements of 2.8 and 6.5 percentage points over the baseline for the highly confusing Red Fuji and Beijing Pinggu Peach categories, respectively, while the recall for the irregularly shaped Nanshui Pear stabilized at 86.5%.

2) High-Precision Internal Quality Models Established via Targeted Preprocessing Optimization
Instead of a one-size-fits-all approach, a targeted optimization strategy for spectral preprocessing was implemented to accommodate distinct fruit peel optical properties (cuticle, pubescence, and stone cells). For Qixia Red Fuji, the IDC-PLSR model was identified as the optimal solution (Rv2: 0.9255; RMSEV: 0.5310). For structurally complex fruits (Peach/Pear), the SNV preprocessing combined with CARS feature selection was proven effective in mitigating strong scattering interference. Independent validation confirmed that the RMSEP for Soluble Solids Content (SSC) across all three varieties was controlled within 0.65%, with an RPD $>$ 2.8, fully satisfying the accuracy requirements for online grading.

3) Multimodal System Integration Validated for Low-Damage, High-Precision Operation
Under continuous conveyor-belt conditions, the integrated system exhibited strong mechanical stability and measurement robustness. The vision module achieved a mean absolute percentage error (MAPE) of 1.7% for fruit diameter measurement, while the flexible gripping unit maintained an average positioning error of 2.67 mm. In terms of dynamic grasping performance, the system achieved a combined success rate of 94.5% for spherical fruits (Qixia Red Fuji Apples and Beijing Pinggu Peaches). For Nanshui Pears, due to their more irregular geometry, the success rate was maintained at 86.0%. Crucially, the single-fruit processing damage rate remained at 0%, confirming that the system satisfies design specifications and provides a reliable engineering solution for the automated processing of high-value fruits.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mukhiddinov, M., Muminov, A., & Cho, J. (2022). Improved classification approach for fruits and vegetables freshness based on deep learning. *Sensors, 22*(21), 8192. [CrossRef]

[2] Arunima, P. L., Gopinath, P. P., Lekshmi, P. G., & Esakkimuthu, M. (2024). Digital assessment of post-harvest Nendran banana for faster grading:

CNN-based ripeness classification model. *Postharvest Biology and Technology, 214*, 112972. [CrossRef]

[3] Ghonimy, M., Alayouni, R., Alshehry, G., Barakat, H., & Ibrahim, M. M. (2025). Integrated Physical–Mechanical Characterization of Fruits for Enhancing Post-Harvest Quality and Handling Efficiency. *Foods, 14*(14), 2521. [CrossRef]

[4] Kang, H., Zhou, H., Wang, X., & Chen, C. (2020). Real-time fruit recognition and grasping estimation for robotic apple harvesting. *Sensors, 20*(19), 5670. [CrossRef]

[5] Tang, Y., Huang, W., Tan, Z., Chen, W., Wei, S., Zhuang, J., ... & Ren, J. (2025). Citrus fruit detection based on an improved YOLOv5 under natural orchard conditions. *International Journal of Agricultural and Biological Engineering, 18*(3), 176-185. [CrossRef]

[6] Wang, X., Huang, Y., Wei, S., Xu, W., Zhu, X., Mu, J., & Chen, X. (2025). ELD-YOLO: A lightweight framework for detecting occluded mandarin fruits in plant research. *Plants, 14*(11), 1729. [CrossRef]

[7] Yu, H., Qian, C., Chen, Z., Chen, J., & Zhao, Y. (2025). Ripe-Detection: A lightweight method for strawberry ripeness detection. *Agronomy, 15*(7), 1645. [CrossRef]

[8] Zhang, Z., Cheng, H., Geng, W., & Guan, J. (2025). Research advances in hyperspectral imaging technology for fruit quality assessment. *Smart Agriculture, 7*(5), 52–66. [CrossRef]

[9] Jiang, X., Zhu, M., Yao, J., Zhang, Y., & Liu, Y. (2022). Calibration of near infrared spectroscopy of apples with different fruit sizes to improve soluble solids content model performance. *Foods, 11*(13), 1923. [CrossRef]

[10] Zong, H., Tian, S., Liu, Z., Guo, B., Wang, Y., Ren, J., ... & Zhang, E. (2025). Dual-branch feature-enhanced neural network for apple SSC estimation from hyperspectral imaging. *Journal of Food Composition and Analysis*, 108700. [CrossRef]

[11] Guo, Z., Zhai, L., Zou, Y., Sun, C., Jayan, H., El-Seedi, H. R., ... & Zou, X. (2024). Comparative study of Vis/NIR reflectance and transmittance method for on-line detection of strawberry SSC. *Computers and Electronics in Agriculture, 218*, 108744. [CrossRef]

[12] He, W., Huang, W., Li, Y., Latinović, N., Zhang, Y., & Zhang, X. (2025). Multimodal information fusion and precision harvesting system for fruit growth driven by flexible optoelectronic sensing and hierarchical attention networks. *Computers and Electronics in Agriculture, 238*, 110784. [CrossRef]

[13] Huang, W., Xia, J., Wang, Y., Jin, X., Zhu, H., & Zhang, X. (2024). Flexible multimode sensors based on hierarchical microstructures enable non-destructive grading of fruits in cold chain logistics. *Materials Today Sustainability, 25*, 100691. [CrossRef]

[14] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence, 39*(6), 1137-1149. [CrossRef]

[15] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[16] Teng, H., Sun, F., Wu, H., Lv, D., Lv, Q., Feng, F., ... & Li, X. (2025). DS-YOLO: A Lightweight Strawberry Fruit Detection Algorithm. *Agronomy, 15*(9), 2226. [CrossRef]

[17] Liao, Y., Li, L., Xiao, H., Xu, F., Shan, B., & Yin, H. (2025). YOLO-MECD: Citrus detection algorithm based on YOLOv11. *Agronomy, 15*(3), 687. [CrossRef]

[18] Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

[19] Liu, Y., Wang, Q., Gao, X., & Xie, A. (2019). Total phenolic content prediction in Flos Lonicerae using hyperspectral imaging combined with wavelengths selection methods. *Journal of Food Process Engineering, 42*(6), e13224. [CrossRef]

[20] Cheng, J. H., & Sun, D. W. (2017). Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food engineering reviews, 9*(1), 36-49. [CrossRef]

Zhenhao Ma is currently pursuing the M.S. degree in Mechanical Engineering with the College of Engineering, China Agricultural University, Beijing, China. Since 2023, he has been focusing on research related to agricultural and livestock product inspection, robotics, and machine vision. (Email: 15238162825@163.com)

Bin Zhang received the B.E. degree from Shenyang Institute of Technology, Shenyang, China, in 1985, and the Ph.D. degree in Engineering from China Agricultural University, Beijing, China, in 1995. He joined the faculty of China Agricultural University in July 1985. Since December 2004, he has been a Professor with the Department of Mechanical Design and Manufacturing Engineering, College of Engineering. His current research interests include agricultural robotics and related fields. (Email: zhangbin64@cau.edu.cn)

Tianzhen Yin is currently pursuing the Ph.D. degree in Mechanical Engineering with the College of Engineering, China Agricultural University, Beijing, China. Since 2022, his research has focused on the non-destructive testing of livestock products. (Email: email@email.com)