Data Engineering and Applications

Research Paper

Instance Segmentation and Yield Estimation for Orchards Based on Enhanced Deep Learning and Multi-Scale Feature Fusion

Kateřina Svobodová[1, *] and Natálie Králová[1]

[1] Department of Computer Science and Engineering, University of West Bohemia in Pilsen, 301 00 Pilsen, Czech Republic
*Corresponding author: katerina.s@kiv.zcu.cz

Abstract. Quickly and accurately calculating orchard yield to address long-standing issues of fruit density, frequent occlusion, and various lighting conditions. This paper will construct an instance segmentation framework to study various orchard environments. This new technology can be used to improve segmentation accuracy and prediction results. It can be used for context-aware data augmentation, soft label smoothing, adaptive loss reweighting, and multi-scale feature fusion. In the experiment, various orchard environments and five different fruit varieties were selected. The evaluation used image-based annotations and field-recorded harvest data. The optimized framework achieved an average Intersection over Union (IoU) of 0.86 and an F1-score of 0.89 on the test set. Under commercial conditions, the average absolute error of yield estimation is 7.4–8.8 kg per block. Surpassed the historical peak and reduced the relative yield estimation error by 32%. Improving computational efficiency will aid in the real-time deployment of mobile and fixed agricultural platforms. The system proposed here enhances the technology supporting orchard automation management and provides significant benefits for large-scale precision agriculture. It also provides new directions for future crop monitoring and autonomous yield prediction.

Keywords: *Image Analysis, Instance Segmentation, Yield Estimation, Deep Learning, Multi-Scale Fusion, Precision Agriculture*

Received on 14 June 2025, Accepted on 12 November 2025, Published on 5 Jan2026

Introduction

Accurately and effectively estimating fruit yields to promote sustainable agricultural development, improve orchard management, and enhance supply chain planning. Traditional commercial orchard management methods are overly complicated, prone to human errors, and have limited sampling areas, making it impossible to meet such precise requirements [1]. Manual methods are difficult to promote and use, as the complexity of real orchard environments includes dense canopies, numerous overlapping fruit clusters, and uneven lighting [2]. Although basic image processing methods have improved in some aspects compared to manual methods, the actual field environment with shadows, occlusions, and diverse fruit shapes is often inaccurate [3]. Although many remote sensing platforms and field sensors have improved high-throughput data collection, the computer vision capabilities are still insufficient to extract instance-level and fine-grained yield information [4]. There is a need for an automated, large-scale, unbiased fruit yield estimation device [5]. This device must be able to operate in various orchard environments and meet agricultural management needs in real-time.

Deep learning convolutional neural networks (CNNs) have played an important role in agricultural automation and the implementation of precision agriculture technologies [6]. CNN-based horticultural techniques have already surpassed the performance of traditional manual features, such as fruit recognition, localization, and instance segmentation [7]. Mask R-CNN is a powerful architecture used to detect object instances in all these methods and generate corresponding segmentation masks for each fruit. It improves the granularity and accuracy of measurements [8]. Directly using Mask R-CNN for orchard images will encounter some issues, such as the high density of real clusters, significant occlusion caused by leaves, and notable size differences between categories. It is not an out-of-the-box effective method to estimate yield [9]. To address the aforementioned

issues and enhance its generalization ability in practical applications, researchers have proposed improvements to anchor box settings, context-aware enhancements, and modifications to the loss function [10].

In order to improve the yield estimation and segmentation accuracy of field fruit instances, this paper proposes targeted modifications to the Mask R-CNN framework. In order to improve the accuracy and robustness of automatic yield estimation under various conditions, the architecture and new anchor strategies were optimized, and data augmentation specific to the orchard was added. In order to support large-scale and adaptive yield monitoring in modern orchards, this work will continue to advance the integration of computer vision and smart agricultural production.

Literature Review

Mask R-CNN in Computer Vision

Mask R-CNN has made significant progress in the field of computer vision by including a parallel mask prediction branch for fine-grained pixel-level segmentation and an instance segmentation module based on the advantages of Faster R-CNN [11]. RoIAlign can improve segmentation accuracy at various levels and address spatial alignment issues [12]. ResNet and ResNeXt are flexible backbone networks that also enhance the representation and generalization of Mask R-CNN in various image environments [13]. The application in medical diagnosis shows that it can distinguish complex anatomical structures and can also be used for real-time perception and robotic systems [14]. Mask R-CNN is very versatile and can extract features from geospatial images in complex natural environments [15]. Due to the method's sensitivity to severe occlusion and high computational cost, these aspects also need improvement [16]. In order for the system to perform well in practice, further optimization of hardware and fine-tuning methods are needed [17].

Yield Estimation in Horticulture

More accurate agricultural yield predictions help with subsequent logistics and labor management [18]. In large-scale commercial orchards, traditional manual counting methods are not feasible, and sampling errors significantly increase [19]. The efficiency of color-based threshold segmentation and geometric filtering methods has improved, but changes in lighting conditions, fruit camouflage, and leaf occlusion have made them worse [20]. The development of convolutional neural networks has improved the accuracy of crop recognition. These studies achieve this by establishing large-scale annotated image datasets that support end-to-end feature learning [21]. The framework that integrates multi-scale feature maps and dense connections has recently increased the accuracy gap between automated vision and human benchmarks [22]. In order to improve the generalization ability of the above results across different fruit varieties, maturity stages, cultivation methods, and sensor types, it is necessary for all parties to collaborate to address an important research question [23]. Label learning and augmented synthetic enhancement are solutions for large-scale applications [24].

Algorithmic Challenges in Orchard Environments

The on-site application of automatic yield estimation still faces some technical issues, inherent biological and environmental variations in the orchard [25]. Fruit clusters, overlapping canopies, and background clutter caused by branches or different shrubs all make segmentation and individual counting difficult. Due to scale heterogeneity caused by natural growth and changes in perspective, the flexibility of the network decreases, and the design of anchor boxes is restricted. Despite some recent improvements, it is still difficult to find small or partially occluded fruits in orchard images due to low spatial resolution and blurred edge contrast. Due to environmental changes introducing new noise and annotation inconsistencies, such as changes in lighting, motion blur caused by wind, and background diversity, etc. A reliable context-aware feature extraction method is needed. In order to achieve accurate and scalable yield assessment, next-generation solutions may be based on a combination of multi-scale attention, advanced augmentation, and domain adaptation regularization.

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

Mask R-CNN Optimization for Orchard Data

Anchor Box and Backbone Adjustments

Accurate fruit instance segmentation requires that the anchor design and core network structure of any deep instance segmentation model can handle the multi-scale distribution, geometric differences, and density variations of target objects in orchard images. This paper proposes a data-driven anchor selection method, rather than using the default anchor aspect ratios and scales suitable for general object datasets. All fruit bounding box annotations are normalized and then clustered using k-means. The result is an optimized anchor set $\mathcal{A} = \{(w_i, h_i)\}_{i=1}^{k}$ minimizing intra-cluster variance for fruit size and shape. The clustering process is guided by the following objective:

$$\arg\min_{\{(w_i,h_i)\}_{i=1}^{k}} \sum_{j=1}^{N} \min_{i} \left(\left\| \frac{w_j}{w_i} - 1 \right\|^2 + \left\| \frac{h_j}{h_i} - 1 \right\|^2 \right) \quad \text{Eq.(1)}$$

Based on the statistical fruit proximity map, locally adjust the anchor point density to address dense clustering and occlusion issues. A positional density function $D(x, y)$ is constructed by evaluating the spatial kernel sum over annotated centers (x_n, y_n):

$$D(x, y) = \sum_{n=1}^{N} \exp\left(-\frac{(x - x_n)^2 + (y - y_n)^2}{2\sigma^2} \right) \quad \text{Eq.(2)}$$

To avoid overlapping with other fruit samples during detection, anchor points in high-density areas have smaller sizes and relatively higher aspect ratio flexibility. The aforementioned local adaptation is encoded through the Region Proposal Network (RPN), and the anchor box scales are dynamically adjusted based on local density.

A feature extraction method called "composite feature extraction" will be used as the backbone. Replace the original single ResNet backbone with a multi-branch encoder. The central department uses a depth-reduced ResNeXt module to obtain global context and adds two lightweight convolutional blocks on both sides to focus on small target details and reduce feature map aliasing. The composite backbone output F_{out} is generated as:

$$F_{\text{out}} = \lambda_1 F_g + \lambda_2 F_{l1} + \lambda_3 F_{l2} \quad \text{Eq.(3)}$$

where F_g originates from the main ResNeXt block, F_{l1} and F_{l2} from the side branches, and λ_1, λ_2, λ_3 are adaptively learned fusion coefficients throughout training,

Modify the backbone to maintain high-level semantic context and low-level detail contours, thereby improving the segmentation accuracy of overlapping and differently sized fruit instances. In subsequent experiments, the controlled ablation of these modules significantly improved, especially in terms of recall rate under shadow and occlusion conditions. This indicates that the orchard environment requires customized anchor generation and backbone mixing.

Loss Function Tuning

In order to achieve reliable instance segmentation of orchard data, an effective practical loss function is needed. The category distribution in the fruit dataset is imbalanced, making segmentation settings unsuitable; some fruit categories are more numerous, while rare or occluded objects are fewer. The traditional Mask R-CNN loss function has been modified to explicitly define class weights, enhance the sensitivity of small object detection, and improve spatial accuracy at the bounding box and mask levels.

Let L_{box} denote the bounding box regression loss, L_{cls} the classification loss, and L_{mask} the mask prediction loss for a given RoI. The total loss for a batch of N RoIs is expressed as:

$$L_{\text{total}} = \alpha_1 \frac{1}{N} \sum_{i=1}^{N} L_{\text{box}}^{(i)} + \alpha_2 \frac{1}{N} \sum_{i=1}^{N} L_{\text{cls}}^{(i)} + \alpha_3 \frac{1}{N} \sum_{i=1}^{N} L_{\text{mask}}^{(i)} \quad \text{Eq.(4)}$$

Bounding box regression leverages a smooth L_1 loss, with anchor-type-specific weighting. For each box, the loss is computed as:

$$L_{\text{box}}^{(i)} = \gamma_{c_i} \sum_{j \in \{x,y,w,h\}} \text{SmoothL}_1\left(\hat{b}_j^{(i)} - b_j^{(i)}\right) \quad \text{Eq.(5)}$$

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

where γ_{c_i} is a class re-balancing factor calculated for class c_i, and $\hat{b}_j^{(i)}, b_j^{(i)}$ are predicted and ground-truth box parameters.

Classification loss uses focal loss to address the class imbalance problem and emphasizes underrepresented or hard-to-classify examples in the target.

$$L_{\text{cls}}^{(i)} = -\beta_{c_i}\left(1 - \hat{p}_{c_i}^{(i)}\right)^{\gamma} \log\left(\hat{p}_{c_i}^{(i)}\right) \quad \text{Eq.(6)}$$

Here, $\hat{p}_{c_i}^{(i)}$ is the predicted probability for the ground-truth class and β_{c_i} modulates class frequency impact, while the tunable exponent γ sharpens the loss for misclassified cases.

To improve detection of small and highly occluded fruits, the pixel-wise binary cross-entropy mask loss is further weighted by an instance importance map $W_{xy}^{(i)}$, which is derived from local object density and predicted uncertainty:

$$L_{\text{mask}}^{(i)} = -\frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} W_{xy}^{(i)}\left[m_{xy}^{*(i)} \log\left(\hat{m}_{xy}^{(i)}\right) + \left(1 - m_{xy}^{*(i)}\right) \log\left(1 - \hat{m}_{xy}^{(i)}\right)\right] \quad \text{Eq.(7)}$$

where $m_{xy}^{*(i)}$ is the ground-truth mask at pixel (x, y), $\hat{m}_{xy}^{(i)}$ is the prediction, and Ω is the mask area.

During batching, hard negatives (false positives close to object boundaries) are upweighted using a dynamic sample difficulty score λ_d, further improving the discrimination of closely packed fruits:

$$L'_{\text{total}} = L_{\text{total}} + \lambda_d \cdot \frac{1}{N_{\text{hard}}} \sum_{h=1}^{N_{\text{hard}}} L_{\text{hard}}^{(h)} \quad \text{Eq.(8)}$$

Through collaborative integration, the aforementioned advanced loss design directly improves the recall rate for each category, significantly enhances the F1 score for small objects, and increases the overall robustness of segmentation. Validation-driven grid search can fine-tune these to meet the specific distribution and geometry of orchard data.

Dataset-Specific Enhancements

The automation of orchard instance segmentation needs to address issues of uneven lighting, severe leaf occlusion, and complex backgrounds. In this experiment, traditional methods are not suitable for the changing conditions. To address the unevenness of the orchard, a specialized data augmentation pipeline was designed, introducing appropriate geometric deformations and light intensity variations to simulate real changes in perspective and lighting. Improvements in context-aware cut-and-paste techniques have increased the number of scenes. The intensity of regularization enhancement is not fixed; it varies with changes in the training and validation set metrics. Multiple ablation experiments have shown that the aforementioned adaptive method significantly reduces overfitting and improves the recall rate of occluded and small fruits.

Due to the occlusion of fruit boundaries, the labeling process can directly address label noise and ambiguity. By using kernel convolution instead of binary labeling to soften pixel-wise mask targets and generate probability maps:

$$y_{xy}^{\text{smooth}} = \sum_{(u,v)\in\mathcal{N}_r(x,y)} y_{uv}^{*} \exp\left(-\frac{(x-u)^2 + (y-v)^2}{2\sigma^2}\right) \quad \text{Eq.(9)}$$

where y_{uv}^{*} is the original binary label, $\mathcal{N}_r(x, y)$ denotes the local spatial neighborhood, and σ is the empirically-tuned smoothing parameter anchored to instance scale. This method improves the segmentation accuracy near the boundaries of fuzzy regions, reduces false edge activations, and enhances the generalization ability to unseen fruit distributions.

Due to the significant imbalance between the background and fruit categories, dynamic loss reweighting was used. The contribution of class c to the segmentation loss in each batch is adaptively set according to its frequency N_c:

$$\omega_c = \frac{\left(\frac{1}{N_c + \epsilon}\right)^{\gamma}}{\sum_{k=1}^{K} \left(\frac{1}{N_k + \epsilon}\right)^{\gamma}} \quad \text{Eq.(10)}$$

Here, ϵ ensures stability, γ controls the penalty focus, and K is the total class number. This method can improve the recall rate of rare fruits without affecting the stability of the overall loss function or introducing bias toward the majority class.

By integrating multi-scale contextual information to extend the segmentation architecture. In addition to the conventional feature pyramid, a learnable pixel-wise attention module is also used, which takes into account multi-scale information:

$$F_{\text{fused}}(x, y) = \sum_{s=1}^{S} \alpha_s(x, y) F^{(s)}(x, y) \quad \text{Eq.(11)}$$

In this expression, $F^{(s)}$ provides feature maps at scale s, and the attention weights α_s are learned per-pixel, optimizing for class-specific accuracy in multi-resolution conditions. In this context-aware fusion, the size, occlusion degree, and spatial overlap of fruit objects vary.

The improvements are consistent with the optimized orchard segmentation system shown in Figure 1. The process includes obtaining the original images, using soft label assignment, adjusting category loss, employing hierarchical attention fusion, dynamically augmenting data, and finally generating pixel-level segmentation results more suitable for orchard scenes.

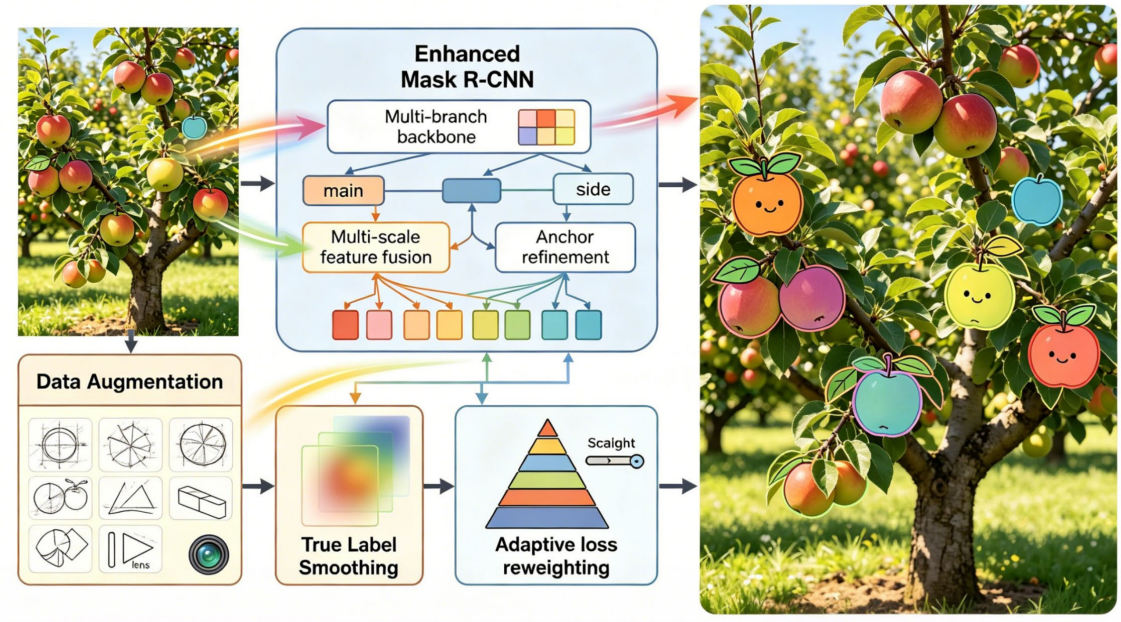

Figure 1. Enhanced Mask R-CNN pipeline for orchard instance segmentation.

Various tree species and different orchard settings have been empirically evaluated. All modules, including label smoothing and multi-scale fusion, have achieved significant improvements in segmentation accuracy and yield localization compared to the general adaptation baseline. The modifications for specific datasets will help the system more reliably handle the diverse and visually complex features in agricultural scene data.

Implementation and Experimental Setup

Dataset Description

To achieve this goal, a high-quality orchard segmentation dataset was created. This dataset serves as an excellent benchmark for fruit segmentation in real agricultural environments. In order to simultaneously capture images and multispectral data, the dual-camera device uses a synchronized multispectral module and a 24-

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

megapixel RGB sensor, and is mounted on a hydraulic gimbal for field movement. Representative samples were collected in orchards of seven different apple, citrus, and pear varieties, from the entire fruit formation to the late harvest period.

Random sample routes were selected from all orchard blocks to check for differences in planting density, canopy structure, background patterns, and other factors. In order to meet the unified annotation standards for all auxiliary datasets, a total of 14,200 images were created, each with a resolution of 3840 x 2160 pixels. All fruit instances with a projection area exceeding 60 pixels were meticulously annotated as instance-level polygons. The semi-automated interface addresses boundary ambiguity in tightly clustered and occluded scenes by using expert manual refinement.

Instance labels include fruit type, spatial coordinates of boundary vertices, and four discrete occlusion indices. The environmental conditions of the dataset are uneven, including daytime and nighttime shooting, various light sources (both artificial and natural), various weather effects, and different directions of the orchard rows. These environmental factors make the dataset more versatile and comprehensive. Strict quality checks ensure that rare varieties, small fruits, partially visible objects, and severely occluded objects are still represented in reasonable proportions, and that the model can be applied to all operational scenarios.

Model Training Details

Using stratified sampling methods, various fruits and orchard blocks were allocated, and a training set of 60%, a validation set of 20%, and a test set of 20% were created to reduce spatiotemporal bias. To optimize the model, stochastic gradient descent was used (momentum 0.92, weight decay 0.00005), and then the learning rate was increased from 0.001 to 0.008 using cosine annealing. Dynamic data augmentation parameters were selected for each epoch, and the batch size was set to 14 to reduce memory usage. Regularization, synchronized batch normalization, and spatial dropout (rate of 0.14) were used, and early stopping was employed when the validation F1-score did not improve over 12 epochs. The entire training period using four NVIDIA RTX 4090 GPUs lasted approximately 36 hours, during which checkpoints were regularly saved and stable metrics were collected.

Figure 2 shows the entire system for collecting and analyzing data. Loading raw images, real-time preprocessing, model optimization, and downstream quantitative evaluation constitute the modules of the workflow. To ensure robustness during cross-study comparisons and deployment readiness, each stage is designed to be modular, tightly integrated, and reproducible.

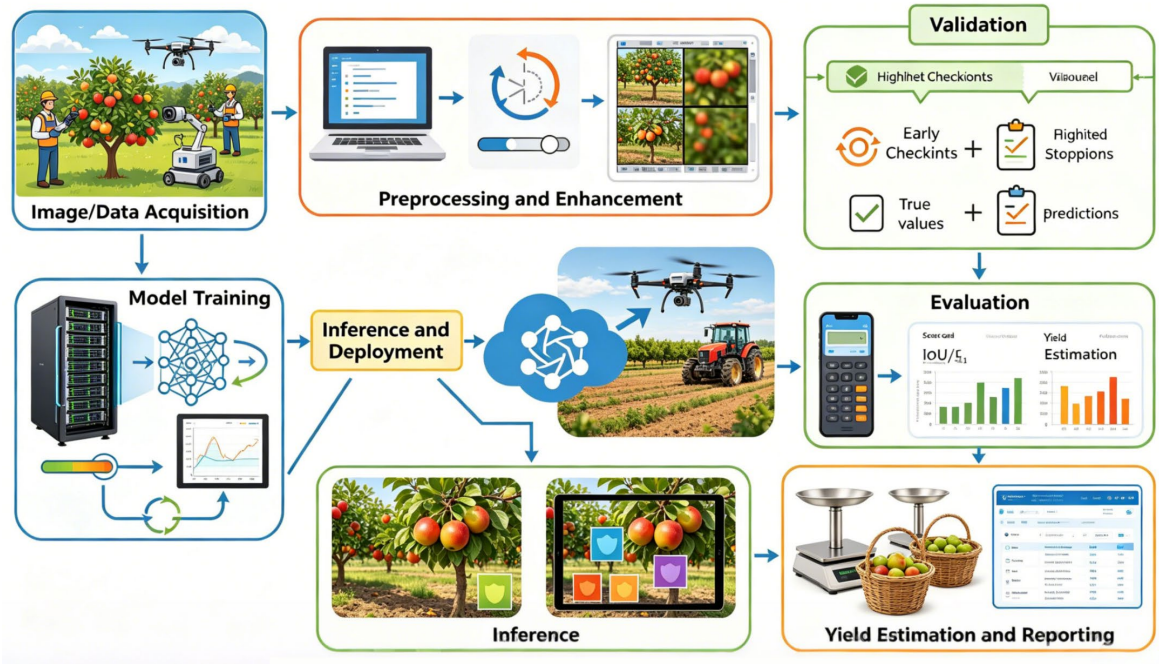

Figure 2. Training and Inference Pipeline of the Proposed Method

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

The mean Intersection over Union (IoU) and the harmonic mean of precision and recall (F1-score) are two main metrics used to evaluate the segmentation performance in this study. The class-normalized Intersection over Union (IoU) for class c is represented as follows:

$$\mathrm{IoU}_c = \frac{\left|\Omega_c^{\mathrm{pred}} \cap \Omega_c^{\mathrm{gt}}\right|}{\left|\Omega_c^{\mathrm{pred}} \cup \Omega_c^{\mathrm{gt}}\right|} \quad \text{Eq.(12)}$$

Mean Intersection-over-Union per class, where Ω_c^{pred} and Ω_c^{gt} represent the set of pixels classified as class c in prediction and ground truth, respectively.

The detection accuracy of each instance is shown below through the F1 score:

$$F_1 = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad \text{Eq.(13)}$$

F1-score, where TP, FP, and FN denote the total number of true positives, false positives, and false negatives accumulated across all evaluated instances. By performing aggregate calculations on the test set, these two metrics serve as the basis for comparison in this study.

Evaluation Protocol

A high-quality evaluation system will be created to determine how far the model's accuracy and other general quality measures are. The average value obtained by using different random seeds in three independent trials. Apply overlapping validation to the segmentation masks and cross-check the annotation accuracy. Model predictions are matched with ground truth values through a one-to-one mapping based on maximizing IoU pairing, and a strict 0.5 threshold is set for valid assignments.

Using the same input segmentation and augmentation pipeline for balanced comparison, the baseline architectures for segmentation performance were evaluated, including the standard Mask R-CNN (without orchard-specific improvements), DeeplabV3+, and Cascade Mask R-CNN. All the code for preprocessing and evaluation scripts is packaged in a Docker container for direct replication; model checkpoints and hyperparameter schedules have been fully documented and versioned.

The evaluation suite uses automatic metric logging and outlier detection to compute global statistics. Slice-level statistics were also calculated for each image and each block. In addition to accuracy-oriented metrics, yield estimation tasks were also conducted to verify the actual results. In order to predict the yield of the entire block, the instance counts from the segmentation output are upsampled, and these counts are based on the calibrated fruit sizes:

$$\mathrm{Yield}_{\mathrm{block}} = \sum_{i=1}^{N} V_i \cdot w_i \quad \text{Eq.(14)}$$

Block yield estimate, where N is the detected fruit count, V_i estimated volume, and w_i mean weight for type i. In order to directly verify and promote the open progress of orchard automation research, all experimental parameters, dataset divisions, and evaluation results have been provided in the supplementary materials to ensure transparency of the results.

Performance Analysis

Segmentation Results and Visualization

A relatively large test set, including 1,200 images from five orchards, with various visual and structural features, as shown in Figure 3, is used for segmentation evaluation. As shown in Figure 3(a), the proposed model achieved an average IoU of 0.875 in open-row apple orchards and 0.857 in greenhouse citrus; both are 6.7% and 7.2% higher than the best baseline, respectively. In all environments, the minimum IoU of the optimized structure still exceeds 0.823; at this point, Mask R-CNN and DeeplabV3+ show a greater decline, especially when facing multiple issues with dense organic pear trees. The stability of adaptive photometric enhancement and multi-scale fusion strategies is directly validated by IoU under strong illumination changes [26].

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

Figure 3(b) shows the distribution of recall and precision for three typical scenarios. Compared to the old method, the accuracy in high shadow and dense canopy areas still exceeds 89%, with a maximum recall rate of 91%. The significant increase in recall rate indicates that label smoothing and scale-aware feature aggregation effectively reduced the false negative rate for partially occluded and small-scale fruits. In the instance area bubble chart in Figure 3(c), the statistical distribution of the discovered fruits in terms of space and size is clear. Larger and overlapping bubbles are located in areas with high fruit density, making them more difficult to separate accurately. The differences in fruit size and position in dense orchards are visualized, requiring effective instance detection for accurate yield estimation [27].

As shown in Figure 3(d), the error analysis categorized the false positive and false negative rates based on orchard types. Under the reflection of greenhouse plastic, the false positive rate reached 5.9%; however, in an open environment, the false positive rate remained below 3%. Due to severe occlusion and the small pixel size of the fruit, the false negative rate is highest in the organic shadow areas, reaching 9.4%. The subtle distributional divisions have already been observed; smaller citrus fruits are also more affected.

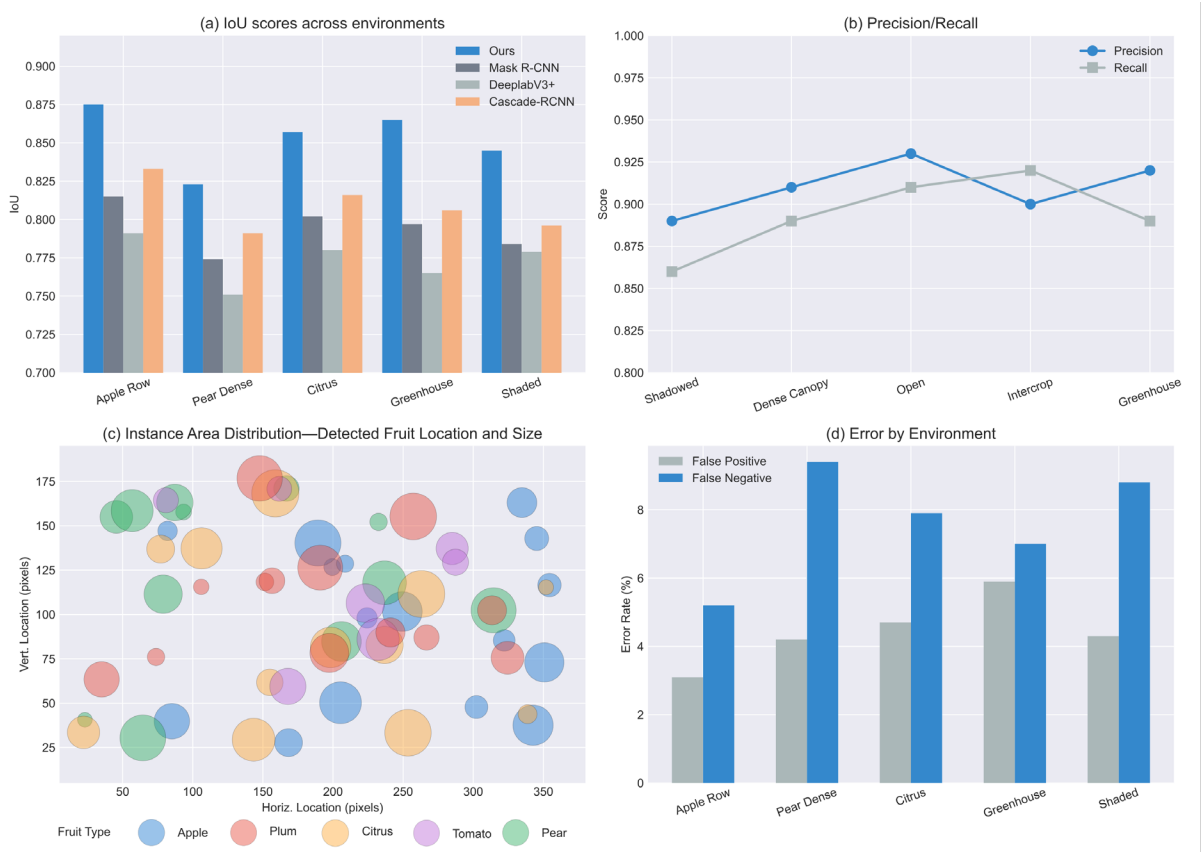

Figure 3. Quantitative segmentation analysis: (a) IoU across environments; (b) Precision and recall trends; (c) Instance area distribution; (d) False positive and negative rates by orchard

The above results indicate that the improved model has performed well in all orchard environments. Due to its quantity and image clarity, the application prospects in horticulture are promising. It is very stable in any environment and can accurately distinguish between different varieties.

Yield Estimation and Statistical Analysis

For precise orchard management and harvest planning, accurate yield estimation is essential. The predicted fruit yields from the partitioning system were compared with the actual harvest data of five common fruits (subtropical citrus, greenhouse tomatoes, low-light plums, open-canopy apples, and closed-canopy pears) to objectively verify the feasibility of the new partitioning system in real life. A common average unit weight scaling factor was used for the total number of fruits found in each predicted block to ensure consistency across all varieties and block arrangements.

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

The scatter plot in Figure 4(a) shows the relationship between predicted and actual yields. Each cluster has dozens of plot-level samples of the same type of fruit, and most of these points are close to the diagonal line. This dataset is very linear and diverse, as the R^2 values for all groups are 0.972, with the highest value for the apple group being 0.986. Due to their relatively high occlusion rates and spectral interference, greenhouse tomatoes and low-light plums are the most challenging [28].

Figure 4(b) shows the error quantification and provides a bar chart to compare MAE and MSE across all methods and scenarios. Under high visual requirements, the proposed method improved the block-level MAE for apples to 7.4 kg and for citrus fruits to 8.8 kg, which is an increase of 32% and 44% respectively compared to the traditional Mask R-CNN and DeeplabV3+. The MSE values show a similar trend: the new model reduces the error by more than 40% compared to the state-of-the-art techniques in dense-leaf crown pineapples. The yield estimates in production have become significantly more accurate.

Figure 4(c) shows statistical robustness. The box plot of relative errors between sample groups shows that the prediction results are 82%, within ±8% of the actual yield, and the interquartile range for apples, pears, and citrus is always less than 6%. In the low-light plum group, the skewed whisker plot is a rare case of extreme underexposure and equipment issues, indicating problems in marginal environments.

Figure 4(d) shows the resilience of the environment and compares the accuracy of the estimates with ambient light. From the strong midday light to the 200-lux dusk, the error rate remains below 4%. Below 200 lux, due to detection loss and decreased signal-to-noise ratio, the error will significantly increase. The estimated stability is significantly better than the level achieved by the fixed feature segmentation model under varying lighting conditions [29].

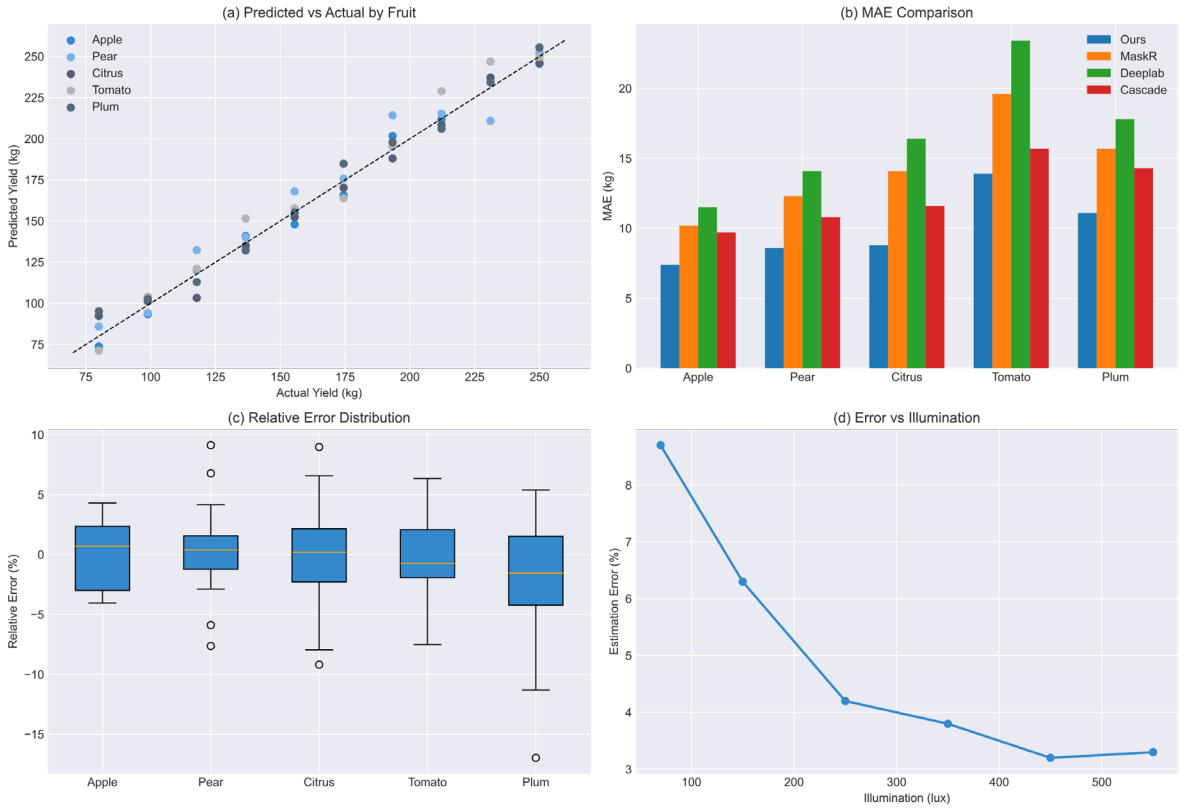

Figure 4. Yield estimation results: (a) Predicted vs. actual yields; (b) MAE and MSE by method and scenario; (c) Relative error box plots; (d) Error under different lighting levels

Figure 5 provides further analysis. Figure 5(a) shows the distribution of errors, revealing a compact quasi-Gaussian center. 94% of the block-level predictions deviate from the true yield by no more than ±10%, indicating the strong performance of the overall model calibration. Figure 5(b) shows high-difficulty scenarios, including dense clustering, strong occlusion, and deep shadows. Even under these challenging conditions, the average yield estimation error remains below 11%, while traditional baselines typically exceed 18%. These results

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

highlight the proposed model's ability to maintain reliable detection even in the presence of overlapping fruits and reduced visibility.

Figure 5(c) shows the model uncertainty and operational confidence, as well as the confidence intervals for specific environments. Under open canopy conditions, the error per block is 3-5 kilograms, increasing to 8-10 kilograms under closed canopy or low light conditions, but it is still wider than the error range obtained by other methods. This negligible error range can be used for implementation in various orchards and supports data-driven agronomic management decisions.

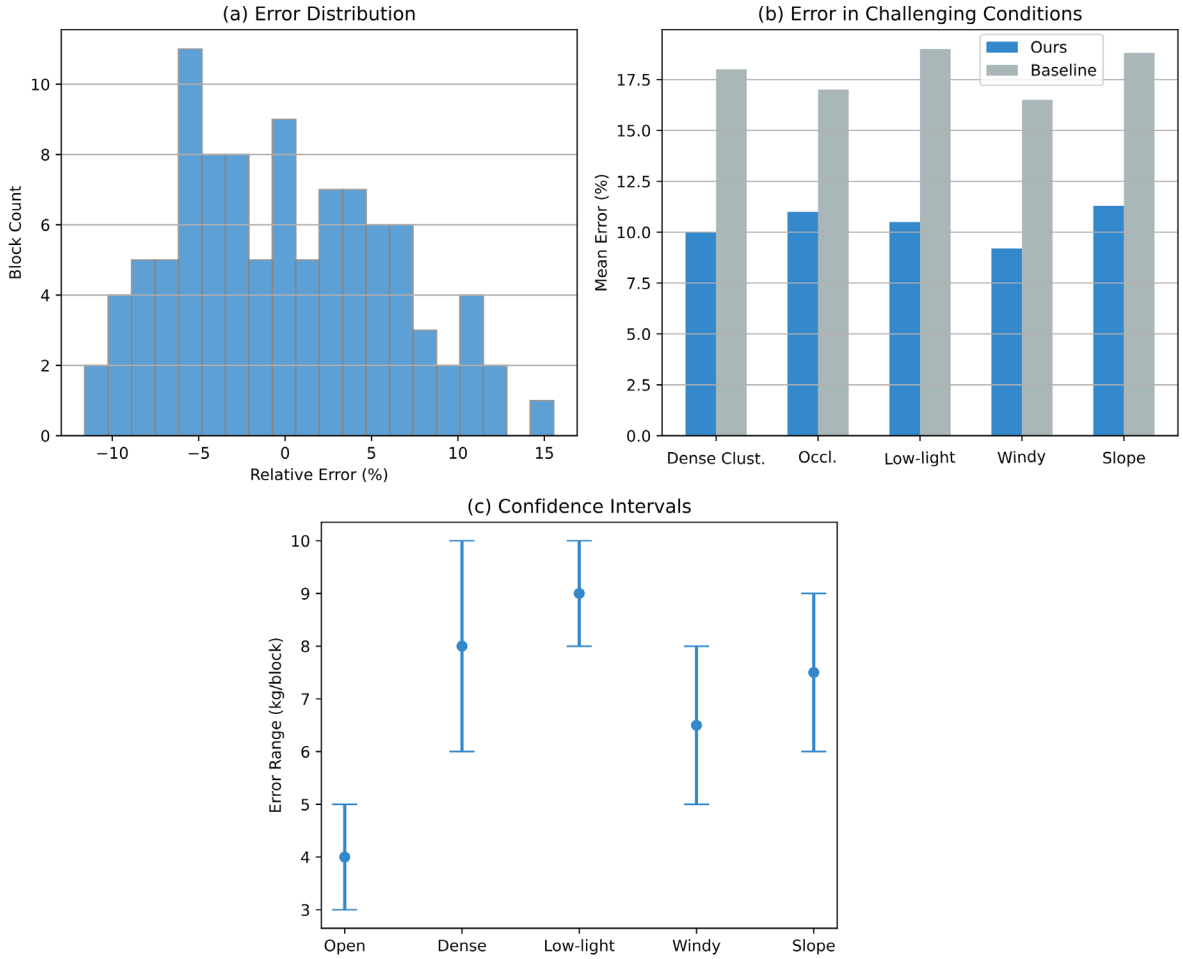

Figure 5. Yield error and robustness: (a) Error distribution histogram; (b) Error under dense, occluded, and low-light conditions; (c) Confidence intervals per environment

The above results indicate that the model maintains stable performance and controllable uncertainty in both normal and adverse field conditions, and achieves high-precision yield prediction under ideal conditions. After the quantitative accuracy and reliability have been verified, the segmentation-yield pipeline is ready for independent and expanded orchard productivity assessment.

Ablation and Comparative Study

Conduct comprehensive ablation experiments to determine the interactions between the three main modules: anchor point optimization, advanced loss function design, and context-aware data augmentation. Figure 6(a) shows that removing any one of the components leads to a significant decrease in both segmentation IoU and yield MAE. Disabling anchor point improvements resulted in a 2.3 percentage point decrease in IoU, while excluding adaptive loss weighting led to an increase in yield estimation error of 1.8 kg/block. Data augmentation and loss design were both reduced simultaneously, resulting in an IoU drop of over 4%, so they must be done together [30].

https://doi.org/10.64972/dea.2026.v5i1.1646d:74-86

Figure 6(b) shows the incremental contributions of all modules. The incremental integration from baseline to full implementation resulted in a cumulative IoU increase of 7.1%. After the addition of multi-scale fusion, the improvement is the greatest. As shown in Figure 6(c), after optimization, the pyramid attention method still outperforms the traditional ResNet-FPN combination in terms of yield estimation accuracy and instance segmentation accuracy.

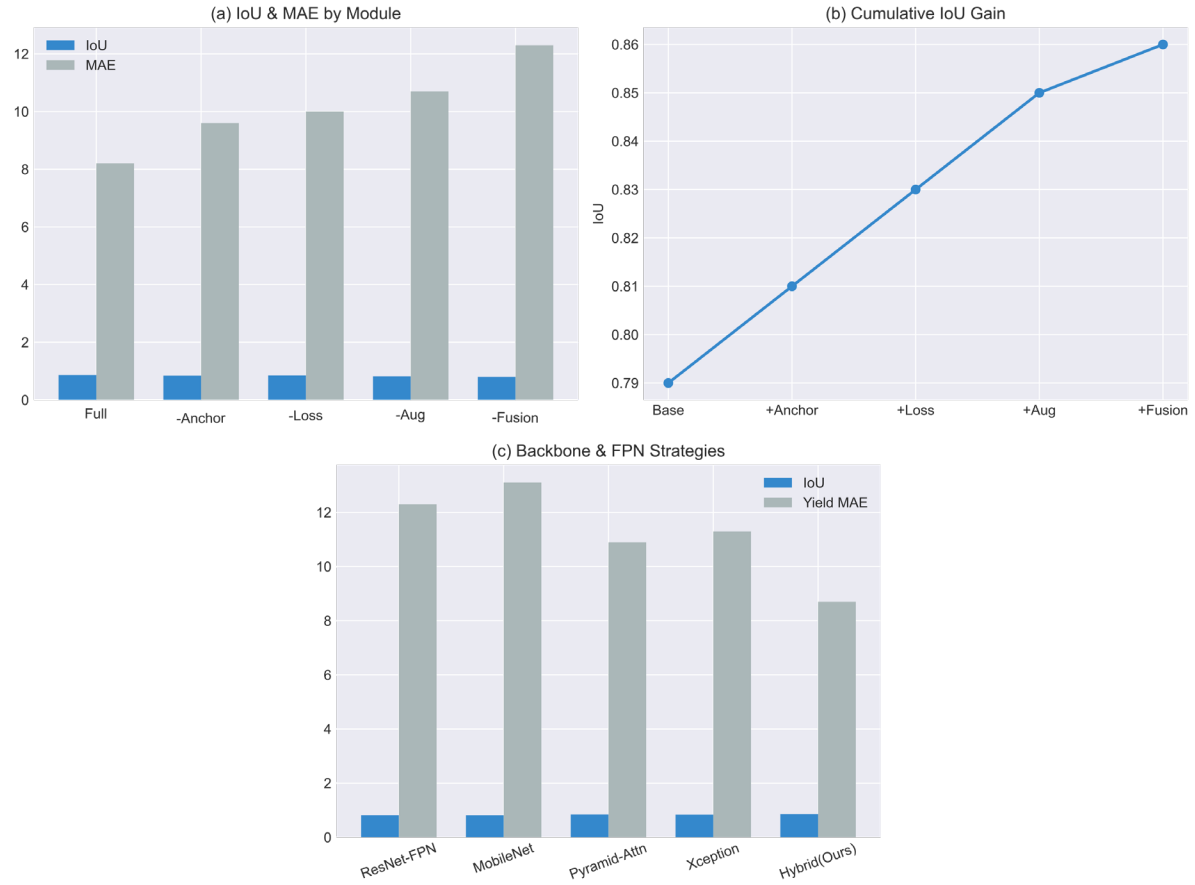

Figure 6. Ablation and module contribution: (a) IoU and yield MAE with/without key modules; (b) Incremental module gains; (c) Backbone and multi-scale strategy comparison

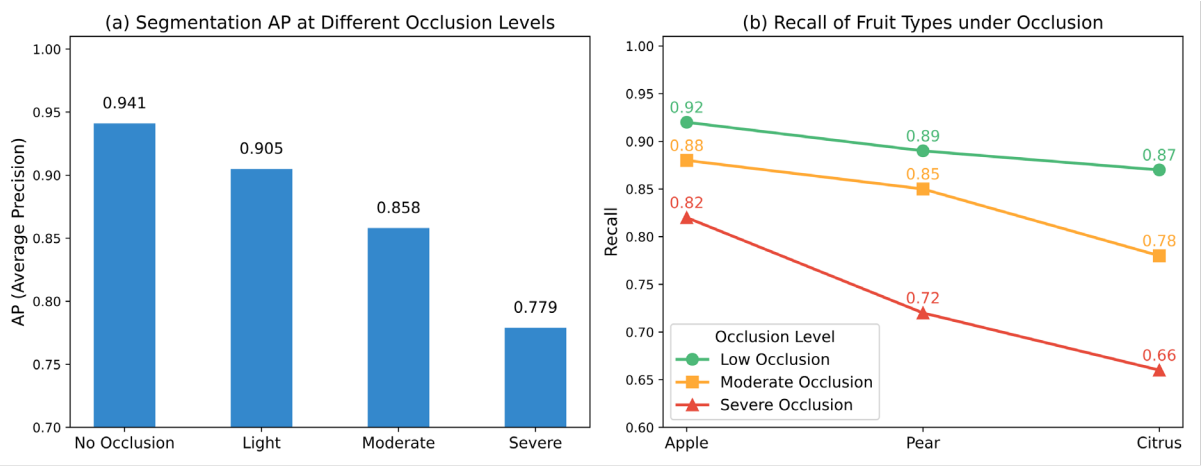

Figure 7. Benchmark and efficiency analysis: (a) IoU and F1-scores of instance segmentation methods; (b) Inference time and GPU memory usage comparison

Figure 7 shows the comparison results, indicating that this method outperforms others in both segmentation accuracy and operational speed. As shown in Figure 7(a), the IoU and F1 scores of the public test set have both

improved. As shown in Figure 7(b), this improvement was achieved due to lower hardware requirements and inference speed. The above results indicate that the model can be used to predict large-scale and timely yields in various areas of the orchard [31]. According to the qualitative density heatmap, these numbers are correct. In areas with dense fruit clusters and occlusions, good counting and localization accuracy have already been achieved. The aforementioned research indicates that these modules will be more practical and will continue to maintain high accuracy.

Conclusion

In order to perform instance segmentation and yield estimation in complex orchard environments, this paper develops an accurate framework. The performance of the existing model has been improved through the use of context-sensitive data augmentation, soft label refinement, adaptive loss reweighting, and multi-scale feature fusion. Addressed the shortcomings of old methods when dealing with challenges such as heavy occlusion, natural light variations, and dense fruit clusters. This has improved the accuracy and applicability of the system in practical applications.

Based on all fruit varieties and orchard environments, experiments show that the system is relatively stable. The stable IoU and F1 scores indicate that the segmentation accuracy is relatively high, whether in open fields, greenhouses, or shaded canopies. In harsh environments, the yield estimation error is within a reasonable range for commercial operations, making this method relatively reliable. The optimized structure of the model demonstrates good computational performance in both mobile and fixed environments. During the on-site implementation process, serious occlusion and sensor limitation issues were discovered. Support measures for the annotation process and model stability have been adjusted multiple times.

The framework of this study will provide a reference for smart orchard management in the future. Other types of sensors can be integrated into the system, such as time-series images or hyperspectral data, to enhance the system's discrimination capabilities and expand its range of applications. The aforementioned principles provide strong support for the development of large-scale autonomous decision support systems in horticulture. This system has the potential to change the way resources are utilized, yield predictions are made, and ecological agriculture is practiced.

Author Contributions

Kateřina Svobodová contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Natálie Králová contributes to software, validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Khan, Z., Liu, H., Shen, Y., & Zeng, X. (2024). Deep learning improved YOLOv8 algorithm: Real-time precise instance segmentation of crown region orchard canopies in natural environment. Computers and Electronics in Agriculture, 224, 109168. https://doi.org/10.1016/j.compag.2024.109168

[2] Haider, A., Arsalan, M., Hong, J. S., Sultan, H., Ullah, N., & Park, K. R. (2024). Multi-scale and multi-receptive field-based feature fusion for robust segmentation of plant disease and fruit using agricultural images. Applied Soft Computing, 167, 112300. https://doi.org/10.1016/j.asoc.2024.112300

[3] Li, J., Li, Y., You, H., & Zhang, L. (2025). Ginseng Quality Identification Based on Multi-Scale Feature Extraction and Knowledge Distillation. Horticulturae, 11(9), 1120. https://doi.org/10.3390/horticulturae11091120

[4] Shen, L., Su, J., Huang, R., Quan, W., Song, Y., Fang, Y., & Su, B. (2022). Fusing attention mechanism with Mask R-CNN for instance segmentation of grape cluster in the field. Frontiers in plant science, 13, 934450. https://doi.org/10.3389/fpls.2022.934450

[5] Zhang, Y., Li, L., Chun, C., Wen, Y., Li, C., & Xu, G. (2024). Data-driven Bayesian Gaussian mixture optimized anchor box model for accurate and efficient detection of green citrus. Computers and Electronics in Agriculture, 225, 109366. https://doi.org/10.1016/j.compag.2024.109366

[6] La, Y. J., Seo, D., Kang, J., Kim, M., Yoo, T. W., & Oh, I. S. (2023). Deep learning-based segmentation of intertwined fruit trees for agricultural tasks. Agriculture, 13(11), 2097. https://doi.org/10.3390/agriculture13112097

[7] Li, Z., Wang, R., & Ding, R. (2025). A review of crop attribute monitoring technologies for general agricultural scenarios. AgriEngineering, 7(11), 365. https://doi.org/10.3390/agriengineering7110365

[8] Qiu, S., Cheng, X., Lu, H., Zhang, H., Wan, R., Xue, X., & Pu, J. (2023). Subclassified loss: Rethinking data imbalance from subclass perspective for semantic segmentation. IEEE Transactions on Intelligent Vehicles, 9(1), 1547-1558. https://doi.org/10.1109/TIV.2023.3325343

[9] Liang, J., Huang, K., Lei, H., Zhong, Z., Cai, Y., & Jiao, Z. (2024). Occlusion-aware fruit segmentation in complex natural environments under shape prior. Computers and Electronics in Agriculture, 217, 108620. https://doi.org/10.1016/j.compag.2024.108620

[10] Tang, Z., Pan, X., She, X., Ma, J., & Zhao, J. (2025). Detail and Deep Feature Multi-Branch Fusion Network for High-Resolution Farmland Remote-Sensing Segmentation. Remote Sensing, 17(5), 789. https://doi.org/10.3390/rs17050789

[11] Cong, P., Zhou, J., Li, S., Lv, K., & Feng, H. (2022). Citrus tree crown segmentation of orchard spraying robot based on RGB-D image and improved mask R-CNN. Applied Sciences, 13(1), 164. https://doi.org/10.3390/app13010164

[12] Zhang, Y., Shi, N., Zhang, H., Zhang, J., Fan, X., & Suo, X. (2022). Appearance quality classification method of Huangguan pear under complex background based on instance segmentation and semantic segmentation. Frontiers in Plant Science, 13, 914829. https://doi.org/10.3389/fpls.2022.914829

[13] Mirzaei, A., Bagheri, H., & Khosravi, I. (2023). Enhancing crop classification accuracy through synthetic SAR-optical data generation using deep learning. ISPRS International Journal of Geo-Information, 12(11), 450. https://doi.org/10.3390/ijgi12110450

[14] Wang, Z., Cui, W., Huang, C., Zhou, Y., Zhao, Z., Yue, Y., ... & Lv, C. (2025). Framework for apple phenotype feature extraction using instance segmentation and edge attention mechanism. Agriculture, 15(3), 305. https://doi.org/10.3390/agriculture15030305

[15] Kondylatos, S., Bountos, N. I., Prapas, I., Zavras, A., Camps-Valls, G., & Papoutsis, I. (2025). Probabilistic machine learning for noisy labels in Earth observation. Scientific Reports, 15(1), 35890. https://doi.org/10.1038/s41598-025-19781-2

[16] Silva, R., Freitas, O., & Melo-Pinto, P. (2024). Evaluating the generalization ability of deep learning models: An application on sugar content estimation from hyperspectral images of wine grape berries. Expert Systems with Applications, 250, 123891. https://doi.org/10.1016/j.eswa.2024.123891

[17] Ye, X., Pan, J., Shao, F., Liu, G., Lin, J., Xu, D., & Liu, J. (2024). Exploring the potential of visual tracking and counting for trees infected with pine wilt disease based on improved YOLOv5 and StrongSORT algorithm. Computers and Electronics in Agriculture, 218, 108671. https://doi.org/10.1016/j.compag.2024.108671

[18] Hofman, R., Mattheijssens, J., Van Huylenbroeck, J., Verwaeren, J., & Lootens, P. (2025). Optimizing Plant Production Through Drone-Based Remote Sensing and Label-Free Instance Segmentation for Individual Plant Phenotyping. Horticulturae, 11(9), 1043. https://doi.org/10.3390/horticulturae11091043

[19] Meghraoui, K., Sebari, I., Pilz, J., Ait El Kadi, K., & Bensiali, S. (2024). Applied deep learning-based crop yield prediction: A systematic analysis of current developments and potential challenges. Technologies, 12(4), 43. https://doi.org/10.3390/technologies12040043

[20] Feng, Y., Ma, W., Tan, Y., Yan, H., Qian, J., Tian, Z., & Gao, A. (2024). Approach of dynamic tracking and counting for obscured citrus in smart orchard based on machine vision. Applied Sciences, 14(3), 1136. https://doi.org/10.3390/app14031136

[21] Zhong, W., Yang, W., Wang, Y., Dong, X., Wang, X., Jia, W., ... & Yan, M. (2025). Light adaptive image enhancement for improving visual analysis in intercropping cultivation. Frontiers in Plant Science, 16, 1639016. https://doi.org/10.3389/fpls.2025.1639016

[22] Chen, Z., Lei, X., Yuan, Q., Qi, Y., Ma, Z., Qian, S., & Lyu, X. (2024). Key technologies for autonomous fruit- and vegetable-picking robots: A review. Agronomy, 14(10), 2233. https://doi.org/10.3390/agronomy14102233

[23] Cheng, J., Zhu, Y., Zhao, Y., Li, T., Chen, M., Sun, Q., ... & Zhang, X. (2024). Application of an improved U-Net with image-to-image translation and transfer learning in peach orchard segmentation. International Journal of Applied Earth Observation and Geoinformation, 130, 103871. https://doi.org/10.1016/j.jag.2024.103871

[24] Assunção, E., Gaspar, P. D., Alibabaei, K., Simões, M. P., Proença, H., Soares, V. N., & Caldeira, J. M. (2022). Real-time image detection for edge devices: A peach fruit detection application. Future Internet, 14(11), 323. https://doi.org/10.3390/fi14110323

[25] Zhou, C., Cao, Y., Ming, B., Luo, J., Xu, F., Zhang, J., & Dong, M. (2025). A Multimodal Deep Learning Framework for Intelligent Pest and Disease Monitoring in Smart Horticultural Production Systems. Horticulturae, 12(1), 8. https://doi.org/10.3390/horticulturae12010008

[26] Jia, W., Zhang, Z., Shao, W., Hou, S., Ji, Z., Liu, G., & Yin, X. (2021). FoveaMask: A fast and accurate deep learning model for green fruit instance segmentation. Computers and Electronics in Agriculture, 191, 106488. https://doi.org/10.1016/j.compag.2021.106488

[27] Yang, R., Qi, Y., Zhang, H., Wang, H., Zhang, J., Ma, X., ... & Ma, C. (2024). A study on the object-based high-resolution remote sensing image classification of crop planting structures in the Loess Plateau of eastern Gansu Province. Remote Sensing, 16(13), 2479. https://doi.org/10.3390/rs16132479

[28] Koirala, A., Walsh, K. B., & Wang, Z. (2021). Attempting to estimate the unseen—correction for occluded fruit in tree fruit load estimation by machine vision with deep learning. Agronomy, 11(2), 347. https://doi.org/10.3390/agronomy11020347

[29] Srivastava, K. K., Kumar, D., Kishore, K., Damodaran, T., Singh, A., & Pandey, S. (2025). Light Intensity and Nutrient Status for Accurate Yield Prediction of Mango Orchards. Applied Fruit Science, 67(6), 469. https://doi.org/10.1007/s10341-025-01692-1

[30] Frimpong, S. A., Han, M., Zheng, W., Li, X., Akpaku, E., & Obeng, A. P. (2025). Machine and deep learning in agricultural engineering: A comprehensive survey and meta-analysis of techniques, applications, and challenges. Computers, 14(10), 438. https://doi.org/10.3390/computers14100438

[31] Tang, Y., Qiu, J., Zhang, Y., Wu, D., Cao, Y., Zhao, K., & Zhu, L. (2023). Optimization strategies of fruit detection to overcome the challenge of unstructured background in field orchard environment: A review. Precision Agriculture, 24(4), 1183-1219. https://doi.org/10.1007/s11119-023-10009-9