Efficient and comprehensive visual solution for a smart apple harvesting robot in complex settings via multi-class instance segmentation

Shiwei Wen[1], Yahao Ge[1], Yingkuan Wang[2,3*], Naishuo Wei[1], Jianguo Zhou[1], Guangrui Hu[4], Liangliang Yang[5], Jun Chen[1*]

(1. *College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China*;
2. *Academy of Agricultural Planning and Engineering, Ministry of Agriculture and Rural Affairs, Beijing 100125, China*;
3. *Chinese Society of Agricultural Engineering, Beijing 100125, China*;
4. *School of Design, Xi'an Technological University, Xi'an 710021, China*;
5. *Laboratory of Bio-Mechatronics, Faculty of Engineering, Kitami Institute of Technology, Hokkaido 090-8507, Japan*)

Abstract: To enable efficient and low-cost automated apple harvesting, this study presented a multi-class instance segmentation model, SCAL (Star-CAA-LADH), which utilizes a single RGB sensor for image acquisition. The model achieves accurate segmentation of fruits, fruit-bearing branches, and main branches using only a single RGB image, providing comprehensive visual inputs for robotic harvesting. A Star-CAA module was proposed by integrating Star operation with a Context-Anchored Attention mechanism (CAA), enhancing directional sensitivity and multi-scale feature perception. The Backbone and Neck networks were equipped with hierarchically structured SCA-T/F modules to improve the fusion of high- and low-level features, resulting in more continuous masks and sharper boundaries. In the Head network, a Segment_LADH module was employed to optimize classification, bounding box regression, and mask generation, thereby improving segmentation accuracy for small and adherent targets. To enhance robustness in adverse weather conditions, a Chain-of-Thought Prompted Adaptive Enhancer (CPA) module was integrated, thereby increasing model resilience in degraded environments. Experimental results demonstrate that SCAL achieves 94.9% AP_M and 95.1% mAP_M, outperforming YOLOv11s by 6.6% and 4.6%, respectively. Under multi-weather testing conditions, the CPA-SCAL variant consistently outperforms other comparison models in accuracy. After INT8 quantization, the model size was reduced to 14.5 MB, with an inference speed of 47.2 frames per second (fps) on the NVIDIA Jetson AGX Xavier. Experiments conducted in simulated orchard environments validate the effectiveness and generalization capabilities of the SCAL model, demonstrating its suitability as an efficient and comprehensive visual solution for intelligent harvesting in complex agricultural settings.

Keywords: apple harvesting, instance segmentation, multi-weather condition, star operation, edge computing device
DOI: 10.25165/j.ijabe.20251804.9619

Citation: Wen S W, Ge Y H, Wang Y K, Wei N S, Zhou J G, Hu G R, et al. Efficient and comprehensive visual solution for a smart apple harvesting robot in complex settings via multi-class instance segmentation. Int J Agric & Biol Eng, 2025; 18(4): 200–215.

1 Introduction

As one of the world's major commercial fruit crops, apples are still predominantly harvested by hand—a process that is both labor-intensive and inefficient[1]. This reliance limits scalability and fails to meet the demands of large-scale commercial production[2,3]. To address these limitations, robotic harvesting technologies have garnered increasing interest[4-6]. Although numerous robotic harvesters have been developed, they often fall short of human performance due to technical constraints. Among these limitations, the lack of reliable visual perception remains a fundamental obstacle to effective robotic harvesting[7]. This limitation is compounded by practical cost and hardware constraints, which preclude the use of complex multi-sensor systems. As a result, achieving comprehensive scene understanding with only a single vision sensor has emerged as a critical technical challenge.

The rapid development of deep learning has significantly accelerated the integration of object detection and image segmentation in agricultural applications[8-10]. Object detection algorithms typically locate objects by generating bounding boxes, serving common tasks such as fruit or branch recognition. However, these methods offer only coarse approximations of target regions, and thus often require additional post-processing to achieve fine localization[11]. In contrast, image segmentation techniques produce precise object masks that can be directly associated with depth information, enabling accurate three-dimensional spatial localization[12,13].

In unstructured orchard environments, the intricate spatial topology of branches and the dynamic morphology of slender fruit-bearing branches present significant challenges for robotic

Received date: 2025-01-10 **Accepted date:** 2025-07-15
Biographies: Shiwei Wen, MS candidate, research interest: image processing, Email: wsw142301@nwafu.edu.cn; **Yahao Ge**, MS candidate, research interest: agricultural engineering, Email: geyahao@nwafu.edu.cn; **Naishuo Wei**, PhD candidate, research interest: image processing, Email: weinaishuo@nwafu.edu.cn; **Jianguo Zhou**, PhD candidate, research interest: harvesting robot, Email: jianguozhou@nwafu.edu.cn; **Guangrui Hu**, PhD, Lecturer, research interest: precision agriculture, Email: hugrui@xatu.edu.cn; **Liangliang Yang**, PhD, Professor, research interest: intelligent agricultural equipment, Email: yang@mail.kitami-it.ac.jp.
***Corresponding author: Jun Chen**, PhD, Professor, research interest: intelligent agricultural equipment. College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China. Tel: +86-13572191773, Email: chenjun_jdxy@nwsuaf.edu.cn; **Yingkuan Wang**, PhD, Researcher, research interest: agricultural engineering. Academy of Agricultural Planning and Engineering, Ministry of Agriculture and Rural Affairs, Beijing 100125, China. Tel: +86-10-59197088, Email: wangyingkuan@163.com.

harvesting. Accurately and rapidly acquiring the spatial positions of apples, main branches, and fruit-bearing branches is essential to improve the performance of vision-guided robotic systems. Precise spatial perception enables optimal grasp planning and path generation, reduces collision risks, and improves harvesting success rates and operational efficiency[14].

A range of deep learning models has been applied to address these challenges. For instance, Wang and He[15] utilized an improved Mask R-CNN to segment apples under complex conditions involving shadows, varied backgrounds, and foliage occlusion, achieving a precision of 97.1% and a segmentation mAP of 91.7%, with an inference time of 250 ms per image. Tong et al.[16] applied a Cascade Mask R-CNN with a Swin-T backbone to segment trunks and branches in dormant orchards, reporting bbox mAP and segm mAP of 94.3% and 94.0%, respectively. Additionally, Sapkota et al.[17] compared YOLOv8 and Mask R-CNN across two scenarios: dormant tree trunk and branch segmentation (Scene 1), and segmentation of unripe apples in leafy conditions (Scene 2). In Scene 1, YOLOv8 achieved 90.6% precision, 74.0% mAP@0.5, and an inference speed of 10.9 ms, while Mask R-CNN reached 81.3%, 70.0%, and 15.6 ms, respectively. In Scene 2, YOLOv8 outperformed Mask R-CNN with 92.9% precision and 90.2% mAP, further highlighting the speed and accuracy advantages of YOLO-based models. Building on this line of research, Yan et al.[18] developed an improved YOLOv8s-based perception model capable of simultaneously detecting apples and segmenting branches and trunks. By embedding SE attention and dynamic snake convolution, the model achieved a precision of 99.6% for apple recognition and an mAP of 81.6% for branch and trunk segmentation.

Nevertheless, most of these studies focus on single-class segmentation, targeting either fruits or branches, and fail to provide the comprehensive multi-object perception required for complex orchard environments. To address this limitation, Rong et al.[19] proposed an enhanced semantic segmentation model based on Swin Transformer V2 for simultaneous segmentation of tomato fruits, calyxes, and stems. By integrating a SeMask module into the encoder, the model achieved improved performance with an inference time of approximately 120 ms. Similarly, Kang and Chen[20] introduced DaSNet-v2, a single-stage detection framework that integrates both instance and semantic segmentation branches. The model achieved 87.3% fruit segmentation accuracy and a branch segmentation IoU of 79.4%, with an average processing time of 70 ms. However, although DaSNet-v2 supports the concurrent segmentation of fruits and branches, it relies exclusively on semantic segmentation for the latter, thus lacking the ability to distinguish individual branch instances. Moreover, its relatively complex architecture poses challenges for deployment on edge computing devices.

While semantic segmentation has been widely applied in agricultural perception tasks, its inability to differentiate between individual instances within the same class limits its effectiveness in multi-target scenarios, particularly in environments characterized by dense foliage or morphologically similar targets[13]. In contrast, instance segmentation distinguishes individual objects within a category and delineates their precise boundaries. When combined with depth information, instance segmentation can assign unique spatial attributes to each object, enabling higher-level reasoning and decision-making in robotic harvesting systems[21]. However, the complexity and computational demands of instance segmentation models pose challenges for real-time deployment on edge devices. Therefore, a balance must be achieved between segmentation accuracy and inference efficiency, highlighting the need for lightweight yet effective visual perception models to support real-time scene understanding in robotic harvesting tasks.

To address the aforementioned challenges, this study proposed a novel multi-class instance segmentation framework, termed SCAL, specifically designed to accurately segment apples, main branches, and fruit-bearing branches in unstructured orchard environments. The main contributions of this work were as follows:

1) The Star-CAA module was designed to enable coordinated modeling between feature and spatial dimensions. This module effectively accommodated scale variation and topological complexity in branching structures, thereby enhancing spatial perception.

2) SCAL operated using only RGB images captured by a single camera, offering a low-cost solution for acquiring detailed scene understanding. The model achieved high segmentation accuracy while maintaining computational efficiency on edge devices, successfully balancing precision and real-time performance. This design significantly improved the capabilities of vision-guided robotic systems in orchard harvesting scenarios.

3) To ensure consistent segmentation under challenging conditions such as rain and fog, a weather-adaptive image augmentation module was incorporated. This enhanced the model's robustness and supported all-weather visual perception, enabling reliable operation in intelligent harvesting systems.

2 Materials and methods

2.1 Dataset construction

2.1.1 Dataset acquisition

The image dataset used in this study was collected from the "Yujia" Orchard Cooperative in Baoji City, Shaanxi Province, China, during the peak apple harvesting season from October to November 2024. The orchard cultivated three commercially significant apple varieties: "Honeycrisp", "Yanfu", and "Ruixianghong". Sampling was conducted in plots managed under a modern high-density dwarf rootstock cultivation system, characterized by row spacing of 2.0-3.5 m and plant spacing of approximately 1.2 m.

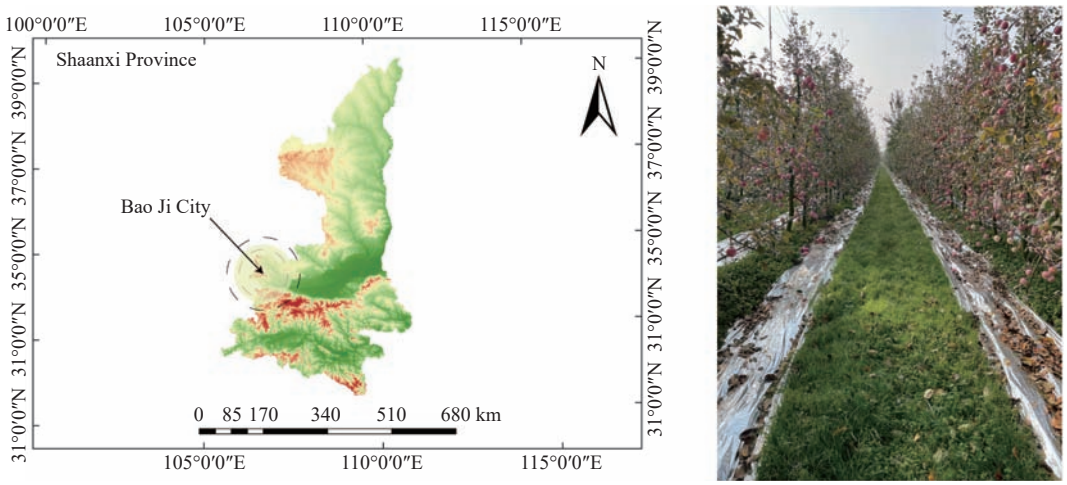
Images were captured using an iPhone 12 Pro, Huawei P40 Pro, and ZED2i depth camera, all positioned directly facing the apple trees at distances ranging from 300 to 600 mm. This configuration simulated the installation of visual sensors on robotic apple harvesters. To improve the model's robustness under varying environmental conditions and enhance its generalization across scenarios, a multi-condition illumination sampling strategy was adopted. Images were systematically acquired under both sunny and overcast weather, at different times of day (morning, midday, and evening), and under both front-lit and backlit lighting conditions. In addition, supplementary samples were collected at night using artificial lighting. The complete data acquisition workflow is shown in Figure 1.

2.1.2 Multi-weather conditions simulation

In real agricultural environments, weather conditions vary significantly, including rainfall, fog, and their combinations. However, such conditions were absent during our image acquisition period, leading to a lack of samples representing these specific meteorological scenarios in the original dataset.

To ensure that the developed visual model is capable of supporting all-weather harvesting operations, a physics-based weather simulation approach was adopted. Specifically, professional rain and fog generation algorithms were applied to the

collected RGB images, thereby constructing a comprehensive dataset of apple images under simulated adverse weather conditions. Representative samples of these simulated images are shown in Figure 2.



a. Location of image acquisition and orchard environment

b. Data acquisition equipment

c. Examples of collected data

Figure 1 Data acquisition workflow diagram

a. Original image b. Simulated rainy condition c. Simulated foggy condition d. Simulated reain-fog combination

Figure 2 Simulated apple images under adverse weather conditions

This strategy enhances the model's generalizability across a wider range of environmental scenarios, enabling it to maintain stable performance even under unfavorable weather conditions.

2.1.3 Dataset annotation

To improve computational efficiency and ensure compatibility with low-resolution image acquisition devices, all dataset images were resized to a uniform resolution of 1024×1024 pixels and saved in JPG format.

In orchard environments, main branches generally grow in horizontal or inclined orientations, whereas fruit-bearing branches extend in more diverse directions. As the objective of this study is apple harvesting, only the main branches and fruit-bearing branches within apple-containing regions were annotated. Due to the distinct morphological differences among apples, fruit-bearing branches, and main branches, distinct annotation strategies were applied. For apples and fruit-bearing branches, a minimum enclosing polygon annotation strategy was employed to minimize background pixels and improve localization accuracy. In contrast, main branches are typically longer and thicker. Annotating the entire structure with a single polygon often introduces excessive background noise and leads to poor boundary alignment, which adversely affects feature extraction. To address this, a segmented

quadrilateral annotation strategy was implemented, wherein multiple rectangular segments were aligned along the primary growth direction of each branch[22].

All annotations were manually created using the Labelme tool. As shown in Figure 3, the annotations were saved in JSON format and then converted to the TXT format compatible with the YOLOv11 framework.

2.1.4 Data augmentation and division

To mitigate the risk of overfitting due to the limited number of training samples, multiple offline data augmentation techniques were applied to the original dataset. These included noise injection, mirroring, rotation, contrast adjustment, brightness variation, and translation, as shown in Figure 4.

By incorporating images captured under diverse weather and lighting conditions, a dataset comprising 7000 images was constructed. The dataset was then partitioned into training, validation, and test sets using a 7:2:1 ratio.

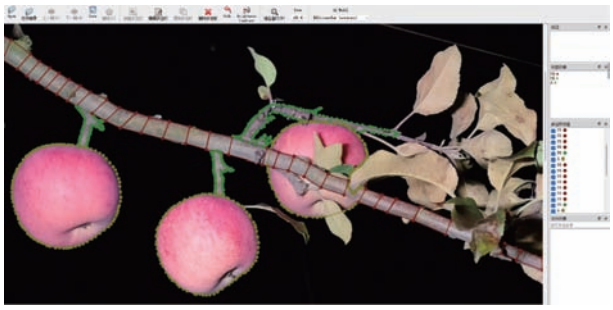


Figure 3 Annotation examples from the dataset

a. Rotation b. Mirroring c. Contrast adjustment

d. Noise injection e. Brightness variation f. Image scaling

Figure 4 Data augmentation for apple recognition

2.2 Design of the Star-CAA and SCA-T/F modules for multi-scale feature fusion

2.2.1 Star Operation

In robotic apple-harvesting tasks, the precise segmentation of main branches, fruit-bearing branches, and fruits directly influences the harvesting success rate, operational efficiency, and the degree of tree structure protection. However, this task presents several challenges: the morphological continuity between main and fruit-bearing branches leads to unclear boundary distinctions; apples of different cultivars often share similar colors and textures; and changing lighting conditions can obscure object edges. These factors are further compounded by the need for both high segmentation accuracy and real-time inference.

Recent studies have shown that the method of feature fusion plays a critical role in determining segmentation performance[23]. Star Operation, as a novel feature fusion mechanism, exhibits unique mathematical properties and strong practical potential[24]. It performs nonlinear, high-order fusion by applying element-wise multiplication to two input features, thereby achieving high-dimensional nonlinear mapping within a low-dimensional space. Given input features $x=[x_1,x_2,\ldots,x_n]\in\mathbb{R}^n$, the two paths are linearly transformed and then multiplied, yielding the output:

$$\Gamma(X)=\mathrm{Relu}\left(\sum_{i=1}^{n}\alpha_i^1 x_i\right)\odot\left(\sum_{j=1}^{n}\alpha_j^2 x_j\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}k_{i,j}x_i x_j \quad (1)$$

where, X denotes the feature set of the input data; $\odot$ denotes the element-wise multiplication; α_i^1 and α_j^2 are learnable parameters; and $k_{i,j}$ denotes the combination coefficient.

Considering the symmetry $x_ix_j=x_jx_i$, the dimensionality of the mapped feature space is approximately:

$$\frac{n(n+1)}{2}\approx\frac{n^2}{2} \quad (2)$$

Through the learnable coefficient $k_{i,j}$, the model can adaptively adjust the feature mapping strategy to suit different structural targets better. Although the theoretical dimensionality of the mapped space is $O\left(\frac{n^2}{2}\right)$, all computations are retained in the original n-dimensional space. When multiple Star Operation layers are cascaded, the dimensionality of the feature space increases exponentially.

$$\dim(\Gamma^t) \approx \left(\frac{n^2}{2}\right)^t \tag{3}$$

where, t is the number of layers. This implicit infinite-dimensionality allows the model to capture fine-grained object variations and complex spatial relationships without substantially increasing computational complexity, providing a solid theoretical foundation for the high-precision segmentation of main branches, fruit-bearing branches, and apples. Additionally, the quadratic term $x_i x_j$ in Equation (1) can be viewed as a nonlinear multi-scale combination, which, when combined with large-kernel depth-wise convolutions, contributes to a multi-scale receptive field mechanism that enhances the model's adaptability to scale changes. For main branches, this helps capture the gradual change in thickness along the trunk. For slender fruit-bearing branches that are only a few pixels wide, it maps spatial features into high-dimensional space, preserving structural continuity and enhancing local contrast. For apples, which vary in size, texture, and illumination, the nonlinear mapping improves robustness.

To preserve structural continuity during segmentation, for adjacent pixels p_i and p_j with corresponding features f_i and f_j, the cosine similarity in the high-dimensional mapped space is given by:

$$\text{Similarity}(p_i, p_j) = \frac{\Gamma(f_i)\Gamma\left(f_j\right)}{\|\Gamma(f_i)\|\left\|\Gamma\left(f_j\right)\right\|} \approx 1 \tag{4}$$

This equation demonstrates the strong aggregation ability of Star Operation in spatially continuous regions. Even under varying illumination or texture conditions in the original feature space, the high-dimensional mapping preserves feature continuity and consistency, thereby reducing segmentation errors arising from structural discontinuities or blurred boundaries.

Notably, Star Operation produces significant gradient enhancement and amplification effects at object boundaries. For a pixel X_{boundary} located on a semantic boundary, the boundary gradient response $\nabla\Gamma(X_{\text{boundary}})$ is defined as:

Let

$$A = \sum_{i=1}^{n} \alpha_i^1 x_i, \quad B = \sum_{j=1}^{n} \alpha_j^2 x_j \tag{5}$$

Then,

$$\nabla\Gamma\left(X_{\text{boundary}}\right) = (\text{Relu}(A)\cdot\nabla B)\odot(B\cdot\nabla A) + \text{Relu}(A)\odot\nabla B \tag{6}$$

This boundary response function integrates three key mechanisms: nonlinear activation, dual-branch feature complementarity, and multiplicative amplification. When the input feature crosses a semantic boundary, the spatial gradients from both paths $\nabla(A)$ and $\nabla(B)$ produce an amplification effect, which is further enhanced when multiplied by the respective feature values. As a result, the boundary signal is significantly boosted. Consequently, even when feature transitions are weak or ambiguous, the Star Operation can generate structurally coherent and highly responsive boundary features. Compared to traditional methods that rely heavily on explicit edge priors or predefined structural assumptions, Star Operation exhibits strong adaptability across diverse target types. It enhances boundary representations by implicitly encoding structural information in high-dimensional space through nonlinear feature composition, eliminating the need for manually designed strategies tailored to specific object classes such as apples or branches.

To encapsulate, the distinctive mathematical characteristics of Star Operation offer a compelling alternative to the traditional deepening or widening of neural network architectures. Leveraging implicit high-dimensional mapping and nonlinear feature interactions, it enables efficient, high-precision perception and segmentation of complex targets, particularly when deployed on resource-constrained edge devices in agricultural environments.

2.2.2 Context anchor attention

In orchard-harvesting scenarios, main branches typically grow in horizontal or oblique orientations, whereas fruit-bearing branches exhibit more irregular and diverse growth patterns. A clear hierarchical structure is present: fruit-bearing branches attach to the main branches, and fruits are primarily located at the distal ends of fruit-bearing branches or directly connected to the main branches. Consequently, the model must be capable of simultaneously perceiving three distinct object types with varying scales and effectively capturing the complex spatial relationships among them.

To address this challenge, the CAA mechanism[25] was introduced to enhance the model's ability to construct the topological structure of branching connections. This module first extracts locally compressed contextual features using global average pooling, expressed as:

$$\boldsymbol{U}_i^{(p)} = \text{Conv}_{1\times1}\left(\text{AvgPool}\left(\boldsymbol{F}_i^{(p)}\right)\right) \tag{7}$$

where, $\boldsymbol{F}_i^{(p)}$ denotes the input feature of the ith module in layer p; and $\boldsymbol{U}_i^{(p)}$ represents its corresponding global contextual feature.

Subsequently, the CAA mechanism separates spatial dependencies along horizontal and vertical directions using depth-wise separable convolutions, forming a cross-decoupled structure as illustrated in Figure 5. This design facilitates the construction of long-range dependencies in both directions:

$$\begin{cases} \boldsymbol{G}_i^{(p)} = \text{DWConv}_{1\times k}\left(\boldsymbol{U}_i^{(p)}\right) \\ \boldsymbol{H}_i^{(p)} = \text{DWConv}_{k\times1}\left(\boldsymbol{G}_i^{(p)}\right) \end{cases} \tag{8}$$

where, $\boldsymbol{G}_i^{(p)}$ and $\boldsymbol{H}_i^{(p)}$ represent the features obtained after horizontal and vertical convolutions, respectively. This process allows the model to effectively capture directional features and gain a better understanding of the orientation and connection patterns of branches. In cases where fruits or branches are partially occluded or missing, CAA can utilize long-range contextual reasoning to infer the continuity of occluded structures. By leveraging both visible local features and their contextual dependencies, the model's robustness is enhanced in complex natural environments.

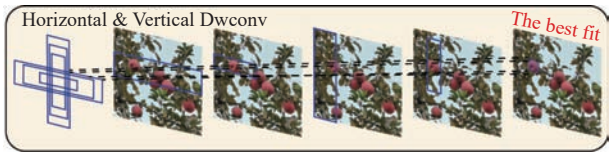


Figure 5 Schematic of horizontal and vertical convolutions in the CAA module

The resulting features are then passed through a convolution layer followed by a Sigmoid activation function to generate a spatial attention map:

$$\begin{cases} \boldsymbol{A}_i^{(p)} = \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(\boldsymbol{H}_i^{(p)}\right)\right) \\ \boldsymbol{A}_i^{(p)} \in [0,1]^{C\times H\times W} \end{cases} \tag{9}$$

This attention map assigns selection weights to each pixel position within the target region (with red indicating weights close to 1), thereby enhancing the model's focus on boundary-relevant regions. As shown in Figure 6, this boundary enhancement mechanism enables the model to accurately locate the connection points between fruit-bearing branches, main branches, and fruits, thereby improving overall segmentation accuracy.

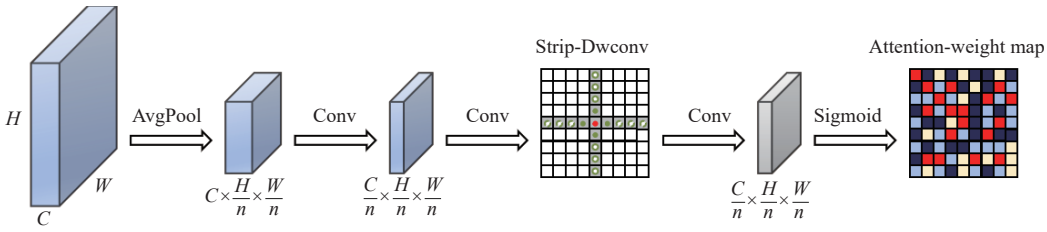


Figure 6 Generation of attention weight maps in the CAA module

2.2.3 Star-CAA: Integrating nonlinear interaction and direction-aware attention

To achieve coordinated modeling between feature and spatial dimensions, this study proposes the Star-CAA module. This module retains the nonlinear feature combination capabilities of the Star Operation while incorporating the CAA mechanism to enhance directional spatial perception of the output feature maps prior to feature dimensionality reduction. This integration significantly improves the network's ability to perceive structural continuity and distinguish ambiguous or overlapping boundary regions.

Let Γ denote the output of the Star Operation (as defined in Equation (1)), which serves as the input to the CAA module. The directionally enhanced features are expressed as:

$$Y = \boldsymbol{A}_i^{(p)} \odot \Gamma \tag{10}$$

Due to the spatial continuity and directional sensitivity of the attention map $\boldsymbol{A}_i^{(p)}$, this multiplicative operation not only enhances the model's response to subtle structural differences in branch connections but also strengthens its ability to capture boundaries in blurred or cluttered regions. Furthermore, the modular structure of this attention mechanism simplifies computation and achieves high efficiency, making it suitable for deployment on resource-constrained edge devices in agricultural settings.

However, in deep networks, if the attention feature map deviates significantly from the original feature distribution, this may suppress useful semantic information and compromise the stability of the network representation. To address this issue, the Star-CAA module introduces a residual connection mechanism to fuse the original input features with the enhanced attention features, thereby mitigating the suppression effect and preserving semantic consistency. The final output is expressed as:

$$Y_{\text{output}} = Y \oplus X \tag{11}$$

This design not only improves the consistency of feature representations between edge and core regions but also facilitates gradient propagation in deep neural networks, thereby enhancing training efficiency.

Compared with traditional attention mechanisms (e.g., SE, which operates solely in the channel dimension, and CBAM, which relies on pooling for attention generation), Star-CAA achieves simultaneous modeling of spatial and channel dimensions with enhanced directional perception. As shown in Figure 7, this makes it particularly suitable for tasks involving objects with explicit directional structures in agricultural environments, such as main branches and fruit-bearing branches.

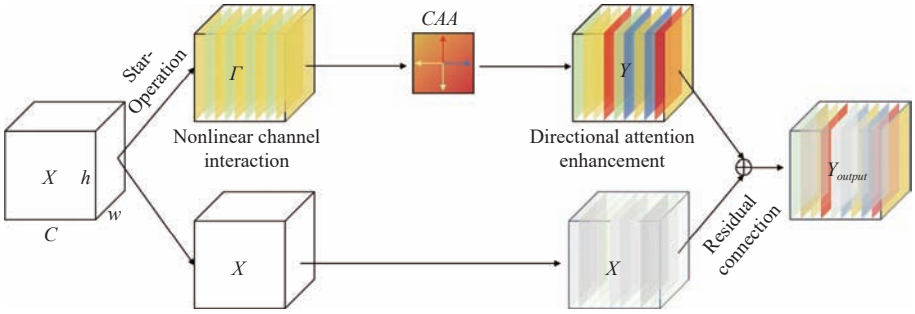


Figure 7 Schematic of spatial-channel joint modeling and directional enhancement fusion in the Star-CAA module

2.2.4 SCA-T/F: Structure-aware multi-scale feature fusion with Star-CAA

SCA-T/F is a newly designed multi-scale feature extraction and fusion module constructed by integrating Star-CAA with C3k2. The architecture offers two complementary configurations adapted to different feature processing demands: SCA-T (a deep cascaded structure) and SCA-F (a lightweight and fast structure), as shown in Figure 8.

When C3k=True, the SCA-T module adopts a more complex CSP (Cross Stage Partial) structure, which splits the input features into two parallel branches. The main path embeds multiple cascaded Star-CAA modules, forming a deep hierarchical transformation stream (NSCA). This layered design expands the effective receptive field of the model. It progressively builds multilevel feature abstraction, from edge textures and geometric structures to high-level semantics, enabling the network to effectively capture the global topological relationships and spatial extension patterns of the target structure.

When C3k=False, SCA-F employs a more straightforward feature processing pathway, embedding only a single Star-CAA module in the main branch. This lightweight design significantly reduces the number of parameters and computational complexity while minimizing redundant transformations and preserving high feature fidelity. Although a single Star-CAA module has a relatively limited receptive field, it exhibits greater sensitivity to local details, enabling effective capture of high-frequency features and boundary information. This sensitivity directly contributes to more explicit boundary representations and stronger modeling of the structural continuity of slender fruit-bearing branches.

In the backbone network design, SCA-T and SCA-F modules are flexibly deployed at different levels according to the semantic depth and spatial resolution of the hierarchical features. In the deep semantic extraction stages, SCA-T is used to enhance structural abstraction capabilities. In contrast, in the shallow detail-preserving

stages, SCA-F is applied to strengthen the response to edge textures and local features. By deploying the two types of modules across semantic levels, a progressively hierarchical structure-aware modeling path is constructed. This collaborative design significantly improves the model's ability to integrate global semantic understanding with local structural modeling, thereby enhancing its capability to perceive multi-class objects with strong structural awareness and spatial topology comprehension.

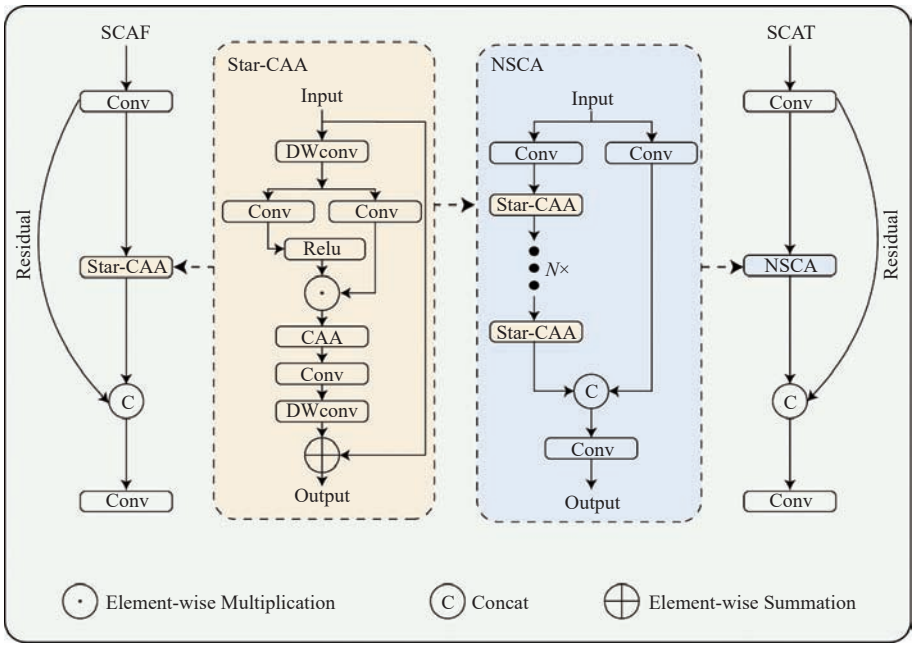


Figure 8 Structural illustration of the SCA-T and SCA-F modules

2.3 Segment_LADH: Triple-branch head with Mask, Classification, and Localization streams

The accurate segmentation of main branches, fruit-bearing branches, and fruits requires the simultaneous extraction of both textural and boundary features. However, conventional coupled head architectures often fail to balance the extraction needs of these heterogeneous features effectively. To address this issue, this study incorporates a Lightweight Asymmetric Dense Head (LADH) into the model's segmentation head. By employing an asymmetric multi-level compression strategy, the decoupled head design introduces task-specific branches that effectively reduce mutual interference among classification, localization, and segmentation tasks within the model[26].

As shown in Figure 9, the Segment_LADH adopts a three-path parallel architecture, with each path dedicated to a specific task, thereby mitigating cross-task interference. The instance segmentation branch further decomposes the segmentation task into two independent sub-processes: mask prototype generation and mask coefficient prediction, which are handled by the ProtoHead and PredictionHead, respectively. This structure enables the model to extract generalized shape representations and multi-scale semantic features separately, leading to precise segmentation of apple contours, stem locations, and connection points between fruits and branches.

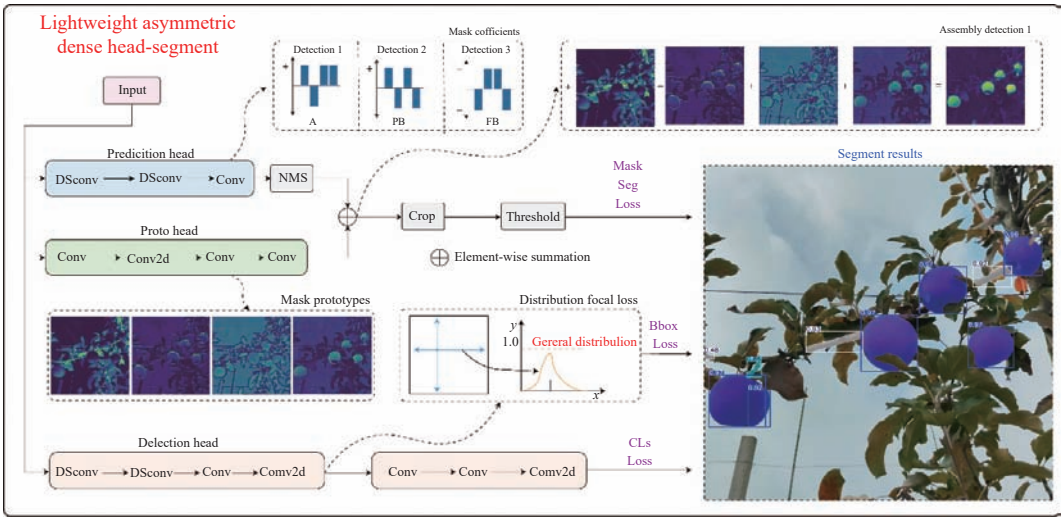


Figure 9 Segment_LADH structure diagram

The classification branch is dedicated to differentiating among the main branches, fruit-bearing branches, and apples. The bounding box prediction path utilizes a DSConv–DSConv–Conv–Conv2d sequence in conjunction with distributed focal loss optimization, in order to achieve accurate localization of all three object categories.

Additionally, the Segment_LADH module extensively replaces standard convolutions with depthwise separable convolutions, significantly reducing both parameter count and computational overhead. This lightweight design facilitates real-time inference on edge computing devices, making it suitable for deployment in orchard robotics and other resource-constrained agricultural environments.

2.4 CPA: Adaptive visual enhancement for multi-condition degradation

Mechanical harvesting in orchard environments faces significant challenges due to highly variable and unstructured natural conditions. In foggy environments, image clarity and contrast are significantly reduced, resulting in blurred boundaries between objects and their surroundings, such as apples and branch structures. Rainy conditions introduce light refraction and water droplet interference, resulting in noisy and complex textures. When both fog and rain coexist, the degradation becomes even more severe, further diminishing image quality. Most conventional visual enhancement techniques are designed for specific types of degradation, making them insufficient for handling diverse and dynamic environmental conditions. Therefore, a visual enhancement approach capable of adaptively addressing multiple types of degradation is urgently needed.

The CPA module addresses this problem through a chain-of-thought prompting-based mechanism, enabling adaptive processing of various environmental degradations[27]. The architecture comprises two core components: the Chain-of-thought Generation Module (CGM) and Content-driven Prompt Block (CPB). Together, they form a cascaded encoder-decoder framework that enables multi-level feature extraction and enhancement, as shown in Figure 10.

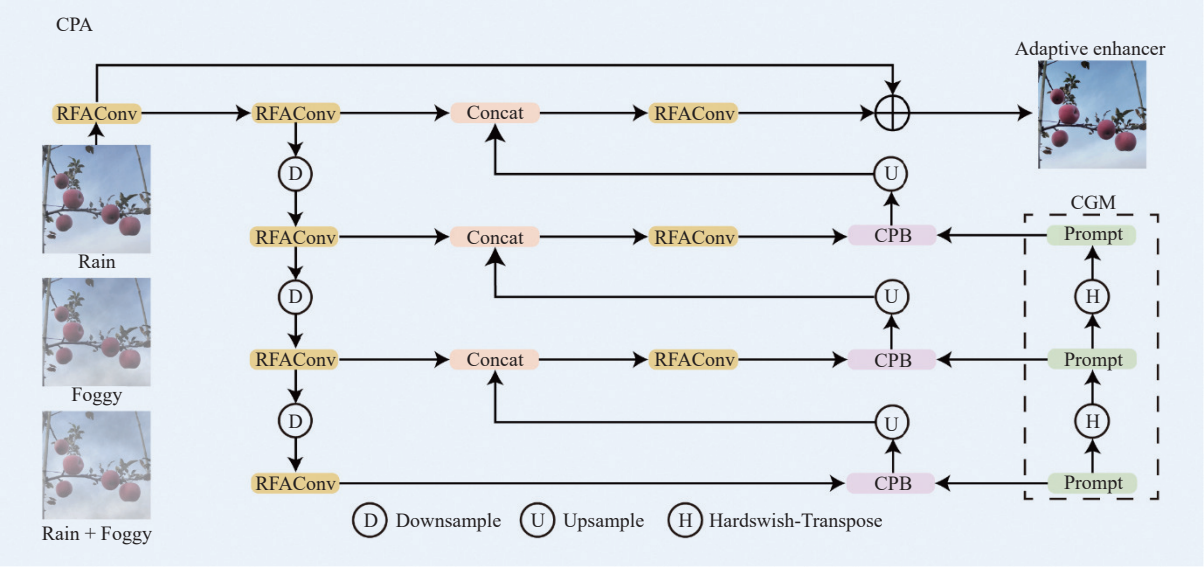


Figure 10 CPA structure diagram

The CGM employs a sequence of transposed convolutional layers to generate prompts that encode multi-scale degradation features. These prompts incorporate environmental clues associated with fog, rain, and mixed weather conditions, allowing the model to reason about the nature and extent of degradation gradually. The CPB divides the input features into multiple segments, each of which is processed by an independent Transformer module. This design allows the encoded degradation information from the prompts to be effectively injected into the input features, ensuring efficient cross-scale interaction.

As a result, the model can dynamically adjust its enhancement strategy in the presence of unknown degradation types and unseen weather conditions, ultimately providing high-quality inputs for the downstream segmentation task.

2.5 SCAL: Instance segmentation model for main branches, fruit-bearing branches, and apples

The instance segmentation framework proposed in this study integrates all previously introduced innovative components to address the multifaceted challenges faced by robotic arms in fruit-picking scenarios. The architecture is composed of three primary components: the Backbone, Neck, and Head, forming a complete pipeline for feature extraction and segmentation. The overall architecture is illustrated in Figure 11.

In the Backbone, the differentiated deployment of SCA-T and SCA-F modules is based on a careful analysis of the feature processing requirements at different hierarchical levels. The overall design follows a progressive receptive field expansion strategy:

1) In the shallow layers, the SCA-F modules are used to preserve high feature fidelity and sensitivity to local details, preventing premature abstraction that could result in the loss of fine-grained information;

2) As the network deepens, the SCA-T modules are introduced to capture morphological structures and spatial context, owing to their larger effective receptive fields and deeper representation capacity. The cascaded Star-CAA units in these modules progressively refine features—from edge textures to structural forms and, finally, semantic information—establishing a complete abstraction hierarchy.

In the Neck, multi-scale feature fusion and enhancement are required to bridge the Backbone outputs with the segmentation head. To this end, we construct a hierarchical fusion structure using both SCA-T and SCA-F modules. Their complementary characteristics facilitate a balanced integration of semantic abstraction and detail preservation during upsampling and feature aggregation. This configuration not only optimizes computational resource allocation but also enables the network to adaptively handle features with different levels of abstraction and spatial resolution, ultimately producing unified and enriched feature representations for the segmentation head.

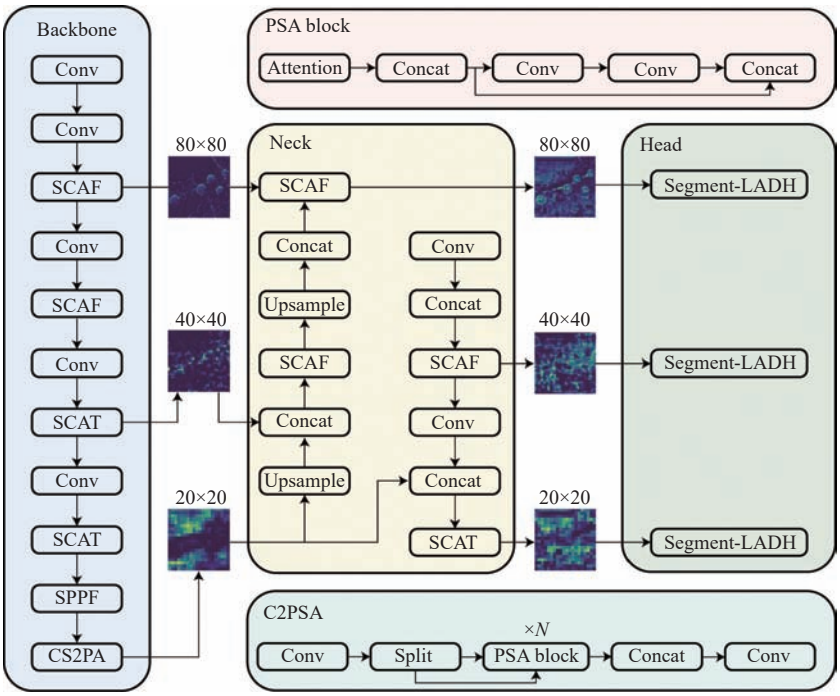


Figure 11 Overall architecture of the proposed instance segmentation model

The Segment_LADH head adopts an asymmetric multi-path architecture that decouples classification, localization, and segmentation tasks. This task-specific branching effectively overcomes mutual interference, thereby enhancing the overall stability of predictions. Its unique mask-generation mechanism allows for the precise delineation of fruit boundaries and key connection points.

Under adverse weather conditions—such as fog, rain, or their combination—the CPA module is activated at the input stage to enhance degraded images, ensuring that high-quality visual information is fed into the network.

Through the collaborative integration of these novel components, the proposed model demonstrates strong performance in addressing critical challenges, such as the continuous transition between branch types, the detection of delicate structures, boundary definition under complex lighting conditions, and the segmentation of attachment points between fruits and branches. This architecture thus establishes a robust and precise visual foundation for intelligent harvesting systems in complex orchard environments.

3 Experimental results and analysis

3.1 Segmentation model training

The model training in this study was conducted on a workstation equipped with an Intel Core i9-14900KF CPU (32 cores), 128 GB of RAM, and an NVIDIA GeForce RTX 4080 SUPER GPU (16 GB memory). The training environment was based on Python 3.10 and the PyTorch 2.1.0 framework, accelerated by CUDA 12.1 and cuDNN 8.8.0.1. The detailed training parameters are listed in Table 1.

Table 1 Training parameters for the recognition model

Training parameters	Values	Training parameters	Values
Optimizer	SGD	Epochs	300
Workers	12	Batch size	8
Mask_ratio	4	Size	640
lr0	0.01	lrf	0.01

3.2 Evaluation metrics

Instance segmentation is an extension of object detection that not only requires locating each target but also achieving pixel-wise segmentation for each instance. Therefore, this study employs multiple metrics to comprehensively evaluate the segmentation model.

To evaluate detection and segmentation accuracy, the following metrics are employed: Average Precision (AP), mean Average Precision (mAP), Average Precision for mask segmentation (AP_M), and its corresponding mean (mAP_M). The mAP metric, which balances precision and recall, is calculated as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\% \tag{12}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\% \tag{13}$$

$$AP = \int_0^1 P(R)\,\mathrm{d}R \times 100\% \tag{14}$$

where, TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{15}$$

where, A represents the predicted bounding box; and B is the ground truth bounding box. In instance segmentation, IoU measures the pixel-level overlap between predicted and ground truth masks.

$$\mathrm{mAP} = \frac{1}{n}\sum_{i=1}^{n} AP_i \times 100\%(\mathrm{IoU} \geq 0.5) \tag{16}$$

where, n refers to the number of target categories, which is 3 in this study.

For evaluating the inference efficiency of the model, this study uses the average processing time per image and estimates the complexity and computational cost through Giga Floating-point Operations (GFLOPs). Model size and the number of parameters is also included for a comprehensive evaluation. The corresponding

calculations are as follows:

$$\text{GFlops} = \lambda \left(\sum_{i=1}^{n} \text{Kernel}_i^2 C_{i-1}^2 C_i + \sum_{i=1}^{n} \theta^2 C_i \right) \quad (17)$$

$$\text{Params} = \lambda \left(\sum_{i=1}^{n} \theta^2 \text{Kernel}_i^2 C_{i-1}^2 C_i \right) \quad (18)$$

where, λ is a constant coefficient; Kernel denotes the kernel size of each convolution layer; n is the number of layers; C is the number of channels; θ represents the input image's height or width; and i is the index of the ith layer.

3.3 Effectiveness validation of Star-CAA

In this study, the proposed Star-CAA module integrates the contextual perception capabilities of the CAA mechanism with the nonlinear high-dimensional feature representation of Star Operation, thereby enabling enhanced structural perception and regional focus in instance segmentation tasks. To evaluate its effectiveness, systematic comparative experiments were conducted. In these experiments, 'CAA' and 'StarBlock' refer to models integrating only the respective components individually. The experimental results are presented in Table 2.

Table 2 Performance comparison of Star-CAA

Model	AP/%	mAP/%	AP_M/%	mAP_M/%
YOLOv11s	88.7	90.3	88.3	90.5
CAA	91.0	91.7	91.0	92.0
StarBlock	89.9	90.4	90.2	90.9
Star-CAA	**92.8**	**93.8**	**93.0**	**93.9**

Compared to YOLOv11s, the Star-CAA module improved AP, mAP, AP_M, and mAP_M by 4.1%, 3.5%, 4.7%, and 3.4%, respectively, indicating notable enhancements in both detection and segmentation performance. Furthermore, relative to the individual integration of either CAA or StarBlock, Star-CAA consistently achieved superior results. Specifically, it outperformed CAA by 1.8%, 2.1%, 2.0%, and 1.9%, and outperformed StarBlock by 2.9%, 3.4%, 2.0%, and 2.0% across the four metrics, validating the complementary integration of combining contextual anchoring with nonlinear high-dimensional feature modeling.

To further evaluate performance across different target categories, a category-level assessment of AP_M was conducted for apples (A), main branches (PB), and fruit-bearing branches (FB), as displayed in Table 3.

Table 3 Category-wise segmentation accuracy

Model	Evaluation metrics	A/%	PB/%	FB/%
YOLOv11s	AP_M	92.2	87.3	85.4
CAA	AP_M	94.7	90.0	88.3
Star-CAA	**AP_M**	**96.5**	**91.5**	**91.0**

As shown, Star-CAA achieves improvements of 4.3%, 4.2%, and 5.6% over YOLOv11s for A, PB, and FB, respectively, demonstrating superior segmentation performance, particularly for the structurally complex and intertwined PB and FB targets. Compared to CAA, Star-CAA achieves further gains of 1.8% for A and 2.7% for FB, suggesting that the Star operation enhances local detail representation when combined with contextual attention mechanisms.

To further evaluate the segmentation performance across categories, Grad-CAM[28] was applied to visualize the heatmaps generated by the three models (Figure 12). Three heatmaps from layers associated with the segmentation head were selected to compare each model's regional focus on the target objects.

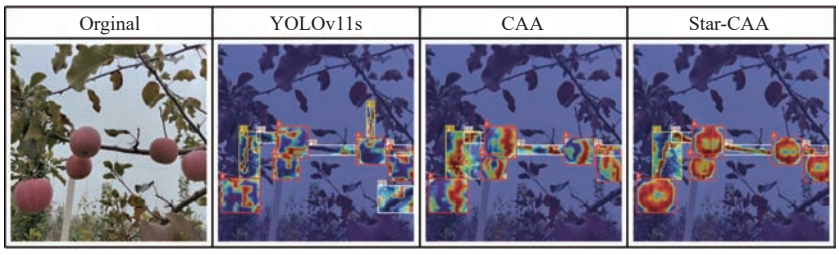


Figure 12 Comparative heatmaps in different models

As observed in Figure 12, the segmentation heatmaps illustrate distinct differences among the three models across target categories.

1) Apple segmentation: YOLOv11s can roughly identify most apple regions but exhibits weak and dispersed activation, indicating limited focus. The CAA model enhances attention through contextual anchoring, resulting in more concentrated activations near fruit boundaries; however, it still fails to achieve full coverage of the fruit areas. In contrast, Star-CAA demonstrates improved edge and contour sensitivity, yielding nearly complete coverage with stronger edge responses and more precise segmentation.

2) Main branch (PB) segmentation: The targets are linear and morphologically diverse. YOLOv11s displays discontinuous and incomplete responses, attending only to fragmented parts of the branches. CAA improves spatial continuity via context modeling and directional enhancement, but the resulting heatmaps remain either fragmented or overly diffuse. Star-CAA generates strong activations at branch junctions and along continuous structures, with heatmaps more closely aligned with the true branch contours, indicating superior spatial modeling and continuity perception.

3) Fruit-bearing branch (FB) segmentation: These structures are small, complex, and often occluded by apples. YOLOv11s shows weak and discontinuous responses with poor edge focus. While CAA increases regional sensitivity, it still struggles to accurately delineate FB contours or differentiate them from nearby PB structures and apples. Star-CAA shows markedly improved responses, with heatmaps that effectively capture fine-scale branches and junctions. This suggests enhanced capability in modeling fine-grained structures and local interactions.

Although the CAA mechanism improves attention focus and direction awareness in complex backgrounds through contextual anchoring, its reliance on linear relationships and spatial attention limits its capacity to capture nonlinear feature interactions and high-dimensional semantics. These limitations are particularly pronounced when processing structurally complex or morphologically variable targets. The Star operation addresses this shortcoming by nonlinearly combining multi-scale features, thereby expanding the representational space and enhancing local semantic expressiveness. This complementary integration allows the Star

operation to compensate for the CAA module's deficiencies in fine-grained modeling. Both quantitative results and visualizations demonstrate that Star-CAA more accurately attends to object boundaries and connection points, facilitating improved modeling of structural continuity, topological relationships, and fine local details.

3.4 Ablation experiment with different improved models

To systematically evaluate the effectiveness of each proposed enhancement, stepwise ablation experiments were conducted. In this setting, 'A' represents the replacement of the original C3k2 module with the SCA-T/F module, while 'B' denotes the substitution of the original segmentation head with the Segment_LADH head.

As listed in Table 4, the integration of SCA-T/F modules at multiple network levels led to notable performance improvements over YOLOv11s. Specifically, AP, mAP, AP_M, and mAP_M improved by 5.3%, 4.4%, 6.0%, and 4.3%, respectively. These results demonstrate that a hierarchical deployment of SCA-T/F enhances the model's capacity for deep feature extraction and improves its adaptability to features across varying levels of abstraction and resolution. Although this integration introduced a slight increase in computational cost, as measured by GFLOPs and parameter count, the performance gains clearly outweighed the added resource demands.

Table 4 Results of the ablation experiments

A	B	AP/ %	mAP/ %	AP_M/ %	mAP_M/ %	GFLOPs/ G	Parameters/ M	Model size/MB	Speed/ ms
×	×	88.7	90.3	88.3	90.5	35.3	100 679 77	20.6	2.2
√	×	94.0	94.7	94.3	94.8	41.6	116 806 97	23.9	3.8
×	√	92.2	93.1	92.5	93.5	32.6	929 501 7	19.1	3.7
√	√	**94.9**	**95.1**	**94.9**	**95.1**	**38.8**	**109 077 37**	**22.4**	**3.3**

When replacing only the segmentation head with the Segment_LADH, a reduction in both parameter count and GFLOPs was observed, primarily due to the use of depthwise separable convolutions in place of standard convolutions. Furthermore, the asymmetric multi-stage compression architecture and the decoupled mask generation mechanism improved segmentation performance for main branches, fruit-bearing branches, and apples. In detail, AP, mAP, AP_M, and mAP_M increased by 3.5%, 2.8%, 4.2%, and 4.3%, respectively, indicating a balanced improvement in both accuracy and computational efficiency. When both modules were integrated to form the complete SCAL model, the performance improved further, achieving the best results across all metrics: AP=94.9%, mAP=95.1%, AP_M=94.9%, and mAP_M=95.1%. Compared to YOLOv11s, these represent absolute improvements of 6.2%, 4.8%, 6.6%, and 4.9%, respectively.

To visually illustrate the performance impact of each module, the evaluation results were normalized and presented in a 3D bar chart (Figure 13). The figure reveals a consistent upward trend across all metrics as each component was incrementally integrated into the model. Notably, the improvements in segmentation-related metrics (AP_M and mAP_M) were particularly pronounced.

These findings confirm the strong complementarity and synergistic effects between the SCA-T/F module and the Segment_LADH head. Their integration enhances the model's robustness and accuracy in addressing critical challenges, including structural continuity modeling of branches, inference of connections between fruits and branches, fine-grained detail extraction, and boundary recognition under complex lighting conditions. This integrated design offers a reliable technical foundation for intelligent visual perception in orchard environments.

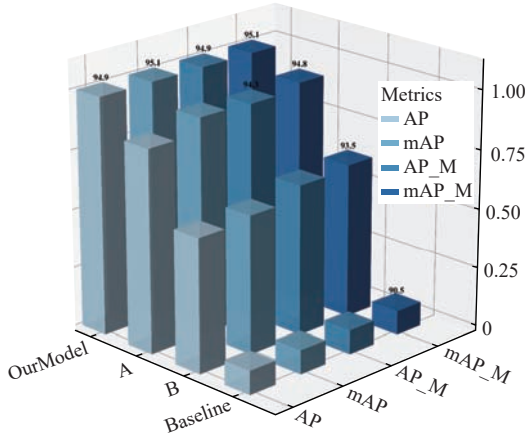


Figure 13 Performance comparison of the ablation models

3.5 Comparison with different models

To comprehensively evaluate the performance advantages of the proposed SCAL model in both detection and segmentation tasks, several mainstream YOLO-based models were selected as baselines, including YOLOv5s[29], YOLOv8s[30], YOLOv10s[31], and YOLOv11s[32]. All models were trained and evaluated on the same unified dataset. The detailed performance metrics are presented in Table 5.

Table 5 Comparative experimental results

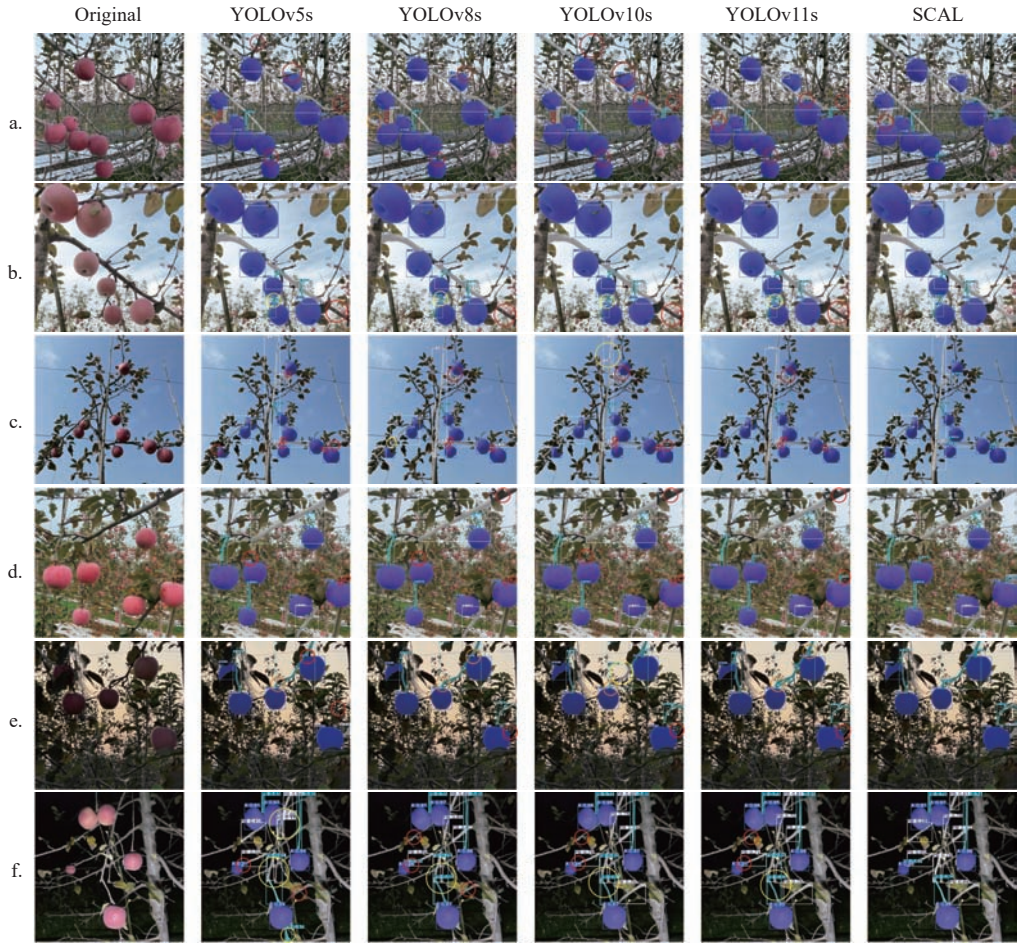
Model	AP/ %	mAP/ %	AP_M/ %	mAP_M/ %	GFLOPs/ G	Parameters/ M	Model size/MB
YOLOv5s	85.3	87.3	85.4	87.7	37.8	976 671 3	19.8
YOLOv8s	88.7	89.7	87.6	89.7	42.4	117 807 61	22.7
YOLOv10s	85.5	88.4	86.0	88.6	40.5	917 109 7	18.7
YOLOv11s	88.7	90.3	88.3	90.5	35.3	100 679 77	20.6
SCAL	**94.9**	**95.1**	**94.9**	**95.1**	**38.8**	**109 077 37**	**22.4**

As listed in Table 5, the SCAL model achieves the highest accuracy across all metrics. Compared to the second-best model, YOLOv11s, SCAL improves mAP by 4.8% and mAP_M by 4.6%, while increasing the computational cost by only 9.9% (from 35.3 to 38.8 GFLOPs). In contrast, YOLOv10s exhibits significantly higher computational overhead (40.5 GFLOPs) with only marginal performance gains, indicating a suboptimal performance-to-cost ratio. YOLOv8s achieves higher accuracy than both YOLOv5s and YOLOv10s, with an AP of 88.7%. However, despite consuming 9.3% more computational resources (42.4 GFLOPs), it underperforms SCAL by 6.2% in AP. These comparisons further validate SCAL's favorable balance between performance and efficiency, demonstrating its ability to deliver substantial accuracy improvements with relatively low computational overhead.

To visually demonstrate the segmentation quality and generalization capability of each model under complex conditions, a set of raw orchard images was used as the test set, and the corresponding segmentation outputs were then visualized for comparative analysis, as shown in Figure 14. The figure presents the segmentation performance of different models on three target categories—main branches, fruit-bearing branches, and apples—in typical orchard scenes.

As shown in the figure, YOLOv5s, YOLOv8s, YOLOv10s, and YOLOv11s exhibit varying degrees of under-segmentation, mis-segmentation, and target adhesion (i.e., the merging of distinct objects into a single instance), particularly under conditions involving occlusion, blurred boundaries, and complex structural overlaps. In contrast, the SCAL model demonstrates enhanced

feature extraction, spatial structure discrimination, and boundary delineation across all three target categories. It consistently shows better adaptability to real-world challenges, including variable lighting conditions, occlusions, and background clutter.



Note: a. Front lighting1; b. Front lighting2; c. Front lighting3; d. Backlighting1; e. Backlighting2; f. Nighttime with auxiliary lighting. In the images, red bounding boxes indicate missed segmentations, yellow bounding boxes represent false segmentations, and orange bounding boxes denote target adhesion, where separate objects are incorrectly merged into a single instance.

Figure 14 Comparison of segmentation results across models in representative orchard scenes

Specifically, YOLOv5s struggles in high-density fruit regions (Figures 14a and 14b), exhibiting blind spots due to limited capacity for handling scale variation and object overlap. YOLOv8s achieves improved performance through an enhanced feature pyramid structure but continues to suffer from boundary blurring and mask adhesion in areas where fruits and branches overlap (Figures 14c and 14d). In low-light environments (Figures 14e and 14f), its ability to distinguish features deteriorates, resulting in poor segmentation. YOLOv10s and YOLOv11s leverage stronger backbone networks and attention mechanisms to enhance structural modeling, but still struggle to differentiate texture-similar targets under challenging lighting conditions and cluttered backgrounds (Figures 14e and 14f). In contrast, SCAL consistently outperforms the other models across all six representative scenes, especially in scenarios with significant scale variance (Figures 14b and 14d). It produces smooth and complete mask boundaries, maintains internal mask consistency, and avoids the fragmented segmentation commonly observed in other YOLO-based models.

This advantage is primarily attributed to SCAL's multi-scale feature fusion mechanism, which enhances global semantic understanding while preserving critical local details. Even under extreme conditions such as uneven illumination or low-light environments (Figures 14e and 14f), SCAL maintains high-quality segmentation and mask coherence, demonstrating strong robustness and adaptability in extracting features under variable field conditions. This coherence and environmental adaptability facilitate the generation of accurate, continuous contour information, significantly enhancing the operational stability and grasping precision of robotic picking systems in complex orchard environments.

3.6 Comparison of models under adverse weather conditions

To comprehensively evaluate the segmentation performance of different models under adverse weather conditions, a composite test set was constructed, encompassing three representative challenging scenarios: Rainy, Foggy, and Mixed (a combination of rain and fog). Based on this dataset, a systematic comparison was conducted among four models: YOLOv11s, SCAL, AirNet-SCAL, and the proposed CPA-SCAL. AirNet-SCAL refers to a SCAL variant integrated with the AirNet module[33], specifically designed for image restoration under adverse weather conditions. This variant serves as a weather-adaptive baseline for evaluating the effectiveness and robustness of the proposed method in complex environmental scenarios.

As listed in Table 6, CPA-SCAL outperforms all comparison models across all key metrics. It achieves an AP of 88.7% and AP_M of 88.0%, with mAP and mAP_M reaching 90.5% and

90.7%, respectively, demonstrating substantial improvements. While AirNet-SCAL shows modest gains over the original SCAL model, it remains inferior to CPA-SCAL, further validating the effectiveness of the proposed weather-aware compensation module.

To visually compare model performance under varying meteorological conditions, Figure 15 presents segmentation results across the three representative scenarios.

Under rainy conditions, reduced brightness and uneven illumination caused by light drizzle significantly degrade segmentation quality. YOLOv11s produces fragmented and incomplete masks, whereas CPA-SCAL, equipped with context-aware enhancement and adaptive feature compensation mechanisms, effectively mitigates these adverse effects, yielding coherent and well-defined segmentation results.

Table 6 Performance comparison under weather variations

Model	AP/ %	mAP/ %	AP_M/ %	mAP_M/ %	Parameters/ M	Model size/MB
YOLOv11s	84.5	87.3	84.5	87.1	100 679 77	20.6
SCAL	85.5	87.6	85.4	87.9	109 077 37	22.4
AirNet-SCAL	84.9	87.6	84.8	87.8	138 844 26	23.1
CPA-SCAL	**88.7**	**90.5**	**88.0**	**90.7**	**144 077 37**	**23.6**

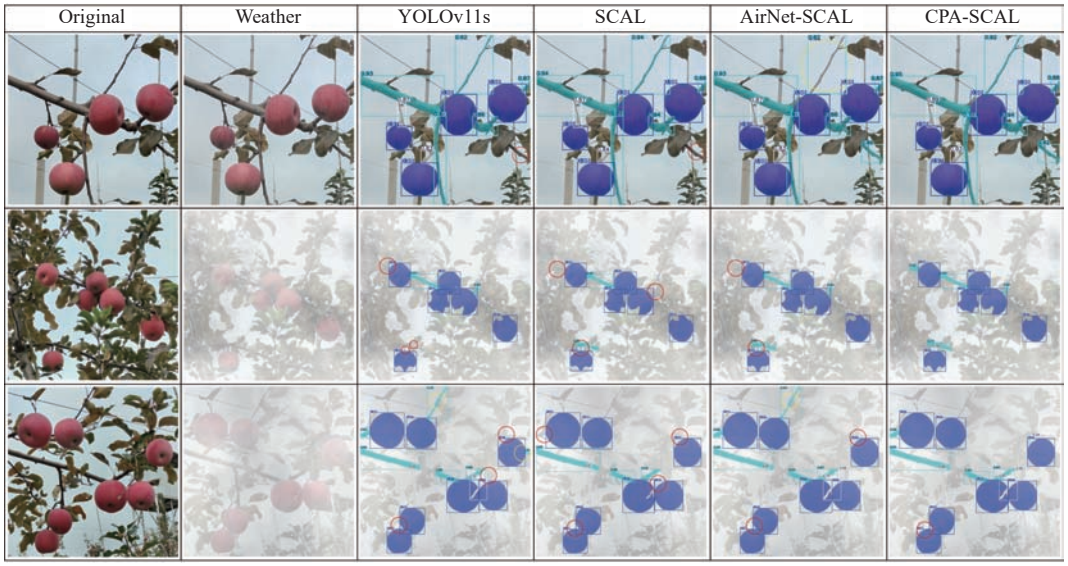


Figure 15 Visual comparison of segmentation results

Under foggy conditions, atmospheric scattering reduces contrast and significantly impairs visibility, particularly for distant objects. YOLOv11s exhibits noticeable missed detections, while SCAL and AirNet-SCAL show acceptable performance in nearby regions but struggle in distant low-visibility zones. In contrast, CPA-SCAL successfully detects and accurately segments all targets, including main branches heavily occluded by dense fog, exhibiting superior adaptability to visibility loss.

Under mixed rain-fog conditions, where compounded degradation causes severe quality loss, target boundaries become increasingly blurred, and structural information is heavily diminished. Most models suffer significant performance drops in this scenario. However, CPA-SCAL maintains high segmentation accuracy, especially in handling multi-target occlusions and boundary ambiguities. The generated masks retain high consistency and shape integrity, highlighting the model's exceptional robustness and generalization capability under extreme environmental conditions.

3.7 Edge deployment and inference optimization of SCAL

The NVIDIA Jetson series represents a class of low-power, GPU-driven edge computing devices that have been widely adopted in various AI applications. To evaluate the inference performance of the SCAL model on resource-constrained platforms, it was deployed on the NVIDIA Jetson AGX Xavier, and experiments were conducted to verify its feasibility within an edge computing environment.

Figure 16 shows the trade-off between inference speed (FPS) and normalized segmentation accuracy, including AP_M and mAP_M, across five models. For clarity, the accuracy values were normalized to the 80%-100% range to highlight the balance between speed and accuracy for each model.

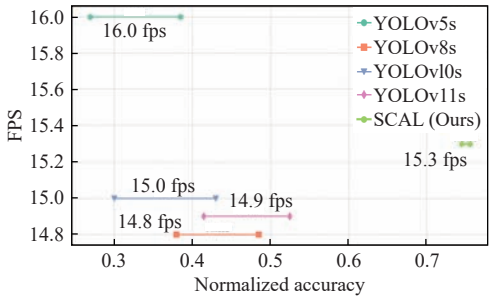


Figure 16 Comparison of inference speed and normalized accuracy across models on NVIDIA Jetson AGX Xavier

As shown in Figure 16, the SCAL model achieves an inference speed of 15.3 fps while maintaining high accuracy, demonstrating strong real-time performance. In contrast, YOLOv5s achieves the highest inference speed (16.0 fps) but exhibits a notable decline in accuracy. Although YOLOv8s and YOLOv11s offer improved accuracy, their relatively slower inference speeds reduce their suitability for real-time applications that require strict latency constraints.

To further enhance inference efficiency and deployment performance, two mainstream low-precision optimization strategies were applied during model export: FP16 (half-precision floating point) and INT8 (8-bit integer quantization). FP16 reduces memory consumption and can accelerate inference on Tensor Core-enabled GPUs, whereas INT8 provides further compression of model size

and increases inference speed. Moreover, the model was optimized using the TensorRT framework to enable hardware-aware acceleration and runtime efficiency.

Applying FP16 quantization reduced the model size to 18 MB and increased the inference speed to 43.2 fps, approximately 2.8 times faster than the unoptimized model. With INT8 quantization, inference speed further improved to 47.2 fps (approximately 3.1 times acceleration), while the model size was reduced to 14.5 MB, significantly lowering storage requirements and computational overhead.

In practical orchard-harvesting scenarios, commonly used depth cameras (e.g., ZED 2i) typically operate at 30.0 fps with 1080P resolution. Thus, the quantized SCAL model is fully capable of processing such image streams in real time. Combined with its high-precision segmentation capability, SCAL is well-suited for deployment on edge computing devices used in fruit-harvesting robots operating under resource constraints and requiring real-time performance.

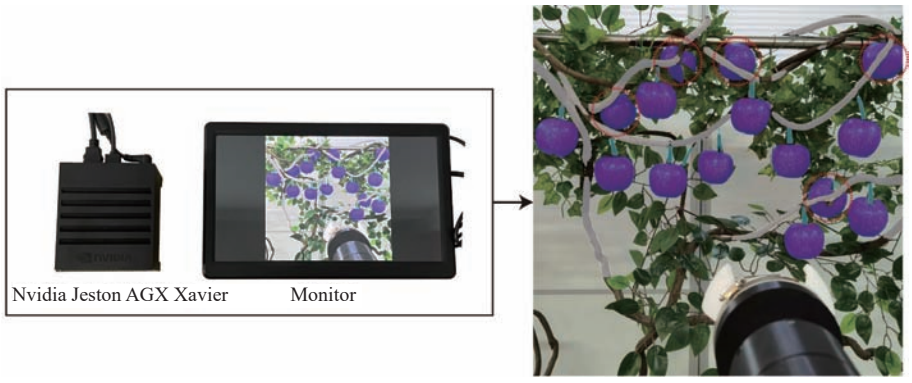
3.8 Laboratory evaluation of SCAL in a simulated orchard environment

To evaluate the effectiveness and generalization capability of the proposed SCAL model in orchard fruit-harvesting scenarios, a simulated orchard environment was constructed in a laboratory setting for real-machine experiments. As shown in Figure 17, the experimental setup includes 15 artificial apple models along with simulated main branches and fruit-bearing branches. Among them, five apples are occluded by main branches to replicate typical complex occlusion conditions observed in real orchard environments. Figure 17b shows an RGB image captured by the ZED 2i depth camera.

a. Setup of the experimental environment b. Images acquired by the ZED2i camera

Figure 17 Experimental environment setup

Within this experimental environment, the SCAL model successfully identified and accurately segmented the contours of all 15 apple samples, including the five occluded by main branches. Additionally, it effectively segmented both the main branches and fruit-bearing branches located in the apple-bearing areas. As illustrated in Figure 18, the model performed exceptionally well in segmenting the boundaries of fruit-bearing branches, even at junctions where they overlap with main branches and apples—an essential capability for improving the success rate of robotic picking operations.



Note: In the images, the red circles indicate the five apples occluded by the main branches.

Figure 18 Segmentation result visualization

This experiment demonstrates the robustness and generalization capability of the SCAL model under complex conditions. It successfully achieved effective segmentation of multiple target categories, providing precise and stable information to support subsequent robotic arm tasks such as obstacle avoidance, path planning, and picking pose estimation. These results further validate the model's potential for effectively utilizing multi-class perception information in practical orchard harvesting applications.

4 Discussion

In complex orchard environments, the proposed model demonstrated the capability to perform instance segmentation of apples, main branches, and fruit-bearing branches using a single visual sensor. It effectively perceived comprehensive visual

information relevant to fruit harvesting tasks, thereby establishing a solid foundation for automated harvesting and offering considerable practical value.

Despite its high segmentation accuracy across multiple target categories, the model was not integrated with depth information captured by visual sensors. As a result, a three-dimensional perception of targets remained absent, which limited the construction of obstacle maps incorporating main and fruit-bearing branches. To enable precise fruit picking and effective obstacle avoidance, future research should incorporate depth cameras to capture real-time spatial information, thereby facilitating scene reconstruction and mapping. In addition, the development of integrated algorithms for picking pose estimation and obstacle-avoidance path planning—taking into account the motion characteristics and kinematics of robotic arms—would contribute to improved harvesting success rates and operational safety[34,35].

Environmental dynamics such as wind or branch vibrations induced by picking actions may also disrupt the visual scene, adversely affecting segmentation stability and potentially resulting in recognition failures or mechanical damage[36]. Therefore, future investigations should examine the impact of such vibrations on visual recognition and localization systems. The incorporation of temporal modeling techniques—such as lightweight temporal convolutions or spatio-temporal feature fusion mechanisms—may enhance the robustness of segmentation under dynamic conditions, thereby improving the model's adaptability in real-world harvesting scenarios.

Experimental results also revealed limitations of the SCAL model under extreme lighting conditions. For instance, in the strongly shadowed region of scene (f) in Figure 14 the model exhibited discontinuous mask generation for main branches, indicating instability in feature representation under uneven illumination. Although the SCAL model outperformed existing models in detecting extremely small or blurred targets (e.g., the slender fruit-bearing branch in scene (a)), this issue was not fully resolved. This suggests that improvements to the hierarchical feature fusion strategy remain necessary. Future studies may explore the integration of multimodal segmentation approaches and the optimization of multi-level feature fusion to enhance the recognition of weak-feature targets[37].

5 Conclusions

This study proposed a multi-category instance segmentation model, SCAL, which achieves accurate segmentation of apples, main branches, and fruit-bearing branches in unstructured orchard environments. The main conclusions are summarized as follows:

1) A Star-CAA module was developed to jointly capture feature and spatial dimensions. By integrating the mathematical properties of the Star operation with the contextual reasoning capability of the CAA attention mechanism, the module enables nonlinear feature fusion, enhances boundary gradient representation, and strengthens directional perception. These advancements collectively improve fine-grained modeling of structural continuity, topological relationships, and local details.

2) Through the integration of the SCA-T/F modules and the Segment_LADH head, the SCAL model demonstrated superior segmentation performance compared with existing algorithms. On the dataset, it achieved a detection/segmentation mAP_M of 95.1%, representing improvements of 4.8% in detection mAP and 4.6% in segmentation mAP_M over the YOLOv11s. Under adverse weather conditions, the CPA-SCAL model achieved a segmentation accuracy of 90.7%, which is 3.6% higher than that of YOLOv11s.

3) To further validate the model's applicability in real-world production scenarios, the SCAL model was deployed on an NVIDIA Jetson AGX Xavier edge device. Following FP16 and INT8 quantization, it achieved real-time segmentation frame rates of 43.2-47.2 fps. Validation experiments in a simulated orchard environment further demonstrated its ability to accurately detect and segment apples under occlusion, providing reliable 3D structural information to support robotic arm path planning and obstacle avoidance in automated harvesting systems.

Acknowledgements

This work was financially supported by the Qinchuangyuan Project of Shaanxi Province (Grant No. 2023KXJ-016).

[References]

[1] Zhou J G, Wang Y K, Chen J, Luo T Y, Hu G R, Jia J L, et al. Research hotspots and development trends of harvesting robots based on bibliometric analysis and knowledge graphs. Int J Agric & Biol Eng, 2024; 17(6): 1–10.

[2] Jia W K, Zhang Y, Lian J, Zheng Y J, Zhao D, Li C J. Apple harvesting robot under information technology: A review. International Journal of Advanced Robotics Systems, 2020; 17(3): 255688461.

[3] Zhang Q. Opinion paper: Precision agriculture, smart agriculture, or digital agriculture. Computers and Electronics in Agriculture, 2023; 211: 107982.

[4] Mhamed M, Zhang Z, Yu J F, Li Y F, Zhang M. Advances in apple's automated orchard equipment: A comprehensive research. Computers and Electronics in Agriculture, 2024; 221: 108926.

[5] Wang Z H, Xun Y, Wang Y K, Yang Q H. Review of smart robots for fruit and vegetable picking in agriculture. Int J Agric & Biol Eng, 2022; 15(1): 33–54.

[6] Nath S. A vision of precision agriculture: Balance between agricultural sustainability and environmental stewardship. Agronomy Journal, 2024; 116(3): 1126–1143.

[7] Du X Q, Meng Z C, Ma Z H, Zhao L J, Lu W W, Cheng H C, et al. Comprehensive visual information acquisition for tomato picking robot based on multitask convolutional neural network. Biosystems Engineering, 2024; 238: 51–61.

[8] Zhang F, Hou Z Y, Gao J, Zhang J X, Deng X. Detection method for the cucumber robotic grasping pose in clutter scenarios via instance segmentation. Int J Agric & Biol Eng, 2023; 16(6): 215–225.

[9] Gao A, Du Y H, Li Y Q, Song Y P, Ren L L. Apple flower phenotype detection method based on YOLO-FL and application of intelligent flower thinning robot. Int J Agric & Biol Eng, 2025; 18(3): 236–246.

[10] Wen S W, Zhou J G, Hu G R, Zhang H, Tao S, Wang Z Y, et al. PcMNet: An efficient lightweight apple detection algorithm in natural orchards. Smart Agricultural Technology, 2024; 9: 100623.

[11] Li T F, Fang W T, Zhao G A, Gao F F, Wu Z C, Li R, et al. An improved binocular localization method for apple based on fruit detection using deep learning. Information Processing in Agriculture, 2023; 10(2): 276–287.

[12] Rong J C, Dai G L, Wang P B. A peduncle detection method of tomato for autonomous harvesting. Complex & Intelligent Systems, 2021; 8: 2955–2969.

[13] Gu W, C Bai S, Kong L X. A review on 2D instance segmentation based on deep neural networks. Image and Vision Computing, 2022; 120: 104401.

[14] Dong L Z, Zhu L C, Zhao B, Wang R X, Ni J P, Liu S C, et al. Semantic segmentation-based observation pose estimation method for tomato harvesting robots. Computers and Electronics in Agriculture, 2025; 230: 109895.

[15] Wang D D, He D J. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. Computers and Electronics in Agriculture, 2022; 196: 106864.

[16] Tong S Y, Yue Y, Li W B, Wang Y X, Kang F, Feng C. Branch identification and junction points location for apple trees based on deep learning. Remote Sensing, 2022; 14(18): 4495.

[17] Sapkota R, Ahmed D, Karkee M. Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. Artificial Intelligence in Agriculture, 2024; 13: 84–99.

[18] Yan B, Liu Y, Yan W H. A novel fusion perception algorithm of tree

branch/trunk and apple for harvesting robot based on improved YOLOv8s. Agronomy, 2024; 14(9): 1895.

[19] Rong Q J, Hu C H, Hu X D, Xu M X. Picking point recognition for ripe tomatoes using semantic segmentation and morphological processing. Computers and Electronics in Agriculture, 2023; 210: 107923.

[20] Kang H W, Chen C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards. Computers and Electronics in Agriculture, 2020; 171: 105302.

[21] Molina J M, Llerena J P, Usero L, Patricio M A. Advances in instance segmentation: Technologies, metrics and applications in computer vision. Neurocomputing, 2025; 625: 129584.

[22] Yang C H, Xiong L Y, Wang Z, Wang Y, Shi G, Kuremot T, et al. Integrated detection of citrus fruits and branches using a convolutional neural network. Computers and Electronics in Agriculture, 2020; 174: 105469.

[23] Karim S, Tong G, Li J, Qadir A, Farooq U, Yu, Y. Current advances and future perspectives of image fusion: A comprehensive review. Information Fusion, 2023; 90: 185–217,

[24] Ma X, Dai X Y, Bai Y, Wang Y Z, Fu Y. Rewrite the stars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle: IEEE, 2024; pp.5694–5703, doi: 10.1109/CVPR52733.2024.00544.

[25] Cai X H, Lai Q X, Wang Y W, Wang W G, Sun Z R, Yao Y Z. Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle: IEEE, 2024; pp.27706–27716. doi: 10.1109/CVPR52733.2024.02617.

[26] Zhang J R, Chen Z H, Yan G X, Wang Y, Hu B. Faster and lightweight: An improved YOLOv5 object detector for remote sensing images. Remote Sensing, 2023; 15(20): 4974.

[27] Zhang Y W, Wu Y, Liu Y M, Peng X Y. CPA-enhancer: Chain-of-thought prompted adaptive enhancer for object detection under unknown degradations. arXiv Preprint, 2024: arXiv: 2403.11220. doi: 10.48550/arXiv.2403.11220.

[28] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 2020; 128: 336–359.

[29] Jocher G, Chaurasia A, Stoken A, Borovec J, NanoCode012, Kwon Y, et al. Ultralytics/YOLOv5. 2020. Available: https://zenodo.org/records/7347926. Accessed on [2024-12-25]. doi: 10.5281/zenodo.3908559.

[30] Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8. 2023. Available: https://github.com/ultralytics/ultralytics. Accessed on [2024-12-28].

[31] Wang A, Chen H, Liu L H, Chen K, Lin Z J, Han J G, et al. YOLOv10: Real-time end-to-end object detection. In: NIPS ’24: Proceedings of the 38th International Conference on Neural Information Processing Systems, 2024; 37: 107984–108011.

[32] Khanam R, Hussain M. YOLOv11: An overview of the key architectural enhancements. arXiv Preprint, 2024: arXiv: 2410.17725. doi: 10.48550/arXiv.2410.17725.

[33] Li B, Liu X, Hu P, Wu Z, Lv J, Peng X. All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; pp.17452–17462.

[34] Hua W J, Zhang Z, Zhang W Q, Liu X H, Hu C, He Y C, et al. Key technologies in apple harvesting robot for standardized orchards: A comprehensive review of innovations, challenges, and future directions. Computers and Electronics in Agriculture, 2025; 235: 110343.

[35] Huang T L, Pan H H, Sun W C, Gao H J. Sine resistance network-based motion planning approach for autonomous electric vehicles in dynamic environments. IEEE Transactions on Transportation Electrification, 2022; 8(2): 2862–2873.

[36] Hu G R, Chen C, Chen J, Sun L J, Sugirbay A, Chen Y, et al. Simplified 4-DOF manipulator for rapid robotic apple harvesting. Computers and Electronics in Agriculture, 2022; 199: 107177.

[37] Yuan J J, Wu F J, Zhao L M, Zhang Q X, Chen Y H. IMFF: A dual-space optimization network via multi-level feature fusion and boundary-aware learning for high-resolution remote sensing scene classification. Expert Systems with Applications, 2025; 296(PartC): 129163.