

## Article

# Lightweight Real-Time Strawberry Volume Estimation Based on Instance Segmentation and Principal-Axis Slicing

Xiang Zhang <sup>1,2</sup>, Quan Gao <sup>1,2</sup> , Yuhai Long <sup>1,2</sup>, Guangchuan Zhang <sup>2,3</sup> and Yun He <sup>1,2,\*</sup><sup>1</sup> College of Big Data, Yunnan Agricultural University, Kunming 650201, China;

2020310460@stu.ynau.edu.cn (X.Z.); gaoq@ynau.edu.cn (Q.G.); 2024240733@stu.ynau.edu.cn (Y.L.)

<sup>2</sup> Key Laboratory for Crop Production and Intelligent Agriculture of Yunnan Province, Yunnan Agricultural University, Kunming 650201, China; 2021210041@stu.ynau.edu.cn<sup>3</sup> College of Mechanical and Electrical Engineering, Yunnan Agricultural University, Kunming 650201, China

\* Correspondence: heyun@ynau.edu.cn

## Abstract

Real-time strawberry volume estimation is a pivotal technology for automated harvesting and precision grading. However, conventional contact methods are prone to damaging fruits, while existing vision-based approaches struggle to balance high accuracy with low computational overhead. To address these challenges, this study proposes a two-stage real-time volume estimation framework coupling a red-green-blue-depth (RGB-D) sensor with an “Instance segmentation–Principal-axis slicing” framework. First, to precisely extract target contours in complex backgrounds, we designed Deformable Feature Aware-YOLO (DFA-YOLO) based on the YOLO11-seg architecture. This model enhances the geometric perception of irregular fruit edges and effectively overcomes the challenges of background noise and multi-scale variations, providing high-precision masks for subsequent spatial mapping. Subsequently, a principal-axis-slicing algorithm extracts the mask’s centroid and principal axis, perpendicularly slicing the mask into infinitesimal micro-slices. By computing and accumulating the pixel-space volume of these slices, the system converts them into precise 3D physical volumes based on RGB-D depth mapping. The entire system was deployed on an NVIDIA Jetson Orin edge computing platform and validated in a greenhouse. Experimental results demonstrate that the estimated volume highly agrees with the true volume, achieving a coefficient of determination ( $R^2$ ) of 0.945 and a mean absolute percentage error (MAPE) of 9.0%. Under typical operating conditions (1–5 targets per field of view), the system maintains an overall frame rate of 8–15 FPS, requiring only 55 ms for single-fruit estimation. This method exhibits favorable stability and lightweight efficiency under the tested greenhouse conditions, offering a reliable solution for real-time non-destructive crop phenotypic monitoring in computationally constrained agricultural environments.



Academic Editor: Dariusz Gozdowski

Received: 19 May 2026

Revised: 24 June 2026

Accepted: 29 June 2026

Published: 1 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** strawberry fruit; volume estimation; instance segmentation; principal-axis slicing; RGB-D sensor; edge computing; non-destructive phenotyping

## 1. Introduction

Strawberries hold a prominent position in the global fruit market due to their distinctive flavor, rich nutritional profile, and high economic return [1]. However, their small size, thin skin, and highly delicate nature make them exceptionally susceptible to mechanical damage [2]. Consequently, traditional contact-based phenotyping is prone to causing injury,

rendering non-contact sensing indispensable for intelligent crop management. Among phenotypic traits, fruit geometry is a crucial indicator; it directly reflects growth status, guides yield prediction and quality grading, and ultimately dictates economic value [3,4]. Furthermore, as agricultural harvesting robots advance, accurate volumetric and morphological data become essential prerequisites for formulating damage-free picking strategies [5]. Therefore, developing a rapid, non-destructive, and high-precision method for real-time volume estimation is of profound practical significance.

Conventional fruit sizing relies predominantly on manual techniques, such as volumetric water displacement or geometric approximation [6]. While water displacement offers high accuracy, it is labor-intensive, time-consuming, and fundamentally incompatible with modern large-scale precision agriculture; moreover, physical contact risks impairing crop development. Conversely, geometric methods estimate volume by fitting measured principal axes to regular solid models. However, strawberries frequently exhibit irregular shapes driven by genetic and environmental factors, leading to substantial estimation errors when applying rigid geometric approximations. For instance, Kurdthongmee et al. [7] employed an ellipsoid fitting method to estimate the volume of mangosteen fruits; under single-view conditions, their method yielded a coefficient of determination ( $R^2$ ) of 0.649 and a mean absolute percentage error (MAPE) of 19.4%. Given that mangosteens possess relatively simple morphological features, such estimation errors would further increase when this approach is applied to crops with highly complex shapes, such as strawberries. Additionally, Liu et al. [8] attempted to match strawberry point clouds using a truncated cone model. Although they achieved reasonable accuracy in estimating the longitudinal, top, and bottom diameters of strawberries, this rigid model failed to accurately fit the irregular boundaries of the fruit, making it unsuitable for precise volume estimation.

To overcome the shortcomings of manual measurements and meet the demands of contemporary precision agriculture, computer vision techniques have been widely introduced to automate fruit size assessment. The fundamental prerequisite for vision-based fruit sizing is the accurate extraction of fruit regions from complex scenes. Early studies predominantly relied on 2D image processing, segmenting targets based on hand-crafted features such as color, shape, and texture. For instance, Liming et al. [9] employed threshold-based segmentation to extract strawberry morphological features, subsequently introducing a multi-attribute decision-making algorithm for quality grading. Similarly, Lu et al. [10] developed an edge-detection method based on color-difference maps to segment mature citrus fruits from cluttered backgrounds. Furthermore, Wu et al. [11] proposed a fusion of color and 3D geometric features for point-cloud segmentation, utilizing a Support Vector Machine (SVM) classifier and Viewpoint Feature Histogram (VFH) filtering to eliminate non-fruit regions. Although these methods achieved accurate segmentation in specific settings, their fundamental reliance on manually designed features limits their ability to automatically extract discriminative information, thereby constraining their robustness in unstructured field conditions [12]. Furthermore, in terms of computational efficiency, traditional multi-step pipelines—especially those involving complex 3D feature extraction and SVM classification—rely heavily on sequential CPU computations. Lacking the parallel acceleration capabilities of modern GPUs, these conventional methods often suffer from high processing latency, making them significantly slower and less suitable for real-time agricultural applications compared to optimized deep learning models [13].

The advent of convolutional neural networks (CNNs) brought revolutionary progress to automated crop phenotyping [14,15]. Numerous studies have applied advanced deep learning models to agricultural tasks, ranging from disease and pest identification [16,17] to maturity determination [18,19] and yield counting [20,21]. However, standard object detectors only generate bounding boxes. While they roughly localize targets, they inevitably



include substantial background clutter and fail to capture precise contour information. To address this limitation, He et al. [22] proposed Mask R-CNN, unifying object detection and instance segmentation to output pixel-wise masks. Applying this to agriculture, Costa et al. [23] utilized Mask R-CNN to segment pecan components, quantifying pixel areas to generate growth curves. Nevertheless, as a two-stage model, Mask R-CNN achieves high accuracy at the cost of substantial computational overhead, limiting its real-time capability.

Conversely, the single-stage YOLO (You Only Look Once) series strikes an elegant balance between speed and accuracy. Starting from YOLOv5, the introduction of instance segmentation variants (YOLO-seg) has shown outstanding performance in diverse agricultural tasks. For example, Chen et al. [24] applied YOLOv9-seg to detect and segment green asparagus, achieving a mean average precision (mAP@50) of 0.952 for segmentation. Likewise, Xue et al. [25] improved YOLO11-seg for tomato defect grading, attaining an 84.1% mAP50, which outperformed mainstream models like Mask R-CNN and earlier YOLO versions.

Beyond segmentation, researchers have extensively investigated the extraction of phenotypic information such as fruit size and volume. One line of research focuses on 2D image analysis. Jaime et al. [26] used Mask R-CNN to segment pomegranate regions and fitted geometric circles to derive relative pixel-level sizes. Qu et al. [27] proposed a reference-object method, automatically utilizing a mature citrus leaf as a dynamic scale to convert pixel size to physical dimensions. Although these 2D approaches avoid tedious camera calibration, the absence of depth information causes inherent scale ambiguity due to varying shooting distances and fruit poses. To resolve this, researchers turned to 3D technologies [28,29]. He et al. [30] developed a multi-view stereo (MVS) system for strawberry phenotyping; however, the requirement for multi-angle capture incurs prohibitive computational and time costs, making real-time performance difficult. While LiDAR provides highly accurate depth information largely insensitive to lighting variations [3,31,32], its historically high hardware cost has restricted large-scale agricultural application. Although recently developed low-cost LiDAR solutions have mitigated this barrier to some extent [33,34], they often lack the natively aligned, high-resolution color textures crucial for fine-grained fruit instance segmentation. Consequently, consumer-grade RGB-D sensors have emerged as a more reliable and cost-effective solution to this dilemma [35]. By hardware-synchronizing optical cameras with depth-sensing modules (e.g., structured light or ToF), these sensors assign a physical distance to each pixel [36]. Providing high-quality multimodal data at a relatively low cost, RGB-D sensors are experiencing rapid expansion in agricultural research [37,38].

In summary, although researchers have explored various technical routes for non-destructive crop phenotyping, existing methods still struggle to achieve an optimal balance among accuracy, efficiency, and cost. Specifically, the fundamental challenges are threefold: 2D analysis and rigid geometric models fail to accurately capture the true 3D volume of irregular fruits; multi-view 3D reconstruction and two-stage segmentation networks suffer from severe computational latency; and traditional depth sensors face hardware cost barriers or modality limitations. To address these issues, this study proposes a low-cost, single-view, real-time strawberry volume estimation framework that integrates an improved YOLO-based instance segmentation model with a principal-axis slicing algorithm. Specifically, the lightweight instance segmentation model accurately delineates irregular strawberry regions in complex scenes, while the subsequent principal-axis slicing algorithm leverages the fruit mask combined with RGB-D depth information to compute the true physical volume, effectively bypassing the limitations of rigid geometric fitting and 2D scale ambiguity. Ultimately, this synergistic approach provides a highly accurate, robust, and real-time solution for computationally constrained agricultural deployments.

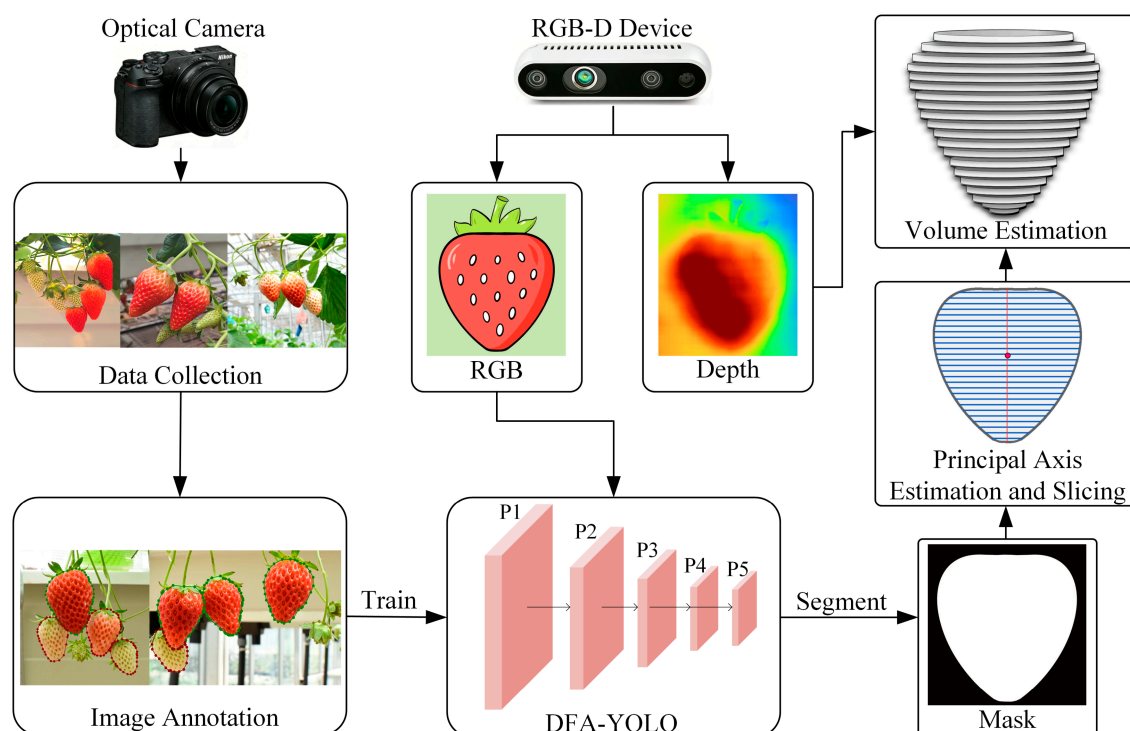
The main contributions of this paper are as follows:

1. To enable rapid volume estimation from a single view in computationally constrained agricultural settings, we introduce a low-overhead real-time estimation algorithm—the principal-axis slicing method. This method fuses geometric features of the 2D instance mask with RGB-D depth data by slicing the mask along its principal axis into micro-slices and accumulating their volumes, achieving a lightweight “2.5D” reconstruction. Compared with conventional 3D reconstruction approaches, this method reduces complex 3D topological operations to a 2D problem, effectively avoiding the high computational complexity of point-cloud processing and the hardware cost of LiDAR, making it highly suitable for multi-target, low-cost, high-throughput strawberry phenotyping in large-scale smart agriculture.
2. We present a systematic geometric and quantitative evaluation of single-view fruit volume estimation methods. By comparing the proposed 2.5D slicing-based dimension-elevation strategy with traditional 3D point-cloud convex-hull reconstruction and 2D ellipsoid approximation, we specifically analyze the inherent biases of conventional baselines in geometric reconstruction, identifying error mechanisms such as the absence of depth information due to single-view occlusion and the loss of edge information caused by rigid model fitting. This comparative analysis clarifies the geometric bottlenecks in current phenotyping methods and verifies the contribution of the proposed approach to morphological fidelity.
3. To precisely extract the strawberry fruit regions, we design an enhanced instance segmentation model, Deformable Feature Aware-YOLO (DFA-YOLO). This model aims to overcome complex agricultural application challenges, such as background interference and target scale variations, providing high-precision fruit masks for the principal-axis slicing volume estimation algorithm.
4. We implemented the principal-axis slicing volume estimation method on an edge platform and conducted agricultural validation. The complete measurement system was deployed on a Jetson Orin edge device and field-tested in a strawberry greenhouse. Results show a coefficient of determination ( $R^2$ ) of 0.945, a mean absolute percentage error (MAPE) of 9.0%, and a single-fruit end-to-end processing time of only 55 ms. Under typical operating conditions (1–5 fruits in the field of view), the system maintains a processing frame rate of 8–15 FPS. The system demonstrates a favorable trade-off between accuracy and computational efficiency, offering a reliable solution for crop phenotyping in computationally constrained environments.

## 2. Materials and Methods

This section details the methodology and overall workflow of the study, covering dataset construction, instance segmentation model optimization, and the principal-axis-slicing volume estimation algorithm. The overarching architecture is illustrated in Figure 1. The workflow proceeds as follows: first, strawberry images are acquired under multiple operating conditions using an RGB camera and meticulously annotated at the pixel level to construct a high-quality dataset for model training and performance evaluation. Second, to address the challenge of strawberry instance segmentation in complex canopy environments, the YOLO11-seg baseline is optimized, trained, and fine-tuned. Finally, volume estimation is performed by fusing the RGB and depth streams obtained synchronously from an RGB-D sensor: the high-precision 2D strawberry masks produced by DFA-YOLO are used to extract the geometric principal axis; micro-slices are generated perpendicular to the principal axis, and each slice is modeled as a 3D micro-cylinder. Their pixel volumes are computed and accumulated, and the spatial mapping relationship provided by

the depth image is then leveraged to achieve accurate conversion from 2D pixels to true physical volume.



**Figure 1.** Overall flow chart. In the pseudo-color depth image, warmer colors indicate regions closer to the camera, whereas cooler colors indicate regions farther from the camera. In the principal-axis estimation and slicing diagram, the red dot denotes the centroid of the strawberry mask, the red line denotes the principal axis, and the blue lines denote the perpendicular slicing lines used for volume estimation.

### 2.1. Dataset Construction and Augmentation

To train and validate the model, a strawberry volume estimation dataset was constructed. Images were collected inside the smart greenhouse of Yunnan Agricultural University, as shown in Figure 2, where cultivars included ‘Zhangji’, ‘Suizhu’, and ‘Benihoppe’. The collection periods were February–March 2025 and February–March 2026. The acquisition device was a Nikon Z30 optical camera (Nikon Corporation, Tokyo, Japan).



**Figure 2.** Greenhouse interior.

It should be explicitly noted that this camera was utilized exclusively for constructing the training, validation, and testing datasets for the DFA-YOLO model. During the actual deployment and final volume estimation experiments, inference was conducted entirely using the RGB images acquired by the RGB-D sensor. Given the differences in resolution, color characteristics, and noise patterns between the two cameras, offline data augmentation was employed during the model training phase to enable the model to adapt to this domain gap. To ensure diversity and quality, the camera was kept approximately level with the fruit at distances of 30–50 cm. Limited by actual environmental conditions and leaf occlusion, some fruits were captured in a single image; however, for fruits with conditions permitting multi-view observation, multiple images were taken from different random viewpoints. Fruit of different maturity stages, sizes, and under varying illumination conditions were sampled to enhance model robustness and generalizability. After discarding invalid and highly repetitive images, 825 high-quality strawberry images were retained.

Annotations were conducted using the open-source labeling tool LabelMe (version 5.11.4). Fruits were categorized into three classes based on color: unripe (entirely bright green or pale-white), half-ripe (commencing reddening or overall pinkish-red), and ripe (entirely bright red). Annotation was performed by personnel with an agricultural engineering background following these principles: outlines closely followed the strawberry contour and excluded the pedicel; for occluded fruit, only the visible portion was labeled, preserving the actual visible silhouette; all clearly visible strawberries in each image were labeled. To prevent data leakage caused by allocating multiple images of the same fruit to different subsets, dataset partitioning was strictly performed at the independent fruit-cluster level. Specifically, all images belonging to the same physical entity were treated as an indivisible unit and assigned entirely to a single subset. Under this rigorous criterion with no biological overlap, the dataset was ultimately split into training (578 images), validation (165 images), and test (82 images) sets at a ratio of 7:2:1. As shown in Table 1, the original dataset contains a total of 5492 annotated strawberry instances (Zhangji 31%, Suizhu 40%, Benihoppe 29%), comprising 1821 unripe, 1714 half-ripe, and 1957 ripe fruits. This relatively balanced distribution minimizes the risk of data distribution shift among the subsets, providing a representative basis for subsequent model training and evaluation.

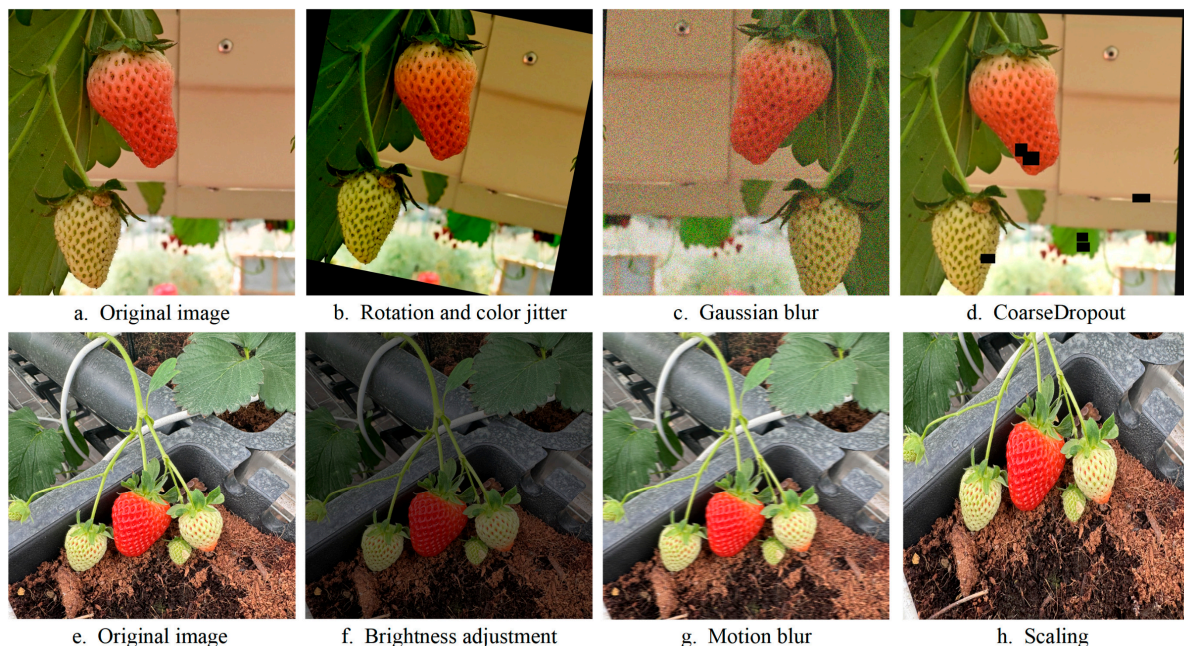
**Table 1.** Distribution of annotated strawberry instances across maturity stages in the dataset subsets.

| Maturity Stage | Train | Validation | Test | Total |
|----------------|-------|------------|------|-------|
| Unripe         | 1268  | 372        | 181  | 1821  |
| Half-ripe      | 1215  | 325        | 174  | 1714  |
| Ripe           | 1358  | 391        | 208  | 1957  |
| Total          | 3841  | 1088       | 563  | 5492  |

To enhance model robustness against complex greenhouse conditions (e.g., variable lighting, occlusion, camera motion), a probabilistic data augmentation strategy using the Albumentations library (version 2.0.3) was applied. Diverse image perturbations were dynamically generated using the following independent execution probabilities: (1) Geometric transformations: Horizontal flipping and spatial manipulations (shifting  $\pm 5\%$ , scaling  $\pm 10\%$ , rotation  $\pm 15^\circ$ ) were applied at a 50% probability to simulate diverse harvesting viewpoints. (2) Lighting and color: Random brightness/contrast adjustments ( $\pm 20\%$ ) and HSV jittering (hue  $\pm 5$ , saturation/value  $\pm 15$ ) were triggered at 50%. The hue shift was strictly limited to preserve sensitive strawberry color features. (3) Image degradation: Gaussian/motion blur and Gaussian noise were each applied at a 20% probability to mimic capture interference. (4) Occlusion: CoarseDropout (30% probability) generated



3–6 random masks to replicate dense leaf coverage. Ultimately, this stochastic strategy expanded the training set to 1284 images, illustrated in Figure 3.



**Figure 3.** Data augmentation example. The black squares in the CoarseDropout image indicate randomly masked regions used to simulate occlusion under greenhouse conditions.

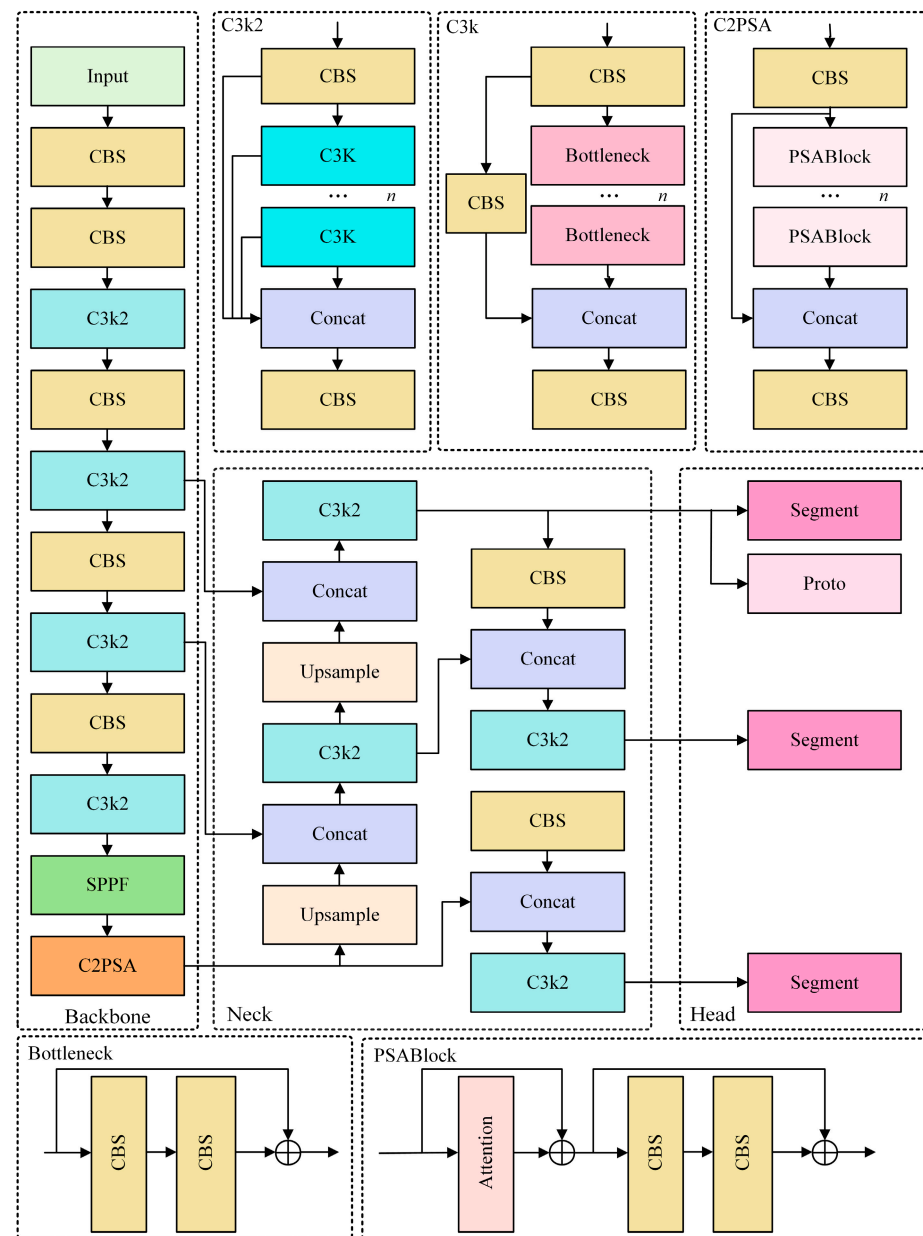
## 2.2. Instance Segmentation Model Construction

YOLO11 is a real-time object detection model released by Ultralytics in 2024 [39]. Building on YOLOv8, it introduces three main improvements: the C3k2 module replaces C2f in the feature extraction stage, the C2PSA attention mechanism enhances detail perception, and a depthwise separable convolution (DWConv)-based lightweight detection head is adopted [40]. These modifications significantly improve detection accuracy and parameter efficiency while maintaining real-time inference speed. YOLO11 also supports multiple vision tasks, including instance segmentation, object classification, pose estimation, and oriented bounding-box detection.

YOLO11-seg shares the same network architecture as YOLO11, which consists of three parts: backbone, neck, and head. However, in the head, YOLO11-seg adds prototype mask (Proto) and mask coefficient branches to generate pixel-wise masks while localizing targets. The architecture of YOLO11-seg is shown in Figure 4. YOLO11-seg offers five size variants (n, s, m, l, x); among them, YOLO11s-seg provides relatively high segmentation accuracy at low computational cost, making it well-suited for deployment in computationally constrained agricultural settings such as fields and greenhouses. Therefore, YOLO11s-seg was selected as the baseline for this study.

Although YOLO11-seg possesses strong segmentation capability, in practical agricultural scenes it faces difficulties such as complex backgrounds, irregular fruit edges, and large target-scale variation, which degrade segmentation accuracy. To address these issues, we propose DFA-YOLO, which improves YOLO11-seg as follows: in the deeper layers of the backbone, the C3k2\_DCNv4 module and the SHSA attention mechanism are introduced to enhance feature extraction under complex conditions; in the neck, a Concat\_BiFPN module is constructed, employing an adaptive weighted feature fusion mechanism to dynamically assign weights to different-scale input features and thereby optimize the feature fusion strategy. The DFA-YOLO architecture is shown in Figure 5.

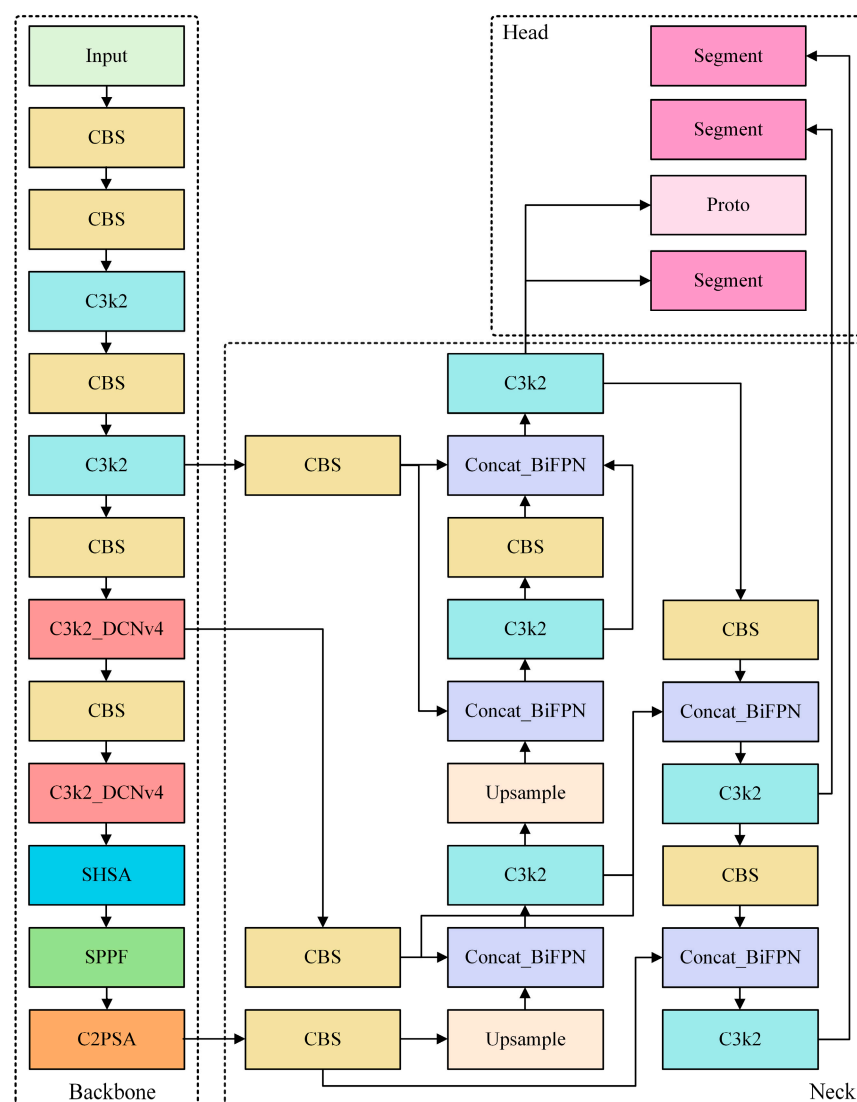




**Figure 4.** YOLO11-seg network architecture.

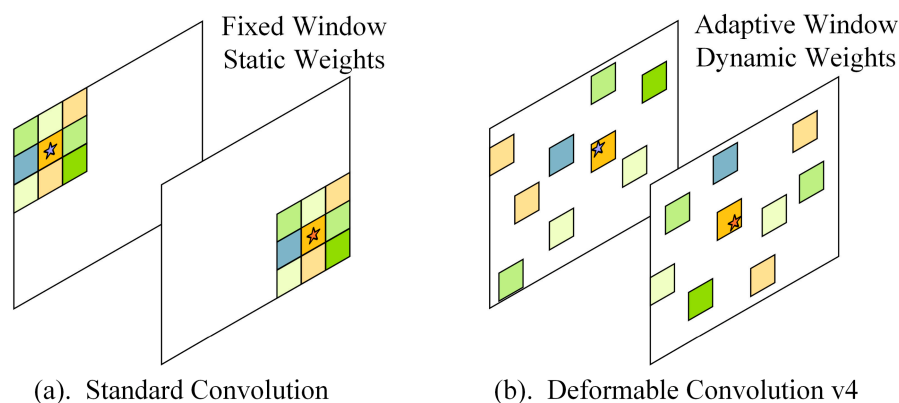
#### 2.2.1. C3k2\_DCNv4

C3k2 is a core functional module in YOLO11, designed based on the previous-generation C2f architecture. By stacking multiple bottleneck layers with residual connections, it achieves rich gradient flow while maintaining computational efficiency. However, its standard convolution is limited by a fixed sampling grid Figure 6a; the receptive field is decoupled from target geometry, and the weights are static, making it difficult to adaptively fit irregular deformations. In strawberry segmentation, fruit often exhibit undulating surfaces, uneven contours, and partial occlusion; standard convolutions frequently yield imprecise boundary responses, leading to masks with jagged edges or missing details. To overcome these limitations, we introduce the fourth generation of Deformable Convolution v4 (DCNv4) to reconstruct the C3k2 module. DCNv4 [41] is the latest in the Deformable Convolution series; its core innovations include the introduction of dynamic weights and learnable spatial offsets for each sampling point, allowing the convolutional kernel to adaptively adjust sampling locations according to target shape and flexibly capture geometric features Figure 6b.



**Figure 5.** DFA-YOLO network architecture.

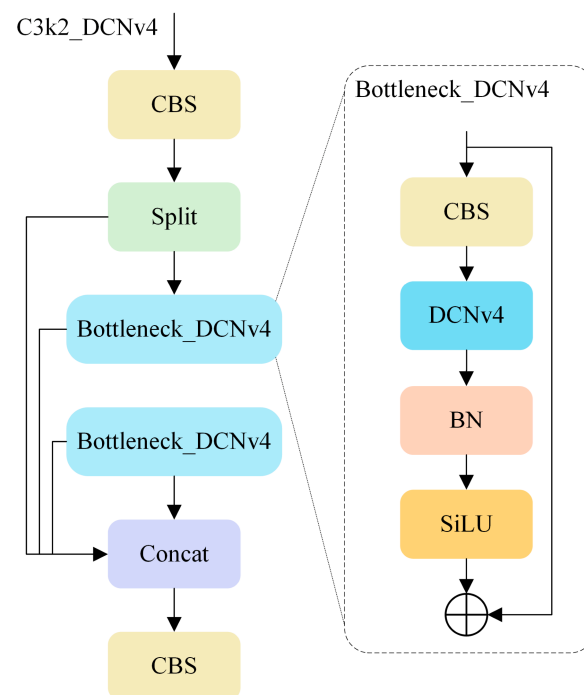
★ ★ query pixels    ■ ■ ■ response pixels



**Figure 6.** Sampling mechanism (redrawn by the authors based on [41]).

Based on the above advantages, we designed the C3k2\_DCNv4 module (Figure 7) and introduced it into the backbone to better adapt to the edge feature extraction of irregular targets. The module adopts a Cross-Stage Partial Network (CSPNet) structure and replaces the standard bottleneck of the original C3k2 with a Bottleneck\_DCNv4 unit.

Each Bottleneck\_DCNv4 contains two convolutional units: the first is a  $1 \times 1$  standard convolution that reduces the channel dimension of the incoming feature map, and the second is an encapsulated DCNv4 operator that reshapes the feature map into a sequence, applies deformable convolution, then restores the dimensions and passes through batch normalization and activation. By replacing two deep-layer C3k2 modules in the YOLO11-seg backbone with C3k2\_DCNv4, the model's ability to perceive irregular strawberry edges is markedly enhanced, providing more precise segmentation masks for subsequent high-accuracy volume estimation.



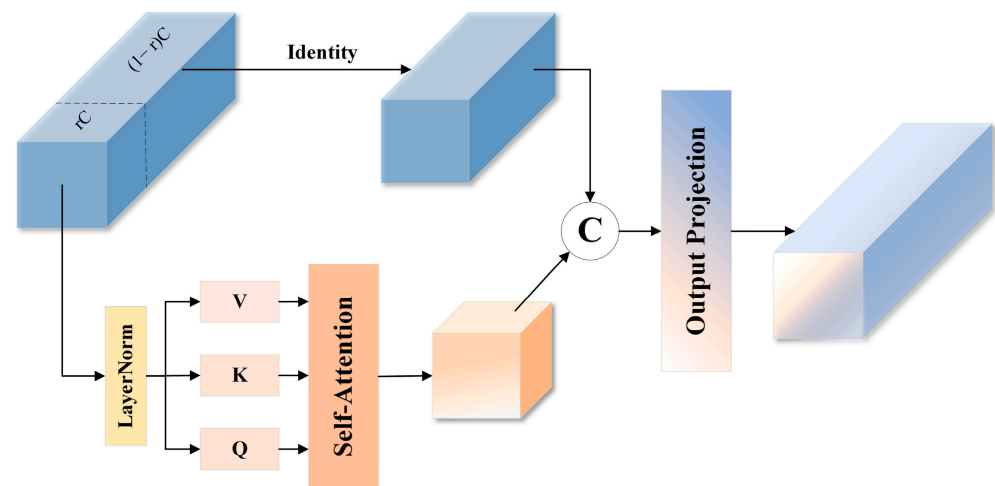
**Figure 7.** Structure of the C3k2\_DCNv4 module.

### 2.2.2. SHSA

SHSA [42] is a novel attention module originally proposed to address the computational redundancy inherent in the Multi-Head Self-Attention (MHSA) mechanism of Vision Transformer models. This study selects SHSA over other common lightweight attention mechanisms (e.g., ECA, CBAM, and SimAM) due to the specific requirements of real-time, high-precision strawberry volume estimation. ECA offers high efficiency but focuses strictly on channel attention, neglecting the spatial boundary details that are critical for our subsequent slicing algorithm. CBAM captures both channel and spatial features, but its sequential structure introduces substantial computational overhead. Furthermore, although SimAM is parameter-free, its complex 3D energy function requires dense mathematical operations, which often bottleneck actual inference speeds on edge devices.

To avoid these bottlenecks and massive computational redundancy, we chose to introduce the SHSA module to improve our model. Its core design is to apply single-head attention to only a subset of channels for spatial feature aggregation, while the remaining channels pass through unchanged; a subsequent full-channel projection layer ensures that attention information effectively propagates to all channels (Figure 8). This mechanism enables SHSA to exhibit an excellent balance in the strawberry fruit segmentation task, retaining sensitivity to complex fruit boundaries while maintaining the ultra-low latency required for real-time agricultural applications. Specifically, the working principle of SHSA can be divided into the following steps:





**Figure 8.** SHSA workflow. The dashed line indicates channel splitting. Different colors distinguish different processing branches/modules. Q, K, and V denote the query, key, and value tensors, respectively, and C denotes concatenation (redrawn by the authors based on [42]).

1. Channel splitting: For an input feature map  $X \in \mathbb{R}^{B \times C \times H \times W}$ , SHSA first splits it along the channel dimension by a ratio  $r$  into an attention branch  $X_{att} \in \mathbb{R}^{B \times rC \times H \times W}$  and an identity branch  $X_{id} \in \mathbb{R}^{B \times (1-r)C \times H \times W}$ , where  $B$  is the batch size,  $C$  the number of channels, and  $W$  and  $H$  the height and width of the feature map. The identity branch is directly retained without participating in attention computation to reduce complexity.
2. Attention computation: Group normalization is applied to  $X_{att}$  to stabilize training, and then a  $1 \times 1$  convolution simultaneously generates the query (Q), key (K), and value (V) tensors:

$$Q, K, V = \text{Split}(\text{Conv}_{1 \times 1}(X_{att})), \quad (1)$$

where  $Q, K \in \mathbb{R}^{B \times D \times H \times W}$ ,  $V \in \mathbb{R}^{B \times rC \times H \times W}$ , and  $D$  is the dimension per head. Subsequently,  $Q, K, V$  are flattened into sequences, and single-head attention weights are computed and normalized:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V. \quad (2)$$

3. Feature fusion and projection: The weighted attention branch and the identity branch are concatenated along the channel dimension, restoring the original channel count  $C$ . Finally, a projection layer consisting of a ReLU activation and a  $1 \times 1$  convolution performs non-linear transformation and inter-channel fusion to produce the output:

$$\text{Output} = \text{Proj}(\text{Concat}(X'_{att}, X_{id})). \quad (3)$$

In this study, the SHSA module is integrated into the backbone network. Through global self-attention, SHSA effectively captures long-range dependencies between strawberry fruits and the background, enhancing the capability to handle occlusion and overlap. Moreover, the module directs the model's focus toward the fruit contour region, helping to generate more accurate segmentation masks, which is critical for the subsequent mask-based volume estimation. Meanwhile, its single-head design avoids the computational redundancy of multi-head attention, boosting performance without significantly increasing inference overhead, thereby well satisfying the real-time requirements of this study.

### 2.2.3. BiFPN-Based Adaptive Multi-Scale Feature Fusion Network

During strawberry greenhouse and field phenotyping, the scale of fruit in images can vary drastically owing to differences in camera viewpoint and fruit growing position. When performing multi-scale feature fusion, the neck of YOLO11 adopts a straightforward channel concatenation (Concat) operation. This undifferentiated fusion approach implicitly assumes that feature maps at all scales contribute equally to the final prediction, which can easily cause semantic conflicts between shallow high-resolution features rich in edge detail and deep low-resolution features rich in semantic information, and may also introduce redundant background noise, ultimately degrading segmentation accuracy.

To address these limitations, we introduce the adaptive weighted feature fusion mechanism of the Bi-directional Feature Pyramid Network (BiFPN) into the neck, constructing Concat\_BiFPN modules to replace the original Concat operations. The core of the adaptive weighted fusion mechanism is that the network dynamically evaluates the contribution of different-scale input features during training, thereby assigning dynamic weights to achieve differential amplification of important features and improving feature quality and effectiveness [43]. In our specific network implementation, each feature map participating in multi-scale fusion is assigned an independent learnable weight  $w_i$ , allowing the network to learn the importance of each input. During the initialization phase, all learnable weights  $w_i$  are initialized to a scalar value of 1.0 to ensure that all feature scales contribute equally at the beginning of the training process. To ensure non-negativity and numerical stability of the fusion weights while avoiding the gradient vanishing problem of conventional ReLU activation for negative values, the Swish function is adopted to non-linearly constrain the initial weights; then a fast normalized fusion strategy computes the dynamically assigned weight  $W_i$ :

$$W_i = \frac{\text{Swish}(w_i)}{\sum_{j=1}^n \text{Swish}(w_j) + \epsilon}, \quad (4)$$

where  $\epsilon$  is a very small constant to prevent division by zero.

After computing the normalized weight for each feature map, the Concat\_BiFPN module performs element-wise weighted summation of the input feature maps  $X_i$  to output the fused feature  $Y$ :

$$Y = \sum_{i=1}^n W_i \cdot X_i, \quad (5)$$

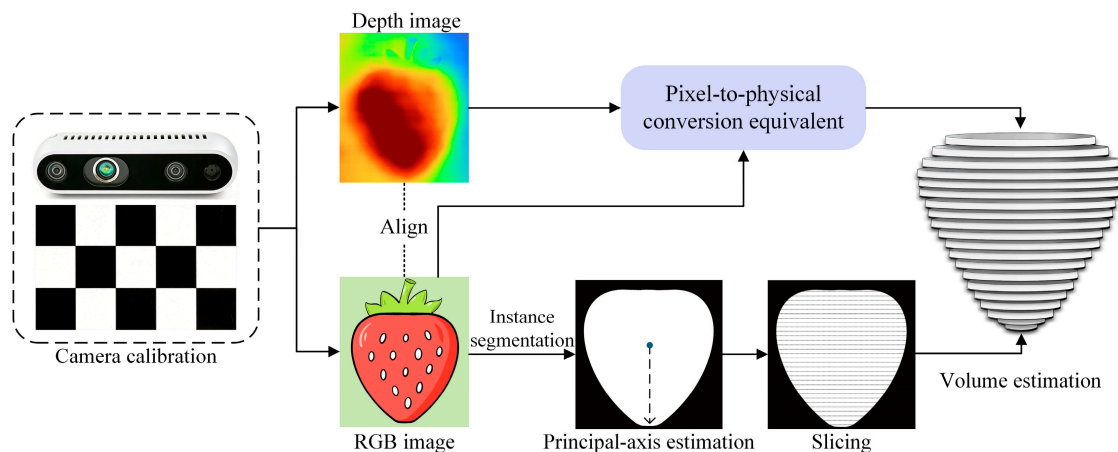
It should be noted that the input feature maps  $X_i$  themselves are not normalized prior to the summation. The fast normalized fusion strategy is applied exclusively to the learnable scalar weights  $w_i$ . Normalizing the scalar weights rather than the high-dimensional feature maps guarantees numerical stability while preserving the real-time inference speed of the network.

This adaptive weighted fusion mechanism readily addresses the challenges of varying morphology and large scale span in strawberry instance segmentation within complex greenhouse and field environments. When segmenting large-scale near-field fruit or severely occluded targets, the model adaptively increases the response weight of deep semantic features to strengthen overall contour recognition; when detecting small-scale distant fruit, it markedly amplifies the weight of shallow high-resolution features to precisely capture local edge texture. This dynamic multi-scale feature fusion strategy effectively suppresses feature interference from complex foliage and canopy background, providing more accurate strawberry masks for the subsequent principal-axis slicing volume estimation.

### 2.3. Volume Estimation Based on Principal-Axis Slicing

Most strawberry cultivars produce fruit that are approximately irregular conical bodies. Based on this characteristic, we designed a principal-axis slicing volume estimation method.

As illustrated in Figure 9, the method leverages the synchronously acquired RGB image and depth map from an RGB-D sensor, combined with the segmentation masks output by the improved instance segmentation model, to implement fast, real-time, and accurate volume estimation through a principal-axis slicing strategy. This approach obviates the need for multi-view stereo and avoids computationally intensive 3D point-cloud reconstruction, making it suitable for real-time processing in greenhouse and field settings.



**Figure 9.** Workflow of the proposed volume estimation method. In the pseudo-color depth image, different colors represent relative depth values. The dashed arrow indicates the estimated principal-axis direction, and the horizontal lines indicate the slicing lines used for volume estimation.

Specifically, the algorithm consists of the following steps:

1. Camera calibration and pre-processing: The RGB-D sensor serves as the image acquisition tool for strawberry volume estimation. Such sensors consist of an RGB module and a depth module; the depth module outputs a depth map that is pixel-wise aligned with the RGB image, where each pixel value represents the spatial distance (in millimeters) between the scene point and the camera optical center. Thus, the RGB-D sensor enables the conversion of the pixel dimensions of a strawberry mask into accurate real-world dimensions. In practice, the RGB module suffers from radial and tangential distortion that can cause image distortion and destroy the spatial consistency between RGB and depth images, impairing the mapping accuracy between pixels and depth values. Therefore, we employed a checkerboard calibration method to calibrate the RGB module. By capturing checkerboard images in various poses and extracting the checkerboard corner coordinates  $(X, Y)$  together with their pixel coordinates  $(u, v)$ , the mapping relationship between the world coordinate system and the image coordinate system is established:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \quad (6)$$

where  $s$  is a scale factor,  $\begin{bmatrix} r_1 & r_2 & t \end{bmatrix}$  are the rotation matrix and translation vector of the camera, and  $K$  is the intrinsic matrix. Using this mapping and the checkerboard corner coordinates, the intrinsic matrix  $K$  can be solved, and the distortion coefficients are obtained through non-linear optimization, completing the calibration.

After acquiring depth and RGB images, the depth map is re-projected onto the RGB image plane to achieve pixel-level alignment, with the RGB image serving as the reference. Simultaneously, multi-level filtering (spatial filtering, temporal filtering, hole filling) is applied to the depth map to reduce noise. The aligned depth map then

provides, for each pixel position  $(u, v)$ , the true depth value  $depth(u, v)$  of the same position in the color image, which is the basis for subsequent true-volume calculation.

2. Instance Segmentation and Mask Processing: The RGB image captured by the depth camera is fed into the trained DFA-YOLO model to perform instance segmentation, yielding accurate pixel-wise strawberry fruit masks. To ensure the accuracy of the subsequent volume calculation, the output masks require processing to suppress edge noise that could cause measurement errors. First, a threshold  $T = 0.5$  is applied to convert the continuous probability map output by the network's Sigmoid activation into a definitive binary mask:

$$M(u, v) = \begin{cases} 1, & P(u, v) \geq 0.5 \\ 0, & \text{Otherwise} \end{cases} \quad (7)$$

where  $P(u, v)$  is the probability score for the strawberry class at pixel  $(u, v)$  as output by the model.  $T = 0.5$  was selected as it represents the standard mathematical decision boundary, effectively balancing the trade-off between over-segmentation and under-segmentation. This threshold ensures the precise extraction of the strawberry's true physical contour. After obtaining the binary strawberry mask, a single erosion operation with a  $3 \times 3$  cross-shaped structuring element is performed to eliminate isolated noise points at the mask edges and make the mask boundary more continuous. At the same time, the spatial alignment between RGB and depth images achieved in step (1) is exploited, and the pixel-wise mapping relationship is used to directly extract the valid depth information for each pixel within the corresponding strawberry mask region.

3. Pixel-to-Physical Conversion Factor and Strawberry Principal-Axis Estimation: After obtaining the high-quality strawberry mask, we propose a fast volume estimation method that combines depth data and 2D geometric features extracted from the strawberry mask. To eliminate background interference and abrupt noise at depth-sensor edges, an effective working depth range of 0.1–1 m is applied as a threshold filter to the depth data within the mask region; then the median of the valid depth set,  $D_{med}$ , is extracted as the global median depth representing the overall depth of the target. Based on the pinhole camera model, the pixel-to-physical conversion factor  $S_p$  is given by

$$S_p = \frac{D_{med}}{f_{avg}}, \quad (8)$$

where  $f_{avg}$  is the average of the horizontal and vertical focal lengths from the camera intrinsic matrix.

Because the strawberry is an irregular conical-like body, the cross-sectional dimension along its principal axis changes most systematically. Hence, we employ Principal Component Analysis (PCA) on the set of pixel coordinates within the binary mask to determine the growth principal axis of each strawberry. Let the set of valid pixel coordinates within the binary mask be  $\Omega = \{(x_i, y_i) | i = 1, 2, \dots, N\}$ , where  $N$  is the total number of pixels. The geometric centroid  $C(\bar{x}, \bar{y})$  of the mask region is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i, \quad (9)$$



Centering the set  $\Omega$  by subtracting the centroid  $C$  yields the centered matrix:

$$X = \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_N - \bar{x} & y_N - \bar{y} \end{bmatrix}, \quad (10)$$

the covariance matrix is then

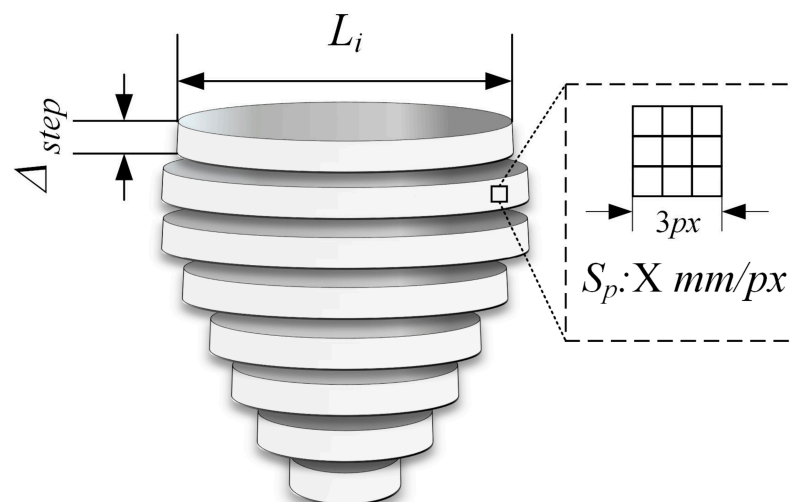
$$\Sigma = \frac{1}{N-1} X^T X = \begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) \\ \text{Cov}(y, x) & \text{Cov}(y, y) \end{bmatrix}, \quad (11)$$

where  $\text{Cov}(x, y)$  denotes the covariance between the  $x$  and  $y$  coordinate directions, reflecting the spread of the target in 2D space. To extract the principal direction, we perform eigendecomposition of the covariance matrix  $\Sigma$  and take the eigenvector corresponding to the largest eigenvalue as the principal-axis direction of the strawberry.

4. Volume Estimation based on micro-slices: After obtaining the principal-axis direction of the strawberry mask, we propose a volume approximation algorithm based on the idea of infinitesimal integration, leveraging the approximately rotationally symmetric spatial geometry commonly exhibited by strawberry fruit. Specifically, the mask is traversed along the principal axis with a pixel step size  $\Delta_{step}$ . At each stepping node on the principal axis, a line perpendicular to the principal axis is drawn; the local 3D space corresponding to this cross-section is approximately regarded as a thin cylindrical slice. The pixel length  $L_i$  of that slice within the mask is computed, and together with the pixel-to-physical conversion factor  $S_p$ , the corresponding physical thickness is  $H = \Delta_{step} \cdot S_p$  and the physical radius is  $R_i = (L_i \cdot S_p) / 2$ . Thus, the volume  $V_i$  of the  $i$ -th cylindrical micro-slice can be expressed as

$$V_i = \frac{\pi}{4} \cdot L_i^2 \cdot S_p^3 \cdot \Delta_{step}, \quad (12)$$

The meaning of the parameters is illustrated in Figure 10.



**Figure 10.** Schematic diagram of volume calculation parameters.

After traversing all  $n$  valid cross-sections along the principal axis, according to the integration philosophy of calculus, the global true physical volume  $V_{strawberry}$  of the strawberry fruit is obtained by summing the volumes of all local micro-slices:

$$V_{strawberry} = \sum_{i=1}^n V_i = \frac{\pi}{4} \cdot S_p^3 \cdot \Delta_{step} \cdot \sum_{i=1}^n L_i^2. \quad (13)$$

Conventional crop phenotyping techniques (e.g., 3D point-cloud reconstruction, multi-view stereo) generally rely on massive spatial topological operations and inefficient multi-view acquisition, inevitably resulting in extremely high computational cost and time consumption that struggle to meet the real-time operational demands of agricultural robots. In contrast, the proposed principal-axis slicing volume estimation method elegantly achieves dimensionality reduction in the complex 3D reconstruction problem. It avoids cumbersome point-cloud generation and model reconstruction steps; relying solely on the 2D high-precision fruit mask and its principal axis extracted by the instance segmentation model, together with the single-depth mapping provided by the RGB-D sensor, the conversion from 2D pixels to 3D physical volume is accomplished based on the micro-integration concept. This innovative volume estimation method not only fundamentally breaks the computational bottleneck of edge devices but also exhibits excellent robustness in measurement accuracy, offering an efficient new paradigm for smart agriculture scenarios with severely limited computational resources.

#### 2.4. Experimental Environment and Parameters

The training platform for the instance segmentation model in this paper runs on Windows 10, with an Intel® Core™ i9-13900K CPU, 128 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM. The programming language used is Python 3.11, and the deep learning framework was PyTorch 2.2.2 with CUDA 12.1. The training hyperparameters are listed in Table 2.

**Table 2.** Training hyperparameters.

| Hyperparameters       | Value     |
|-----------------------|-----------|
| Input image size      | 640 × 640 |
| Training epoch        | 300       |
| Batch size            | 32        |
| Number of workers     | 8         |
| Initial learning rate | 0.001     |
| Optimizer             | SGD       |
| Momentum factor       | 0.937     |
| Weight decay          | 0.0005    |

To investigate the performance of the entire strawberry volume estimation method in a real agricultural setting, deployment experiments were conducted in a strawberry greenhouse. The RGB-D sensor used was an Intel RealSense D435i (Intel Corporation, Santa Clara, CA, USA), and the lightweight edge deployment device was an NVIDIA Jetson Orin Nano Super (NVIDIA Corporation, Santa Clara, CA, USA), as shown in Figure 11.

Under the hardware configuration and hyperparameter settings used in this study, the final DFA-YOLO model required 2.24 h (8064 s) to complete 300 training epochs, corresponding to an average of 26.88 s per epoch.



**Figure 11.** Edge deployment device and RGB-D sensor used in the experiment.

The NVIDIA Jetson Orin Nano Super features an NVIDIA Ampere-architecture GPU with 1024 CUDA cores and 8 GB of memory, delivering up to 67 TOPS of AI performance with a peak power consumption of only 25 W. Its developer kit measures just 100 mm × 79 mm × 21 mm and weighs approximately 180 g, making it exceptionally compact and suitable for agricultural deployment. The Intel RealSense D435i integrates an RGB camera, two infrared cameras, and an infrared projector, and perceives depth through active stereo infrared technology. Its effective measurement range is 0.3 m–3.0 m, and its dimensions are only 90 mm × 25 mm × 25 mm, facilitating integration into various edge-vision solutions.

### 2.5. Evaluation Metrics

To comprehensively assess the performance of the proposed method in the strawberry volume estimation task, we employ the following metrics. For segmentation mask accuracy, precision (P), recall (R), and mean average precision (mAP) are used to evaluate the strawberry instance segmentation quality in complex scenes. For phenotypic reconstruction, the coefficient of determination ( $R^2$ ), absolute percentage errors (APE), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) are utilized to verify the reliability of the principal-axis slicing volume estimation algorithm. Additionally, model parameters (Params), floating-point operations (GFLOPs), and frames per second (FPS) are introduced to evaluate the real-time deployment efficiency of both the model and the entire system.

## 3. Results

### 3.1. Performance Evaluation of Instance Segmentation Models

#### 3.1.1. Ablation Experiment

To validate the effectiveness of the proposed improvement strategies for strawberry instance segmentation and to deeply analyze the independent contribution of each module as well as their synergistic effects, we conducted an ablation study using YOLO11s-seg as the baseline. The three improved modules were progressively introduced and combined in different configurations to quantitatively assess their impact on model performance and their interactions. The ablation results are presented in Table 3.

**Table 3.** Results of Ablation Experiments. Note: “√” indicates module inclusion.

| BiFPN | C3k2_DCNv4 | SHSA | P (%) | R (%) | mAP (%) |           | Params (M) | GFLOPs |
|-------|------------|------|-------|-------|---------|-----------|------------|--------|
|       |            |      |       |       | mAP@50  | mAP@50-95 |            |        |
| √     | √          | √    | 87.6  | 84.4  | 90.5    | 80.5      | 10.08      | 35.6   |
|       |            |      | 84.3  | 90.5  | 92.2    | 81.5      | 10.44      | 38.2   |
|       |            |      | 88.5  | 87.3  | 92.4    | 83.4      | 9.92       | 35.4   |
| √     | √          | √    | 88.3  | 85.7  | 92.0    | 82.4      | 10.38      | 35.8   |
|       |            |      | 86.2  | 85.8  | 91.0    | 82.1      | 10.27      | 38.0   |
|       |            |      | 87.6  | 87.7  | 92.4    | 82.9      | 10.73      | 38.4   |
| √     | √          | √    | 89.7  | 86.3  | 93.5    | 83.8      | 10.27      | 35.4   |
|       |            |      | 88.7  | 89.8  | 94.4    | 83.7      | 10.57      | 38.2   |

Table 3 shows that replacing the original C3k2 module in the backbone with C3k2\_DCNv4 improved the model’s mAP@50 and mAP@50-95 by 1.9 and 2.9 percentage points, respectively, while reducing Params and GFLOPs by 0.16M and 0.2G. These results indicate that the introduction of C3k2\_DCNv4 optimizes the network parameter structure, reduces computational complexity, and refines the spatial sampling logic of convolutional layers, thereby enhancing the model’s ability to model irregular objects and dense strawberry scenes and producing masks with more accurate boundaries. After incorporating the weighted feature fusion mechanism of BiFPN into the neck network, the model’s capacity for cross-scale integration of shallow edge details and deep semantics was substantially improved. Experimental results show that this modification increased R and mAP@50 by 6.1 and 1.7 percentage points, respectively, fully demonstrating that the dynamic feature weighting of BiFPN effectively bridges the semantic gap between features at different scales and strengthens the network’s ability to capture small-scale fruits and occluded targets under complex greenhouse canopies. Introducing the SHSA module into the deep layers of the backbone enabled the establishment of global dependencies, which helps improve feature perception under noisy backgrounds. Without significantly increasing computational complexity, the network was able to focus more on strawberry targets while attenuating interference from irrelevant background information. With the addition of the SHSA module, the model’s mAP@50 and mAP@50-95 reached 92.0% and 82.4%, respectively, confirming that the self-attention mechanism plays a positive role in filtering out background noise.

In multi-module combination experiments, the joint application of C3k2\_DCNv4 and SHSA exhibited a clear synergistic effect. C3k2\_DCNv4 flexibly adapts to the irregular geometric features of strawberry during feature extraction and captures abundant edge details, but the produced feature maps tend to retain similarly shaped interference items such as leaves and twigs. When these geometrically rich feature maps are subsequently fed into SHSA, the self-attention mechanism evaluates the features semantically across a global receptive field; through spatial weight redistribution, the network identifies and preserves geometric features belonging to the strawberry body while suppressing edge responses from background clutter. The experimental data fully substantiate this “extract-then-filter” cooperative mechanism: this combination raised precision to 89.7% and mAP@50-95 to its highest value of 83.8%, demonstrating that the synergy of C3k2\_DCNv4 and SHSA effectively enhances the robustness of strawberry segmentation in complex occlusion scenarios.

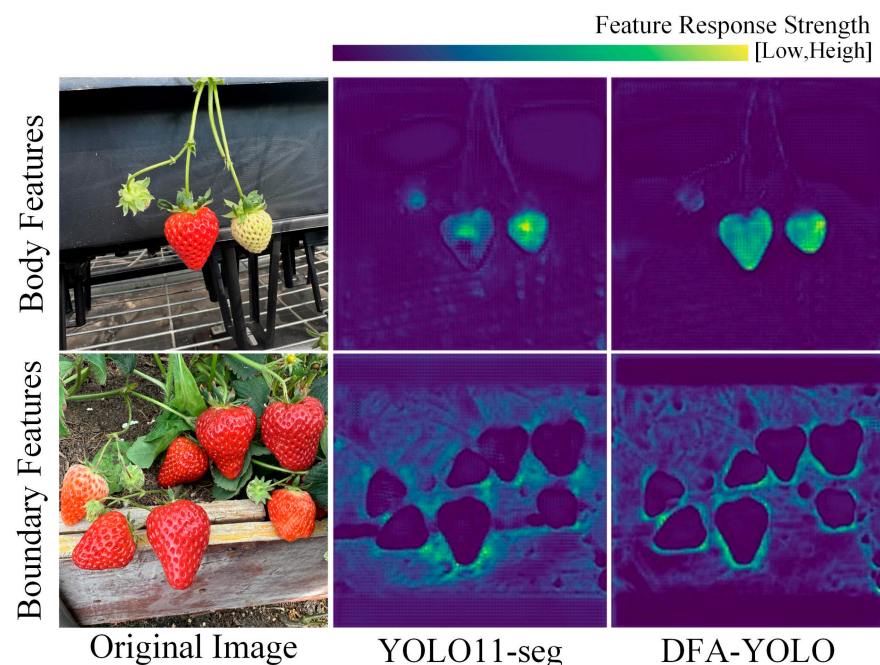
Although the combination of C3k2\_DCNv4 and SHSA achieved the highest mAP@50-95 (83.8%), the complete architecture integrating all three modules is ultimately selected as the optimal configuration. This decision is primarily driven by the significant advantages in recall and primary segmentation accuracy brought by the BiFPN module. Compared to the dual-module combination, the full model yields a 3.5 percentage point increase in recall and



a 0.9 percentage point increase in mAP@50. For the strawberry volume estimation task, it is crucial to minimize missed detections of multi-scale fruits spanning the unripe, half-ripe, and ripe stages. The negligible 0.1% reduction in the strict boundary refinement metric is an acceptable trade-off for a substantial enhancement in overall detection robustness. Compared to the baseline, the proposed full model attained a precision of 88.7%, a recall of 89.8%, and a mAP@50 of 94.4%, representing improvements of 1.1, 5.4, and 3.9 percentage points, respectively. Regarding model complexity, the improved model has only 10.57 M parameters, a mere 4.8% increase over the baseline, and GFLOPs are controlled at 38.2 G. This confirms that the improved model retains strong lightweight characteristics while achieving a significant boost in segmentation accuracy, thus possessing practical feasibility for deployment on edge devices.

To further evaluate whether the improvement in segmentation performance was statistically meaningful, instance-level IoU values on the test set were extracted from the proposed DFA-YOLO model and the baseline YOLO11s model, and non-parametric statistical analyses were conducted. The Shapiro–Wilk tests indicated that the IoU distributions of both models significantly deviated from normality ( $p < 0.05$ ); therefore, the Wilcoxon signed-rank test was used for paired comparison. The results showed that DFA-YOLO achieved a higher median IoU than YOLO11s (0.938 vs. 0.932), with a smaller interquartile range (0.088 vs. 0.125). The Wilcoxon signed-rank test further indicated a statistically significant difference between the two models ( $p = 0.021$ , effect size  $r = 0.17$ ). These results suggest that DFA-YOLO provides a statistically supported improvement in segmentation accuracy and stability over YOLO11s.

To provide a visual interpretation of the performance advantages of DFA-YOLO, we visualized and compared the prototype masks (Proto), which represent object and edge features, between the baseline and the improved model. The results are shown in Figure 12.



**Figure 12.** Feature prototype masks of the baseline (YOLO11-seg) and the proposed DFA-YOLO.

It can be visually observed that in the object-feature activation comparison, the baseline YOLO11-seg exhibits a clear defocusing phenomenon in the highlight activation area, failing not only to cover the fruit body completely but also displaying abundant high-frequency noise activations in the background. In contrast, the proposed DFA-YOLO model achieves extremely high feature response over the fruit body, fully covering the fruit area, with a



clean background devoid of obvious noise. In terms of edge features, YOLO11-seg's edge activation band is relatively broad and blurred, making it difficult to precisely locate the irregular geometric edges of the strawberry; DFA-YOLO generates an extremely sharp edge feature response map with activations highly converged onto the fruit edge, tightly delineating the real fruit boundary. In summary, compared with the baseline, DFA-YOLO not only effectively suppresses background false activations to achieve complete object-feature focusing, but also precisely and sharply delineates the irregular physical boundary of the fruit, significantly enhancing the model's geometric representation of target topology.

### 3.1.2. Effectiveness Analysis of Data Augmentation

To quantitatively evaluate the actual effectiveness of the data augmentation strategy, this study trained the DFA-YOLO model using the original dataset and the augmented dataset, respectively. The comparative experimental results are shown in Table 4.

**Table 4.** Comparison of model performance before and after data augmentation.

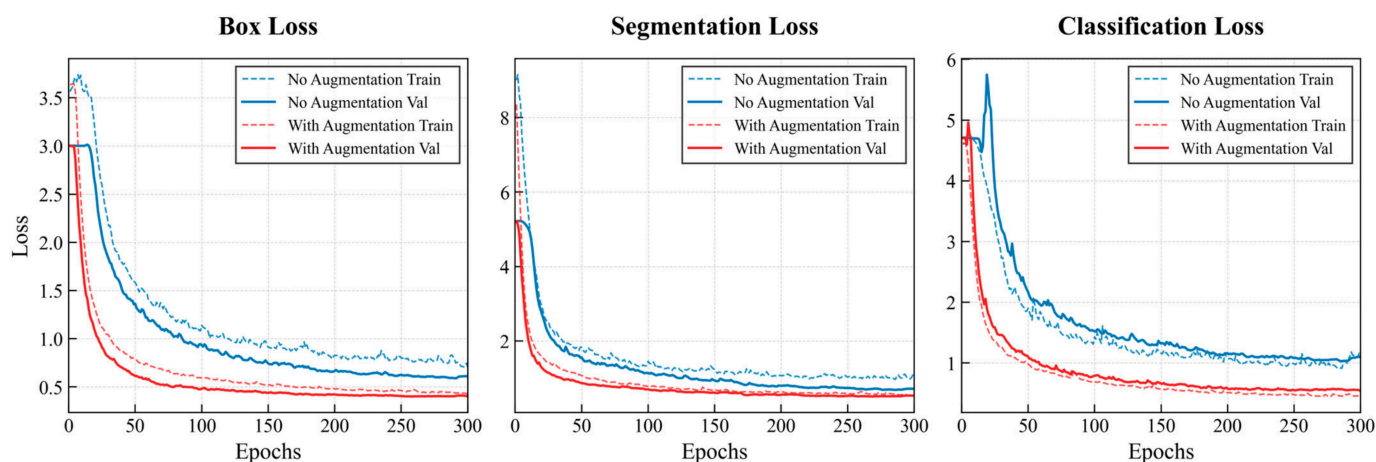
| Data Augmentation | P (%) | R (%) | mAP@50 (%) | mAP@50-95 (%) |
|-------------------|-------|-------|------------|---------------|
| No                | 86.8  | 83.1  | 90.8       | 77.3          |
| Yes               | 88.7  | 89.8  | 94.4       | 83.7          |

The experimental results indicate that the data augmentation strategy yielded a significant positive gain in model performance. Compared to the model trained on the original dataset, the DFA-YOLO trained on the augmented dataset achieved a substantial leap of 6.7 percentage points in recall (R), while mAP@50 and mAP@50-95 also increased by 3.6 and 6.4 percentage points, respectively. Specifically, the dramatic jump in recall demonstrates that augmentation techniques such as brightness adjustment, random occlusion, and geometric transformations effectively simulated the complex real-world agricultural environment, greatly reducing the missed detection of strawberries. Notably, the improvement margin of mAP@50-95 is considerably larger than that of mAP@50. This phenomenon provides compelling evidence that data augmentation not only improved basic object localization capability but also significantly enhanced the boundary-fitting precision of the segmentation masks at higher IoU thresholds. Furthermore, under the premise of a massive increase in recall, precision (P) did not decline but instead simultaneously increased by 1.9 percentage points. This indicates that while effectively reducing the missed detection rate, the model successfully suppressed the increase in background false positives, breaking the conventional precision–recall trade-off limitation and achieving a comprehensive optimization of the overall segmentation quality.

To intuitively demonstrate the beneficial effects of data augmentation, Figure 13 compares the loss trajectories of the models with and without data augmentation over 300 epochs.

Regarding convergence behavior, the model trained on the augmented dataset exhibits faster convergence and higher stability. During the initial training phase, the validation loss of the augmented model decreases more rapidly, indicating that the introduced sample perturbations assist the network in extracting global features more efficiently. As training progresses to the later stages, the loss curves of the augmented model show minimal fluctuations, with no divergence or oscillation observed. Furthermore, while the baseline model displays a noticeable generalization gap between its training and validation losses, the data augmentation strategy significantly narrows this gap. It aligns the validation curve more closely with the training curve, converging to a lower final validation loss. These quantitative results demonstrate that data augmentation accelerates the convergence

process, mitigates overfitting, and enhances the model's generalization and segmentation performance on greenhouse environment images.



**Figure 13.** Comparison of training and validation loss curves before and after data augmentation.

Furthermore, the data augmentation strategy employed in this study serves as a crucial domain generalization technique, effectively bridging the domain gap between different hardware devices. The DFA-YOLO model was trained using high-resolution images from the Nikon Z30, whereas the final deployment and volume estimation relied on the RealSense D435i sensor. Compared to the images captured by the Nikon Z30, the RGB images acquired by the RealSense D435i sensor have lower resolution and exhibit inherent differences in color profiles and noise patterns, resulting in a lower overall image quality. To overcome the discrepancy in image quality between the two sensors, operations such as Gaussian noise, motion blur, and HSV color jittering were applied to intentionally degrade the training images. This effectively simulated the spectral response and noise characteristics of the RGB-D sensor, enabling the model to acquire strong cross-domain adaptability during the training phase and providing a robust foundation for the subsequent strawberry volume estimation.

### 3.1.3. Comparative Experiments

To comprehensively verify the superiority of DFA-YOLO for real-time strawberry volume estimation, eight other mainstream instance segmentation models were compared on our self-built dataset. The experimental results are presented in Table 5.

**Table 5.** Results of Comparative Experiment.

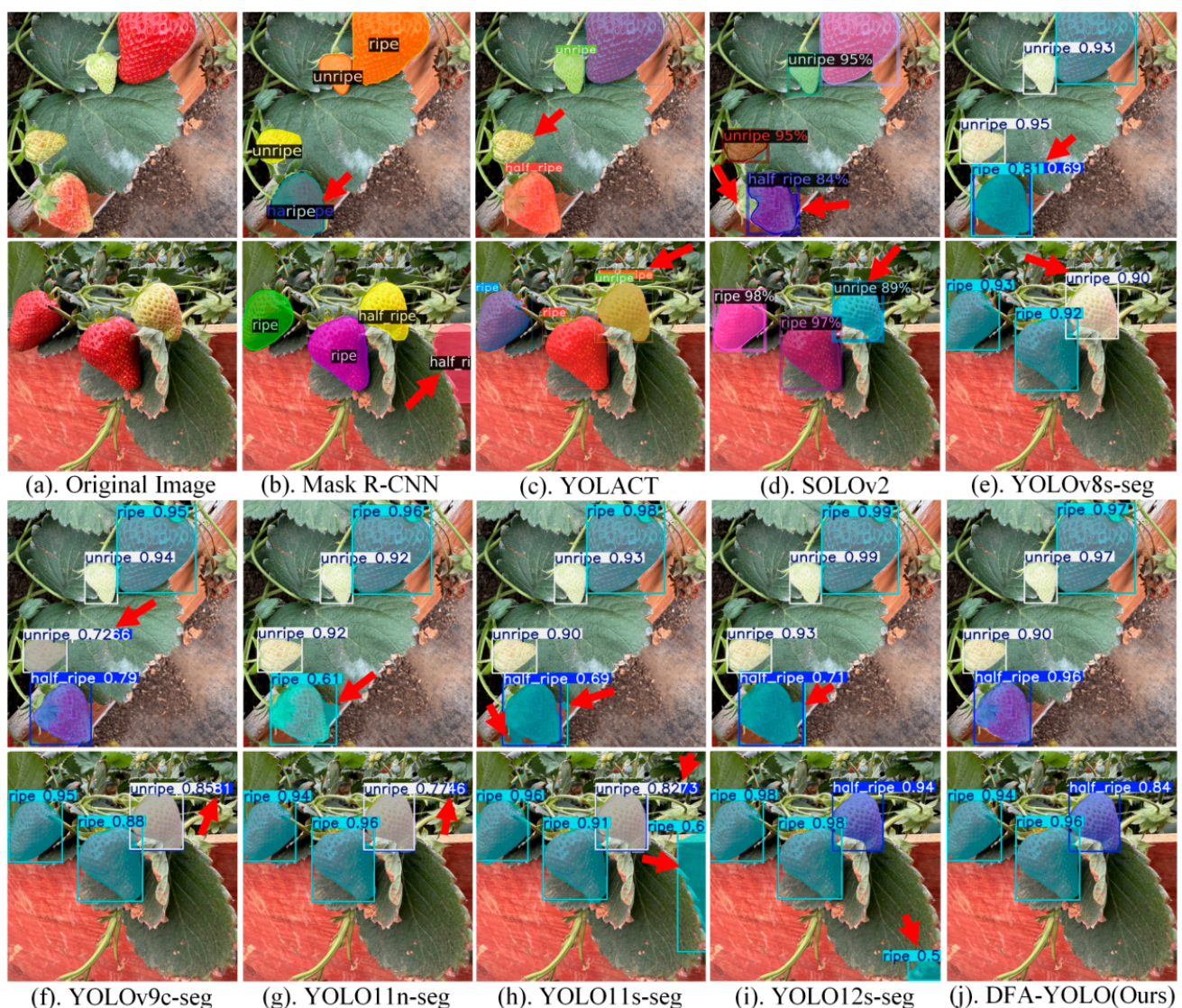
| Model       | mAP@50 (%) | mAP@50-95 (%) | Params (M) | FPS |
|-------------|------------|---------------|------------|-----|
| Mask R-CNN  | 86.6       | 75.9          | 43.9       | 49  |
| YOLACT      | 83.6       | 71.2          | 31.4       | 73  |
| SOLOv2      | 81.3       | 65.5          | 46.2       | 54  |
| YOLOv8s-seg | 91.8       | 82.5          | 11.8       | 132 |
| YOLOv9c-seg | 92.4       | 84.5          | 27.8       | 105 |
| YOLO11n-seg | 86.2       | 75.8          | 2.8        | 182 |
| YOLO11s-seg | 90.5       | 80.5          | 10.1       | 154 |
| YOLO12s-seg | 89.4       | 79.8          | 9.9        | 137 |
| DFA-YOLO    | 94.4       | 83.7          | 10.6       | 143 |

The comparative data in Table 5 demonstrate that the proposed model achieves an excellent balance between segmentation accuracy and lightweight design. Compared with



the new-generation, comparably sized YOLO12s-seg, mAP@50 and mAP@50-95 increased by 5 and 3.9 percentage points, respectively, while FPS was slightly higher. Regarding the discrepancy at stricter thresholds, although the larger-scale YOLOv9c-seg achieves the highest mAP@50-95 (84.5%), this is an expected trade-off. YOLOv9c-seg possesses a parameter count 2.6 times larger, granting it a higher network capacity to refine extreme pixel-level boundaries. However, this marginal boundary gain causes its inference frame rate to drop to 105 FPS. Against the two-stage Mask R-CNN and the classical single-stage segmentation models YOLACT [44] and SOLOv2 [45], DFA-YOLO achieved comprehensive superiority in segmentation accuracy as well as a substantial lead in inference speed. Overall, our model, with a modest parameter count of 10.6 M, attains the highest mAP@50 (94.4%) among all lightweight models and a high real-time speed of 143 FPS, satisfying the demands of agricultural edge computing devices for high-precision, real-time strawberry volume estimation.

To intuitively illustrate the performance advantages and robustness of DFA-YOLO, visual comparisons of the aforementioned models on our self-built dataset are presented in Figure 14, with erroneous regions indicated by red arrows.



**Figure 14.** Comparison of segmentation results across different models. Red arrows indicate regions with segmentation errors, and overlapping labels indicate typical segmentation failure cases.

In the first scenario (top row), Mask R-CNN, YOLOv8s-seg, YOLOv9c-seg, YOLO11s-seg, and YOLO12s-seg all exhibited semantic confusion and duplicate segmentation, incorrectly detecting the half-ripe strawberry in the bottom-left corner as multiple classes. Additionally, YOLACT failed to detect an unripe strawberry on the left, while SOLOv2 and YOLO11n-seg suffered from background confusion. In the second scenario (bottom row), YOLACT, SOLOv2, YOLOv8s-seg, YOLOv9c-seg, YOLO11n-seg, and YOLO11s-seg misclassified the half-ripe strawberry in the top-right corner as the unripe category. Furthermore, Mask R-CNN and YOLO12s-seg generated false-positive errors by misidentifying the red wooden frame in the background as a fruit. In contrast, DFA-YOLO not only correctly identifies all target classes but also produces complete masks that tightly hug the physical boundaries of the fruit. This indicates that DFA-YOLO, by virtue of its superior multi-scale feature fusion and deformable spatial perception capabilities, successfully overcomes semantic ambiguity challenges in complex scenes, demonstrating significantly better robustness and fine-grained geometric representation than current mainstream segmentation networks in unstructured agricultural environments. This experiment provides direct evidence of the performance advantage of the proposed model, delivering high-precision strawberry masks for subsequent volume estimation.

### 3.2. Experiments on Principal-Axis-Slicing-Based Volume Estimation

#### 3.2.1. Volume Estimation Algorithm Comparison

To objectively evaluate the comprehensive performance of the proposed principal-axis slicing volume estimation algorithm, two baseline comparison methods were selected: the traditional ellipsoid fitting method [7] and a 3D convex-hull reconstruction algorithm [46]. For the ellipsoid fitting baseline, the algorithm calculates the optimal fitting ellipse based on the 2D mask's contour moments to extract the major and minor axes, estimating the volume under an idealized prolate spheroid assumption. Typically, 3D convex-hull reconstruction requires point clouds from multiple viewpoints for accurate 3D reconstruction. To ensure a fair comparison, we made adaptive modifications for single-view applicability. Because self-occlusion from a single view restricts the RGB-D sensor to capturing only the camera-facing surface of the strawberry point cloud, the back-side point cloud is missing; therefore, the acquired single-surface point cloud was duplicated and mirror-flipped to complete the missing rear side, then down-sampled to reduce density while preserving overall geometry and lowering computational overhead. The 3D convex-hull algorithm then computed the smallest convex polyhedron enclosing all discrete points to form the geometric shell of the strawberry, and the volume of the strawberry was estimated by summing the volumes of tetrahedral meshes inside the closed convex polyhedron. To guarantee fairness, all methods acquired data at the same shooting distance and angle under the same lighting conditions. Fifty strawberry fruits of diverse size and shape were selected, and each sample was measured five times for statistical evaluation. The comparative results are shown in Table 6.

**Table 6.** Accuracy comparison of different volume estimation methods.

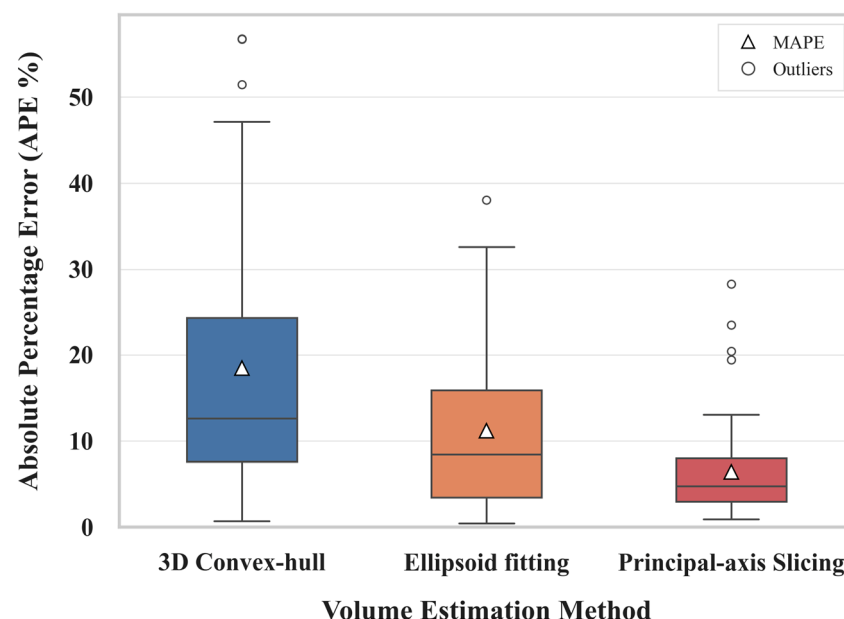
| Method                        | RMSE (cm <sup>3</sup> ) | MAPE (%) |
|-------------------------------|-------------------------|----------|
| Single-view 3D convex-hull    | 4.7                     | 18.4     |
| Ellipsoid fitting             | 3.2                     | 11.2     |
| Principal-axis slicing (Ours) | 2.4                     | 6.3      |

The experimental results demonstrate that the proposed principal-axis slicing method significantly outperforms both baseline methods in volume estimation accuracy. Specifically, thanks to the accurate representation of localized geometric fluctuations by the slicing



strategy, our method achieved the lowest MAPE (6.3%) and RMSE (2.4 cm<sup>3</sup>). Compared with the traditional ellipsoid fitting and the single-view 3D convex-hull reconstruction, the MAPE of our method decreased by 4.9 and 12.1 percentage points, respectively. Correspondingly, the RMSE dropped markedly from 3.2 cm<sup>3</sup> and 4.7 cm<sup>3</sup> to 2.4 cm<sup>3</sup>, achieving a substantial leap in overall estimation accuracy and robustness.

Beyond the conventional point metrics, non-parametric inferential statistics were introduced to further corroborate the superiority of the principal-axis slicing algorithm. Initially, a Shapiro–Wilk test indicated that the absolute percentage errors (APE) were not normally distributed ( $p < 0.05$ ). Consequently, the Friedman test was employed for a global evaluation, revealing a highly significant statistical difference in estimation errors among the three methods ( $\chi^2(2) = 15.06$ ,  $p < 0.001$ ). Subsequently, a post hoc analysis utilizing the Wilcoxon signed-rank test with Bonferroni correction was conducted. The results demonstrated that the APE of the principal-axis slicing method (Median = 4.78%, interquartile range (IQR) = 5.02%) was significantly lower than that of both the ellipsoid fitting (Median = 8.44%, IQR = 12.40%,  $p < 0.05$ , effect size  $r = 0.25$ ) and the single-view 3D convex-hull reconstruction (Median = 12.61%, IQR = 16.71%,  $p < 0.001$ , effect size  $r = 0.39$ ). Specifically, compared to the baseline methods, the notably smaller IQR (5.02%) of the slicing algorithm robustly proves that the geometric fidelity attributed to the proposed method yields not only higher overall estimation accuracy but also exceptional measurement stability and robustness across varying fruit sizes and morphologies. As visually demonstrated in the boxplot (Figure 15), the error distribution further corroborates these statistical findings.

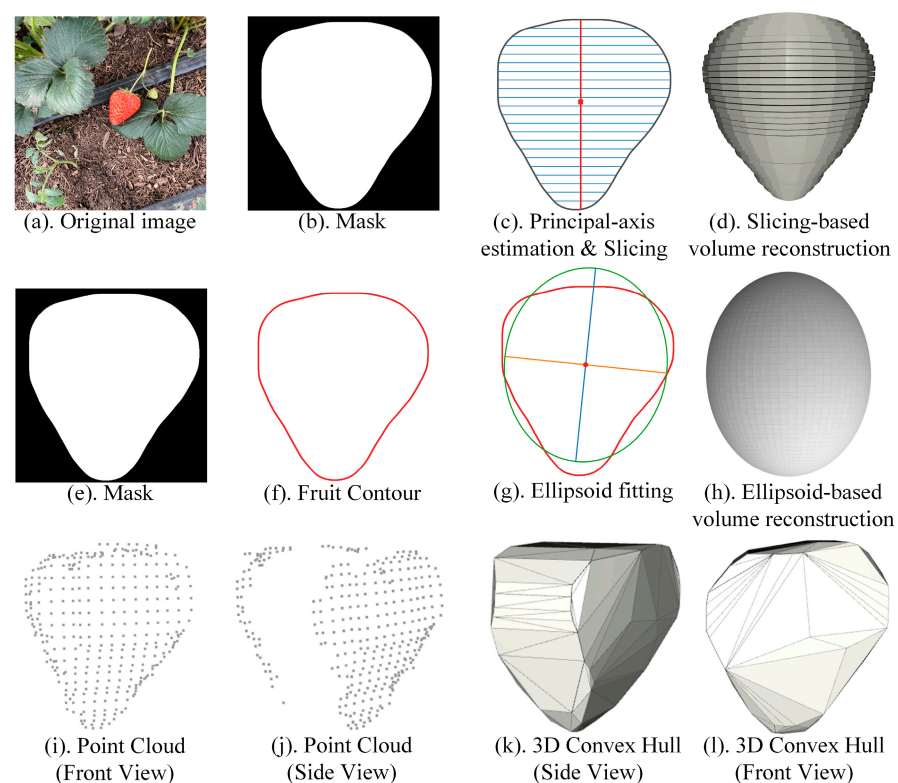


**Figure 15.** Boxplots of the absolute percentage error (APE) for the three volume estimation methods.

Furthermore, Figure 16 visually depicts the topological differences and morphological reconstruction quality among the proposed principal-axis slicing method, the traditional ellipsoid fitting, and the single-view convex-hull method. As shown in Figure 16b–d, the proposed method builds the morphology reconstruction directly on the high-quality 2D instance segmentation mask. By extracting the symmetric principal axis of the mask and applying micro-slicing (Figure 16c), the algorithm maps the discrete 2D pixel widths into 3D micro-cylinders under the prior assumption of rotational symmetry, perfectly preserving the geometric characteristics of the strawberry fruit and achieving an accurate recovery of its true 3D contour (Figure 16d). Furthermore, Figure 16e–h illustrate the geometric limita-



tions of the traditional ellipsoid fitting baseline. Although the mask provides high-precision authentic boundaries (Figure 16f), the algorithm simplifies the rich edge information into merely two parameters—the major and minor axes—discarding local geometric fluctuations. By imposing a global orthogonal symmetry constraint (Figure 16g), the model struggles to adapt to the naturally irregular morphology of the strawberry. Influenced by the fruit’s asymmetric “broad-shoulder, narrow-apex” mass distribution and the algorithm’s least-squares optimization characteristics, the best-fit ellipse produces geometric deviations: it expands outward at the shoulders, incorporating non-fruit regions, and shrinks inward at the apex, omitting portions of the true fruit volume (Figure 16g). Ultimately, as a mathematical approximation, the ellipsoid-based 3D reconstruction (Figure 16h) fails to accurately reflect the true surface curvature and shape of the fruit, thereby limiting the full utilization of the shape information provided by high-quality instance segmentation. In contrast, Figure 16i–l expose the systemic bottleneck of directly using 3D point clouds for volume computation: due to the single-view constraint, the raw scanned point cloud appears complete in the front view (Figure 16i) yet reveals total absence of depth information on the back side in the side view (Figure 16j). After mirror-flipping to fill the missing back side, the convex-hull algorithm forcibly creates rough spatial bridges between the two point sets (Figure 16k). Additionally, the limited physical sampling precision of the RGB-D sensor produces substantial outliers at the fruit edges (Figure 16j). These factors ultimately lead to a reconstructed strawberry model from the 3D convex-hull method that cannot accurately reproduce the fruit’s geometric features, resulting in large volume estimation errors. In summary, compared to the rigid geometric assumptions of traditional ellipsoid fitting and the single-view physical occlusion inherent in 3D point-cloud reconstruction, the proposed principal-axis slicing algorithm effectively leverages the shape details of high-precision 2D masks while circumventing the problem of missing back-side data via an axial-symmetry prior, thereby achieving higher morphological fidelity and estimation stability.



**Figure 16.** Comparison of 3D geometry reconstruction methods: (a) original image; (b–d) proposed principal-axis slicing method; (e–h) ellipsoid fitting method; (i–l) 3D convex-hull reconstruction method.

### 3.2.2. Effect of Shooting Distance on Volume Estimation Accuracy

To investigate the applicability of the proposed strawberry volume estimation method at different shooting distances, 10 standard-shaped and 10 misshapen strawberries of large, medium, and small sizes were selected for multiple volume estimation trials at distances ranging from 30 to 70 cm with a 10 cm interval. The results are presented in Table 7.

**Table 7.** Strawberry volume estimation accuracy at different shooting distances.

| Shooting Distance<br>(cm) | MAPE (%) |        |       |           |      | RMSE<br>(cm <sup>3</sup> ) |
|---------------------------|----------|--------|-------|-----------|------|----------------------------|
|                           | Small    | Medium | Large | Misshapen | ALL  |                            |
| 30                        | 9.8      | 8.4    | 7.6   | 12.7      | 9.6  | 2.6                        |
| 40                        | 5.1      | 4.5    | 3.6   | 12.5      | 6.4  | 2.2                        |
| 50                        | 11.3     | 9.2    | 7.9   | 17.2      | 11.4 | 3.6                        |
| 60                        | 18.6     | 11.3   | 9.8   | 21.5      | 15.3 | 4.1                        |
| 70                        | 21.3     | 13.1   | 10.9  | 23.8      | 17.3 | 4.9                        |

The results indicate that 40 cm is the optimal working distance for the proposed method. At this distance, for regularly shaped strawberry fruit, MAPE remains consistently around 5%, the overall MAPE reaches its minimum of 6.4%, and RMSE stays within a reasonable range. At 30 cm, the camera is near the minimum depth limit of the RealSense D435i, where the depth map is prone to holes, causing missing depth data and increased error. As the shooting distance increases (50–70 cm), MAPE rises from 11.4% to 17.3%, mainly because of the depth sensor's precision decay and the edge-expansion effect of instance segmentation: the model-output strawberry mask inevitably expands relative to the true fruit boundary, mixing background depth noise and leading to overestimated volume. For smaller fruit, the estimation error rises rapidly with distance because small fruit occupy fewer pixels at longer distances; insufficient resolution prevents the algorithm from accurately removing edge details, causing more severe mask expansion and geometric distortion. Meanwhile, the relatively larger errors observed for misshapen fruit objectively indicate the inherent limitation of the proposed method when dealing with highly irregular morphology.

### 3.2.3. Effect of Depth Feature Sampling Strategies on Volume Estimation Performance

Accurately mapping the 2D pixel scale to the 3D physical scale is a prerequisite for the precise estimation of strawberry volume using the principal-axis slicing method. Given the computational constraints of edge hardware and the inherent noise of depth sensors, exploring a depth sampling strategy that balances accuracy and efficiency is particularly crucial. To this end, a comparative experiment was conducted using 50 strawberry samples of varying sizes and shapes at a shooting distance of 40 cm. Three distinct depth sampling methods were systematically evaluated: local pixel-wise depth (extracting the independent depth of each pixel within the mask for calculation), micro-slice average depth (extracting the average depth of the local region corresponding to each micro-slice), and global median depth (extracting the median depth of the entire mask area to represent the main body depth of the fruit). The comparative results of these sampling strategies regarding estimation error and computational overhead are presented in Table 8.

The experimental results in Table 8 indicate that among the three depth sampling strategies, the global median depth achieves the highest and most robust estimation accuracy, with a MAPE of 6.3% and an RMSE of 2.4 cm<sup>3</sup>. Conversely, the local pixel-wise depth and micro-slice average depth yield relatively larger errors. The fundamental reason why an increase in sampling density leads to accuracy degradation lies in the hardware limitations of consumer-grade RGB-D sensors (such as the RealSense D435i utilized in this

study). When capturing small targets with reflective curved surfaces like strawberries, the sensor is highly susceptible to generating flying pixels and abrupt depth variations at the fruit boundaries. The local pixel-wise and micro-slice strategies inadvertently incorporate these edge noises into the calculation process, resulting in significant estimation deviations. In contrast, the global median depth leverages the statistical median to act as a robust statistical filter, effectively eliminating edge outliers caused by the precision limits of the RGB-D sensor. Furthermore, since the natural depth variation in the strawberry surface is extremely small relative to the shooting distance of approximately 40 cm, adopting a global depth approximation does not introduce significant scale projection errors. Regarding computational overhead, the global median depth maintains a low computational complexity of only 9 ms, as it utilizes a single depth median to represent the depth features of the fruit region. In comparison, the micro-slice average depth takes 13 ms due to the necessity of local matrix traversals. The local pixel-wise depth incurs the highest computational complexity—taking 20 ms—as it involves coordinate mapping from 2D pixels to a 3D point cloud and dense distance computations.

**Table 8.** Effect of depth sampling strategies on strawberry volume estimation performance.

| Sampling Strategy         | Estimation Algorithm Time (ms) | MAPE (%) | RMSE (cm <sup>3</sup> ) |
|---------------------------|--------------------------------|----------|-------------------------|
| Local pixel-wise depth    | 20                             | 12.6     | 4.2                     |
| Micro-slice average depth | 13                             | 11.2     | 3.4                     |
| Global median depth       | 9                              | 6.3      | 2.4                     |

In conclusion, under agricultural edge deployment scenarios characterized by complex lighting conditions and limited computational resources, the global median depth sampling strategy effectively balances estimation accuracy and computational efficiency, verifying its rationality and applicability within the proposed algorithm.

#### 3.2.4. Effect of Slicing Step Size on Volume Estimation Performance

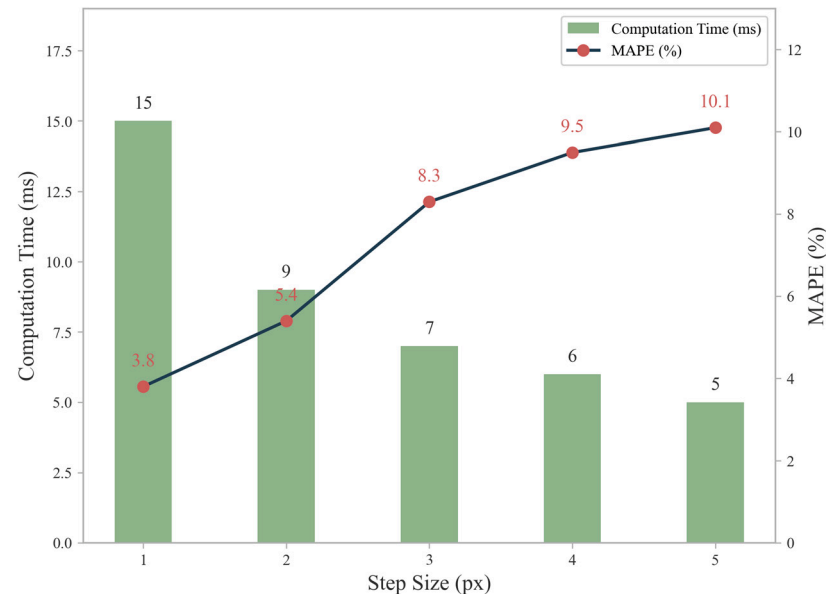
In the proposed principal-axis slicing volume estimation algorithm, the slicing step size  $\Delta_{step}$  is a core parameter that determines the trade-off between volume estimation accuracy and computational overhead. To investigate the impact of different step sizes on overall system performance and to identify the optimal balance between accuracy and real-time capability, performance experiments were conducted with varying step sizes. Each experimental sample was measured 10 times and the average values are reported in Table 9.

**Table 9.** Impact of slicing step size on volume estimation performance.

| Step Size (px) | Number of Slices | Estimation Algorithm Time (ms) | MAPE (%) |
|----------------|------------------|--------------------------------|----------|
| 1              | 64               | 15                             | 3.8      |
| 2              | 32               | 9                              | 5.4      |
| 3              | 22               | 7                              | 8.3      |
| 4              | 16               | 6                              | 9.5      |
| 5              | 13               | 5                              | 10.1     |

The experimental results demonstrate that slicing step size has a significant bi-directional impact on computational efficiency and estimation accuracy. As the step size increases from 1 px to 5 px, the number of slices decreases sharply, and the single-fruit estimation time drops from 15 ms to 5 ms, effectively releasing the computational resources

of the edge device. However, a larger step size also substantially weakens the algorithm's ability to fit irregular strawberry edges, amplifying staircase truncation errors, and MAPE rises from 3.8% to 10.1%. To intuitively evaluate the overall system performance under different step sizes and determine the optimal parameter, the data were visualized, as shown in Figure 17.



**Figure 17.** Trends in computation time and MAPE under different slicing step sizes.

From Figure 17, it can be clearly observed that when the step size increases from 1 px to 2 px, the algorithm time is reduced by 40%, while MAPE only slightly increases to 5.4%, still remaining within a very high accuracy range. However, when the step size continues to increase to 3 px and above, the marginal benefit in time reduction diminishes substantially, yet the error curve shows a steep upward trend. Considering the requirements of both constrained computational resources in agricultural scenarios and high-precision strawberry volume estimation, a step size of 2 px represents the optimal balance point; therefore, this step size was adopted as the standard operating parameter in the subsequent experiments.

### 3.2.5. Edge Deployment Performance Overhead

For a strawberry volume estimation method, estimation speed is an extremely important indicator directly determining its practical feasibility. Hence, the proposed real-time strawberry volume estimation system was deployed on an NVIDIA Jetson Orin Nano Super to conduct performance overhead experiments, measuring the time consumed by each pipeline component. For different sample sizes, the time spent in each stage was collected 10 times and averaged, and the FPS value was taken as the average of the real-time frame rate printed by the system during sampling. The results are given in Table 10.

**Table 10.** Edge deployment time cost analysis.

| Number of Fruits | Step 0 * (ms) | Step 1 * (ms) | Step 2 * (ms) | FPS |
|------------------|---------------|---------------|---------------|-----|
| 0                |               | 18            | -             | 21  |
| 1                |               | 26            | 8             | 15  |
| 2                | 21            | 30            | 18            | 13  |
| 5                |               | 48            | 45            | 8   |
| 10               |               | 68            | 85            | 5   |

\* Step 0: Data pre-processing and depth-to-RGB alignment time; Step 1: Instance segmentation inference time of DFA-YOLO; Step 2: Volume estimation time via principal-axis slicing.



The results show that the proposed method imposes a very small computational burden. Even when deployed on an edge device with limited computational power, for a single strawberry, the times for data pre-processing, strawberry segmentation, and volume estimation are only 21 ms, 26 ms, and 8 ms, respectively, and the overall system frame rate reaches 15 FPS—fully verifying that the method offers excellent execution speed on edge devices and can satisfy real-time operational demands in agricultural scenarios. As the number of targets in the scene increases, the overall system latency rises accordingly. Specifically, when many targets are present in the field of view, the overhead of instance segmentation post-processing and volume integration escalates sharply, causing the frame rate to drop from 15 FPS for a single target to 5 FPS in dense scenes. This objectively reflects the computational bottleneck and limitation of the system when handling high-density scenarios. However, in practice, due to strawberry growth characteristics and the optimal working distance constraint, the number of strawberry fruit in the camera's field of view is typically 1–5. Under these typical 1–5 fruit conditions, the proposed method meets the real-time requirements of agricultural harvesting robots for visual feedback and volume measurement, confirming good prospects for practical deployment. To more intuitively demonstrate the performance advantage of the proposed algorithm, the baseline method (Single-view 3D convex-hull and Ellipsoid fitting) was also deployed on the same edge device for a performance overhead comparison, as shown in Table 11.

**Table 11.** Performance overhead comparison of different volume estimation methods.

| Method                        | Step 0 * (ms) | Step 1 * (ms) | Step 2 * (ms) | Total Time |
|-------------------------------|---------------|---------------|---------------|------------|
| Single-view                   | 31            |               | 14            | 71         |
| 3D convex-hull                |               | 26            |               |            |
| Ellipsoid fitting             | 21            |               | 5             | 52         |
| Principal-axis slicing (Ours) | 21            |               | 8             | 55         |

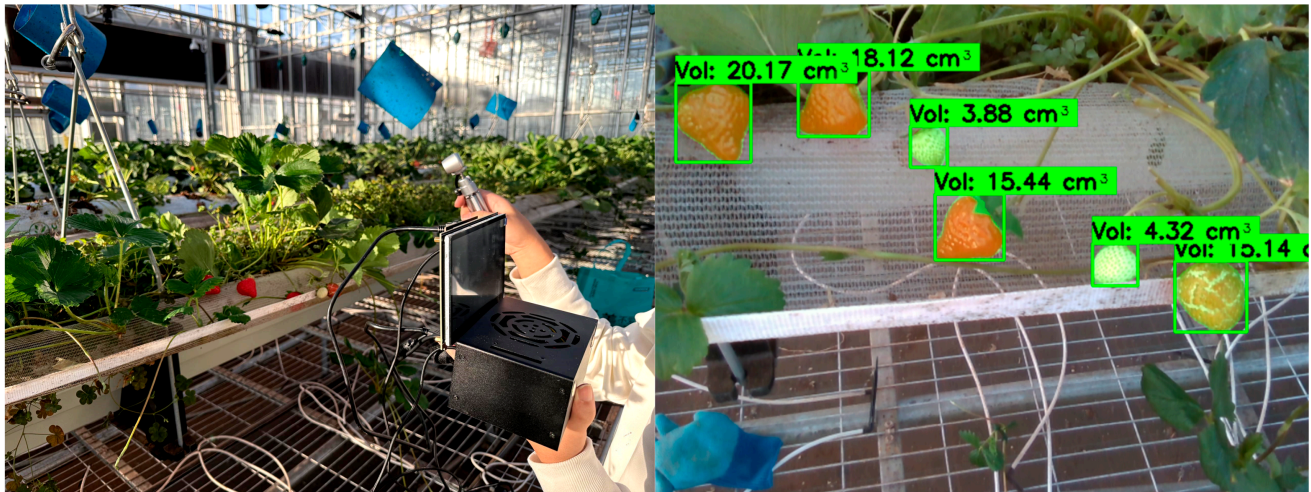
\* Step 0: Data pre-processing and depth-to-RGB alignment time; Step 1: Instance segmentation inference time of DFA-YOLO; Step 2: Volume estimation algorithm time.

The results in Table 11 objectively illustrate the performance overhead of the different volume estimation methods. During pre-processing, the single-view convex-hull method processes a large number of raw depth points, increasing the computational load of operations such as filtering and down-sampling (31 ms). In contrast, both the traditional ellipsoid fitting and the proposed principal-axis slicing method operate directly on 2D masks, bypassing 3D point-cloud generation and reducing pre-processing time to 21 ms. For the core volume estimation stage, the 3D convex-hull reconstruction relies on spatial geometric topology operations of higher complexity, consuming 14 ms. The ellipsoid fitting method requires the least time at this stage (5 ms) because it reduces the volume calculation to a basic analytical formula based on limited parameters; however, as previously analyzed, this computational efficiency comes at the cost of a significant loss in morphological fitting accuracy. The proposed method transforms the 3D reconstruction problem into the discrete accumulation of 2D micro-slice volumes, effectively constraining the computational overhead to 8 ms. Overall, the total processing time per fruit for our method is 55 ms, representing an approximate 22.5% reduction compared to the 3D convex-hull baseline (71 ms). This demonstrates that the proposed algorithm achieves a favorable balance between computational efficiency and volume estimation accuracy, showing potential for lightweight and real-time applications.

### 3.2.6. Greenhouse Validation of Volume Estimation Accuracy

To further validate the reliability and robustness of the proposed volume estimation system under real agricultural conditions, 218 strawberry samples spanning different

ripeness stages and sizes were randomly selected from the smart greenhouse of Yunnan Agricultural University. Using the edge-computing system built in the previous section and maintaining the optimal working distance (40 cm), volume estimation was performed on these samples. The sampling process is shown in Figure 18. The absolute volume measured by water displacement was used as the ground truth, and the volume output by the edge system was taken as the estimated value. The estimation error statistics for different maturity categories are shown in Table 12.



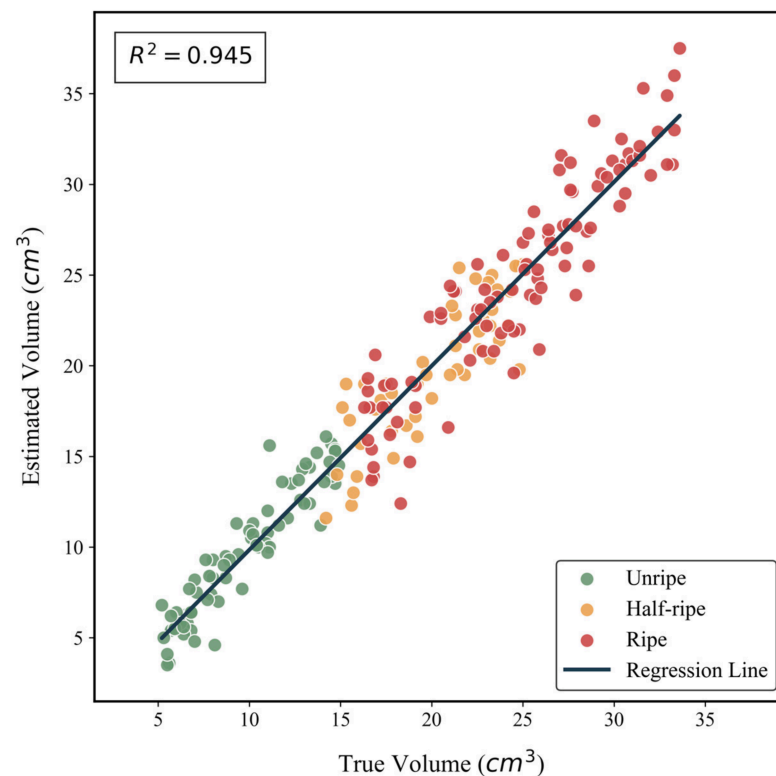
**Figure 18.** On-site volume estimation. Green boxes indicate detected strawberries, and the corresponding estimated volumes are shown in the image. From left to right, the estimated volumes are 20.17, 18.12, 3.88, 15.44, 4.32, and 15.14 cm<sup>3</sup>.

**Table 12.** Estimation error statistics by maturity stage.

| Maturity  | Sample Size | MAE (cm <sup>3</sup> ) | MAPE (%) |
|-----------|-------------|------------------------|----------|
| Unripe    | 71          | 1.0                    | 11.2     |
| Half-ripe | 48          | 1.6                    | 8.5      |
| Ripe      | 99          | 1.8                    | 7.6      |
| ALL       | 218         | 1.5                    | 9.0      |

The data indicate that the system's volume estimation accuracy improves markedly with fruit maturity: MAPE decreases from 11.2% for unripe fruit to 7.6% for ripe fruit. This is primarily because unripe fruit are small and exhibit a greenish or pale-white color that tends to blend with the background, amplifying the relative computational error of edge segmentation; ripe fruit, with their larger volume basis and strong color contrast, effectively dilute pixel-level jitter and enhance mask sharpness. Meanwhile, the MAE shows a reasonable upward trend from 1.0 cm<sup>3</sup> to 1.8 cm<sup>3</sup>, consistent with the physical law of machine vision: ripe fruit, despite higher relative accuracy, accumulate slightly larger absolute pixel deviations due to their greater physical volume. Overall, the system achieves an overall MAPE of 9.0% and exhibits particularly high robustness within the core operational range of harvesting robots (half-ripe to fully ripe stages), fully meeting the demand for non-destructive fruit phenotyping in complex greenhouse environments.

Having clarified the gradient change in error with maturity, a linear regression model was introduced to comprehensively quantify the overall consistency of the system in the "2D-mask-to-3D-physical-volume" mapping process. The physical true volume was taken as the independent variable and the system-estimated volume as the dependent variable. The regression analysis results and the spatial distribution of samples across maturity stages are shown in Figure 19.



**Figure 19.** Linear regression analysis between estimated and true volume.

Figure 19 intuitively reveals that the 218 test samples of different maturities are tightly converged on both sides of the fitted regression line. The global coefficient of determination  $R^2$  reaches 0.945. This high degree of fit convincingly demonstrates that the system possesses not only excellent performance for specific sizes, but also extremely high and stable linear predictability across the entire strawberry growth cycle, from small unripe fruit to large ripe fruit.

#### 4. Discussion

Traditional computer vision-based fruit volume measurement schemes highly rely on multi-view 3D reconstruction and high-beam Light Detection and Ranging (LiDAR). However, these methods face severe computational bottlenecks and hardware cost constraints in practical orchard environments. The “Instance segmentation–Principal-axis slicing” two-stage volume estimation framework proposed in this study essentially introduces the geometric prior of strawberry fruits as approximate solids of revolution, replacing high-cost 3D spatial sampling with high-dimensional feature extraction. Compared with traditional 3D point cloud methods, this study reduces the dimensionality of complex spatial polygon fitting into a 2D mask micro-slice volume accumulation problem, achieving real-time processing at 8–15 FPS on an edge computing platform (NVIDIA Jetson Orin) and thoroughly breaking through the computational constraints under low-power hardware. This strategy of trading hardware cost for lightweight algorithms provides a highly cost-effective solution for agricultural robots regarding fruit volume feature perception.

Although the system achieved an extremely high linear fitting performance overall ( $R^2$ ), the estimation errors exhibited minor fluctuations when confronting fruits of different maturity stages (the MAPEs for unripe, half-ripe, and ripe fruits were 11.2%, 8.5%, and 7.6%, respectively). This phenomenon profoundly reveals the disturbance mechanism of complex agricultural environments on computer vision perception: On the one hand, a portion of the system’s error increment originates from the “mask dilation” effect under

small-scale targets. Limited by the physical resolution limit of the depth camera, the edge pixels of small-volume targets are highly prone to spatial aliasing with the background, causing minor positive deviations to be locally amplified during the volume accumulation process. On the other hand, the high spectral overlap between leaves and unripe fruits under complex canopy backgrounds exacerbates the distortion of instance segmentation boundaries. In contrast, ripe red fruits gain extremely strong visual saliency by virtue of their intense color contrast with the background, ensuring the sharpness of 2D mask extraction and thereby converging the estimation error to the lowest level. This further demonstrates that in unstructured agricultural scenes, the target's own color contrast and phenotypic scale exert a decisive influence on the accuracy of the proposed strawberry volume estimation method. This conclusion also profoundly confirms the necessity of designing the improved DFA-YOLO model at the front end of this study. Faced with the aforementioned visual challenges in unstructured environments, traditional segmentation networks often struggle to maintain the integrity of mask boundaries. In contrast, the deformable feature perception and adaptive feature fusion mechanisms equipped in DFA-YOLO can effectively enhance the model's ability to capture local irregular geometric edges and aggregate global context information to suppress the interference of background noise. It is exactly this robustness of the front-end segmentation network in the dimension of feature extraction that effectively ensures the quality of mask generation under harsh lighting and severe occlusion, thereby providing a reliable input baseline for the subsequent principal-axis slicing volume estimation algorithm.

Limited by single-view observational information, the estimation method of this system still possesses certain limitations. The elevated errors observed for misshapen fruits (MAPE 12.5–23.8%) expose the inherent limitations of the principal-axis slicing algorithm, which fundamentally relies on a prior assumption of rotational symmetry. In real-world agricultural environments, complex fruit morphologies intertwine with environmental disturbances to degrade estimation accuracy. Specifically, when encountering severe morphological anomalies such as flattened or fasciated structures, the symmetry assumption is directly violated, rendering the algorithm unable to accurately infer the occluded back-half volume. Beyond inherent malformations, extreme shooting perspectives can obscure key contour features, causing the model to extract incomplete segmentation masks that subsequently misalign the estimated principal axis. Furthermore, partial occlusions by canopy foliage or adjacent fruits physically truncate the target mask, forcing an early termination of the micro-element slicing and triggering systematic volume underestimation. Despite these limitations, considering that malformed strawberries typically account for only 3–15% of the total yield during a standard harvest season [47–49], the proposed method retains high robustness for the vast majority of standard and mildly deformed fruits.

Furthermore, due to the dataset constraints regarding environmental and cultivar variability in the current study, the generalization capability of the proposed fruit volume estimation method requires further comprehensive evaluation. First, regarding environmental diversity, data collection was restricted to a single greenhouse. While this provided a stable baseline for validating the core algorithm, it inherently lacks the illumination fluctuations, background complexities, and varied canopy structures typical of different cultivation sites. Second, concerning cultivar generalization capability, this study evaluated three specific cultivars (Zhangji, Suizhu, and Benihoppe). Although these represent the dominant commercial varieties with typical conical and long-conical morphologies, they cannot encapsulate the entire genetic diversity of all strawberry cultivars (e.g., white strawberries). Nevertheless, the core principal-axis slicing algorithm fundamentally relies on the macroscopic geometric property of rotational symmetry, whereas generalization challenges arising from cultivar-specific textures or site-specific lighting remain confined



to the instance segmentation stage. Therefore, the algorithm theoretically possesses the capability to generalize to other standard-shaped cultivars under varying environments. Collecting multi-site datasets and incorporating a broader spectrum of cultivars will be a primary focus of future work to further evaluate the environmental and cultivar-level generalization capability of the proposed method.

## 5. Conclusions

In summary, to effectively balance measurement accuracy, computational overhead, and hardware cost in non-destructive crop phenotyping, this study proposes a lightweight, single-view, real-time strawberry volume estimation framework. At the perception front-end, an improved DFA-YOLO instance segmentation model is deployed to extract high-precision 2D fruit masks. This model effectively overcomes challenges such as edge blurring and multi-scale variations in complex greenhouse environments, successfully isolating the generalization challenges caused by illumination and cultivar-specific textures at the segmentation stage. Building upon this, the back-end leverages a principal-axis slicing algorithm combined with RGB-D depth information. By transforming the complex 3D volume modeling into a micro-slice accumulation process, the algorithm fundamentally bypasses the computational bottlenecks of dense point cloud reconstruction and the estimation inaccuracies of rigid geometric fitting.

Greenhouse field tests deployed on edge devices demonstrate that, under the tested greenhouse conditions, the proposed system achieves stable volume estimation performance with low computational load, yielding a coefficient of determination ( $R^2$ ) of 0.945 between the estimated and true volumes. These results suggest that the proposed system has practical potential for in situ strawberry grading and real-time visual feedback in controlled greenhouse environments.

While the system exhibits robust performance for standard-shaped fruits in controlled agricultural environments, its volume inference is inherently bounded by the prior assumption of rotational symmetry. Moving forward, future work will focus on collecting multi-site datasets across a broader spectrum of cultivars to comprehensively evaluate the model's environmental and biological generalization capabilities. Furthermore, targeted field experiments will be conducted to empirically quantify error boundaries under extreme morphological anomalies, alongside deepening the parallelized deployment of the algorithm to promote its commercial application in large-scale orchard scenarios.

**Author Contributions:** Conceptualization, X.Z. and Y.H.; methodology, X.Z.; software, Q.G.; validation, X.Z., Y.L. and G.Z.; formal analysis, X.Z.; investigation, X.Z. and Q.G.; resources, Y.H.; data curation, X.Z. and Y.L.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z. and Y.H.; visualization, X.Z. and G.Z.; supervision, Y.H.; project administration, Y.H.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 32560410; the Special Project for Young Talents of the Rejuvenating Yunnan Talents Support Plan, grant number XDYC-QNRC-2023-0401; the Major Science and Technology Projects of Yunnan Province, grant number 202502AE090019; and the Special Fund Project for Guiding Local Scientific and Technological Development by the Central Government, grant number 202407AB110024.

**Data Availability Statement:** The data presented in this study are available on reasonable request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [\[CrossRef\]](#)
2. Wang, C.; Pan, W.; Zou, T.; Li, C.; Han, Q.; Wang, H.; Yang, J.; Zou, X. A review of perception technologies for berry fruit-picking robots: Advantages, disadvantages, challenges, and prospects. *Agriculture* **2024**, *14*, 1346. [\[CrossRef\]](#)
3. Jeong, H.; Moon, H.; Jeong, Y.; Kwon, H.; Kim, C.; Lee, Y.; Yang, S.; Kim, S. Automated technology for strawberry size measurement and weight prediction using AI. *IEEE Access* **2024**, *12*, 14157–14167. [\[CrossRef\]](#)
4. Chuquimarca, L.E.; Vintimilla, B.X.; Velastin, S.A. A review of external quality inspection for fruit grading using CNN models. *Artif. Intell. Agric.* **2024**, *14*, 1–20. [\[CrossRef\]](#)
5. Magistri, F.; Pan, Y.; Bartels, J.; Behley, J.; Stachniss, C.; Lehnert, C. Improving robotic fruit harvesting within cluttered environments through 3d shape completion. *IEEE Robot. Autom. Lett.* **2024**, *9*, 7357–7364. [\[CrossRef\]](#)
6. Neupane, C.; Pereira, M.; Koirala, A.; Walsh, K.B. Fruit sizing in orchard: A review from caliper to machine vision with deep learning. *Sensors* **2023**, *23*, 3868. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Kurdthongmee, W.; Sukkuea, A.; Alahi, M.E.E.; Zeng, Q. Non-Destructive Mangosteen Volume Estimation via Multi-View Instance Segmentation and Hybrid Geometric Modeling. *J. Imaging* **2025**, *12*, 1. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Liu, X.; Song, Z.; Yang, S.; Cui, H.; Ma, Y. Estimating strawberry size using improved YOLOv8s-seg and semi-truncated cone models. *Trans. Chin. Soc. Agric. Eng.* **2026**, *42*, 243–253. [\[CrossRef\]](#)
9. Liming, X.; Yanchao, Z. Automated strawberry grading system based on image processing. *Comput. Electron. Agric.* **2010**, *71*, S32–S39. [\[CrossRef\]](#)
10. Lu, J.; Sang, N. Detecting citrus fruits and occlusion recovery under natural illumination conditions. *Comput. Electron. Agric.* **2015**, *110*, 121–130. [\[CrossRef\]](#)
11. Wu, G.; Li, B.; Zhu, Q.; Huang, M.; Guo, Y. Using color and 3D geometry features to segment fruit point cloud and improve fruit recognition accuracy. *Comput. Electron. Agric.* **2020**, *174*, 105475. [\[CrossRef\]](#)
12. Huang, Q.; Huang, J. Comprehensive review of edge and contour detection: From traditional methods to recent advances. *Neural Comput. Appl.* **2025**, *37*, 2175–2209. [\[CrossRef\]](#)
13. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [\[CrossRef\]](#)
14. Jiang, X.; Ji, K.; Jiang, H.; Zhou, H. Research progress of non-destructive detection of forest fruit quality using deep learning. *Trans. Chin. Soc. Agric. Eng.* **2024**, *40*, 1–16. [\[CrossRef\]](#)
15. Ariza-Sentís, M.; Vélez, S.; Martínez-Peña, R.; Baja, H.; Valente, J. Object detection and tracking in Precision Farming: A systematic review. *Comput. Electron. Agric.* **2024**, *219*, 108757. [\[CrossRef\]](#)
16. Guan, B.; Wu, Y.; Zhu, J.; Kong, J.; Dong, W. GC-Faster RCNN: The object detection algorithm for agricultural pests based on improved hybrid attention mechanism. *Plants* **2025**, *14*, 1106. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wang, Y.; Wang, H.; Peng, Z. Rice diseases detection and classification using attention based neural network and bayesian optimization. *Expert Syst. Appl.* **2021**, *178*, 114770. [\[CrossRef\]](#)
18. Wang, Y.; Jin, X.; Zheng, J.; Zhang, X.; Wang, X.; He, X.; Polovka, M. An energy-efficient classification system for peach ripeness using YOLOv4 and flexible piezoelectric sensor. *Comput. Electron. Agric.* **2023**, *210*, 107909. [\[CrossRef\]](#)
19. Viveros Escamilla, L.D.; Gómez-Espinosa, A.; Escobedo Cabello, J.A.; Cantoral-Ceballos, J.A. Maturity recognition and fruit counting for sweet peppers in greenhouses using deep learning neural networks. *Agriculture* **2024**, *14*, 331. [\[CrossRef\]](#)
20. Psota, E.T.; Luc, E.; Pighetti, G.; Schneider, L.; Fryxell, R.T.; Keele, J.; Kuehn, L. Development and validation of a neural network for the automated detection of horn flies on cattle. *Comput. Electron. Agric.* **2021**, *180*, 105927. [\[CrossRef\]](#)
21. Wu, Z.; Yang, J.; Zhang, H.; Fang, C. Enhanced methodology and experimental research for caged chicken counting based on YOLOv8. *Animals* **2025**, *15*, 853. [\[CrossRef\]](#) [\[PubMed\]](#)
22. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
23. Costa, L.; Ampatzidis, Y.; Rohla, C.; Maness, N.; Cheary, B.; Zhang, L. Measuring pecan nut growth utilizing machine vision and deep learning for the better understanding of the fruit growth curve. *Comput. Electron. Agric.* **2021**, *181*, 105964. [\[CrossRef\]](#)
24. Chen, C.; Li, J.; Liu, B.; Huang, B.; Yang, J.; Xue, L. A robust vision system for measuring and positioning green asparagus based on YOLO-seg and 3D point cloud data. *Comput. Electron. Agric.* **2025**, *230*, 109937. [\[CrossRef\]](#)
25. Xue, S.; Li, Z.; Wang, D.; Zhu, T.; Zhang, B.; Ni, C. YOLO-ALDS: An instance segmentation framework for tomato defect segmentation and grading based on active learning and improved YOLO11. *Comput. Electron. Agric.* **2025**, *238*, 110820. [\[CrossRef\]](#)
26. Giménez-Gallego, J.; Martínez-del-Rincon, J.; González-Teruel, J.D.; Navarro-Hellín, H.; Navarro, P.J.; Torres-Sánchez, R. On-tree fruit image segmentation comparing Mask R-CNN and Vision Transformer models. Application in a novel algorithm for pixel-based fruit size estimation. *Comput. Electron. Agric.* **2024**, *222*, 109077. [\[CrossRef\]](#)
27. Qu, H.; Du, H.; Tang, X.; Zhai, S. Citrus fruit diameter estimation in the field using monocular camera. *Biosyst. Eng.* **2025**, *252*, 47–60. [\[CrossRef\]](#)

28. Ghose, P.; Bashir, A.; Zahid, A. Advancements in 3D Reconstruction for Plant Phenotyping: Technologies, Applications, Challenges, and Future Directions. *Sensors* **2026**, *26*, 2730. [[CrossRef](#)] [[PubMed](#)]
29. Li, Y.; Zhuo, L.; Liu, B.; Lijuan, Y.; Fanghao, W.; Wanqiang, Q.; Xi, Q. Applications of 3D reconstruction techniques in crop canopy phenotyping: A review. *Agronomy* **2025**, *15*, 2518. [[CrossRef](#)]
30. He, J.Q.; Harrison, R.J.; Li, B. A novel 3D imaging system for strawberry phenotyping. *Plant Methods* **2017**, *13*, 93. [[CrossRef](#)] [[PubMed](#)]
31. Zhang, G.; Wang, Y.; Yang, Y.; Zhang, F.; Wang, X.; Liu, L.; Chen, L.; Li, Z.; Yang, X.; Qiao, Y. High-throughput extraction of individual plant height in rapeseed based on LiDAR-Camera data fusion. *Comput. Electron. Agric.* **2026**, *240*, 111168. [[CrossRef](#)]
32. Rivera, G.; Porras, R.; Florencia, R.; Sánchez-Solís, J.P. LiDAR applications in precision agriculture for cultivating crops: A review of recent advances. *Comput. Electron. Agric.* **2023**, *207*, 107737. [[CrossRef](#)]
33. Murcia, H.F.; Tilaguy, S.; Ouazaa, S. Development of a Low-Cost System for 3D Orchard Mapping Integrating UGV and LiDAR. *Plants* **2021**, *10*, 2804. [[CrossRef](#)] [[PubMed](#)]
34. Wang, H.; Lin, Y.; Wang, Z.; Yao, Y.; Zhang, Y.; Wu, L. Validation of a low-cost 2D laser scanner in development of a more-affordable mobile terrestrial proximal sensing system for 3D plant structure phenotyping in indoor environment. *Comput. Electron. Agric.* **2017**, *140*, 180–189. [[CrossRef](#)]
35. Fu, L.; Gao, F.; Wu, J.; Li, R.; Karkee, M.; Zhang, Q. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. *Comput. Electron. Agric.* **2020**, *177*, 105687. [[CrossRef](#)]
36. Zollhöfer, M. Commodity RGB-D sensors: Data acquisition. In *RGB-D Image Analysis and Processing*; Rosin, P.L.L., Lai, Y.-K., Shao, L., Liu, Y., Eds.; Springer: Cham, Switzerland, 2019; pp. 3–13.
37. Li, Y.; Wang, W.; Guo, X.; Wang, X.; Liu, Y.; Wang, D. Recognition and positioning of strawberries based on improved YOLOv7 and RGB-D sensing. *Agriculture* **2024**, *14*, 624. [[CrossRef](#)]
38. Zhu, K.; Li, J.; Zhang, K.; Arunachalam, C.; Bhattacharya, S.; Lu, R.; Li, Z. Foundation model-based apple ripeness and size estimation for selective harvesting. *Comput. Electron. Agric.* **2025**, *236*, 110407. [[CrossRef](#)]
39. Wang, L.; Ye, L.; Chen, X.; Chu, N. MSFN-YOLOv11: A novel multi-scale feature fusion recognition model based on improved YOLOv11 for real-time monitoring of birds in wetland ecosystems. *Animals* **2025**, *15*, 3472. [[CrossRef](#)] [[PubMed](#)]
40. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
41. Xiong, Y.; Li, Z.; Chen, Y.; Wang, F.; Zhu, X.; Luo, J.; Wang, W.; Lu, T.; Li, H.; Qiao, Y. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2024; pp. 5652–5661.
42. Yun, S.; Ro, Y. Shvit: Single-head vision transformer with memory efficient macro design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2024; pp. 5756–5767.
43. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
44. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9157–9166.
45. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
46. Qi, Y.; Dong, X.; Chen, P.; Lee, K.-H.; Lan, Y.; Lu, X.; Jia, R.; Deng, J.; Zhang, Y. Canopy volume extraction of *citrus reticulata* blanco cv. shatangju trees using UAV image-based point cloud deep learning. *Remote Sens.* **2021**, *13*, 3437. [[CrossRef](#)]
47. Abrol, D.P.; Gorka, A.K.; Ansari, M.J.; Al-Ghamdi, A.; Al-Kahtani, S. Impact of insect pollinators on yield and fruit quality of strawberry. *Saudi J. Biol. Sci.* **2019**, *26*, 524–530. [[CrossRef](#)] [[PubMed](#)]
48. Cao, Z.; He, G.; Mu, S.; Qu, H. Effects of Bee Density and Hive Distribution on Pollination Efficiency for Greenhouse Strawberries: A Simulation Study. *Agronomy* **2023**, *13*, 731. [[CrossRef](#)]
49. Cui, M.; Pham, M.D.; Hwang, H.; Chun, C. Flower development and fruit malformation in strawberries after short-term exposure to high or low temperature. *Sci. Hortic.* **2021**, *288*, 110308. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.