agronomy

MDPI

Article

A Peduncle-Aware Instance Segmentation and Geometric Localization Framework for Cluster Tomato Harvesting

Heng Zhang, Xiaoming Fu and Shengxue Zhao *

College of Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China; zhangheng4816@163.com (H.Z.); dqfxm@byau.edu.cn (X.F.)
* Correspondence: zhaoshengxue@163.com

Abstract

Reliable segmentation of cherry tomato peduncles is difficult in greenhouse harvesting because peduncles are slender, occupy few pixels, resemble shoots and petioles, and are frequently occluded by clustered fruit. This study presents a peduncle-aware instance-segmentation and geometric-consistency framework for cluster-tomato harvesting perception. From the public 2022 Shanxi Agricultural Valley Tomato Town greenhouse cluster-tomato dataset, 1383 images with visible peduncles were selected and re-annotated in Labelme with polygon masks for fruit-cluster regions and peduncles. The framework combines boundary-weighted mask optimization, Coordinate Attention, and Soft-NMS. Evaluation on 208 held-out images showed a class-averaged mask AP50 of 0.9606 under the high-recall configuration; fruit-cluster and peduncle AP50 values were 0.9900 and 0.9313, respectively. At a score threshold of 0.50, overall precision, recall, F1 score, and instance-level Jaccard score were 0.8440, 0.9495, 0.8937, and 0.8078. In the 2D mask-geometric consistency evaluation, centroid-derived proxy points yielded matched-pair mean errors of 8.49 pixels for the peduncle and 5.94 pixels for the fruit-cluster region. These values quantify consistency between predicted-mask-derived and ground-truth-mask-derived points and do not represent physical robotic positioning accuracy.

Keywords: cherry tomato; harvesting robot; peduncle recognition; Mask R-CNN; instance segmentation; mask-geometric consistency

1. Introduction

Cherry tomato is an economically important crop for intensive protected cultivation. As agricultural labor becomes less available and greenhouse environments become more complex, manual identification, grasping, and harvesting increasingly struggle to meet requirements for stable, efficient, and low-damage production. Visual perception is fundamental to target recognition, harvestability assessment, and end-effector localization. Reviews of harvesting robotics consistently identify recognition, localization, and system integration as principal constraints on harvesting success [1–3]. Early tomato-specific research showed that non-color-coded visual features can improve robustness to illumination variation, fruit adhesion, and foliage occlusion [4]. More recent studies have advanced real-time instance segmentation, fruit-bunch and pedicel perception, RGB-D localization, and integrated tomato-harvesting systems [5–13]. Nevertheless, reliable identification of the peduncle and an executable harvesting target remains difficult in dense greenhouse canopies.

Ripe tomatoes are comparatively easy to detect because they have salient color and occupy relatively large image regions. Peduncle recognition and harvesting-point localization are substantially more difficult. Peduncles are slender, occupy few pixels, and are

Check for updates

Academic Editor: Dimitrios Kateris

Received: 13 August 2026
Revised: 2 October 2026
Accepted: 5 October 2026
Published: 1 November 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) license.

visually similar to shoots, petioles, sepals, and supporting structures. In dense clusters or under occlusion, their boundaries may be fragmented, merged with adjacent structures, or completely invisible; a bounding box therefore cannot adequately represent the harvestable region. For cluster and cherry tomato harvesting, a robot must identify both a peduncle harvesting point and a fruit grasping point. Pixel-level instance segmentation and geometric point estimation are consequently more appropriate than bounding-box detection alone [6,14–23].

For clustered crops, published perception pipelines combine target recognition or segmentation with geometric or keypoint estimation. Ning et al. segmented grape stems using an improved Mask R-CNN before locating a picking point near the stem centroid [14]. Ma and Li estimated grape picking and grasping points from minimum-area rectangles [15], whereas Liang et al. studied image-based tomato-cluster picking-point information [16]. Recent tomato-specific systems include SwinS-YOLACT for real-time instance segmentation, MTA-YOLACT for multitask fruit-bunch and peduncle analysis, and YOLOv8-based RGB-D localization [5–7]. These studies, together with work on grape stems, keypoints, and harvesting systems [17–27], show that harvesting-robot vision must balance boundary quality, instance separation, interpretable localization, and deployability. Because these studies use different datasets, target definitions, and physical validation protocols, their numerical results provide context rather than a direct performance ranking. Improved Mask R-CNN has also been applied to greenhouse tomato maturity-stage segmentation, while visual instance segmentation has supported fruit detection and segmentation for apple harvesting [28,29].

Compared with grape bunches, cluster-tomato images contain more tightly packed fruit instances, whereas the peduncle is shorter and is often occluded by sepals, fruit, or shoots. The Shanxi Agricultural Valley cluster-tomato dataset covers different times of day, illumination directions, and viewing angles; however, its original LabelImg annotations consist of YOLO bounding boxes for mature, raw, and occluded mature fruit and cannot directly support pixel-level fruit–peduncle segmentation [30]. We therefore performed task-oriented image selection and polygon-mask annotation and developed a peduncle-aware framework based on Mask R-CNN [31]. The framework addresses three harvesting-specific perception challenges: (i) a boundary-weighted BCE–Dice objective emphasizes narrow peduncle regions and contour pixels [32,33]; (ii) Coordinate Attention retains directional positional information for elongated structures [34]; and (iii) Soft-NMS reduces erroneous suppression of densely overlapping fruit and peduncle candidates during inference [35].

The contributions of this study are task-specific rather than the invention of new standalone network components. First, a visible-peduncle subset was re-annotated with fruit-cluster and peduncle polygon masks for harvesting-oriented perception. Second, established boundary-sensitive supervision, Coordinate Attention, and overlap-aware candidate re-scoring were integrated and evaluated within Mask R-CNN for slender peduncle and dense cluster scenes. Third, standard and high-recall operating configurations, component ablations, fixed-budget cross-model comparisons, and mask-derived geometric rules were analyzed on the same held-out test set. Fourth, favorable and failure cases were used to trace how fruit-cluster merging, cross-instance mismatch, peduncle false detection, and fragmentation propagate from instance masks to deterministic image-plane proxy points.

2. Materials and Methods

2.1. Data Source, Sample Selection, and Polygon-Mask Annotation

The original images were obtained from the 2022 Shanxi Agricultural Valley Tomato Town greenhouse cluster-tomato fruit dataset [30]. The images were acquired in a glass greenhouse in Gezitoucun, Fancun Town, Taigu District, Jinzhong, Shanxi Province, China,

from July to August 2022. The dataset includes sunny and cloudy conditions; morning, midday, afternoon, and dusk; front, back, and overhead lighting; and multiple viewing angles. After curation, the public dataset contains 3665 cluster-tomato images. The original study used LabelImg to annotate mature, raw, and occluded mature fruit as YOLO-format bounding boxes. These labels were designed for maturity detection and do not include peduncle instance masks, so they cannot directly support pixel-level peduncle segmentation or harvesting-point evaluation.

For the present task, the public images were manually screened and annotated for harvesting-oriented perception. Images were excluded when the peduncle was completely invisible or severely blurred, the target was too distant, motion blur prevented reliable delineation, or the fruit-peduncle connection could not be determined. Images were retained when the target cluster was sufficiently clear and at least part of its peduncle was visible. This process yielded 1383 images. Polygon masks were created in Labelme [36]. The annotation identifier 'fruit' denotes the visible region of one target fruit cluster as a single harvesting-task instance rather than each individual tomato. Sepals and individual-fruit pedicels belonging to that cluster were included, whereas neighboring clusters were excluded. Occluded portions that were not visible were not inferred. The identifier 'stem' denotes only the visible principal peduncle connecting the target cluster to the supporting main branch; individual-fruit pedicels, sepals, petioles, unrelated branches, and greenhouse supports were excluded. The manuscript uses the anatomical term peduncle consistently and retains 'stem' only for the historical code label. The dataset should therefore be interpreted as a visible-peduncle subset rather than an unrestricted sample of greenhouse harvesting conditions.

The polygon-mask annotations provide pixel-level supervision for instance segmentation and harvesting-point geometry. They describe irregular fruit boundaries, slender peduncle contours, and partially occluded visible regions more accurately than bounding boxes and support subsequent calculations such as centroids, minimum-area rectangles, skeleton midpoints, and projection centers. Fruit instances generally occupy larger regions and exhibit more distinctive color and continuous boundaries. Peduncle instances are smaller, elongated, and visually confounded with sepals, shoots, and petioles. Accordingly, the proposed loss and attention design focuses on peduncle boundaries and directional structure.

The first author, a master's student in mechanical engineering with practical image-annotation experience, produced the polygon masks after self-study of LabelMe and the agreed ontology. Professor Shengxue Zhao, an agricultural-machinery specialist, reviewed the completed annotations and made the final decision when class or boundary disagreements arose. The reviewer did not independently re-annotate the images. Consequently, no independently double-annotated subset was available and no inter-annotator IoU or agreement rate was calculated. The absence of an independent annotation-agreement measurement is treated as a limitation.

2.2. Overall Framework of Fruit–Peduncle Segmentation and Localization

The proposed framework consists of four stages (Figure 1). First, greenhouse cluster-tomato images containing visible fruit–peduncle structures are selected and annotated at pixel level. Second, the peduncle-aware Mask R-CNN performs fruit and peduncle instance segmentation through boundary-sensitive optimization and coordinate-aware feature enhancement. Third, Soft-NMS re-scores densely overlapping detections to reduce the loss of adjacent fruit and peduncle candidates. Finally, geometric information extracted from the predicted masks is used to estimate image-plane fruit grasping points and peduncle

harvesting points. Performance is assessed using instance-segmentation and geometric-localization metrics.

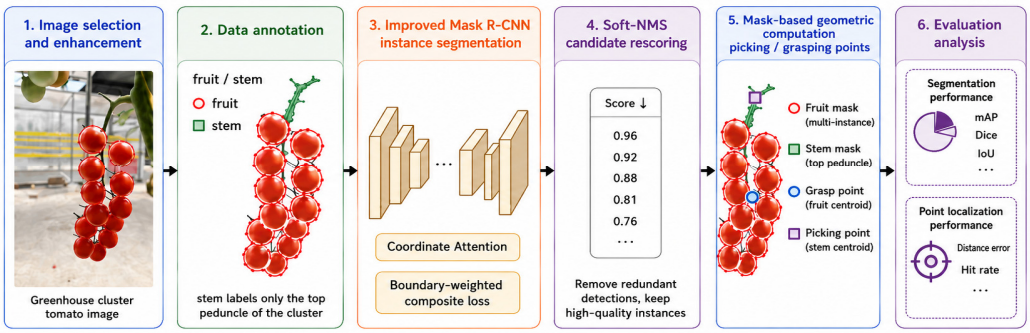


Figure 1. Workflow for fruit–peduncle recognition and harvesting-point localization in cluster tomatoes.

2.3. Peduncle-Aware Instance-Segmentation Framework

Mask R-CNN consists of a backbone, feature pyramid network (FPN), region proposal network (RPN), ROIAlign, and classification, bounding-box regression, and mask branches [31,37,38]. This two-stage architecture is suitable for agricultural scenes that require separation of multiple instances of the same class and accurate boundary extraction. In this study, the architecture was adapted to the perception requirements of cluster-tomato harvesting by introducing boundary-sensitive supervision, coordinate-aware feature enhancement, and overlap-aware instance selection. Each component corresponds to a specific scene-level difficulty rather than serving as an independent, general-purpose module substitution.

Specifically, the boundary-weighted compound loss strengthens supervision of narrow contours and endpoints; Coordinate Attention enhances directional and positional features before prediction; and Soft-NMS retains plausible candidates in densely overlapping regions during inference.

The first two components affect training and feature extraction, whereas Soft-NMS is applied only during inference and does not alter the trained network. This task-driven design retains the interpretability of two-stage instance segmentation while explicitly addressing slender peduncles and densely overlapping fruit instances.

Figure 2a presents the task-level pipeline, including image selection, Labelme polygon annotation, deterministic aspect-ratio-preserving resizing, model training, instance-mask output, and proxy-point calculation. The archived proposed-model loader did not apply Copy-Paste, Mosaic, HSV perturbation, random flipping, or random-scale augmentation. Figure 2b shows the model-level architecture, including the backbone and FPN, RPN, ROIAlign, classification and bounding-box branches, mask branch, Soft-NMS, and geometric post-processing.

2.3.1. Boundary-Weighted Compound Loss

Peduncle-mask errors are concentrated near the boundary. Because peduncles occupy only a small fraction of an image, conventional pixel-wise binary cross-entropy (BCE) can be dominated by background pixels and may underweight slender edges and endpoints. Dice and boundary losses are commonly used to alleviate class imbalance and boundary insensitivity [32,33]. We therefore combined boundary-weighted BCE with Dice loss to form the compound mask loss:

$$B_{ni} = [D_3(Y_n) - E_3(Y_n)]_i, w_{ni} = 1 + \gamma B_{ni} \quad (1)$$

$$L_{\mathrm{BWBCE}} = -\frac{\sum_{n=1}^{N}\sum_{i} w_{ni}\left[y_{ni}\ln(p_{ni}) + (1-y_{ni})\ln(1-p_{ni})\right]}{\sum_{n=1}^{N}\sum_{i} w_{ni}} \tag{2}$$

$$L_{\mathrm{Dice}} = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{2\sum_{i} p_{ni}y_{ni} + \varepsilon}{\sum_{i} p_{ni} + \sum_{i} y_{ni} + \varepsilon} \tag{3}$$

$$L_{\mathrm{mask}} = \alpha L_{\mathrm{BWBCE}} + (1-\alpha)L_{\mathrm{Dice}} \tag{4}$$

In Equations (1)–(3), n indexes the N positive mask samples and i indexes pixels. Y denotes the binary target mask; D and E with subscript 3 denote 3 × 3 morphological dilation and erosion; B is the boundary indicator; w is the pixel weight; p and y are the predicted probability and target label; gamma = 4.0 is the boundary-pixel amplification factor; alpha = 0.6 weights the boundary-weighted BCE term; and epsilon = 1×10^{-6} is the Dice smoothing constant. Equation (2) is normalized by the sum of pixel weights, consistent with the implementation.

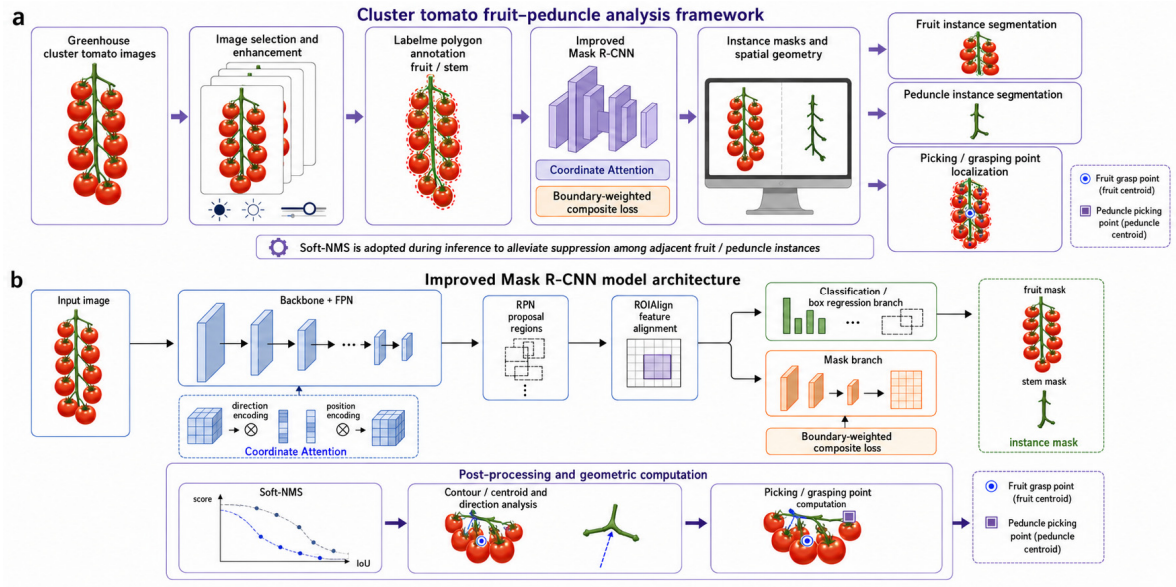


Figure 2. Fruit–peduncle analysis framework and architecture of the proposed peduncle-aware instance-segmentation model. (**a**) Task-level pipeline from image selection and polygon annotation to instance-mask output and mask-derived proxy-point localization; (**b**) architecture of the improved Mask R-CNN, including its prediction branches, Soft-NMS, and geometric post-processing.

The boundary-weight map increases the contribution of pixels near the ground-truth contour, whereas Dice loss constrains global overlap between the predicted and reference masks. Their combination improves learning of peduncle contours, endpoints, and narrow structures while reducing the influence of foreground–background imbalance.

2.3.2. Coordinate Attention

Conventional channel attention models the importance of feature channels but does not explicitly preserve directional positional information. Coordinate Attention separately aggregates features along the horizontal and vertical directions, enabling the attention weights to encode both channel responses and coordinates [34]. Peduncles have pronounced orientation and a relatively stable spatial relationship with the fruit and sepals. Coordinate Attention was therefore used to improve discrimination between peduncles and visually similar background shoots while maintaining sensitivity to peduncle location in dense fruit regions.

$$z_c^h(h) = \frac{1}{W}\sum_{w=1}^{W} x_c(h,w) \tag{5}$$

$$z_c^w(w) = \frac{1}{H}\sum_{h=1}^{H} x_c(h,w) \tag{6}$$

$$f = \delta(F_1([z^h, z^w])) \tag{7}$$

$$A_c^h(h) = \sigma\left(\left[F_h\left(f^h\right)\right]_c\right), A_c^w(w) = \sigma\left(\left[F_w\left(f^w\right)\right]_c\right) \tag{8}$$

$$y_c(h,w) = x_c(h,w)A_c^h(h)A_c^w(w) \tag{9}$$

Equations (5) and (6) define the channel-wise coordinate descriptors obtained by pooling along width and height, respectively. In Equation (7), the shared transformation consists of a 1 × 1 convolution, batch normalization, and the *h*-swish activation. Its output is split into horizontal and vertical tensors. In Equation (8), two direction-specific 1 × 1 transformations followed by the sigmoid function generate channel- and position-dependent attention weights. Equation (9) applies both weights to the input feature at the same channel and spatial position. Coordinate Attention uses a reduction ratio of 32 and is applied to the five 256-channel FPN outputs.

Coordinate Attention does not change the output heads; instead, it enhances relevant features before prediction. In the ablation experiment, the module produced only a small improvement when used alone but yielded a larger increase in peduncle AP50 when combined with the boundary-weighted loss, indicating complementary effects between positional encoding and boundary supervision.

2.3.3. Soft-NMS Post-Processing

Fruit and peduncles in a cluster are frequently adjacent or overlapping. Conventional non-maximum suppression (NMS) removes candidates with high overlap with a higher-scoring detection and can therefore discard valid occluded peduncles or neighboring fruit. Soft-NMS instead decays candidate scores as a continuous function of overlap and retains more potentially correct instances in dense scenes [35]. In this study, Soft-NMS was used in the high-recall configuration. Because a low pre-filtering score threshold retains more low-confidence candidates, it increases recall and AP50 but can reduce precision and instance-level accuracy. Deployment thresholds should therefore be selected according to the relative costs of false harvesting and missed harvesting.

$$\hat{s}_i = s_i \exp\left(-\frac{\mathrm{IoU}(b_m, b_i)^2}{\sigma_s}\right) \tag{10}$$

In Equation (10), b with subscript m is the current highest-scoring box, b with subscript i is another candidate, s with subscript i is its original score, the hatted score is its decayed score, and sigma with subscript s is the Gaussian Soft-NMS decay parameter. Soft-NMS is applied independently within each class using bounding-box IoU, sigma = 0.5, an internal retention threshold of 0.001, and an external score threshold of 0.03 in the high-recall configuration.

2.3.4. Overall Training Objective and Algorithm

The overall objective combines classification, bounding-box regression, and mask supervision and is expressed in Equation (11).

$$L_{\text{total}} = L_{\text{RPN,obj}} + L_{\text{RPN,box}} + L_{\text{ROI,cls}} + L_{\text{ROI,box}} + \lambda_{\text{mask}} L_{\text{mask}} \tag{11}$$

In Equation (11), the first two terms are the RPN objectness and proposal-regression losses, the next two terms are the ROI classification and box-regression losses, and the final term is the compound mask loss defined in Equation (4). The mask-branch coefficient is 1.0 because the implementation sums the returned loss terms without an additional multiplier.

2.4. Two-Dimensional Mask-Geometric Consistency Evaluation from Instance Masks

The instance masks define the visible fruit-cluster and peduncle regions. Independently annotated physical cutting and grasping locations and RGB-D calibration were not available. For each geometric rule, the reference point was generated by applying that same rule to the matched ground-truth mask, while the predicted point was obtained from the predicted mask. The experiment therefore measures the stability of a mask-to-point transformation under segmentation error. It does not determine which point is physically optimal or executable for harvesting. The final reported rule used the peduncle centroid as a harvesting-point proxy and the fruit-cluster centroid as a grasping-point proxy.

For a processed instance mask M, let the foreground-pixel set contain all image coordinates at which the binary mask equals one. The centroid coordinates are calculated as follows:

$$x_c = \frac{1}{|\Omega_M|} \sum_{(x,y)\in\Omega_M} x \tag{12}$$

$$y_c = \frac{1}{|\Omega_M|} \sum_{(x,y)\in\Omega_M} y \tag{13}$$

$$p_{\text{harvest}} = c\left(M'_{\text{peduncle}}\right), p_{\text{grasp}} = c\left(M'_{\text{fruit}}\right) \tag{14}$$

$$e_k = \sqrt{\left(x_{\text{pred},k} - x_{\text{ref},k}\right)^2 + \left(y_{\text{pred},k} - y_{\text{ref},k}\right)^2} \tag{15}$$

$$\text{SR}_\tau = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(e_k \le \tau) \times 100\% \tag{16}$$

In Equations (12)–(16), the foreground set is denoted by uppercase omega with mask subscript M, and its cardinality is the number of foreground pixels. The function c returns the centroid of a processed mask. The prime denotes a post-processed mask. For matched pair k, the predicted-mask-derived and reference-mask-derived coordinates define the Euclidean error. N is the number of matched pairs, tau is the image-plane distance threshold, and the indicator function equals one when the error does not exceed the threshold and zero otherwise. Conditional errors exclude unmatched targets; coverage and end-to-end success over all ground-truth targets are reported separately.

For a peduncle, the centroid is less sensitive than skeleton or endpoint methods to fragmentation and endpoint noise. For a fruit cluster, the mask centroid is generally more stable than a local distance-transform maximum under occlusion or irregular visible masks. Alternative rules were also evaluated by applying each rule separately to the predicted mask and its matched ground-truth mask. Accordingly, differences among methods quantify transformation stability with respect to mask perturbation and do not rank the physical suitability of cutting or grasping targets.

Figure 3 details the point-estimation procedure: predicted and ground-truth masks are morphologically processed, peduncle and fruit centroid proxies are calculated using the same rules, and their 2D mask-geometric consistency is assessed using Euclidean distance and threshold-based success rates.

2.5. Experimental Settings and Reproducibility

The curated dataset contained 1383 images and 2772 exported polygon instances. This was the dataset used throughout the original training and evaluation; no images were added during revision. The earlier value of 1300 was a reporting error. The fixed training, validation, and test subsets contained 968 images with 1940 instances, 207 images with 416 instances, and 208 images with 416 instances, respectively. The test set contained 208 fruit-cluster and 208 peduncle ground-truth instances and was held out from model development. The archived split generator shuffled image-annotation pairs with seed

2026 and then applied the 70%/15%/15% proportions. Plant, acquisition-sequence, and greenhouse-area identifiers were not enforced as grouping constraints. The split is therefore described as a fixed image-level split, and strict group independence cannot be claimed.

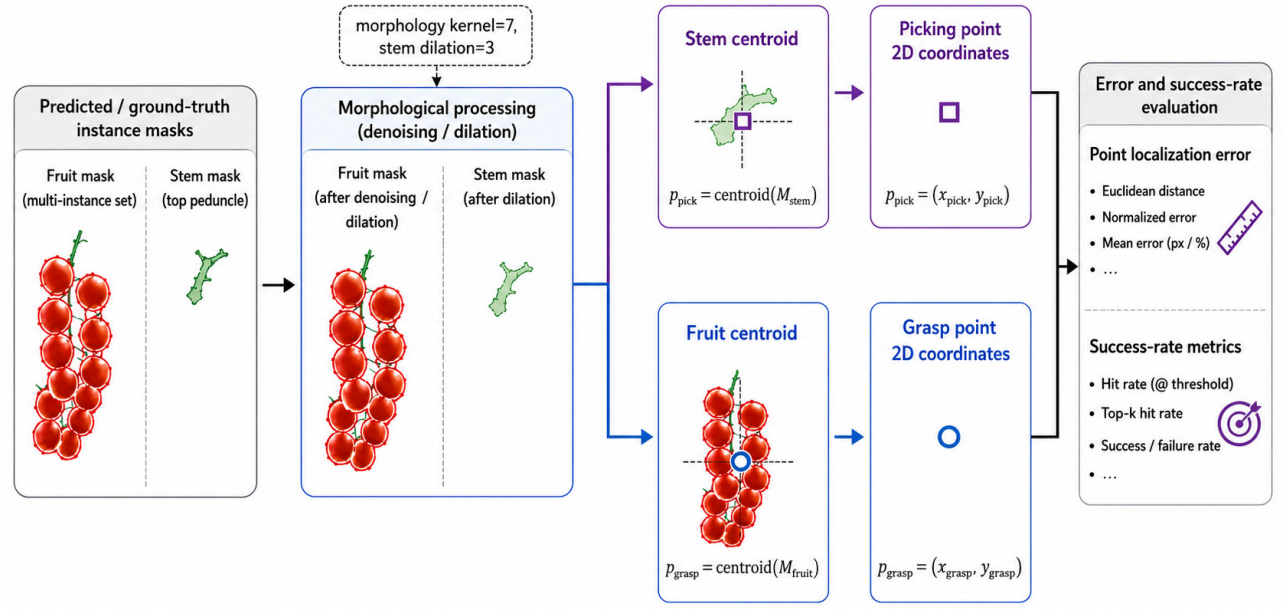


Figure 3. Two-dimensional mask-geometric consistency evaluation of peduncle and fruit proxy points.

The proposed model used TorchVision Mask R-CNN with a ResNet-50-FPN backbone and default COCO-pretrained weights [39]. Images retained their aspect ratio; the dataset loader limited the longest side to 1280 pixels, and the model transform used a 640-pixel minimum side and a 1280-pixel maximum side. Training used SGD with an initial learning rate of 0.0025, momentum 0.9, weight decay 1×10^{-4}, batch size 2, four workers, automatic mixed precision, and 20 epochs. MultiStepLR reduced the learning rate by 0.1 at epochs 10 and 15. No Copy-Paste, Mosaic, HSV, random-flip, or random-scale augmentation was used. The best checkpoint maximized validation peduncle AP50 plus validation mean matched peduncle mask IoU; the archived A7 checkpoint was selected at epoch 4. Three complete training runs used seeds 42, 1234, and 3472. Table 1 reports results from the archived A7 checkpoint. In Table 2, only the A0, A3, A6, and A7 training-component rows are reported as three-run means, whereas the threshold and Soft-NMS rows use the archived A7 checkpoint. The cross-model results in Table 3 and the geometric analyses likewise use the archived A7 checkpoint. Run-wise values and standard deviations were not retained and are not reported. Default TorchVision 0.17.1 RPN and ROI settings were retained.

The author-confirmed environment used an NVIDIA GeForce RTX 4090 vGPU allocation with 32 GB available memory, 12 allocated cores of an Intel Xeon Platinum 8260 CPU, and 62 GB RAM; the GPU driver was 580.105.08. The archived software environment records Python 3.10.20, PyTorch 2.2.1+cu121, TorchVision 0.17.1+cu121, CUDA Runtime 12.1, cuDNN 8.9.2, and OpenCV 4.10.0. The final model contained 44.01 million parameters. The batch-size-1 timer began after CPU-to-GPU transfer and included forward inference and Soft-NMS but excluded file loading, decoding, preprocessing, visualization, proxy-point calculation, and robot control. The recorded 25.50 ms per image (39.21 FPS) is therefore model-pipeline throughput rather than end-to-end robot throughput.

Table 1. Instance-segmentation performance of the archived A7 checkpoint under two inference protocols.

Configuration	Class	COCO Mask AP	Mask AP50 (AUC)	Mask AP75	Precision	Recall	F1 Score	Instance Jaccard	Mean Mask IoU
our_model/std (score threshold = 0.50; no Soft-NMS)	Fruit	0.8511	0.9817	0.9702	0.8874	0.9856	0.9339	0.8761	0.9222
our_model/std (score threshold = 0.50; no Soft-NMS)	Peduncle	0.4688	0.8981	0.4306	0.8017	0.9135	0.8539	0.7451	0.7754
our_model/std (score threshold = 0.50; no Soft-NMS)	Overall	0.6793	0.9412	0.7454	0.8440	0.9495	0.8937	0.8078	0.8515
final_model (score threshold = 0.03 + Soft-NMS)	Fruit	0.8557	0.9900	0.9743	0.7667	0.9952	0.8661	0.7638	0.9207
final_model (score threshold = 0.03 + Soft-NMS)	Peduncle	0.4830	0.9313	0.4396	0.4727	0.9567	0.6328	0.4628	0.7679
final_model (score threshold = 0.03 + Soft-NMS)	Overall	0.6899	0.9625	0.7535	0.5876	0.9760	0.7335	0.5792	0.8458

Table 2. Training-component ablation and controlled inference analysis. The first four rows are three-run means; the final three rows use the archived A7 checkpoint.

Model and Inference Configuration	Fruit Mask AP50	Peduncle Mask AP50	Mask mAP50	FPS	Interpretation
Mask R-CNN baseline	0.9782	0.8719	0.9250	43.22	Reference
Baseline + boundary-weighted loss	0.9785	0.8805	0.9295	38.22	Modest peduncle gain
Baseline + Coordinate Attention	0.9692	0.8751	0.9222	35.34	No overall gain alone
Boundary loss + Coordinate Attention, threshold 0.50	0.9817	0.8981	0.9399	38.87	A7 standard output
A7 + Soft-NMS, threshold 0.50	0.9818	0.9003	0.9411	n.r.	Soft-NMS gain 0.0012
A7, threshold 0.03	0.9898	0.9275	0.9587	n.r.	Threshold effect
A7 + Soft-NMS, threshold 0.03	0.9900	0.9313	0.9606	39.21	High-recall setting

The comparison models used the same image split but model-specific preprocessing and optimizers [40–42]. Cascade Mask R-CNN used ResNet-50-FPN, 1333 × 800 aspect-ratio-preserving inputs, SGD with learning rate 0.02, batch size 2, and a score threshold of 0.05. RTMDet-Ins-S used CSPNeXt-S, 640 × 640 inputs, AdamW with learning rate 0.004, batch size 2, and a score threshold of 0.05. YOLACT used ResNet-50, 550 × 550 inputs, SGD with learning rate 0.001, batch size 2, and a score threshold of 0.05. YOLOv8n-seg used 640 × 640 inputs, batch size 8, seed 0, and Ultralytics 8.4.66 defaults. All were limited to 20 epochs. Because RTMDet-Ins-S and YOLACT were derived from longer recommended schedules, this experiment is a fixed-budget comparison rather than an architecture-optimal benchmark.

Table 3. Fixed-budget comparison using the archived checkpoint of each model.

Method	COCO Mask AP	Mask AP50 (AUC)	Precision	Recall	F1 Score	Instance Jaccard
our_model/std (score threshold = 0.50; no Soft-NMS)	0.6793	0.9412	0.8440	0.9495	0.8937	0.8078
final_model (score threshold = 0.03 + Soft-NMS)	0.6899	0.9625	0.5876	0.9760	0.7335	0.5792
Cascade Mask R-CNN	0.6810	0.9510	0.8282	0.9615	0.8899	0.8016
RTMDet-Ins-S	0.6060	0.8970	0.3388	0.9447	0.4987	0.3322
YOLACT	0.6090	0.9120	0.8939	0.9111	0.9024	0.8221
YOLOv8n-seg	0.7320	0.9659	0.8837	0.9495	0.9154	0.8440

2.6. Evaluation Metrics

Instance-level precision, recall, F1 score, and the instance-level Jaccard score were calculated after one-to-one matching at mask IoU ≥ 0.50. The Jaccard score is TP/(TP + FP + FN) and is not conventional classification accuracy based on background true negatives. COCO AP is the mean AP over IoU thresholds from 0.50 to 0.95, whereas AP50 is the area under the precision-recall curve at IoU = 0.50. The standard proposed-model operating point used score threshold 0.50 without Soft-NMS. The high-recall operating point used score threshold 0.03 with Gaussian Soft-NMS.

The archived geometric-consistency analysis used one-to-one greedy mask-IoU matching with a permissive acceptance threshold of 0.05. This threshold was used only to establish tentative correspondence for a diagnostic mask-to-point analysis and was not used for the main instance-segmentation metrics, which use IoU ≥ 0.50. Point errors and conditional success rates were computed only for matched pairs. Coverage and end-to-end success over all 208 ground-truth targets were therefore also reported, counting unmatched targets as unsuccessful. The 20-pixel and 30-pixel criteria are sensitivity thresholds in the resized image coordinate system, not physical robot tolerances. A separate threshold-sensitivity analysis at mask IoU ≥ 0.25 and ≥ 0.50 was performed using the archived direct-centroid pair records.

3. Results

3.1. Final Instance-Segmentation Performance

Table 1 summarizes the proposed model under the two explicitly separated inference protocols. Under our_model/std (score threshold = 0.50 without Soft-NMS), the overall F1 score was 0.8937, indicating a favorable precision–recall balance. Fruit mask AP50 was 0.9817, whereas peduncle mask AP50 was 0.8981. The difference reflects the smaller area, weaker texture contrast, and less distinct boundaries of peduncles. Fruit also achieved high AP75 and mean mask IoU. By contrast, peduncle COCO AP was 0.4688 despite an AP50 of 0.8981, demonstrating that small boundary displacements have a pronounced effect at stricter IoU thresholds.

Under final_model (score threshold = 0.03 with Soft-NMS), overall mask AP50 (AUC) was 0.9625, and the class-averaged mask mAP50 was 0.9606; peduncle mask AP50 reached 0.9313. Recall increased substantially, whereas precision and instance accuracy decreased because more low-confidence candidates were retained. The two protocols therefore represent different inference operating points and are reported separately throughout the revised manuscript.

3.2. Ablation Study

The training-component ablation and controlled inference analysis are summarized in Table 2 and Figure 4. The first four rows of Table 2 report three-run means for seeds 42, 1234, and 3472. Boundary-weighted loss increased peduncle mask AP50 from 0.8719 to 0.8805 and mean mask AP50 from 0.9250 to 0.9295. Coordinate Attention alone increased peduncle AP50 slightly to 0.8751 but reduced mean mask AP50 to 0.9222. Combining the two training components yielded peduncle mask AP50 of 0.8981 and mean mask AP50 of 0.9399. The last three rows are post hoc inference analyses of the archived A7 checkpoint rather than three-run means. Because run-wise values and standard deviations were not retained, the training-component differences are not claimed to be statistically significant.

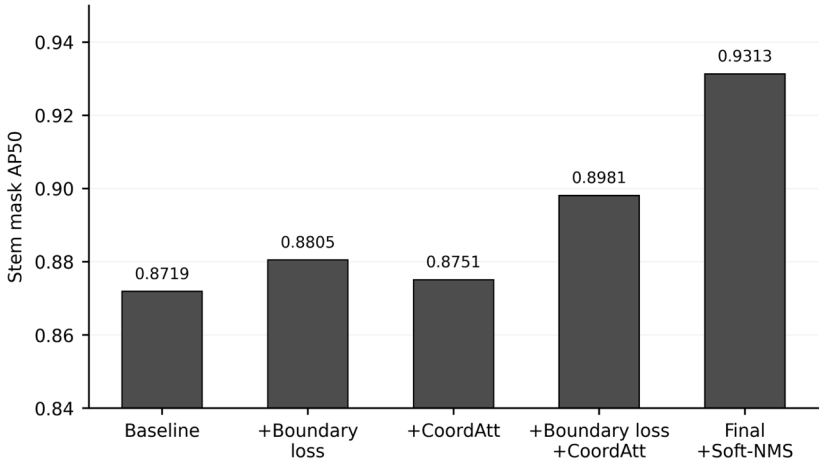


Figure 4. Peduncle mask AP50 under different ablation configurations.

The controlled 2 × 2 comparison separates score-threshold and Soft-NMS effects. At threshold 0.50, Soft-NMS increased mean mask AP50 from 0.9399 to 0.9411. At threshold 0.03, it increased mean mask AP50 from 0.9587 to 0.9606. In contrast, lowering the threshold from 0.50 to 0.03 without Soft-NMS increased mean mask AP50 by 0.0188. Most of the apparent high-recall gain therefore resulted from retaining lower-confidence candidates rather than Soft-NMS alone. The 39.21 FPS value applies only to the low-threshold Soft-NMS operating point under the timing boundary defined in Section 2.5.

3.3. Comparison with Other Instance-Segmentation Models

Table 3 presents a fixed 20-epoch-budget comparison with Cascade Mask R-CNN, RTMDet-Ins-S, YOLACT, and YOLOv8n-seg. At the standard operating point, the proposed model and Cascade Mask R-CNN were numerically similar: the proposed model had slightly higher precision and F1 score, whereas Cascade Mask R-CNN had slightly higher COCO mask AP, AP50, recall, and instance-level Jaccard score. Confidence intervals were not available, so no statistical-superiority or equivalence claim is made. YOLOv8n-seg achieved the highest COCO mask AP, AP50, F1 score, and instance-level Jaccard score in the recorded comparison. The proposed high-recall configuration achieved the highest recall but substantially lower precision and F1 score because it retained many low-confidence candidates.

The comparison therefore demonstrates operating trade-offs rather than general state-of-the-art superiority. It is also not an architecture-optimal benchmark because all models were limited to 20 epochs, while RTMDet-Ins-S and YOLACT were inherited from longer recommended schedules. The proposed contribution is instead evaluated through task-specific integration, transparent component analysis, fruit-cluster and peduncle annotation, and error propagation from masks to geometric proxy points.

Figure 5 provides qualitative comparisons on four representative test images. Yellow circles indicate ground-truth regions of interest, green circles indicate regions retained by the proposed model, and red circles mark weak segmentation by comparison models near peduncles or fruit–peduncle connections. The proposed model preserved fruit contours and produced more continuous masks in narrow peduncle regions and connection zones. Cascade Mask R-CNN, RTMDet-Ins-S, and YOLACT more frequently exhibited weak peduncle responses, omissions, fragmented connections, confusion with background shoots, or unstable instance matching. These observations are consistent with the intended roles of boundary supervision and coordinate-aware feature modeling.

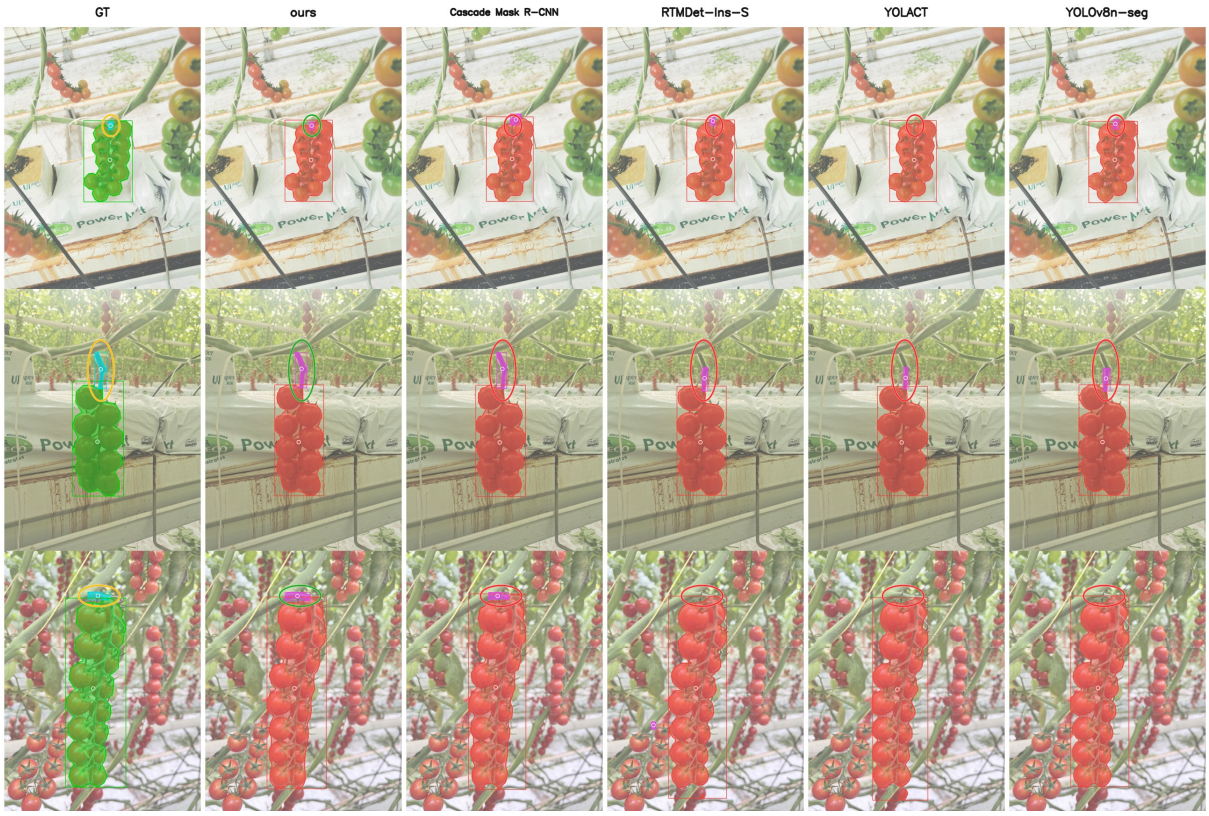


Figure 5. Instance-mask comparison on the same test images, with highlighted fruit–peduncle connection regions.

3.4. Two-Dimensional Mask-Geometric Consistency Evaluation

The 2D mask-geometric consistency evaluation was performed using the predicted instance masks. Reference points were generated automatically from ground-truth masks using the same geometric rules applied to predicted masks; no independently annotated physical harvesting or grasping points and no three-dimensional closed-loop robot experiments were used. The reported values therefore quantify consistency between predicted-mask-derived and ground-truth-mask-derived image-plane points and should not be interpreted as physical end-effector positioning error.

Table 4 reports the main post-processed centroid results at the archived matching threshold of 0.05. The analysis produced 189 peduncle pairs and 206 fruit-cluster pairs from 208 targets per class. Mean matched-pair error was 8.49 pixels for the peduncle centroid and 5.94 pixels for the fruit-cluster centroid. Conditional 20-pixel success was 94.18% and 94.66%, whereas end-to-end success was 85.58% and 93.75% when unmatched targets were counted as failures. In the threshold-sensitivity check based on archived direct-centroid pair records, all 206 fruit-cluster pairs already had IoU $\geq$ 0.50, so fruit coverage (99.04%),

mean error (5.86 pixels), conditional 20-pixel success (94.66%), and end-to-end 20-pixel success (93.75%) were unchanged at thresholds of 0.05, 0.25, and 0.50. For peduncles, coverage was 90.87%, 90.38%, and 88.94%; mean direct-centroid error was 8.52, 8.36, and 7.81 pixels; conditional 20-pixel success was 94.18%, 94.68%, and 95.68%; and end-to-end 20-pixel success was 85.58%, 85.58%, and 85.10%, respectively. Thus, stricter matching slightly reduced peduncle coverage but did not reverse the geometric-consistency conclusion.

Table 4. Conditional and end-to-end image-plane mask-geometric consistency of the centroid proxy strategy.

Metric	Peduncle Centroid Proxy	Fruit-Cluster Centroid Proxy
Matched pairs/GT targets	189/208	206/208
Match coverage (%)	90.87	99.04
Mean error among matched pairs (pixels)	8.49	5.94
Conditional success within 20 pixels (%)	94.18	94.66
Conditional success within 30 pixels (%)	96.30	97.57
End-to-end success within 20 pixels (%)	85.58	93.75
End-to-end success within 30 pixels (%)	87.50	96.63

3.5. Comparison of Mask-Derived Geometric Point Methods

As shown in Table 5, the peduncle centroid produced a mean matched-pair error of 8.492 pixels, slightly lower than the 8.749-pixel error of the minimum-area rectangle center, and achieved a 20-pixel conditional success rate of 94.180%. Skeleton midpoints, polynomial-fit midpoints, and connection-point offsets produced larger errors because peduncle masks frequently contained local fragmentation, curvature, or incomplete endpoints. These results compare stability under each rule's own ground-truth-mask-derived reference and do not identify a physically optimal cutting location.

Table 5. Comparison of peduncle mask-derived proxy-point methods among matched pairs.

Method	Mean Error (Pixels)	Within 20 Pixels (%)	Within 30 Pixels (%)
Minimum-area rectangle center	8.749	93.651	96.296
Centroid (proposed)	8.492	94.180	96.296
Horizontal midpoint near centroid	8.801	92.063	95.238
Skeleton midpoint	10.940	86.243	94.709
Polynomial-fit midpoint	11.484	85.185	94.709
Connection-point offset	12.281	82.540	89.947

As shown in Table 6, the fruit-cluster mask centroid yielded a mean matched-pair error of 5.941 pixels and was more stable than the bounding-box center, minimum-area rectangle center, and largest-visible-component center. The distance-transform center produced an error of 71.133 pixels because a local distance maximum can shift between visible tomatoes when the cluster mask is occluded or asymmetric. The previous combined normalized-success column was removed because its permissive definition produced a ceiling of approximately 99.5% to 100% for most methods and therefore contributed little discrimination.

Table 6. Comparison of fruit-cluster mask-derived proxy-point methods among matched pairs.

Method	Mean Error (Pixels)	Within 20 Pixels (%)	Within 30 Pixels (%)
Bounding-box center	9.577	91.748	95.631
Minimum-area rectangle center	10.877	85.437	92.718
Centroid (proposed)	5.941	94.660	97.573
Projection centroid	6.242	94.660	97.573
Distance-transform center	71.133	51.456	57.767
Hybrid center	31.824	62.621	69.417
Largest visible-component center	12.654	82.524	89.806

3.6. Qualitative Case Analysis

Representative test cases were visualized to relate segmentation quality to 2D mask-geometric consistency error. Ground-truth fruit regions are shown in green and predicted fruit regions in red; ground-truth peduncle regions are shown in cyan and predicted peduncle regions in purple. Open circles denote proxy reference points derived from ground-truth masks, solid markers denote proxy points derived from predicted masks, and yellow or white line segments indicate the Euclidean discrepancy between the two mask-derived points. This representation jointly displays boundary quality, instance matching, and image-plane geometric consistency.

Figure 6 illustrates representative localization results. When the cluster boundary was complete and the fruit–peduncle connection was visible, predicted fruit masks closely overlapped the ground truth and both localization points were near their references. Even when the peduncle occupied a small area, a mask covering its principal visible region was sufficient for stable centroid estimation.



Figure 6. Visualization of fruit-cluster masks and mask-derived peduncle/fruit proxy points. (**a**) Close-view fruit cluster with ground-truth and predicted masks and their peduncle/fruit proxy points; (**b**) vertically oriented fruit cluster under background clutter with the same mask and proxy-point overlays.

3.6.1. Favorable Cases

Favorable samples generally had continuous cluster contours, visible fruit-peduncle connections, and sufficient background contrast. Table 7 summarizes the three favorable

cases shown in Figure 7. In these examples, small local mask differences produced little centroid displacement. The cases illustrate the behavior of the combined model but do not isolate the causal contribution of an individual module.

Table 7. Instance segmentation and mask-geometric consistency results for favorable cases.

Case	Sample Characteristics	Fruit IoU	Peduncle IoU	Grasping-Point Error (Pixels)	Harvesting-Point Error (Pixels)	Interpretation
Case 1	Complete cluster contour; clear peduncle	0.9362	0.8721	0.90	0.81	Fruit and peduncle regions matched well; points were close to references
Case 2	Vertical cluster; stable centroid geometry	0.9426	0.8253	0.74	1.15	Regular masks were insensitive to small boundary fluctuations
Case 3	Partial occlusion; accurate principal region	0.9501	0.7032	0.62	1.33	Principal-region overlap preserved stable centroid localization

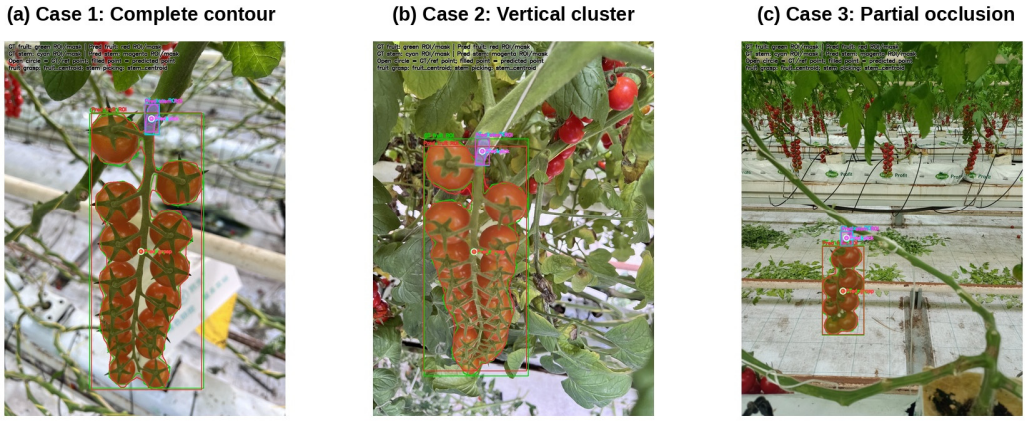


Figure 7. Regions of interest, masks, and mask-derived proxy points for favorable prediction cases.

Case 1 contained a complete fruit contour and a clear peduncle. The predicted fruit boundary nearly coincided with the ground truth (fruit IoU = 0.9362), and the grasping-point error was 0.90 pixels. The predicted peduncle mask covered the principal ground-truth region (peduncle IoU = 0.8721), producing a harvesting-point error of 0.81 pixels.

Case 2 showed a relatively regular, vertically oriented cluster. Fruit IoU was 0.9426 and grasping-point error was 0.74 pixels; peduncle IoU was 0.8253 and harvesting-point error was 1.15 pixels. Because both points depend on the overall mask geometry, small local gaps and protrusions had little influence on the centroids.

Case 3 contained partial foliage and peduncle occlusion. The predicted fruit mask still covered the principal cluster region (fruit IoU = 0.9501; grasping-point error = 0.62 pixels). Although peduncle IoU was lower (0.7032), the principal region remained correctly positioned and the harvesting-point error was only 1.33 pixels.

3.6.2. Failure Cases

Failure cases exposed two principal mechanisms, summarized in Table 8 and visualized in Figure 8. Fruit-cluster masks could merge with neighboring clusters or be matched to the wrong instance, shifting the centroid. Peduncle omission, false detection, or fragmentation could produce similar displacement. These examples show why matched-pair geometric error should be interpreted together with detection coverage and false-positive statistics.

Table 8. Error types and sources in failure cases for segmentation and mask-geometric consistency.

Case	Error Type	Fruit/Peduncle IoU	Grasping/Harvesting Error (Pixels)	Principal Cause	Potential Improvement
Case 4	Fruit-instance merging	0.8176/—	65.12/—	Mask expanded toward an adjacent cluster and shifted the centroid	Strengthen dense-instance separation and boundary constraints
Case 5	Cross-instance fruit mismatch	0/—	593.45/—	Prediction matched another similar cluster in the same image	Improve one-to-one target matching and target-selection constraints
Case 6	Small-peduncle false detection	Favorable/0	5.01/418.85	Shoot, petiole, or support structure was classified as peduncle	Add hard negatives and fruit–peduncle connectivity/topology constraints
Case 7	Fragmented or incomplete peduncle	0.9323/0.4117	9.07/78.67	Predicted peduncle covered only part of the target or an adjacent branch	Use slender-target augmentation and continuity constraints

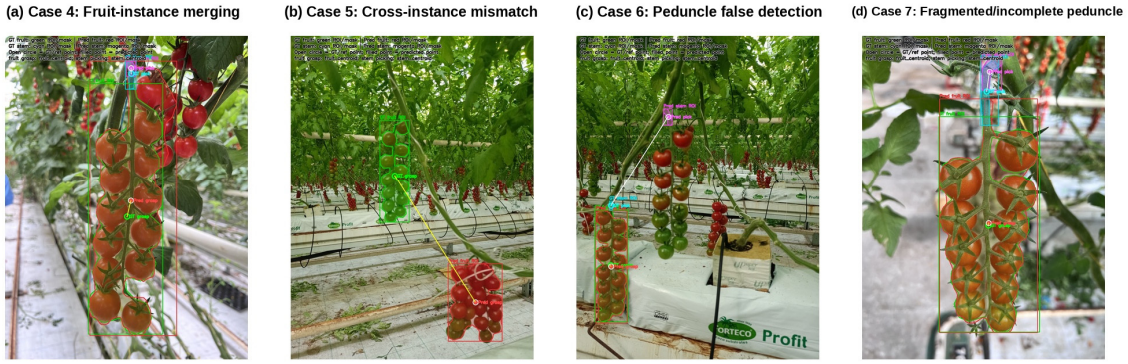


Figure 8. Regions of interest, masks, mask-derived proxy points, and geometric-consistency error vectors for failure cases.

In Case 4, the fruit mask expanded toward an adjacent cluster. Fruit IoU remained 0.8176, but inclusion of neighboring fruit enlarged the predicted region and shifted its centroid, producing a grasping-point error of 65.12 pixels.

Case 5 was a cross-instance mismatch in an image containing multiple similar clusters. Fruit IoU was 0 because the selected prediction corresponded to another cluster, and the grasping-point error reached 593.45 pixels. Localization evaluation must therefore be based on one-to-one mask matching rather than simply selecting the highest-scoring instance.

In Case 6, a shoot or petiole was falsely classified as the peduncle. The fruit prediction remained accurate (grasping-point error = 5.01 pixels), but peduncle IoU was 0 and the predicted peduncle region occurred above the true fruit–peduncle connection, producing a harvesting-point error of 418.85 pixels. Similar color and elongated shape make shoots, petioles, and greenhouse structures difficult negatives.

In Case 7, the fruit instance remained stable (fruit IoU = 0.9323; grasping-point error = 9.07 pixels), but the predicted peduncle mask covered only part of the target or extended toward an adjacent branch. Peduncle IoU was 0.4117 and the harvesting-point error was 78.67 pixels, illustrating the sensitivity of small peduncles to occlusion and boundary fragmentation.

4. Discussion

The results indicate that harvesting-oriented perception should be evaluated using segmentation quality, operating-point behavior, match coverage, and image-plane geometric consistency. The proposed model did not uniformly outperform all comparison architectures. Cascade Mask R-CNN produced numerically similar results to the proposed standard setting on several aggregate metrics, and YOLOv8n-seg achieved stronger aggregate segmentation results under the recorded fixed-budget experiment. Because confidence intervals were not available, no statistical-equivalence or superiority claim is made. The contribution of the present work therefore lies in task-specific fruit-cluster and peduncle annotation, an interpretable Mask R-CNN adaptation, controlled diagnosis of threshold and Soft-NMS effects, and analysis of how mask errors propagate to deterministic proxy points.

Fruit localization failures were dominated by instance merging and cross-instance mismatch. Because entire clusters were annotated as fruit instances, expansion of a predicted mask toward a neighboring cluster could preserve AP50 while markedly shifting the centroid. In multi-cluster images, incorrect target selection produced much larger errors than moderate boundary inaccuracies.

Peduncle localization failures were dominated by omission, false detection, and fragmentation of small targets. Peduncles are much smaller than fruit and resemble shoots, petioles, and support structures. Increasing the number of hard negative examples, applying targeted augmentation for slender structures, and incorporating fruit–peduncle connectivity or topology constraints are promising directions.

The results also show that a high mask AP50 does not necessarily imply a small localization error. AP50 tests whether an instance exceeds an IoU threshold, whereas harvesting- and grasping-point errors also depend on mask geometry, one-to-one instance matching, and consistency of the target definition. Reporting COCO AP, AP50, instance-level F1 score, mean mask IoU, mean localization error, and threshold-based success rates therefore provides a more complete assessment for harvesting applications.

Limitations and Future Work

Several limitations should be considered. The evaluation cohort was intentionally restricted to images in which at least part of the peduncle could be annotated reliably; completely invisible or severely blurred peduncles and other indeterminate cases were excluded. This selection can overestimate robustness in unrestricted greenhouse scenes. The archived implementation used a fixed image-level random split and did not enforce group-disjoint partitioning by plant, acquisition sequence, or greenhouse area. Related images may therefore occur in different subsets. One primary annotator produced the masks and an agricultural-machinery expert reviewed them, but no independently double-annotated subset or inter-annotator agreement metric was available. Three seeded runs were completed for the training-component ablation, but run-wise values and dispersion were not retained; no statistical-significance claim is made. The cross-model comparison used one 20-epoch budget rather than the complete recommended schedule of each architecture.

The geometric analysis has additional limits. References were generated automatically from ground-truth masks by the same rule applied to predicted masks. The archived 0.05 IoU threshold was deliberately permissive and is retained only as a diagnostic correspondence rule. The stricter-threshold sensitivity analysis showed stable conditional geometric errors but slightly lower peduncle coverage. The method comparison therefore evaluates transformation stability rather than physical target suitability. Practical harvesting still requires independent physical-point annotation, depth sensing, hand-eye calibration, end-effector constraints, collision checking, and closed-loop experiments.

5. Conclusions

This study developed a peduncle-aware Mask R-CNN and mask-geometric analysis framework for greenhouse cluster-tomato harvesting perception. A visible-peduncle annotation set of 1383 images was constructed from a public detection dataset by labeling fruit-cluster regions and visible peduncles with polygon masks. Because difficult cases with unannotatable peduncles were excluded, the reported results apply specifically to this curated subset.

Across the three seeded training runs, boundary-weighted loss and Coordinate Attention produced modest mean changes, and the combined training configuration improved peduncle AP50 relative to the baseline. The threshold and Soft-NMS analyses and the geometric evaluation were performed with the archived A7 checkpoint. Because run-wise values and dispersion were not retained, the training-component differences are not interpreted as statistically significant. Controlled inference analysis showed that most of the high-recall AP50 increase came from lowering the candidate threshold, while Soft-NMS produced only a small additional gain at a fixed threshold.

Centroid-based proxy points produced matched-pair mean image-plane errors of 8.49 pixels for peduncles and 5.94 pixels for fruit-cluster regions. Match coverage was 90.87% and 99.04%, respectively, and unmatched targets reduced end-to-end 20-pixel success to 85.58% and 93.75%. These values quantify agreement between predicted-mask-derived and ground-truth-mask-derived points and do not represent physical robotic positioning accuracy.

The practical value of the framework is as an interpretable perception and diagnostic layer. It exposes whether errors arise from segmentation boundaries, missed instances, false peduncle detections, cross-instance matching, or unstable geometric transformations. Future work should preserve machine-auditable group assignments and run-wise statistics, report confidence intervals and boundary-sensitive metrics, include unrestricted difficult cases and independent physical-point annotations, and perform RGB-D transformation and closed-loop harvesting tests before stronger deployment claims are made.

Author Contributions: Conceptualization, H.Z. and S.Z.; Methodology, H.Z.; Software, H.Z.; Validation, H.Z.; Formal analysis, H.Z.; Data curation, H.Z.; Writing—original draft, H.Z.; Writing–review and editing, H.Z.; Supervision, S.Z.; Project administration, X.F.; Funding acquisition, X.F. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Heilongjiang Province "Double First-Class" Discipline Collaborative Innovation Achievement Project, grant number LJGXCG2024-F18, and the Daqing "San Yan" Talent Support Program Project ("Xin Yan" Talent Team): "Daqing Precision Agriculture Technology Innovation Team".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original image dataset is publicly available as described in Ref. [30], subject to the source license. The fruit-cluster and peduncle polygon annotations, fixed split identifiers, model configurations, trained weights, evaluation scripts, environment specification, and reproduction instructions will be deposited in a permanent public repository upon acceptance. The repository DOI will be added when the archival record is minted.

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

References

1. Bac, C.W.; van Henten, E.J.; Hemming, J.; Edan, Y. Harvesting Robots for High-Value Crops: State-of-the-Art Review and Challenges Ahead. *J. Field Robot.* **2014**, *31*, 888–911. [CrossRef]
2. Zhao, Y.; Gong, L.; Huang, Y.; Liu, C. A Review of Key Techniques of Vision-Based Control for Harvesting Robot. *Comput. Electron. Agric.* **2016**, *127*, 311–323. [CrossRef]

3. Tang, Y.; Chen, M.; Wang, C.; Luo, L.; Li, J.; Lian, G.; Zou, X. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. *Front. Plant Sci.* **2020**, *11*, 510. [CrossRef]
4. Zhao, Y.S.; Gong, L.; Zhou, B.; Huang, Y.X.; Niu, Q.L.; Liu, C.L. Non-Color-Coded Target Recognition Algorithm for Tomato Harvesting Robots. *Trans. Chin. Soc. Agric. Mach.* **2016**, *47*, 1–7. (In Chinese with English abstract) [CrossRef]
5. Ni, J.P.; Zhu, L.C.; Dong, L.Z.; Cui, X.Z.; Han, Z.H.; Zhao, B. Real-Time Instance Segmentation Algorithm for Tomato Picking Robot Based on SwinS-YOLACT. *Trans. Chin. Soc. Agric. Mach.* **2024**, *55*, 18–30. (In Chinese with English abstract) [CrossRef]
6. Li, Y.; Feng, Q.; Liu, C.; Xiong, Z.; Sun, Y.; Xie, F.; Li, T.; Zhao, C. MTA-YOLACT: Multitask-Aware Network on Fruit Bunch Identification for Cherry Tomato Robotic Harvesting. *Eur. J. Agron.* **2023**, *146*, 126812. [CrossRef]
7. Fan, X.P.; Zhang, Y.Q.; Zhou, S.; Ren, M.F.; Wang, Y.W.; Chai, X.J. Recognition and Position Method of Tomato Picking Robot Based on Improved YOLOv8s and RGB-D Information Fusion. *Trans. Chin. Soc. Agric. Eng.* **2025**, *41*, 106–116. (In Chinese with English abstract) [CrossRef]
8. Jun, J.; Kim, J.; Seol, J.; Kim, J.; Son, H.I. Towards an Efficient Tomato Harvesting Robot: 3D Perception, Manipulation, and End-Effector. *IEEE Access* **2021**, *9*, 17631–17640. [CrossRef]
9. Lei, Z.L.; Liu, C.; Wang, Q. Recognition-Localization and Picking Methods of Facility Tomato Picking Robots. *Trans. Chin. Soc. Agric. Mach.* **2025**, *56*, 219–226. (In Chinese with English abstract) [CrossRef]
10. Zhao, B.; Liu, S.C.; Zhang, W.P.; Zhu, L.C.; Han, Z.H.; Feng, X.G.; Wang, R.X. Performance Optimization of Lightweight Transformer Architecture for Cherry Tomato Picking. *Trans. Chin. Soc. Agric. Mach.* **2024**, *55*, 62–71, 105. (In Chinese with English abstract) [CrossRef]
11. Li, X.X.; Chen, W.B.; Wang, Y.Q.; Yang, S.; Wu, H.R.; Zhao, C.J. Design and Experiment of an Automatic Cherry Tomato Harvesting System Based on Cascade Vision Detection. *Trans. Chin. Soc. Agric. Eng.* **2023**, *39*, 136–145. (In Chinese with English abstract) [CrossRef]
12. Yu, F.H.; Zhou, C.Q.; Yang, X.; Guo, Z.H.; Chen, C.L. Design and Experiment of a Tomato Harvesting Robot for Solar Greenhouses. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 41–49. (In Chinese with English abstract) [CrossRef]
13. Xie, X.X.; Kong, D.Y.; Pan, B.; Shan, J.H.; Xie, H.B. Design of a ROS-Based Tomato Harvesting Robot Control System for Plant Factories. *J. Agric. Mech. Res.* **2025**, *47*, 82–89. (In Chinese with English abstract) [CrossRef]
14. Ning, Z.T.; Luo, L.F.; Liao, J.X.; Wen, H.J.; Wei, H.L.; Lu, Q.H. Recognition and the Optimal Picking Point Location of Grape Stems Based on Deep Learning. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 222–229. (In Chinese with English abstract) [CrossRef]
15. Ma, Z.J.; Li, Z.Y. Table Grape Picking Point and Grasping Point Positioning Method Based on Machine Vision. *J. Agric. Mech. Res.* **2025**, *47*, 191–197. (In Chinese with English abstract)
16. Liang, X.F.; Jin, C.Q.; Ni, M.D.; Wang, Y.W. Acquisition and Experiment on Location Information of Picking Point of Tomato Fruit Clusters. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 163–169. (In Chinese with English abstract) [CrossRef]
17. Jin, Y.C.; Yu, C.C.; Yin, J.J.; Yang, S.X. Detection Method for Table Grape Ears and Stems Based on a Far–Close-Range Combined Vision System and Hand–Eye-Coordinated Picking Test. *Comput. Electron. Agric.* **2022**, *202*, 107364. [CrossRef]
18. Kalampokas, T.; Vrochidou, E.; Papakostas, G.A.; Pachidis, T.; Kaburlasos, V.G. Grape Stem Detection Using Regression Convolutional Neural Networks. *Comput. Electron. Agric.* **2021**, *186*, 106220. [CrossRef]
19. Wu, Z.; Xia, F.; Zhou, S.; Xu, D. A Method for Identifying Grape Stems Using Keypoints. *Comput. Electron. Agric.* **2023**, *209*, 107825. [CrossRef]
20. Du, W.; Jia, Z.; Sui, S.; Liu, P. Table Grape Inflorescence Detection and Clamping Point Localisation Based on Channel-Pruned YOLOV7-TP. *Biosyst. Eng.* **2023**, *235*, 100–115. [CrossRef]
21. Zhang, Q.; Pang, Y.S.; Li, B. Visual Localization and Harvesting-Pose Estimation of Tomato Clusters Based on Instance Segmentation. *Trans. Chin. Soc. Agric. Mach.* **2023**, *54*, 205–215. (In Chinese with English abstract) [CrossRef]
22. 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]
23. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and Segmentation of Overlapped Fruits Based on Optimized Mask R-CNN Application in Apple Harvesting Robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]
24. Jung, D.H.; Kim, C.Y.; Lee, T.S.; Park, S.H. Depth Image Conversion Model Based on CycleGAN for Growing Tomato Truss Identification. *Plant Methods* **2022**, *18*, 83. [CrossRef]
25. Liu, T.H.; Nie, X.N.; Wu, J.M.; Zhang, D.; Liu, W.; Cheng, Y.F.; Zheng, Y.; Qiu, J.; Qi, L. Pineapple (*Ananas comosus*) Fruit Detection and Localization in Natural Environment Based on Binocular Stereo Vision and Improved YOLOv3 Model. *Precis. Agric.* **2023**, *24*, 139–160. [CrossRef]
26. 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]
27. Arad, B.; Balendonck, J.; Barth, R.; Ben-Shahar, O.; Edan, Y.; Hellström, T.; Hemming, J.; Kurtser, P.; Ringdahl, O.; Tielen, T.; et al. Development of a Sweet Pepper Harvesting Robot. *J. Field Robot.* **2020**, *37*, 1027–1039. [CrossRef]

28. Long, J.H.; Zhao, C.J.; Lin, S.; Guo, W.Z.; Wen, C.W.; Zhang, Y. Segmentation of Tomato Fruits at Different Maturity Stages in Greenhouse Environments Using an Improved Mask R-CNN. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 100–108. (In Chinese with English abstract) [CrossRef]
29. Kang, H.; Chen, C. Fruit Detection and Segmentation for Apple Harvesting Using Visual Sensor in Orchards. *Sensors* **2019**, *19*, 4599. [CrossRef]
30. Song, G.Z.; Shi, Y.; Wang, J.; Jing, C.; Luo, G.F.; Sun, S.; Wang, X.L.; Li, Y.N. A Dataset of Greenhouse Cluster Tomato Fruit in the Tomato Town of Jinzhong National Agricultural High-Tech Industries Demonstration Zone in 2022. *China Sci. Data* **2023**, *8*, 1–5, (In Chinese with English metadata). [CrossRef]
31. 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. [CrossRef]
32. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2017; pp. 240–248. [CrossRef]
33. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ben Ayed, I. Boundary Loss for Highly Unbalanced Segmentation. *Med. Image Anal.* **2021**, *67*, 101851. [CrossRef]
34. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722. [CrossRef]
35. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object Detection with One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5562–5570. [CrossRef]
36. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
38. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755. [CrossRef]
40. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. [CrossRef]
41. 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. 9156–9165. [CrossRef]
42. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**, arXiv:2212.07784. [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.