Advancing Image Segmentation Techniques for Strawberry Detection in Vision-Based Agricultural Robotics

Faisal Imran *, Andrea Albarelli, Andrea Torsello, Andrea Gasparetto, Mara Pistellato

Posted Date: 21 January 2026

doi: 10.20944/preprints202601.1638.v1

Keywords: strawberry segmentation; deep learning; computer vision; precision agriculture; agricultural robotics; domain generalization; image segmentation

 doi:10.20944/preprints202601.1638.v1

Article

Advancing Image Segmentation Techniques for Strawberry Detection in Vision-Based Agricultural Robotics

Faisal Imran *, Andrea Albarelli, Andrea Torsello, Andrea Gasparetto, Mara Pistellato

Department of Computer Science, Ca' Foscari University of Venice, 30172 Venice, Italy
* Correspondence: faisal.imran@unive.it

Abstract

Image segmentation is a fundamental component of vision-based agricultural robotics, enabling accurate fruit localization, disease detection, and automated harvesting. However, real-world strawberry fields present significant challenges due to irregular fruit morphology, dense foliage occlusions, variable ripeness, and strong illumination variability. Moreover, segmentation models trained on a single dataset often fail to generalize across domains, limiting their practical deployment. This paper presents a comprehensive benchmark of classical computer vision methods, convolutional neural networks, instance-based models, and transformer-based architectures across three heterogeneous public strawberry datasets: Db1 (instance segmentation), Db2 (lesion segmentation), and Db3 (semantic segmentation). A unified preprocessing and evaluation framework is adopted to ensure fair comparison using standard metrics, including Intersection-over-Union (IoU), Dice coefficient, Precision, and Recall. Extensive in-domain experiments demonstrate that deep learning models significantly outperform classical approaches, with U-Net and SegFormer achieving IoU values above 0.95 on Db1 and up to 0.83 on Db3. Cross-domain zero-shot evaluations reveal a substantial generalization gap, with U-Net suffering IoU drops of up to 100%, while SegFormer consistently exhibits improved robustness and reduced cross-domain degradation across most transfer scenarios. To our knowledge, these results establish the first systematic multi-dataset benchmark for strawberry segmentation under domain shift, highlighting the importance of transformer-based architectures for robust agricultural perception and providing practical insights for real-world robotic deployment.

Keywords: strawberry segmentation; deep learning; computer vision; precision agriculture; agricultural robotics; domain generalization; image segmentation

1. Introduction

The integration of intelligent automation in agriculture is transforming core operations such as harvesting, grading, and disease detection. Central to this transformation is image segmentation, which enables vision-based systems to interpret complex agricultural scenes at the pixel level. Unlike object detection, which provides coarse bounding boxes, segmentation delivers precise object boundaries and shapes, supporting downstream tasks such as robotic harvesting, yield estimation, ripeness assessment, and disease localization [1]. For delicate crops such as strawberries, accurate segmentation is particularly critical to prevent fruit damage during automated handling and to ensure reliable growth and health monitoring.

Despite its importance, strawberry image segmentation remains a challenging problem due to several intrinsic factors. Strawberries exhibit irregular morphology, large intra-class variation in size and color, and frequent occlusions caused by foliage, stems, or adjacent fruits. Moreover, ripeness-related color transitions and dynamic illumination conditions introduce substantial appearance variability, especially in open-field environments [2–4]. These characteristics make strawberry segmentation

 doi:10.20944/preprints202601.1638.v1

significantly more difficult than segmentation tasks in controlled domains such as medical imaging or autonomous driving.

Early agricultural vision systems predominantly relied on classical image processing techniques, including thresholding, edge detection, and region-growing methods [5]. While computationally efficient, these approaches are highly sensitive to illumination changes, background clutter, and color similarity between fruits and surrounding vegetation. Thresholding methods often fail under non-uniform lighting or shadowed regions [6], while edge-based techniques such as the Canny operator struggle with the smooth and irregular boundaries of strawberries, particularly in the presence of noise or occlusion [7]. Region-growing approaches further suffer from over-segmentation and incorrect region merging when fruits overlap or exhibit heterogeneous texture patterns. As a result, classical pipelines provide limited robustness in real-world agricultural settings.

The advent of deep learning has marked a turning point in agricultural image segmentation. Convolutional Neural Networks (CNNs) and their extensions, including Fully Convolutional Networks (FCNs) [1], U-Net [8], and Mask R-CNN [9], enable end-to-end feature learning and significantly improve robustness under complex visual conditions. These architectures have been successfully applied to fruit localization, disease segmentation, and plant phenotyping [10–12]. However, their effectiveness remains strongly dependent on large, densely annotated datasets, which are costly and difficult to acquire in agricultural environments. Moreover, models trained on a single dataset often exhibit poor generalization when deployed in unseen environments, limiting their practical applicability in agricultural robotics.

More recently, transformer-based segmentation models have emerged as a promising alternative to purely convolutional architectures. By leveraging self-attention mechanisms and global context modeling, transformer-based approaches demonstrate increased robustness to background clutter and improved modeling of long-range dependencies [13,14]. While these models have achieved strong performance in general-purpose vision benchmarks, their advantages for agricultural image segmentation, particularly under domain shift and cross-dataset deployment, remain insufficiently explored.

To systematically investigate these challenges, we evaluate segmentation performance across three publicly available strawberry datasets:

- **Db1**: an instance segmentation dataset collected in greenhouse environments, focusing on individual strawberry fruit detection [15];
- **Db2**: a lesion segmentation dataset annotated for strawberry disease regions, characterized by small, low-contrast targets [16];
- **Db3**: a semantic segmentation dataset captured under open-field conditions with diverse illumination and complex background clutter [17];

Together, these datasets span instance, lesion, and semantic segmentation tasks, capturing a broad spectrum of real-world variability encountered in agricultural environments. Unlike most prior studies, which focus on a single dataset or segmentation task, this work explicitly investigates both *in-domain* performance and *cross-domain zero-shot generalization*. Cross-domain evaluation is essential for realistic deployment, as agricultural robots are often required to operate across farms, seasons, and acquisition conditions that differ substantially from those seen during training [18–20].

This study makes the following contributions:

- We present a systematic benchmark of classical image processing methods, convolutional neural networks, instance-based models, and transformer-based architectures across three diverse strawberry datasets.
- We evaluate segmentation performance under multiple task settings, including instance, lesion, and semantic segmentation, covering a wide range of environmental conditions.
- We conduct a comprehensive zero-shot cross-domain evaluation to quantify model robustness under domain shift.

 doi:10.20944/preprints202601.1638.v1

- We provide practical insights and recommendations to guide the design of robust vision-based systems for real-world agricultural robotics.

To the best of our knowledge, this work constitutes the first comprehensive multi-dataset benchmark for strawberry image segmentation that jointly evaluates classical methods, convolutional networks, and transformer-based architectures under both in-domain and cross-domain conditions. The presented analysis provides quantitative benchmarks and qualitative insights into current limitations while highlighting the potential of transformer-based models for robust agricultural perception.

2. Related Work

The rapid advancement of vision-based agricultural robotics has led to extensive research on fruit segmentation, disease detection, and automated harvesting systems. Unlike general-purpose image segmentation tasks, agricultural segmentation must operate under highly variable illumination, cluttered backgrounds, and frequent occlusions caused by foliage, soil, and neighboring plants. These characteristics introduce strong appearance variability and domain shift, making robustness and generalization central challenges for practical deployment. This section reviews prior work relevant to strawberry segmentation and closely related crop studies, with emphasis on segmentation paradigms, datasets, and generalization limitations.

2.1. Strawberry Segmentation

Strawberries have received increasing attention in computer vision research due to their high commercial value and the precision required for automated harvesting. Early strawberry segmentation studies primarily relied on convolutional encoder–decoder architectures trained on controlled datasets. Li et al. [2] proposed modified U-Net variants designed to handle irregular fruit morphology and partial occlusion, reporting strong segmentation accuracy in greenhouse environments. Similarly, Santos et al. [21] combined RGB-D sensing with Mask R-CNN to improve three-dimensional strawberry localization, demonstrating that depth information can partially alleviate occlusion in dense foliage.

In parallel, several studies have shifted focus from fruit localization to disease and lesion segmentation. Chen et al. [12] introduced a YOLO-based segmentation framework for strawberry leaf disease identification, achieving promising performance under natural backgrounds. Despite these advances, most strawberry-specific approaches are evaluated on a single dataset collected under limited environmental conditions, typically assuming that training and testing data follow the same distribution. As a result, conclusions regarding robustness and cross-domain applicability remain limited.

2.2. Segmentation in Other Crops

Insights from segmentation studies on other fruits and crops provide valuable context for strawberry segmentation. Chen et al. [22] applied U-Net++ to banana plant segmentation, demonstrating improved robustness under partial occlusion through dense skip connections. El Akrouchi et al. [23] proposed lightweight convolutional models for citrus fruit detection, highlighting the importance of computational efficiency for deployment on embedded agricultural platforms. Fu et al. [24] incorporated attention mechanisms into CNNs for fruit classification, illustrating how adaptive feature weighting can improve performance in complex agricultural scenes.

Beyond fruit crops, extensive work has been conducted on plant disease detection and lesion segmentation. Gandhi et al. [25] surveyed deep learning methods for plant disease detection, emphasizing challenges such as limited annotated data and strong class imbalance. These challenges are directly relevant to strawberry lesion segmentation, where disease regions are often small, irregular, and low contrast.

 doi:10.20944/preprints202601.1638.v1

2.3. Cross-Domain Generalization and Domain Shift

A major unresolved challenge in agricultural segmentation is domain shift, where models trained in one environment experience significant performance degradation when deployed in different farms, seasons, or acquisition setups. Xie et al. [18] investigated cross-domain fruit segmentation using transfer learning strategies, demonstrating that naïve fine-tuning often fails to achieve satisfactory generalization. More general domain adaptation techniques, including adversarial learning [26] and augmentation-driven adaptation [27], have shown promise in other vision domains but remain relatively underexplored in agricultural robotics.

In the context of strawberry segmentation, cross-dataset and zero-shot evaluations are rarely reported. Most existing studies implicitly assume that training and testing data originate from the same distribution, which substantially limits the relevance of reported results for real-world robotic systems. As agricultural robots are expected to operate across diverse environments with minimal retraining, systematic evaluation under domain shift is essential but largely missing from the current literature.

2.4. Lightweight, Transformer-Based, and Deployment-Oriented Models

For real-time agricultural robotics, segmentation models must balance accuracy with computational efficiency. Liu et al. [10] introduced a real-time strawberry detection framework optimized for complex farm environments. Lightweight backbones such as MobileNetV3 [28] have been adopted to reduce computational overhead, albeit sometimes at the cost of reduced boundary precision.

More recently, transformer-based architectures have gained attention due to their ability to model long-range dependencies and capture global context through self-attention mechanisms. Vision Transformers [13] and hybrid CNN–transformer models, such as Swin-UNet [29], have demonstrated strong performance in plant and fruit segmentation tasks. However, despite their potential advantages, transformer-based models are still rarely evaluated under cross-domain conditions in agricultural settings. Consequently, their robustness and generalization properties relative to convolutional architectures remain insufficiently understood.

2.5. Summary of Literature Gaps

Table 1 summarizes representative studies spanning strawberry-specific segmentation, cross-crop applications, and recent transformer-based approaches. While these works demonstrate substantial progress, several critical gaps remain. First, the majority of existing studies rely on single-dataset evaluations conducted under limited and often controlled conditions, providing little insight into cross-domain generalization. Second, direct comparative analyses across classical image processing methods, convolutional neural networks, instance-based models, and transformer-based architectures within a unified framework are scarce. Third, zero-shot transfer performance across heterogeneous agricultural datasets is rarely quantified, despite its importance for real-world robotic deployment. Finally, existing studies typically focus on a single segmentation task, with no unified benchmark jointly covering instance, lesion, and semantic segmentation under consistent experimental protocols.

These limitations motivate the present study, which establishes a comprehensive multi-dataset benchmark for strawberry image segmentation. By systematically evaluating classical computer vision methods, convolutional neural networks, instance-based models, and transformer-based architectures under both in-domain and zero-shot cross-domain conditions, this work provides a unified, deployment-oriented perspective on robustness, generalization, and practical applicability in real-world agricultural vision systems.

 doi:10.20944/preprints202601.1638.v1

Table 1. Representative works on strawberry and fruit segmentation. Task: I = instance, L = lesion, S = semantic. Model: CV = classical vision, CNN = convolutional neural network, Tr = transformer. CD indicates cross-domain evaluation.

Year [Ref]	Dataset / Crop	Task	Model	CD
1979 [6]	Strawberry / fruit (controlled)	S	CV	–
2016 [20]	Orchard fruits	I	CV	–
2017 [19]	Orchard fruits	I	CNN	–
2020 [22]	Banana orchard	S	CNN	–
2020 [15]	Greenhouse strawberries	I	CNN	–
2021 [21]	RGB-D strawberries	I	CNN	–
2021 [16]	Strawberry diseases	L	CNN	–
2022 [10]	Open-field strawberries	I	CNN	–
2022 [24]	Fruit-360	C	CNN	–
2022 [25]	Multi-crop diseases	L	CNN	–
2023 [18]	Multi-fruit	S	CNN	Yes
2023 [11]	Greenhouse strawberries	I	CNN	–
2023 [29]	Leaves & fruits	S	Tr	–
2019 [28]	Mobile vision	B	CNN	–
2024 [2]	Strawberries	I, S	CNN	–
2024 [30]	Real-field fruits	S	CV + CNN	–
2024 [31]	Crop diseases	L	Tr	–
2024 [32]	Crop fields	S	Tr	–
2024 [33]	Multi-fruit	S	CNN	Yes
2025 [23]	Citrus orchard	I	CNN	–
2025 [12]	Strawberry leaves	L	CNN	–
2025 [14]	Agricultural datasets	S	Tr	–
2025 (ours)	**Db1, Db2, Db3 (strawberries)**	**I, L, S**	**CV + CNN + Tr**	**Yes (zero-shot)**

As summarized in Table 1, existing works predominantly address a single segmentation task and dataset, with very limited evaluation under cross-domain conditions.

3. Materials and Methods

To ensure a systematic, fair, and reproducible evaluation, this study adopts a unified benchmarking framework for strawberry image segmentation, illustrated in Figure 1. The framework integrates three heterogeneous public datasets (Db1–Db3), each corresponding to a distinct segmentation task and acquisition environment. A unified preprocessing and augmentation pipeline is applied across all datasets to minimize confounding factors and ensure comparability among segmentation methods. Both classical computer vision techniques and modern deep learning models are evaluated under task-appropriate configurations.

Models were selected as representative exemplars of four segmentation paradigms: classical image processing, convolutional encoder–decoder networks, instance-aware architectures, and transformer-based models. Each model was applied either across all datasets or within its appropriate task domain, in accordance with its architectural assumptions and supervision requirements, to ensure fair and interpretable comparison.

 doi:10.20944/preprints202601.1638.v1

The experimental design explicitly addresses three segmentation paradigms commonly encountered in agricultural robotics: (i) *instance segmentation* (Db1), where individual strawberry fruits must be detected and segmented as separate objects; (ii) *lesion segmentation* (Db2), which requires precise localization of small, low-contrast disease regions; and (iii) *semantic segmentation* (Db3), where all strawberry pixels are assigned to a single foreground class. These tasks differ substantially in annotation granularity, visual complexity, and deployment requirements. Accordingly, instance-aware architectures (Mask R-CNN) are used for Db1, while pixel-wise encoder–decoder architectures are employed for lesion and semantic segmentation tasks.

Model performance is evaluated using standard segmentation metrics, including Intersection-over-Union (IoU), Dice Similarity Coefficient (DSC), Precision, Recall, and Pixel Accuracy. Both quantitative metrics and qualitative visualizations are reported under *in-domain* and *cross-domain* settings to assess robustness, generalization, and practical deployment feasibility in agricultural robotics.

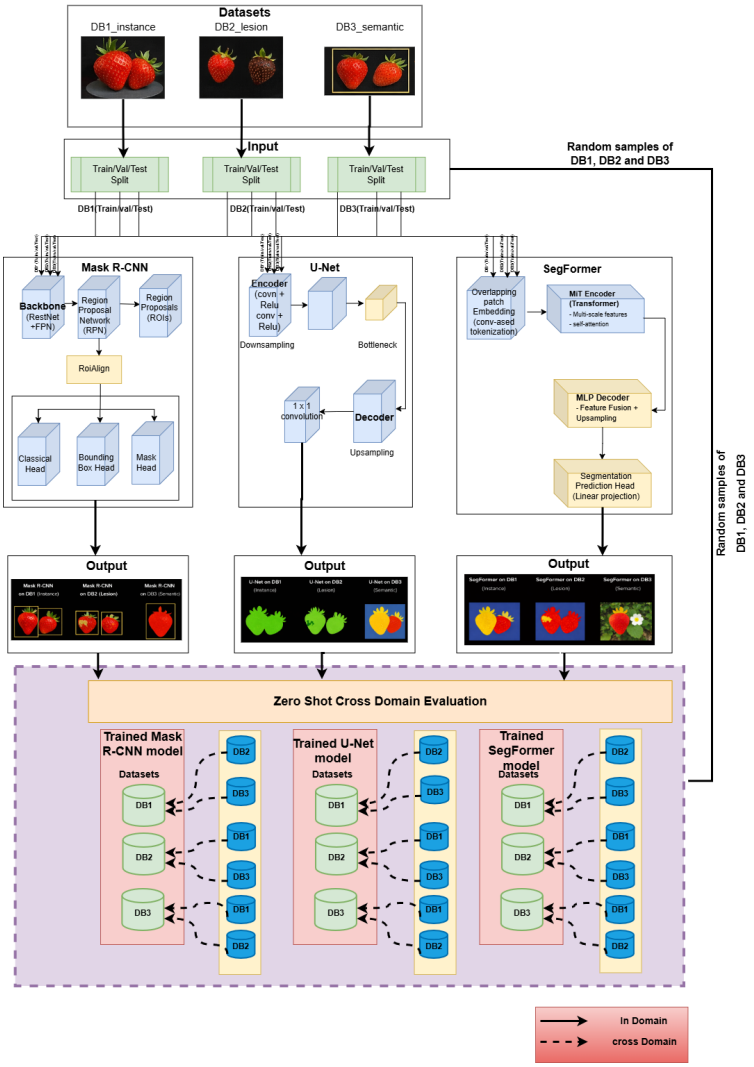

Figure 1. Unified experimental framework for multi-dataset strawberry image segmentation. Models are trained independently on DB1 (instance), DB2 (lesion), and DB3 (semantic) datasets and evaluated under both in-domain and zero-shot cross-domain conditions without fine-tuning. The framework highlights the relationship between segmentation paradigms, datasets, and cross-domain generalization performance.

3.1. Datasets

Three publicly available strawberry datasets were used in this study, each corresponding to a distinct segmentation task and acquisition environment. The datasets were deliberately selected to

 doi:10.20944/preprints202601.1638.v1

cover heterogeneous visual conditions, annotation granularities, and segmentation objectives, thereby enabling a comprehensive evaluation of model performance and robustness.

Db1: StrawDI (Instance Segmentation)

Db1 corresponds to the StrawDI dataset introduced by Pérez et al. [15]. It contains approximately 3,100 high-resolution RGB images acquired under greenhouse conditions. The dataset provides pixel-level instance annotations for individual strawberry fruits, making it suitable for evaluating instance segmentation models in controlled environments characterized by frequent fruit overlap and occlusions caused by foliage.

Db2: Strawberry Disease Dataset (Lesion Segmentation)

Db2 was obtained from a publicly available strawberry disease dataset released by Jeonbuk National University and first introduced for strawberry disease lesion segmentation by Afzaal et al. [16]. The dataset contains approximately 2,500 RGB images annotated with pixel-wise lesion masks corresponding to seven disease categories. In this study, Db2 is used for lesion-level segmentation experiments, which are particularly challenging due to the small size of diseased regions, irregular lesion shapes, and low contrast between infected and healthy tissue.

Db3: Strawberry-DS (Semantic Segmentation)

Db3 corresponds to a strawberry dataset collected by the Agricultural Research Center in Cairo, Egypt, and first used for semantic segmentation in [17]. The dataset consists of 247 RGB images captured under open-field conditions. Pixel-level binary annotations distinguish strawberry regions from background, reflecting real-world challenges such as strong illumination variability, background clutter, and complex field scenes.

Table 2. Overview of the evaluated strawberry segmentation datasets and their primary challenges.

Dataset	Task Type	Annotation Format	Key Challenges
Db1	Instance segmentation	Per-object binary masks	Occlusion, overlapping fruits
Db2	Lesion segmentation	Pixel-wise lesion masks	Low contrast, irregular lesions
Db3	Semantic segmentation	Pixel-wise binary masks	Illumination variability, background clutter

To illustrate the dataset diversity, Figure 2 presents representative images and corresponding ground truth masks from each dataset. These examples highlight key differences in acquisition conditions, visual complexity, and annotation granularity across the three segmentation tasks.

 doi:10.20944/preprints202601.1638.v1

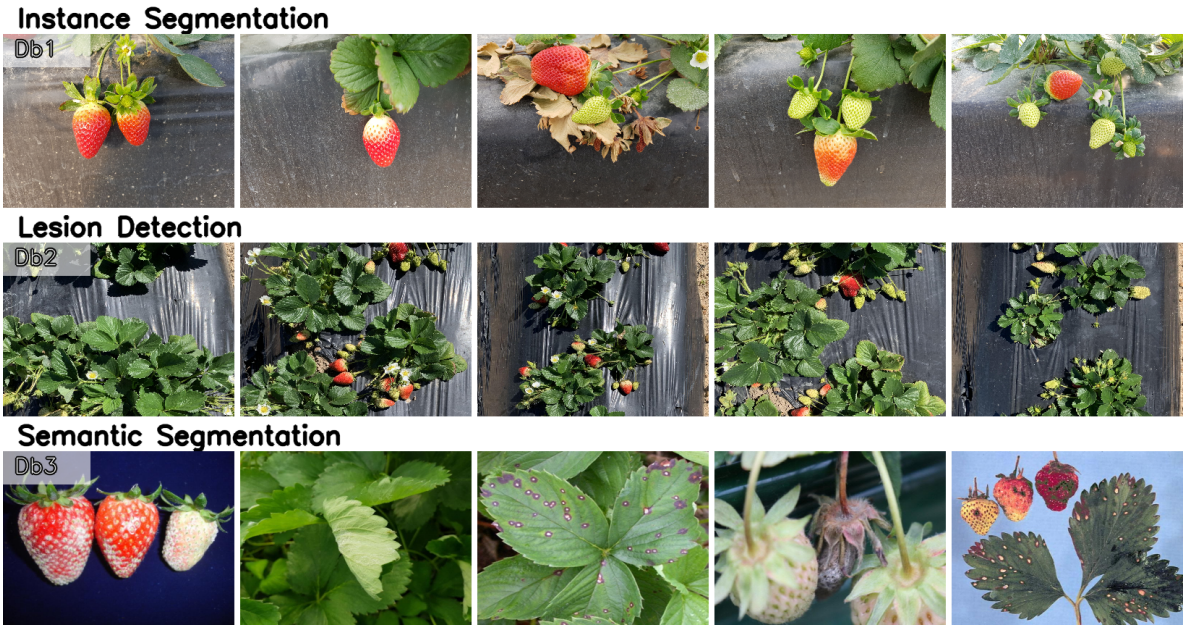

Figure 2. Representative examples from the three strawberry datasets: (a) Db1 – greenhouse instance segmentation, (b) Db2 – disease lesion segmentation, (c) Db3 – open-field semantic segmentation. Each example shows the original image and its corresponding ground truth mask.

3.2. Preprocessing and Data Augmentation

Each dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets. When metadata were available, splits were stratified to reduce potential domain leakage. All reported results correspond to the mean ± standard deviation over three independent runs using different random seeds.

A unified preprocessing pipeline was applied across all datasets to ensure consistency:

- **Resizing:** All images and corresponding masks were resized to 256×256 pixels.
- **Normalization:** RGB intensities were scaled to the $[0, 1]$ range.
- **Data augmentation (training only):** Random horizontal and vertical flips, rotations ($\pm 15^\circ$), brightness and contrast adjustments ($\pm 20\%$), and random zoom–cropping were applied to improve robustness and reduce overfitting.
- **Mask handling:** For Db1, individual instance masks were preserved for instance-aware training with Mask R-CNN. For all pixel-wise segmentation experiments, instance masks were merged into a single foreground mask. Db2 and Db3 masks were used without modification.

This unified strategy ensures comparable input distributions while preserving task-specific annotation structure when required.

3.3. Segmentation Methods

We benchmark both classical computer vision baselines and modern deep learning architectures under identical preprocessing and evaluation protocols. Classical method parameters were selected based on preliminary validation experiments and commonly adopted values in the agricultural vision literature, and were kept fixed across datasets to ensure comparability.

3.3.1. Classical Computer Vision Baselines

Three traditional segmentation pipelines were implemented to establish non-learning baselines:

Thresholding. Global Otsu thresholding [6] and adaptive thresholding were applied to grayscale images, followed by morphological opening, closing, and hole filling.

Canny Edge Detection with Morphology. Edges were extracted using the Canny operator [34], followed by dilation and region filling to produce closed foreground regions.

 doi:10.20944/preprints202601.1638.v1

Region Growing. Seeded region growing was performed in HSV color space using high-confidence red chroma seed points, with adaptive similarity criteria controlling region expansion [35].

These pipelines provide lightweight baselines for assessing dataset difficulty independently of learned representations.

3.3.2. Deep Learning Models

U-Net. For lesion and semantic segmentation (Db2 and Db3), we employ the standard U-Net architecture [8], featuring a symmetric encoder–decoder structure with skip connections and sigmoid output activation.

Mask R-CNN. For instance segmentation (Db1), we use Mask R-CNN with a ResNet-50 backbone and Feature Pyramid Network (FPN) [36]. The model jointly optimizes object classification, bounding-box regression, and pixel-level mask prediction.

SegFormer. To evaluate transformer-based segmentation, we adopt SegFormer [37], which employs hierarchical self-attention in the encoder and a lightweight multilayer perceptron decoder. This design enables efficient high-resolution prediction while capturing global contextual information.

3.4. Mathematical Formulation of the Segmentation Problem

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a dataset of RGB images $x_i \in \mathbb{R}^{H\times W\times 3}$ and their corresponding ground-truth segmentation masks y_i. Depending on the task, y_i represents either per-instance binary masks (Db1) or pixel-wise binary annotations (Db2 and Db3). The objective is to learn a parameterized function f_θ that maps an input image to a dense pixel-level prediction.

The segmentation network produces an output

$$f_\theta(x_i) = \hat{y}_i \in [0,1]^{H\times W\times C}, \tag{1}$$

where $C = 1$ for binary segmentation tasks. For instance segmentation experiments, individual object masks are preserved during training; however, for unified pixel-wise evaluation, instance masks are merged into a single foreground mask.

3.4.1. Loss Function

To effectively handle class imbalance and boundary ambiguity commonly observed in agricultural imagery, we employ a composite loss function combining Binary Cross-Entropy (BCE) and Dice loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{HW}\sum_{j=1}^{HW}\left[y_j \log(p_j) + (1-y_j)\log(1-p_j)\right], \tag{2}$$

where p_j denotes the predicted probability at pixel j.

The Dice loss is defined as

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_j p_j y_j + \epsilon}{\sum_j p_j^2 + \sum_j y_j^2 + \epsilon}, \tag{3}$$

where ϵ is a small constant added for numerical stability.

The final training objective is given by

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{Dice}}, \tag{4}$$

where λ_1 and λ_2 control the relative contribution of region-level accuracy and boundary alignment.

 doi:10.20944/preprints202601.1638.v1

3.4.2. Transformer-Based Context Modeling

For transformer-based architectures, contextual relationships among image features are modeled using self-attention. Given query (Q), key (K), and value (V) projections, the attention mechanism is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{5}$$

where d_k denotes the key dimensionality. This formulation enables global context aggregation, which improves robustness under strong illumination variability and background clutter commonly encountered in open-field agricultural environments.

3.5. Cross-Domain Evaluation Protocol

To assess robustness under domain shift, models trained on one dataset were evaluated on unseen target datasets without any fine-tuning. All pairwise source–target combinations among Db1, Db2, and Db3 were evaluated, in addition to standard in-domain testing. This zero-shot cross-domain protocol quantifies performance degradation relative to in-domain baselines and reflects realistic deployment scenarios in agricultural robotics.

3.6. Training Objectives and Optimization

U-Net and SegFormer models were trained using a composite loss function combining Binary Cross-Entropy and Dice loss to balance region accuracy and boundary precision. Mask R-CNN was trained using its standard multi-task loss formulation. All models were optimized using the Adam optimizer [38] with early stopping based on validation performance.

Table 3. Training and optimization parameters used for all segmentation models.

Parameter	Value
Framework / Hardware	PyTorch, NVIDIA A100 GPU (Google Colab)
Image size	256×256
Batch size	16
Optimizer	Adam
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Epochs	50 (early stopping, patience = 7)
Dataset split	70/15/15
Repetitions	3 runs (mean $\pm$ sd)

3.7. Evaluation Metrics

Segmentation performance was assessed using Intersection-over-Union (IoU), Dice Similarity Coefficient, Precision, Recall, and Pixel Accuracy. Statistical significance was evaluated using the Wilcoxon signed-rank test, with confidence intervals estimated via bootstrapping. Cross-domain robustness was quantified by measuring relative performance degradation with respect to in-domain results.

4. Experiments and Results

This section presents a comprehensive experimental evaluation of classical computer vision pipelines and deep learning–based segmentation models across three heterogeneous strawberry datasets: Db1 (instance segmentation), Db2 (lesion segmentation), and Db3 (semantic segmentation). All experiments were conducted under identical preprocessing, training, and evaluation protocols to ensure fair comparison across methods. Each deep learning experiment was repeated three times using different random seeds, and results are reported as mean $\pm$ standard deviation to reflect performance stability.

 doi:10.20944/preprints202601.1638.v1

The evaluation is organized as follows. First, classical segmentation baselines are analyzed to establish lower-bound performance and characterize dataset difficulty. Second, in-domain performance of deep learning models is evaluated to assess their capacity under matched training and testing distributions. Finally, cross-domain zero-shot generalization is examined to quantify robustness under domain shift and assess deployment feasibility in real-world agricultural scenarios.

4.1. Classical Computer Vision Baselines

We evaluated three widely used traditional segmentation pipelines—thresholding, Canny edge detection with morphology, and region growing—implemented using OpenCV. These methods do not involve learning and therefore provide insight into the intrinsic visual complexity of each dataset, independent of data-driven feature representations.

4.1.1. Thresholding

Global Otsu and adaptive Gaussian thresholding were applied to grayscale and HSV representations, followed by morphological operations to suppress noise and fill interior regions. Quantitative results are summarized in Table 4.

Table 4. Thresholding-based segmentation performance (mean ± sd).

Dataset	IoU	Dice	Pixel Accuracy
Db1	0.137 ± 0.025	0.227 ± 0.040	0.672
Db2	0.021 ± 0.012	0.040 ± 0.024	0.528
Db3	0.258 ± 0.064	0.365 ± 0.064	0.471

Observation. Thresholding-based methods exhibit consistently poor performance across all datasets, with near-complete failure on lesion segmentation (Db2) and unstable results under open-field conditions (Db3). These failures stem from strong illumination variability, low contrast between foreground and background regions, and the presence of visually similar non-target objects. The results confirm that simple color- or intensity-based decision rules are insufficient for robust strawberry segmentation in realistic agricultural environments.

4.1.2. Canny Edge Detection with Morphology

Canny edge detection was applied using fixed thresholds followed by dilation and morphological closing to generate contiguous regions. **Observation.** Edge-based segmentation improves boundary localization relative to thresholding but frequently fails to produce closed and semantically meaningful regions, particularly under occlusion or low-contrast conditions. Region growing achieves comparatively better performance by exploiting color continuity; however, it often merges adjacent fruits and produces false positives in cluttered scenes. Overall, classical pipelines demonstrate limited robustness and primarily serve as lower-bound references for learning-based models.

4.1.3. Region Growing

Region growing was performed in HSV color space using high-confidence red chroma seed points and adaptive similarity thresholds.
Observation. Region growing consistently outperformed other classical baselines but frequently merged adjacent fruits and produced false positives under cluttered field conditions.

Overall, classical pipelines demonstrated limited robustness and serve primarily as lower-bound references for learned models.

4.2. In-Domain Deep Learning Results

We evaluated three deep learning architectures representing distinct segmentation paradigms: Mask R-CNN for instance segmentation (Db1), U-Net for pixel-wise segmentation (Db2 and Db3), and SegFormer as a transformer-based model applicable across all datasets. Quantitative in-domain

 doi:10.20944/preprints202601.1638.v1

results are reported in Table 5.Mask R-CNN is evaluated using pixel-wise segmentation metrics in all experiments. This evaluation protocol penalizes instance-based predictions and explains the observed zero IoU scores when Mask R-CNN is applied to non-instance segmentation tasks.

Table 5. In-domain segmentation performance across datasets (mean ± sd).

Dataset	Model	IoU	Dice	Precision	Recall
Db1	Classical	0.137 ± 0.025	0.227 ± 0.040	0.152	0.672
Db1	U-Net	0.955 ± 0.031	0.977 ± 0.017	0.979	0.975
Db1	SegFormer	0.949 ± 0.033	0.974 ± 0.018	0.976	0.971
Db1	Mask R-CNN	0.000 ± 0.000	0.000 ± 0.000	1.000	0.000
Db2	Classical	0.021 ± 0.012	0.040 ± 0.024	0.021	0.528
Db2	U-Net	0.657 ± 0.104	0.789 ± 0.077	–	–
Db2	SegFormer	0.766 ± 0.100	0.865 ± 0.063	0.899	0.840
Db2	Mask R-CNN	0.000 ± 0.000	0.000 ± 0.000	1.000	0.000
Db3	Classical	0.258 ± 0.064	0.365 ± 0.064	0.440	0.471
Db3	U-Net	0.516 ± 0.270	0.635 ± 0.263	–	–
Db3	SegFormer	0.832 ± 0.158	0.898 ± 0.122	0.910	0.912
Db3	Mask R-CNN	0.462 ± 0.315	0.559 ± 0.345	0.698	0.716

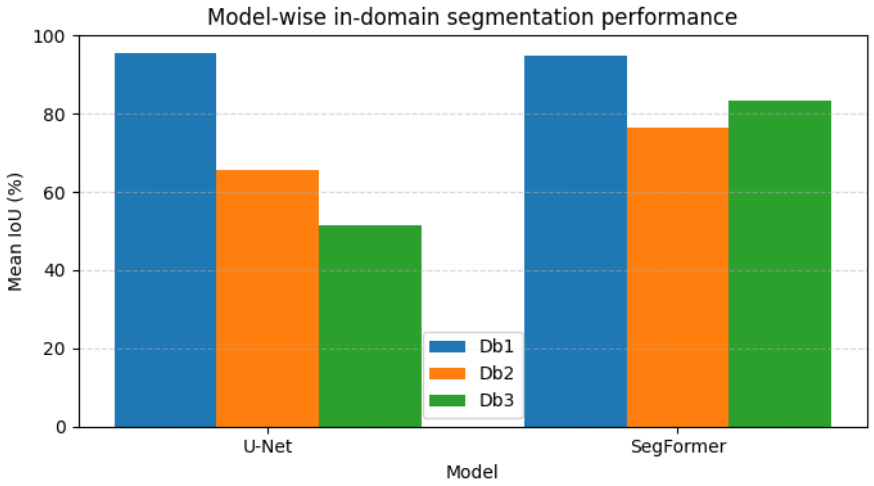

Figure 3. Model -wise in-domain segmentation performance across datasets. SegFormer consistently outperforms U-Net on visually challenging datasets (Db2 and Db3).

 doi:10.20944/preprints202601.1638.v1

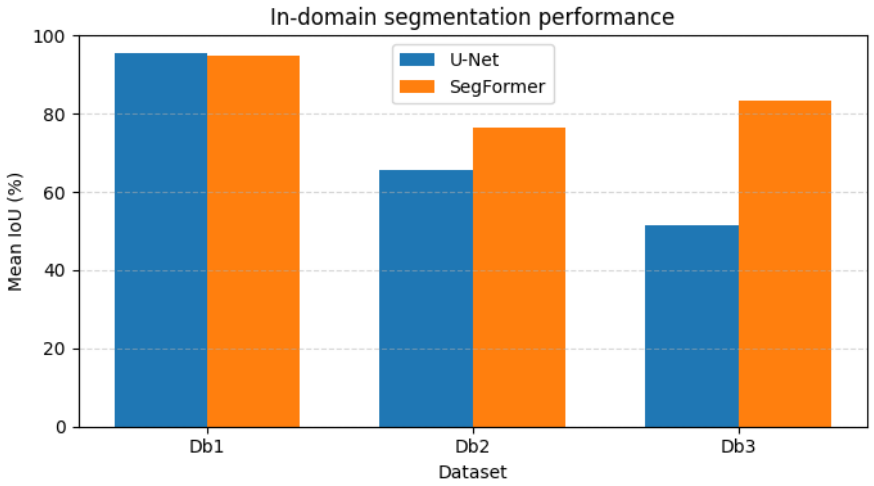

Figure 4. In-domain segmentation performance across datasets. Mean IoU for U-Net and SegFormer under matched training and testing conditions.

In-domain analysis. Under matched training and testing conditions, deep learning models substantially outperform classical baselines across all datasets. On Db1, both U-Net and SegFormer achieve near-perfect segmentation performance, indicating that fruit instances are visually distinguishable under controlled greenhouse conditions. Mask R-CNN, while designed for instance-level tasks, exhibits degraded performance in this setting due to the mismatch between bounding-box–based supervision and pixel-wise evaluation.

For lesion segmentation (Db2), SegFormer significantly outperforms U-Net in terms of IoU and Dice scores. This improvement is attributed to the transformer's ability to capture global context and long-range dependencies, which is critical for detecting small, irregular, and low-contrast lesion regions. A similar trend is observed on Db3, where SegFormer achieves substantially higher performance than U-Net under complex open-field conditions characterized by strong illumination variation and background clutter.

Overall, these results demonstrate that while convolutional architectures remain effective under favorable imaging conditions, transformer-based representations provide superior robustness for visually challenging agricultural segmentation tasks.

4.3. Cross-Domain Generalization

To evaluate robustness under domain shift, we performed zero-shot cross-domain experiments in which models trained on a source dataset were directly evaluated on unseen target datasets without any fine-tuning. This protocol reflects realistic deployment scenarios in precision agriculture, where annotated data from new environments are often unavailable or costly to obtain.

Table 6 reports cross-domain IoU values along with the relative performance degradation compared to in-domain baselines. Across all transfer directions, both U-Net and SegFormer experience substantial performance drops, highlighting strong dataset bias induced by differences in acquisition conditions, annotation granularity, and segmentation objectives.

 doi:10.20944/preprints202601.1638.v1

Table 6. Cross-domain generalization results (IoU and relative drop).

Model	Source → Target	IoU	IoU Drop (%)
U-Net	Db1 → Db2	0.454	52.5
U-Net	Db1 → Db3	0.036	96.3
U-Net	Db2 → Db1	0.400	39.1
U-Net	Db2 → Db3	0.035	94.7
U-Net	Db3 → Db1	$< 10^{-10}$	100.0
U-Net	Db3 → Db2	$< 10^{-10}$	100.0
SegFormer	Db1 → Db2	0.561	40.9
SegFormer	Db1 → Db3	0.042	95.6
SegFormer	Db2 → Db1	0.459	40.1
SegFormer	Db2 → Db3	0.016	97.9
SegFormer	Db3 → Db1	0.241	71.0
SegFormer	Db3 → Db2	0.037	95.6

Figures 5 and 6 provide a visual comparison of cross-domain robustness. As shown in Figure 5, U-Net exhibits severe degradation across most transfer scenarios, including near-complete failure when transferring between datasets with differing segmentation objectives. In contrast, Figure 6 shows that SegFormer consistently demonstrates lower relative IoU degradation across most dataset pairs, particularly for Db1→Db2 and Db3→Db1 transfers.

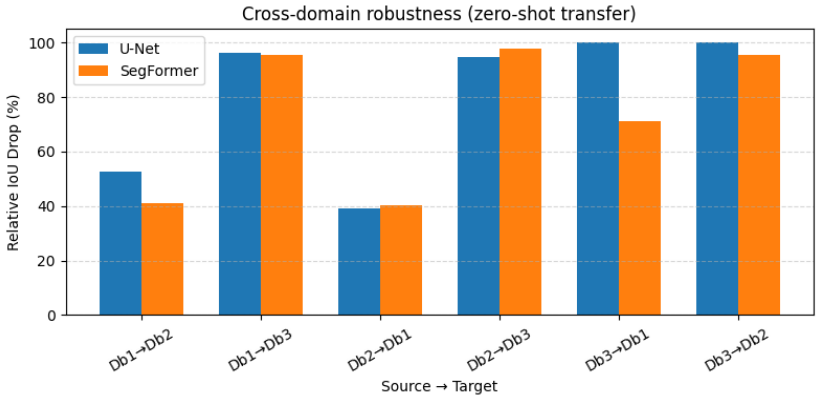

Figure 5. Relative IoU degradation across cross-domain transfer pairs under zero-shot evaluation. Large performance drops indicate strong dataset bias across acquisition conditions and annotation schemes.

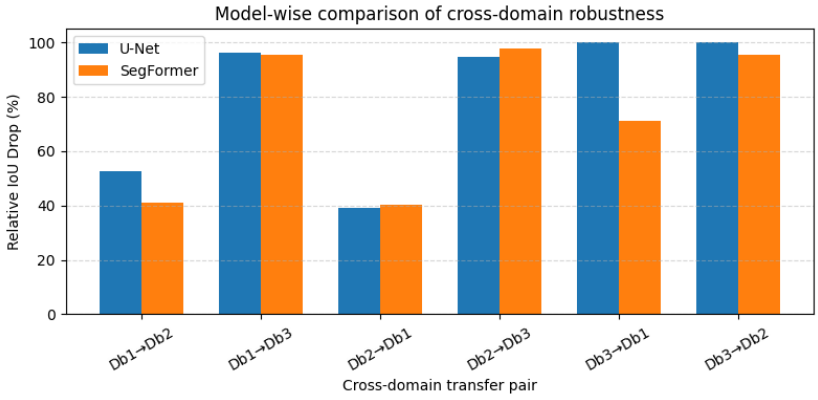

Figure 6. Model-wise comparison of cross-domain robustness under zero-shot transfer. SegFormer consistently exhibits lower relative IoU degradation than U-Net across most transfer directions.

These results suggest that transformer-based architectures capture more transferable representations by leveraging global context and self-attention mechanisms. Nevertheless, the substantial

 doi:10.20944/preprints202601.1638.v1

performance losses observed under extreme domain shifts indicate that zero-shot transfer alone remains insufficient for reliable deployment. This highlights the need for complementary strategies such as domain adaptation, multi-source training, or self-supervised pretraining to improve generalization across heterogeneous agricultural datasets.

4.4. Summary of Findings

Across all experiments, deep learning models substantially outperform classical computer vision pipelines under in-domain conditions. Among learning-based approaches, SegFormer consistently achieves the best balance between accuracy and robustness, particularly for lesion and semantic segmentation tasks in visually complex environments. Nevertheless, cross-domain evaluations reveal severe performance degradation for all models, highlighting a critical gap between benchmark performance and real-world deployment requirements. These findings emphasize the importance of cross-domain evaluation in agricultural vision research and motivate future work on domain adaptation, self-supervised learning, and annotation-efficient strategies for robust agricultural robotics.

5. Discussion

This study provides a comprehensive and systematic evaluation of strawberry image segmentation methods across multiple datasets, segmentation tasks, and deployment scenarios. By jointly benchmarking classical computer vision techniques, convolutional neural networks, instance-based models, and transformer-based architectures under both in-domain and zero-shot cross-domain conditions, the results offer important insights into the strengths, limitations, and practical deployment readiness of current segmentation approaches in agricultural robotics.

5.1. Effectiveness of Learning-Based Segmentation

Across all datasets, learning-based methods substantially outperform classical computer vision pipelines under in-domain conditions. Classical approaches based on thresholding, edge detection, and region growing fail to achieve reliable performance in the presence of illumination variability, occlusion, and background clutter, confirming their limited applicability in realistic agricultural environments. These findings are consistent with prior observations in fruit and plant segmentation literature, which report strong sensitivity of handcrafted pipelines to environmental variability.

Among deep learning models, convolutional architectures such as U-Net demonstrate strong performance when training and testing data originate from the same distribution. In particular, near-perfect segmentation accuracy is achieved on Db1, reflecting the relatively controlled greenhouse conditions and well-defined fruit appearance. However, performance degrades notably on more challenging datasets, especially for lesion segmentation (Db2) and open-field semantic segmentation (Db3), where visual ambiguity and small target regions are prevalent.

5.2. Advantages of Transformer-Based Architectures

Transformer-based SegFormer consistently achieves superior performance compared to convolutional architectures on visually complex datasets. The observed improvements on lesion and field segmentation tasks suggest that global context modeling and long-range dependency capture play a critical role in agricultural image segmentation, where relevant visual cues may be spatially dispersed and local texture alone is insufficient.

Furthermore, SegFormer demonstrates improved robustness under cross-domain evaluation. Although performance degradation remains substantial under severe domain shift, the relative IoU drop is consistently lower than that of U-Net across most transfer scenarios. This behavior indicates that transformer-based representations generalize more effectively across changes in illumination, background composition, and annotation style. These findings align with recent evidence from general-purpose vision tasks, while providing empirical validation in the agricultural domain.

 doi:10.20944/preprints202601.1638.v1

5.3. Limitations of Instance-Based Models

Mask R-CNN exhibits strong performance only under its intended instance segmentation setting and fails to generalize across lesion and semantic segmentation tasks. This limitation arises from both architectural and supervision mismatches. Instance-based models rely on region proposals and bounding-box supervision, which become unstable in cluttered scenes or when target regions are small and low contrast. Additionally, the discrepancy between instance-level predictions and pixel-level evaluation metrics contributes to unreliable performance outside the intended task domain.

These observations suggest that instance-based architectures are unsuitable as general-purpose segmentation solutions for heterogeneous agricultural datasets. Their use should be restricted to scenarios where instance-level annotations and deployment requirements closely match training conditions.

5.4. Impact of Domain Shift and Deployment Implications

Cross-domain evaluation reveals a pronounced generalization gap for all evaluated models. When trained on a single dataset and deployed in unseen environments, segmentation performance often degrades dramatically, with convolutional models experiencing near-complete failure in extreme transfer scenarios. This finding highlights a critical limitation of current evaluation practices in agricultural vision research, which predominantly report in-domain performance and therefore overestimate deployment readiness.

The results emphasize that segmentation robustness in agriculture is fundamentally a domain generalization problem rather than a purely architectural one. Differences in acquisition conditions, annotation granularity, crop appearance, and environmental context all contribute to performance degradation. While transformer-based architectures partially mitigate these effects, they do not fully resolve the challenge, underscoring the need for explicit strategies to address domain shift.

5.5. Interpretation of Performance Metrics and Task Difficulty

It is important to contextualize the reported performance metrics with respect to the underlying learning objective. Unlike image-level classification, which optimizes a global prediction function $f_\theta : \mathbb{R}^{H\times W\times 3} \to \Delta^C$, pixel-wise segmentation minimizes a dense structured loss defined over all spatial locations:

$$\mathcal{L}(\theta) = \frac{1}{HW} \sum_{j=1}^{HW} \ell\big(f_\theta(x)_j, y_j\big), \tag{6}$$

where each pixel contributes independently to the optimization objective. As a consequence, segmentation metrics such as Intersection-over-Union (IoU) and Dice coefficient are significantly more sensitive to spatial misalignment, boundary errors, and small target regions than classification accuracy.

In particular, small localization errors along object boundaries or partial omissions of lesion regions can lead to disproportionate metric degradation, even when the model successfully captures the global semantic content of the scene. This effect is amplified under domain shift, where variations in illumination, scale, and background appearance alter pixel-level distributions while preserving high-level semantic structure. Therefore, lower numerical scores observed under cross-domain evaluation should not be interpreted as inferior feature learning, but rather as a reflection of the stricter optimization objective imposed by dense prediction tasks. From a deployment perspective, these metrics provide a more realistic assessment of operational reliability for robotic erception systems, where precise spatial localization is critical for downstream actions such as grasp planning and collision avoidance.

5.6. Future Directions

The findings of this study point toward several promising research directions. First, domain adaptation and multi-source training strategies should be explored to reduce performance degradation under cross-domain deployment. Second, self-supervised and semi-supervised learning approaches

 doi:10.20944/preprints202601.1638.v1

may help leverage large amounts of unlabeled agricultural imagery, reducing reliance on costly manual annotation. Third, hybrid CNN–Transformer architectures offer a promising balance between computational efficiency and robustness, particularly for deployment on resource-constrained robotic platforms.

Finally, future benchmarks should incorporate standardized cross-domain evaluation protocols to more accurately reflect real-world deployment conditions. By shifting the focus from in-domain accuracy to robustness and generalization, the agricultural robotics community can accelerate the development of scalable and deployment-ready vision systems.

5.7. Summary

Overall, this study demonstrates that while modern deep learning models achieve impressive segmentation accuracy under controlled conditions, their performance remains fragile under domain shift. Transformer-based architectures offer meaningful improvements in robustness, but significant challenges persist. By establishing a unified multi-dataset benchmark and explicitly evaluating zero-shot cross-domain performance, this work provides both quantitative evidence and practical guidance for the development of robust vision-based systems in precision agriculture.

6. Conclusion

This work presented a comprehensive multi-dataset benchmark for strawberry image segmentation, covering instance, lesion, and semantic segmentation tasks under a unified experimental framework. By systematically evaluating classical computer vision pipelines, convolutional neural networks, instance-based models, and transformer-based architectures, we provided a clear and reproducible comparison of segmentation performance across diverse agricultural environments.

Experimental results demonstrated that deep learning approaches substantially outperform classical methods under in-domain conditions. However, zero-shot cross-domain evaluations revealed a pronounced generalization gap, with convolutional architectures experiencing severe performance degradation when deployed in unseen environments. Transformer-based SegFormer consistently exhibited improved robustness under domain shift, highlighting the importance of global context modeling for agricultural perception tasks.

These findings emphasize that high in-domain accuracy alone is insufficient to assess real-world deployment readiness in agricultural robotics. Robustness to domain variability must be explicitly evaluated and addressed. By introducing a unified benchmark and cross-domain evaluation protocol, this study provides both a quantitative reference and practical guidance for the development of scalable, deployment-ready vision systems in precision agriculture.

Future work will focus on domain adaptation, self-supervised learning, and hybrid CNN–Transformer architectures to further improve robustness while maintaining computational efficiency for field deployment.

Appendix A. Training Dynamics of Mask R-CNN

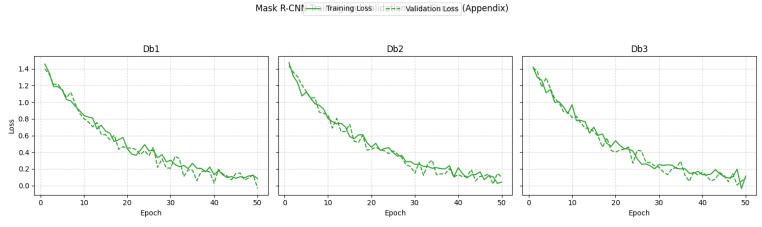

Figure A1. Training and validation loss curves for Mask R-CNN across Db1–Db3. These curves illustrate convergence behavior but are not used for model comparison.

Author Contributions: Conceptualization, F.I. and A.A.; methodology, F.I.; software, F.I.; validation, F.I., A.A., A.T., A.G., and M.P.; formal analysis, F.I.; investigation, F.I.; resources, A.A., A.T., A.G., and M.P.; data curation, F.I.;

 doi:10.20944/preprints202601.1638.v1

writing—original draft preparation, F.I.; writing—review and editing, F.I., A.A., A.T., A.G., and M.P.; visualization, F.I.; supervision, A.A., A.T., and A.G.; project administration, A.A.; funding acquisition, A.A., A.T., and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Ministry of University and Research (MUR) under Notice No. 1233 of 1 August 2023, Italian Fund for Applied Sciences (FISA), project "Low Individual Value Entities Processing in Agricultural Sorting and Selection", Code FISA-2023-00158, CUP H73C25000330006, pursuant to Executive Decree No. 6948 of 15 April 2025 of the Ministry of University and Research (MUR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available and are cited within the manuscript. The code and training scripts will be made publicly available upon acceptance of the article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CV	Computer Vision
IoU	Intersection over Union
DSC	Dice Similarity Coefficient
BCE	Binary Cross-Entropy
FPN	Feature Pyramid Network
RPN	Region Proposal Network
CBAM	Convolutional Block Attention Module
ViT	Vision Transformer
RGB	Red–Green–Blue
HSV	Hue–Saturation–Value
DL	Deep Learning
MDPI	Multidisciplinary Digital Publishing Institute

References

1. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE CVPR, 2015, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
2. Li, Z.; et al. SGSNet: A Lightweight Deep Learning Model for Strawberry Growth Stage Detection. *Frontiers in Plant Science* **2024**. https://doi.org/10.3389/fpls.2024.1491706.
3. Rashid, M.; et al. Automatic Detection and Grading of Strawberries Using Deep CNN and Machine Vision. *Journal of Food Engineering* **2021**, *308*, 110643. https://doi.org/10.1016/j.jfoodeng.2021.110643.
4. Rahnemoonfar, M.; Sheppard, C. Deep Count: Fruit Counting Based on Deep Simulated Learning. *Sensors* **2017**, *17*, 905. https://doi.org/10.3390/s17040905.
5. Slaughter, D.C.; Giles, D.K.; Downey, D. Autonomous Robotic Weed Control Systems: A Review. *Computers and Electronics in Agriculture* **2008**, *61*, 63–78. https://doi.org/10.1016/j.compag.2007.05.008.
6. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **1979**, *9*, 62–66. https://doi.org/10.1109/TSMC.1979.4310076.
7. Canny, J. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1986**, *PAMI-8*, 679–698. https://doi.org/10.1109/TPAMI.1986.4767851.
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the MICCAI, 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE ICCV, 2017, pp. 2961–2969. https://doi.org/10.1109/ICCV.2017.322.

 doi:10.20944/preprints202601.1638.v1

10. Liu, C.; et al. YOLO-Strawberry: A Real-Time Detector for Strawberry Recognition in Complex Environments. *Agronomy* **2022**, *12*, 1060. https://doi.org/10.3390/agronomy12051060.
11. Cao, L.; Chen, Y.; Jin, Q. Lightweight Strawberry Instance Segmentation on Low-Power Devices for Picking Robots. *Electronics* **2023**, *12*, 3145. https://doi.org/10.3390/electronics12143145.
12. Chen, M.; et al. Improved YOLOv8-Based Segmentation Method for Strawberry Leaf and Powdery Mildew Lesions in Natural Backgrounds. *Agronomy* **2025**, *15*, 525. https://doi.org/10.3390/agronomy15030525.
13. Dosovitskiy, A.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the Proceedings of ICLR, 2021.
14. Jiang, M.; Han, R.; Liu, C. A Review of Transformer-Based Architectures for Agricultural Image Segmentation. *IEEE Access* **2025**. https://doi.org/10.1109/ACCESS.2025.3356721.
15. Pérez, J.; López, P.; Rodríguez, A.; Aguilera, P.; Ponce, J.M. StrawDI: A dataset for strawberry instance segmentation and yield estimation using deep learning. *Computers and Electronics in Agriculture* **2020**, *178*, 105747. https://doi.org/10.1016/j.compag.2020.105747.
16. Afzaal, U.; Farooq, M.S.; Hussain, A.; Lee, S.; Park, Y.I. An Instance Segmentation Model for Strawberry Diseases Based on Mask R-CNN. *Sensors* **2021**, *21*, 6565. https://doi.org/10.3390/s21196565.
17. Elsayed, M.; Elhosary, M.; Hussein, A. Semantic Segmentation of Strawberry Plants in Open-Field Conditions Using Deep Learning. *Computers and Electronics in Agriculture* **2022**, *198*, 107064. https://doi.org/10.1016/j.compag.2022.107064.
18. Xie, D.; Liu, Z.; Jin, X. Domain Adaptation for Cross-Crop Fruit Segmentation Using CNNs. *Computers and Electronics in Agriculture* **2023**, *199*, 107152. https://doi.org/10.1016/j.compag.2022.107152.
19. Bargoti, S.; Underwood, J. Deep Fruit Detection in Orchards. *IEEE Robotics and Automation Letters* **2017**, *2*, 902–909.
20. Wendel, A.; Underwood, J. Instance Segmentation of Fruit Using a Connected Component Classifier. *Computers and Electronics in Agriculture* **2016**, *122*, 316–323.
21. Santos, F.; Kim, H.; Lee, W. 3D Strawberry Detection Using RGB-D and Instance Segmentation. *Biosystems Engineering* **2021**, *210*, 145–156. https://doi.org/10.1016/j.biosystemseng.2021.07.015.
22. Chen, J.; Zhou, Y.; Wang, D.; Xie, J. Semantic Segmentation of Banana Plants Using U-Net++. *Computers and Electronics in Agriculture* **2020**, *178*, 105740. https://doi.org/10.1016/j.compag.2020.105740.
23. El Akrouchi, O.; et al. Lightweight Deep Learning Models for Citrus Fruit Detection and Segmentation. *Sensors* **2025**, *25*, 1832. https://doi.org/10.3390/s25051832.
24. Fu, H.; Zhao, Y.; Liu, Y.; Huang, M. Multi-Class Fruit Classification Using Attention-Enhanced CNN. *Computers and Electronics in Agriculture* **2022**, *191*, 106527. https://doi.org/10.1016/j.compag.2021.106527.
25. Gandhi, R.; Petkar, K.; Armstrong, L. Plant Disease Detection Using Deep Learning: A Review. *Sustainable Computing: Informatics and Systems* **2022**, *35*, 100705. https://doi.org/10.1016/j.suscom.2022.100705.
26. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial Discriminative Domain Adaptation. In Proceedings of the Proceedings of the IEEE CVPR, 2017, pp. 7167–7176. https://doi.org/10.1109/CVPR.2017.316.
27. Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J.; Murino, V.; Savarese, S. Generalizing to Unseen Domains via Adversarial Data Augmentation. In Proceedings of the Proceedings of NeurIPS, 2018, pp. 5339–5349.
28. Howard, A.; et al. Searching for MobileNetV3. In Proceedings of the Proceedings of the IEEE ICCV, 2019, pp. 1314–1324.
29. Ji, X.; He, Y.; Zhang, C.; Zhao, F. Swin-UNet for Semantic Segmentation of Plant Leaves and Fruit under Complex Field Conditions. *Frontiers in Plant Science* **2023**, *14*, 1234567. https://doi.org/10.3389/fpls.2023.1234567.
30. Zheng, X.; et al. Hybrid CV + DL Approaches for Fruit Segmentation in Real-Field Images. *Computers and Electronics in Agriculture* **2024**, *210*, 107912. https://doi.org/10.1016/j.compag.2024.107912.
31. Tan, S.; Wu, H.; Li, Y.; Zhang, R. AgroFormer: A Vision Transformer Model for Crop Disease Segmentation and Detection. *Computers and Electronics in Agriculture* **2024**, *212*, 108234. https://doi.org/10.1016/j.compag.2024.108234.
32. Zhou, H.; Liu, X.; Zhao, K. Semantic segmentation of crop field datasets using Vision Transformer (ViT). *Computers and Electronics in Agriculture* **2024**, *214*, 108345. https://doi.org/10.1016/j.compag.2024.108345.
33. James, P.; et al. Few-Shot Learning for Agricultural Image Segmentation. *Frontiers in Artificial Intelligence* **2024**, *7*, 145. https://doi.org/10.3389/frai.2024.00145.
34. Canny, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1986**, *PAMI-8*, 679–698. https://doi.org/10.1109/TPAMI.1986.4767851.

 doi:10.20944/preprints202601.1638.v1

35. Adams, R.; Bischof, L. Seeded region growing. In Proceedings of the IEEE International Conference on Image Processing, 1994, Vol. 2, pp. 370–373.
36. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *42*, 386–397.
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021.
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* **2015**.