

Article

High-Precision Peach Fruit Segmentation under Adverse Conditions Using Swin Transformer

Dasom Seo ¹, Seoul Ki Lee ², Jin Gook Kim ³ and Il-Seok Oh ^{1,4,*}

¹ Department of Computer Science & Artificial Intelligence, Jeonbuk National University, Jeonju-si 54896, Republic of Korea; dasomseo@jbnu.ac.kr

² Fruit Research Division, National Institute of Horticultural & Herbal Science, Wanju-gun 55365, Republic of Korea; lsk0729@korea.kr

³ Institute of Agriculture and Life Sciences, Gyeongsang National University, Jinju 52157, Republic of Korea; jgkim119@gnu.ac.kr

⁴ Center for Advanced Image and Information Technology (CAIIT), Jeonbuk National University, Jeonju-si 54896, Republic of Korea

* Correspondence: isoh@jbnu.ac.kr

Abstract: In the realm of agricultural automation, the efficient management of tasks like yield estimation, harvesting, and monitoring is crucial. While fruits are typically detected using bounding boxes, pixel-level segmentation is essential for extracting detailed information such as color, maturity, and shape. Furthermore, while previous studies have typically focused on controlled environments and scenes, achieving robust performance in real orchard conditions is also imperative. To prioritize these aspects, we propose the following two considerations: first, a novel peach image dataset designed for rough orchard environments, focusing on pixel-level segmentation for detailed insights; and second, utilizing a transformer-based instance segmentation model, specifically the Swin Transformer as a backbone of Mask R-CNN. We achieve superior results compared to CNN-based models, reaching 60.2 AP on the proposed peach image dataset. The proposed transformer-based approach specially excels in detecting small or obscured peaches, making it highly suitable for practical field applications. The proposed model achieved 40.4 AP for small objects, nearly doubling that of CNN-based models. This advancement significantly enhances automated agricultural systems, especially in yield estimation, harvesting, and crop monitoring.

Keywords: agricultural automation; computer vision; fruit segmentation; transformer



Citation: Seo, D.; Lee, S.K.; Kim, J.G.; Oh, I.-S. High-Precision Peach Fruit Segmentation under Adverse Conditions Using Swin Transformer. *Agriculture* **2024**, *14*, 903. <https://doi.org/10.3390/agriculture14060903>

Academic Editors: Jibo Yue, Chengquan Zhou, Haikuan Feng, Ning Zhang and Yanjun Yang

Received: 30 April 2024

Revised: 30 May 2024

Accepted: 4 June 2024

Published: 7 June 2024



Copyright: © 2024 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The usual tasks in an orchard, such as yield counting, predicting the appropriate harvest time, and picking fruits, are time-consuming and require significant labor. Various automated farming systems and robots have been studied through fusion with other fields to enhance the efficiency of such tasks [1–4]. For automated farming systems in an orchard, especially agricultural robots, computer vision algorithms are required to obtain information on fruit [4,5]. Typically, fruit objects are detected as bounding boxes in orchard images for growth monitoring and yield estimation and to localize positions for harvesting. However, to obtain and use more detailed information, fruits should be segmented at the pixel level. Pixel-level results extend monitoring tasks to estimate size or ripeness. Within the segmentation domain, two main techniques are employed: semantic segmentation, which involves the separation of the foreground and background, and instance segmentation, which identifies and differentiates individual object instances. In this study, we assume that instance segmentation is more suitable for orchard images because many fruits are captured, and robot tasks require separate information for each fruit.

Conventional algorithms have typically been studied using crop images with minimal occlusion or focusing on specific zoomed-in crops [6–9]. However, real field scenes in

agricultural environments are often complex, with significant occlusions and variations in crop sizes. Algorithms should be tested on more dense and cluttered images to reflect the real world, and this challenge was acknowledged by [3]. This issue is also crucial for enabling diverse tasks in the application of vision systems or robots. Some tasks, such as monitoring or spraying, also require a wider field of view than that required for traditional crop images. In the agricultural field, there is a tendency to prefer local datasets that mostly comprise zoomed-in images focused on specific targets to open datasets [5]. The MinneApple dataset [10] is unique in that it provides images in which an entire tree is visible, along with individual apple instances annotated at the pixel level.

Several comprehensive peach image datasets have been proposed, although datasets with such images and pixel-level labeling are rare. The authors of [11] presented a dataset for peach image analysis, focusing on instance segmentation and ripeness classification. However, most images in [11] were captured at close range, featuring large object areas and relatively simple backgrounds. Further, the annotation masks provided by [11] covered only visible regions, excluding obscured areas. The authors of [12] presented a dataset with naturally cluttered scenes comprising small objects, but the annotations were bounding boxes only because the dataset was for detection tasks. The authors of [13] provided a large dataset comprising complex peach orchard images to detect immature fruits, so ground truths were annotated with bounding boxes only. In this study, we present a dataset containing various image types, ranging from fruit-focused to whole-tree visible images. For each image, the corresponding fruit instance masks are also annotated. The dataset is specially collected in the peach (*Prunus persica* (L.) Batsch) orchard because peach tree images are highly cluttered with long leaves, lots of bunches, and occlusions.

CNN-based segmentation models have shown significant achievements, but the difficulty in effectively segmenting small or occluded objects exists because of several issues. Typical CNN models can handle small areas well in their initial layers, but the receptive field increases when the layers go deeper, causing a problem in detecting small objects and resulting in inaccurate segmentation due to missing fine details. CNN models usually use pooling layers, such as max pooling, to reduce feature map size, resulting in the loss of information related to small objects. Therefore, CNN-based segmentation models have limitations in accurate segmentation for small or occluded objects. Referencing [4,14,15], most traditional fruit detection methods typically utilize well-known CNN-based models such as YOLO [16], Faster R-CNN [17], mask R-CNN [18], YOLACT [19], SSD [20], and SegNet [21]. These methods commonly face limitations, including difficulty in detecting small fruits, partially hidden or overlapping fruits, and performance degradation under complex natural conditions.

Transformer-based models have been researched and have demonstrated promising performance in solving the aforementioned problems. Transformer-based models use a self-attention mechanism that enables broader context consideration than CNN. Transformers process an entire input sequence at once through self-attention. This feature also enables an enhanced understanding of the global context. Notably, the global context of specific objects is essential in small object segmentation [22]. Despite these advantages, the application of transformer-based models in the agricultural field is currently slow. However, active study is required to develop efficient automated systems (or robots) for agricultural applications. Ref. [15] also points out that transformer models have not been widely explored in fruit analysis within the existing literature, despite the strong representation capabilities of visual transformers. In this study, we specifically apply the Swin Transformer to build a hierarchical model based on shifted windows that contain patches with diverse sizes [23]. Patches in the initial layer windows are small, which is helpful for detecting small objects. While there is example of applying the Swin Transformer for grape detection, note that this application is primarily focused on bounding boxes only [24].

In this paper, we present a peach image dataset and an application of a transformer-based instance segmentation model to accurately detect peach fruits in a cluttered orchard scene. This paper includes the following contributions:

- A novel peach image dataset for instance segmentation, specifically designed to include obscured regions for detailed information and more accurate performance in cluttered orchard environments similar to real orchards.
- An application of a transformer-based instance segmentation model, specially the Swin Transformer, to detect small and obscured peaches more accurately.
- The proposed model achieves an average precision (AP) of 60.2, outperforming CNN-based models, and shows an AP of 40.4 for small objects, nearly doubling that of CNN-based models.

2. Materials and Methods

2.1. Instance Segmentation Based on Transformer

Conventional object detection and instance segmentation models lack detailed segmentation accuracy with small objects or complex background images. To handle diverse situations, the authors of [25] introduced a feature pyramid network (FPN), which serves as the backbone for conventional instance segmentation models, such as mask R-CNN [18] and YOLACT [19]. However, there is a loss of high-resolution feature details, and the potential omission of small or obscured objects still occurs in the down-sampling process of the top-down pathway in the FPN. Because the transformer-based approach can address this issue with large context consideration through self-attention, as mentioned in Section 1, we apply the Swin Transformer [23], a well-known transformer-based instance segmentation method.

Figure 1 depicts the Swin Transformer architecture, shifted window, Swin Transformer block, and vision transformer (ViT) block. First, an input image is partitioned into nonoverlapping patches using a patch split module; 4×4 patches are basically used, and the feature dimension for each patch is $4 \times 4 \times 3$. After the patch partition, the architecture is separated into four stages. Stage 1 comprises a linear embedding layer and Swin Transformer blocks. The linear embedding layer is applied to project features to arbitrary dimension C . The output tokens from the linear embedding layer pass through the SwinTransformer blocks, and the block keeps the number of tokens as $\frac{H}{4} \times \frac{W}{4}$. Stage 2 comprises a patch merging layer and Swin Transformer blocks for feature transformation. The patch merging layer concatenates 2×2 neighborhood patches, and the dimension is increased to $4C$. To reduce the dimension by half, a linear layer is applied. The output tokens pass through the Swin Transformer blocks, and then the output resolution is $\frac{H}{8} \times \frac{W}{8}$. Stages 3 and 4 repeat the process of Stage 2, and the output resolutions become $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively. The feature dimensions increase to $2C$, $4C$, and $8C$ for Stages 2, 3, and 4, respectively. In summary, as tokens pass each stage, the number of patches increases and the dimension of each token doubles.

The Swin Transformer blocks in each stage are transformer modules based on ViT [22] (Figure 1a). Although ViT uses a multi-head self-attention (MSA) module, Swin-T applies shifted window-based MSA modules (W-MSA and SW-MSA in Figure 1c). Similar to ViT, LayerNorm (LN) layers come before W-MSA, multilayer perceptron (MLP) blocks are added after W-MSA blocks, and GELU activation is applied between two MLP blocks. The basic transformer architecture computes self-attention globally between a token and all other tokens. However, global computation has quadratic complexity with respect to the number of tokens, posing challenges for tasks requiring large tokens, such as dense prediction and high-resolution images. To address this issue, self-attention between local windows has been proposed [23], where windows are divided evenly into even-sized patches to ensure nonoverlapping windows within an image (Figure 1b). The W-MSA module has a limitation in modeling power due to a lack of connectivity between windows. To address this issue, a method called “Shifted Window Partitioning” has been proposed [23].

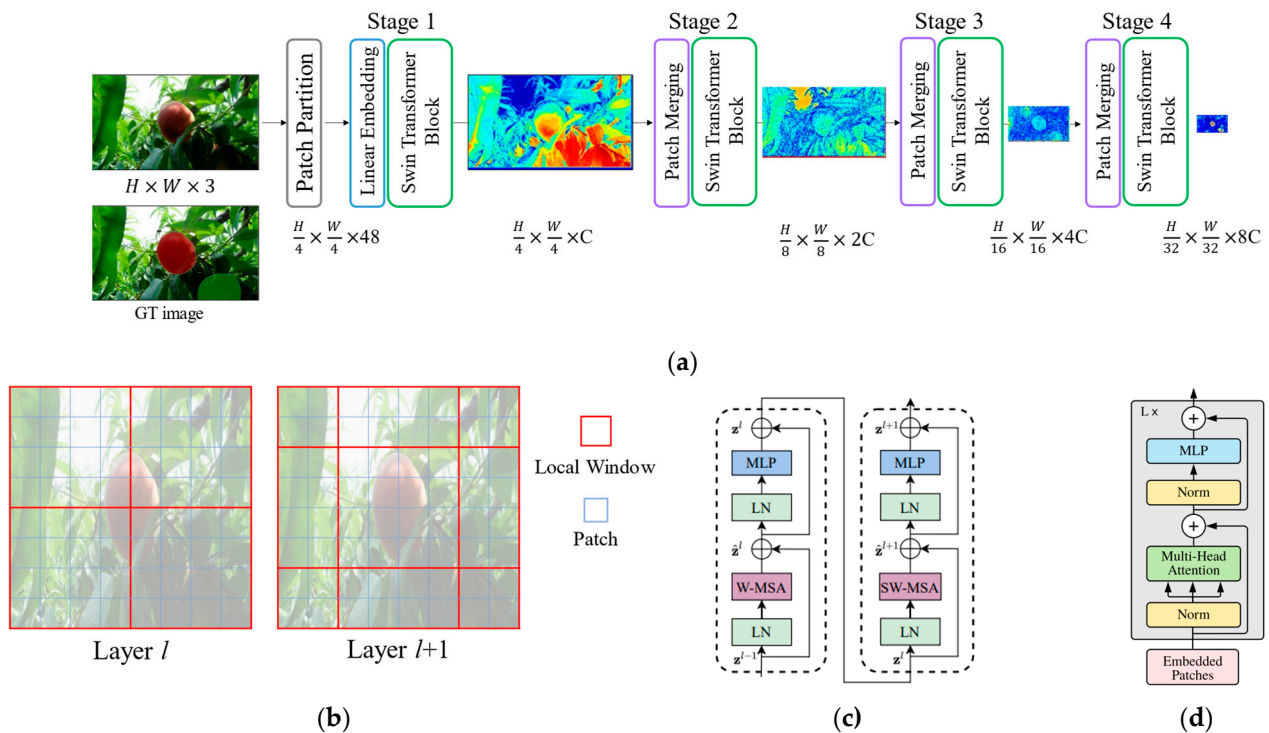


Figure 1. Component illustrations of Swin Transformer. The maps following Swin Transformer blocks are attention maps of each stage. (a) Swin Transformer architecture; (b) shifted window approach; (c) Swin Transformer block; (d) ViT block. (c,d) were extracted from [23] and [22], respectively.

2.2. Dataset Comprising Cluttered Scenes for Peach Segmentation

In this study, we introduce a peach image dataset comprising highly cluttered images. This dataset will serve as a valuable resource for training and evaluating segmentation models. In this subsection, we provide a comprehensive description of the dataset, including its composition, data collection process, and key characteristics.

The dataset comprises 125 RGB images depicting peach trees bearing fruits, each accompanied by corresponding ground-truth annotations in the form of masks for instance segmentation. For experimentation, we partitioned the dataset into three subsets, allocating an 8:1:1 ratio, resulting in 99, 13, and 13 images for training, validation, and testing, respectively. The dataset also includes 1077 peach fruit objects, with an average of eight peaches per image.

The images were captured around the harvest time of 90 days after full bloom, June 2021, in a peach (*Prunus persica* (L.) Batsch) orchard. All images were taken in the morning on the same day under clear weather condition. Canon IXY DIGITAL 220 IS and Samsung MV800 digital cameras were used to acquire images, and the image resolutions encompassed 1600×1200 , 1920×1080 , 1440×1080 , and 4032×1024 or their respective inverses. We used various equipment and resolutions to enhance the generalization of the model and robustness in real-world scenarios. The images were captured freely at the standard distances because it was difficult to set a static camera position or location in an orchard environment.

The images were acquired in two types of view: tree- and fruit-bunch-focused. Four trees and nine bunches of peaches were captured; each tree was captured in approximately 20 styles within 1–5 m and each bunch was captured in 5 styles within 0.3 m. Examples of the captured images are shown in the first row of Figure 2. Figure 2a,b show tree-focused images, and Figure 2c shows a fruit-bunch-focused image. Owing to the widely sprawled feature of peach trees, the images could not be captured at a fixed distance, and tree-focused images contained an entire tree or only part of a tree. Fruit-bunch-focused images comprised a few peach fruits, about one or two, in the middle of the image. The

numbers of peach fruit objects for each type are listed in Table 1. There are 80 tree-focused and 45 fruit-bunch-focused images, with an average of 12.0 and 2.7 peach objects per image, respectively.

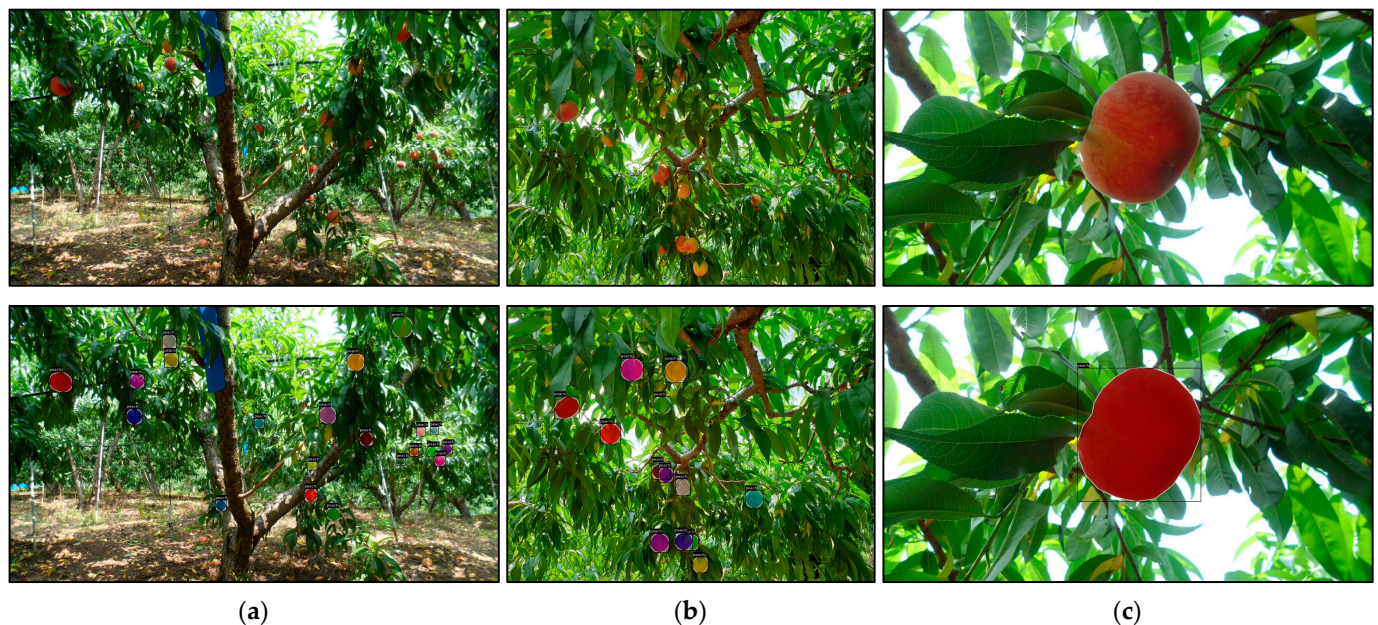


Figure 2. Examples of image data and corresponding ground-truth masks: (a,b) tree-focused images; (c) fruit-bunch-focused image. The top row presents original images, and the bottom row shows labeled masks.

Table 1. Number of objects and images in the dataset.

	Tree	Fruit Bunch	Total
Objects	957	120	1077
Images	80	45	125
Average number of objects	12.0	2.7	8.6

Instance masks were manually annotated using LabelMe 5.0.1 [26] and labeled as “peach” class, except for dropped fruits. The masks were basically labeled by polygons, and a circle was used if more than 50% of the target was occluded by leaves or other peaches. All masks were annotated, including occluded areas, to apply the dataset to various tasks, such as fruit location and size estimation. Examples of ground-truth masks are shown in the second row of Figure 2.

To clarify clutter and diversity similar to the real field, unlike other datasets, we analyze two characteristics: occlusion and object size. Because the images were captured in an orchard, fruits are often obscured by leaves, branches, and neighborhood fruits. We flagged obscured fruits if the peach edge could not be precisely labeled, and 64.1% of peaches were flagged in total. Figure 3 depicts an image with significant occlusion, where 93.8% of the peaches are occluded.

We analyze the characteristics of object size separately for tree- and fruit-bunch-focused images. Figure 4 shows histograms of the object area ratios of the images. Object size characteristics were analyzed using ratios because each image had a different resolution. Tree-focused images include 957 peach objects; 887 of them, i.e., approximately 92.7% of the total, occupy area ratios smaller than 1% in an image. The minimum and maximum area ratios are 0.0015% and 5.46%, respectively. Images including objects with minimum and maximum area ratios are shown in the first row of Figure 5. Target objects are boxed in a yellow line. Fruit-bunch-focused images include 120 peach objects; 50 peaches are

smaller than 1% of the image size. The minimum ratio of fruit-bunch-focused images is also small (0.022%); however, the maximum ratio is 12.22%, and 58.33% of peaches occupy an area larger than 1.0%. Examples of images for each ratio are shown in the second row of Figure 5.



Figure 3. (a) Image with significant occlusion; (b) the corresponding ground truth. Most peaches are obscured by leaves.

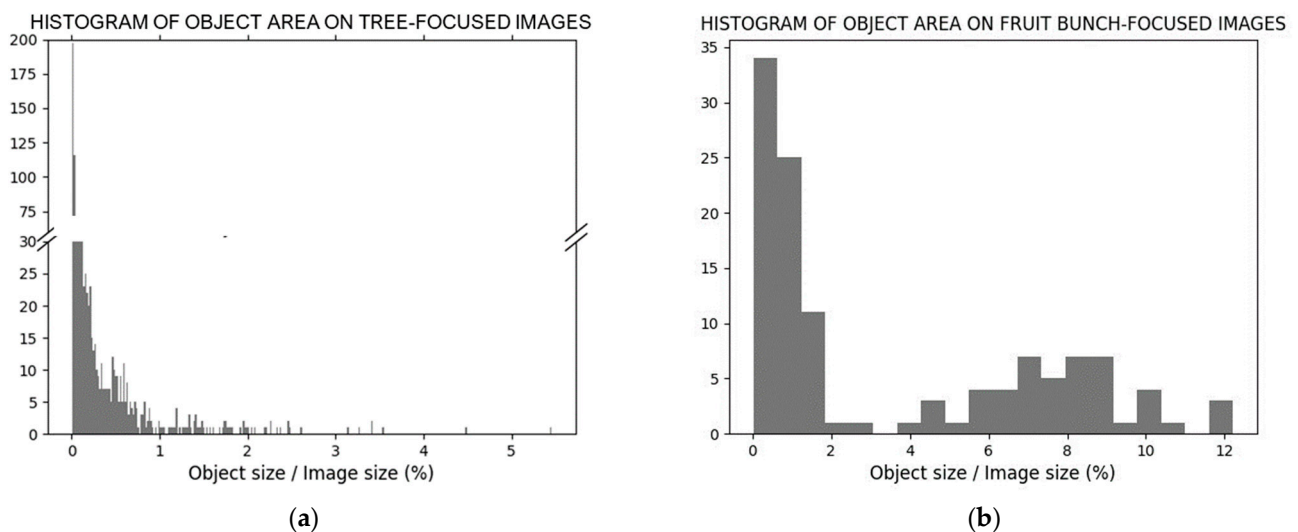


Figure 4. (a) Histograms of object area for tree-focused images; (b) and fruit-bunch-focused images.

Statistics about occlusion and object size show how the dataset represents a real field environment not highly controlled. An analysis of object size also shows the diversity of the scenes. We used this dataset to train segmentation models to reflect the real field well; the results are described in Section 3.

Currently, there is a significant shortage of segmentation datasets for agricultural purposes that reflect real field environments. We aim to contribute to the academic community by releasing our dataset to the public. It is available at <https://github.com/ssomda21/Dataset-for-Instance-Segmentation-of-Fruits-in-Peach-Tree-Image> (accessed on 23 February 2023).



Figure 5. Examples of images containing objects with minimum and maximum areas. Each target object is represented by a yellow box.

2.3. The Metric for Instance Segmentation

To evaluate and compare performance, the standard evaluation metric, average precision (AP), is used. The AP score is derived from the intersection over union (IoU). A prediction is classified as a “true positive (TP)” if IoU between the ground truth and predicted masks exceeds a certain threshold. Typically, an IoU threshold of 0.5 is employed, implying that a prediction is considered TP if the masks overlap by more than half. AP is a metric related to precision–recall, which computes the average value of precision at every recall position, the same as the area under the precision–recall curve. Precision and recall are single-valued metrics based on the prediction results returned by models. The detailed equations are described in Equation (1). Notably, FP and FN mean false positive and false negative, respectively. Mean AP (mAP), which computes the mean value for all categories, is the commonly used metric for object detection/segmentation; however, we use AP because our target category is only “peach.”

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad (1)$$

3. Results

In this section, we present the experimental results on peach fruit segmentation within orchard scenes. We trained the Swin-T pretrained model from the Swin Transformer. We trained CNN-based models, specifically Mask R-CNN with a ResNet-50 backbone [18], YOLACT [19], and SOLOv2 [27], for comparative analysis. The experiments were run on an NVIDIA RTX 2080 Ti GPU and an Intel(R) Xeon(R) Gold 6126 CPU. We used the MMDetection 2.25.1 library, supported by OpenMM, to implement the segmentation models [28]. Each model was fine-tuned with a pretrained model from MMDetection and trained on our dataset. We trained all models for 100 epochs, with specific training parameters detailed in Table 2.

Table 2. Parameters of each segmentation model.

Backbone Type	Model	Parameters		
		Optimizer	Learning Rate	Batch Size
CNN	Mask R-CNN (ResNet-50) [18]	SGD	0.02	2
	YOLOACT [19]	SGD	0.001	8
	SOLOv2 [27]	SGD	0.001	4
Transformer	Mask R-CNN (Swin-T)	AdamW	0.0001	2

3.1. Evaluation of Peach Fruit Instance Segmentation

The detailed evaluation results for each model are presented in Table 3. Notably, the term “AP” in the table refers to the overall AP score, which is an average across IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05. Mask R-CNN with ResNet-50 backbone [hereinafter, Mask R-CNN(ResNet-50)] recorded an overall AP of 55.1. At IoU thresholds of 0.5 and 0.75, it achieved APs of 83.9 and 56.0, respectively, indicating that while the model exhibits high precision for general overlaps, its performance slightly diminishes at stricter IoU thresholds. YOLOACT recorded an overall AP of 52.6. At IoU thresholds of 0.5 and 0.75, it recorded APs of 81.9 and 58.5, respectively. This model’s performance tended to decrease as the IoU threshold increased. SOLOv2 recorded an overall AP of 53.2 and recorded 79.6 AP and 55.8 AP at IoU thresholds of 0.5 and 0.75.

Table 3. Evaluation results on the testing set (best performance is shown in bold).

Model	AP	AP@0.5	AP@0.75	Inference Time (ms)
Mask R-CNN(ResNet-50)	55.1	83.9	56	19.88
YOLOACT	52.6	81.9	58.5	25.92
SOLOv2	53.2	79.6	55.8	27.38
Mask R-CNN(Swin-T)	60.2	87.1	65.2	20.92

Mask R-CNN with the Swin-T backbone (hereinafter, Mask R-CNN(Swin-T)), which incorporates a transformer architecture, notably outperformed the other models. It achieved an overall AP of 60.2. Impressively, when evaluated at an IoU threshold of 0.75, it recorded an AP of 65.2, indicating its ability to maintain high precision, especially in situations that require a close overlap between predictions and ground truth. Although Mask R-CNN(Swin-T) showed the best segmentation performance on the provided dataset, it is essential to note the inference time. Mask R-CNN(ResNet-50) was the swiftest, processing an image in just 19.88 ms. Meanwhile, YOLOACT, SOLOv2, and Mask R-CNN(Swin-T) took 25.92, 27.39, and 20.92 ms per image.

Figures 6 and 7, respectively, show the confusion matrix and result images for each model, including only the detection results with a confidence score greater than 0.5. In Figure 6, the order of the boxes is TP, FN, true negative (TN), and FP. Because we consider an instance segmentation problem in this study, we concentrate on TP, FN, and FP. TN should be zero because the models cannot predict background objects. Notably, FN indicates that the model misses existing objects, and FP indicates that the model predicts the wrong region. This analysis was performed on the testing set; the total number of ground-truth objects was 110. Each ratio is calculated based on the number of ground truths, and the numbers in parentheses indicate the real number of predictions. Mask R-CNN(Swin-T) recorded the best performance with 92% TP, while Mask R-CNN (ResNet-50) and YOLOACT recorded 87% and 76% TP, respectively. Notably, Mask R-CNN(Swin-T) missed only eight peaches in the entire test dataset. FP predictions often occurred in diseased leaves or dense peach groups. For the first and fifth examples in Figure 7, Mask R-CNN(ResNet-50) predicted duplicate objects in dense peach groups. It also predicted leaves as peaches. YOLOACT predicted the least FP among the three models; however, some masks were localized well,

but the area was significantly different compared with the ground truth. SOLOv2 misses peach objects similar to YOLACT; however, a FP occurred at least for the examples in Figure 7. In Figure 7, specific examples of FNs and FPs are highlighted by cyan and yellow boxes, respectively.

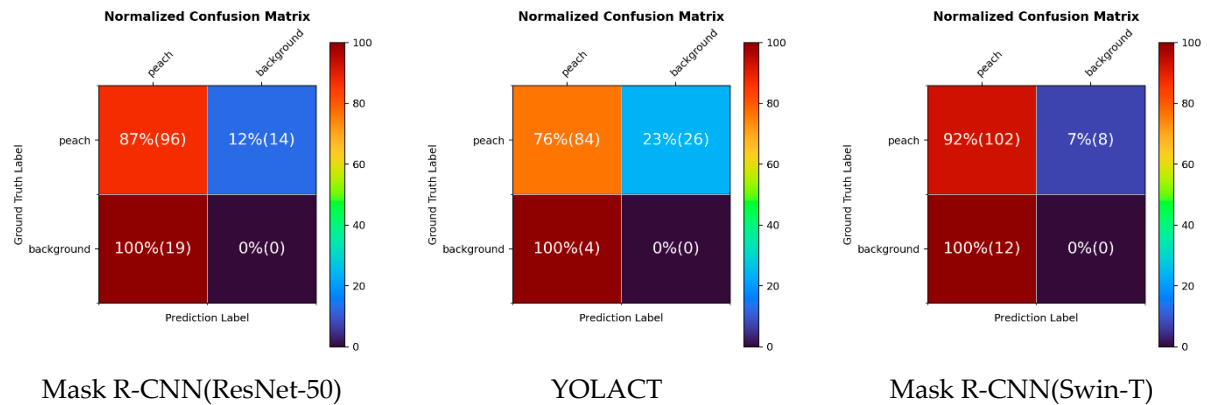


Figure 6. Confusion matrix for each model. Starting from the top left and moving clockwise, the order is TP, FN, TN, and FP. Each value in the box indicates the ratio with respect to the ground truth, and the values in parentheses indicate the real number of predictions. Mask R-CNN(Swin-T) has the fewest peach misses.

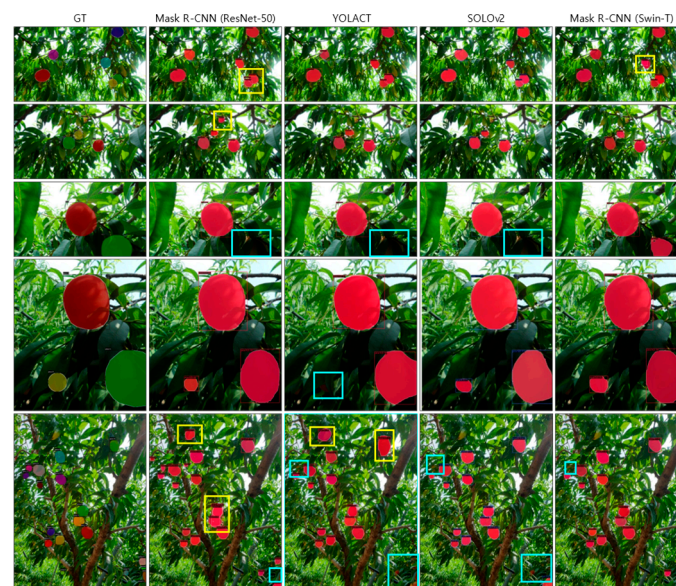


Figure 7. Examples of test images from each model. From left to right, ground-truth images and output images from Mask R-CNN(ResNet-50), YOLACT, SOLOv2, and Mask R-CNN(Swin-T) are shown. FP and FN predictions are highlighted as yellow and cyan boxes, respectively. Mask R-CNN(Swin-T) yields the fewest FPs and FNs.

3.2. Evaluation of Challenging Cases

To find peaches well in cluttered images, the model should be robust to small or occluded objects. We specifically analyze the results by focusing on small and occluded objects. Notably, the “small object” is determined by the condition used by the MS-COCO dataset [29], i.e., smaller than 32×32 pixels. Table 4 shows the AP and average recall (AR) scores of the small objects. Mask R-CNN(Swin-T) achieved outstanding performance, scoring 40.4 AP and 40.0 AR, which were more than 1.6 times higher than those of Mask R-CNN(ResNet-50), YOLACT, and SOLOv2. To analyze the robustness of the models for occluded objects, we specifically evaluated only images in the testing set where most objects,

exceeding 50%, were obscured. Mask R-CNN(Swin-T) achieved the best performance, 59.2 AP and 63.2 AR, indicating its suitability for cluttered images.

Table 4. Small object segmentation performance.

	Mask R-CNN (ResNet-50)	YOLOACT	SOLOv2	Mask R-CNN (Swin-T)
AP	25.2	12.6	20.2	40.4
AR	25	25	20	40

Additional inferences were performed on unlabeled images; qualitative comparisons are shown in Figure 8. We selected examples that were challenging to count manually, underscoring the difficulty for human observers. YOLOACT and SOLOv2 showed lower performance in detecting peaches obscured by leaves in all examples. SOLOv2 also detected significant duplications of peaches. Mask R-CNN(ResNet-50) effectively detected small and obscured peaches; however, some peaches were missed, and the edges of the detected regions were not smooth. These CNN-based models also tended to predict multiple groups of peaches as a single object. Meanwhile, Mask R-CNN(Swin-T) predicted well with lots of objects and well-localized highly obscured objects.

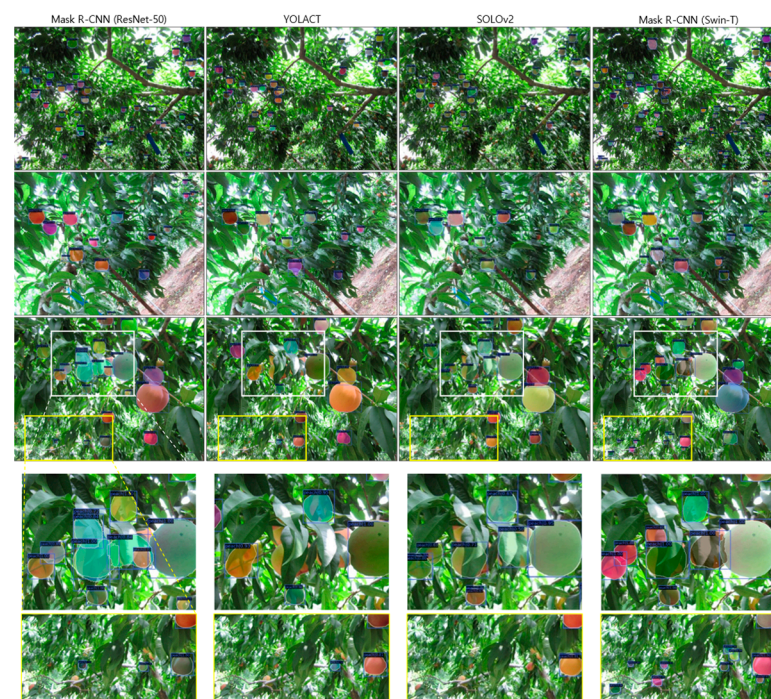


Figure 8. Images from unlabeled inputs. Highlighted regions from the third image are magnified under the images. The region in the white box shows the difference in region separation between Mask R-CNN(ResNet-50) and Mask R-CNN(Swin-T). Mask R-CNN(ResNet-50) detects a group of peaches as a single peach with a high confidence score of 0.72, while Mask R-CNN(Swin-T) detects each peach individually. The white box also shows that YOLOACT misses about four peaches and does not separate the peach group on the left side of the image as individual objects. SOLOv2 also misses obscured peaches in the white box and shows significant duplicated detections. The yellow box shows that Mask R-CNN(Swin-T) detects tiny peaches better than other models.

4. Discussion

4.1. Necessity and Limitations of the Dataset

In this study, we present a specialized dataset for peach instance segmentation that is robust to challenging cases, such as occluded and small objects. The ground-truth masks are

annotated at the pixel level, including obscured regions, referred to as “amodal masks” [30]. This annotation method helps to develop advanced models for instance segmentation, size estimation, yield estimation, and crop monitoring in the agricultural field.

The primary advantage of the presented dataset is the focus on detailed and complex segmentation scenarios that reflect real field environments because almost all existing datasets for segmentation provide simpler images with large and distinct fruits [3,5,11,31]. This level of detail is unprecedented and offers an invaluable resource for training more sophisticated fruit segmentation algorithms. By providing a dataset that reflects real field environments, we anticipate significant improvements in the real-world applicability of various agricultural technologies. The necessity for such a comprehensive dataset stems from the current limitations in the application of computer vision in agriculture due to the lack of datasets that truly reflect the complexity and variability of field conditions. Our dataset addresses this gap, thereby meeting a critical need for the development of robust peach segmentation models. The implications for agricultural infrastructure are profound, as better segmentation can lead to optimized resource allocation and improved crop management. Unlike other fields, such as medical imaging, which has a wealth of public datasets and challenges [32], the agricultural field faces a shortage of such resources [5]. This deficiency in data availability directly impacts the pace of research progress. To contribute to the agricultural field’s growth and overcome these challenges, we have made our dataset openly accessible to facilitate research and foster advancements within the field.

However, the presented dataset has limitations in terms of environmental conditions: images were captured on the same day under the same light and weather conditions. These limitations may reduce the variance but also limit the dataset’s ability to generalize across different conditions. In addition, the focus on a single cultivar of peach, although beneficial for in-depth analysis, may not capture the diversity of real-world agricultural scenarios. Future iterations of data acquisition should aim to include a broader variety of cultivars, diverse environmental conditions, and different stages of growth to enhance the generalizability of corresponding segmentation models.

4.2. Robustness for Challenging Cases with Potential for Broad Extension

The transformer-based segmentation model used in this study is robust to challenging cases of agricultural images under adverse conditions [3] and shows a potential for broad extension in automated farming systems. Recently, the agricultural field has consistently shown an interest in restoring obscured areas in detection and segmentation tasks, as this could significantly impact the enhancement of size estimation and growth monitoring performance [33,34]. The proposed model, Mask R-CNN(Swin-T), shows better performance for obscured peaches (see Table 5 and Figures 7 and 8). Mask R-CNN(Swin-T) also outperformed the comparative CNN-based models in terms of occluded area reconstruction. Based on this capability, the transformer-based model can be extended to fruit quality assessments that evaluate quality features, such as size, shape, and color.

Table 5. Segmentation performance on images with more than 50% occlusion.

	Mask R-CNN (ResNet-50)	YOLOACT	SOLOv2	Mask R-CNN (Swin-T)
AP	53.9	50.8	51.7	59.2
AR	60	56.4	58.9	63.2

As shown in Table 4 and Figure 8, Mask R-CNN(Swin-T) achieved outstanding performance in segmenting small peaches and a significant number of peaches, indicating that it effectively localized and identified peach regions even when images were captured from a distance. The region in the yellow box in the third image from Figure 8 shows that Mask R-CNN(Swin-T) detected tiny peaches significantly more than the comparative CNN-based models. This advantage eliminates the need to consider shooting distance when developing agricultural robots. This robustness to small objects also helps shape separations, such

as detecting each peach in a fruit bunch. Mask R-CNN(ResNet-50) occasionally detected a group of peaches as a single peach (Figures 7 and 8), whereas Mask R-CNN(Swin-T) detected each peach well. Particularly, the region highlighted with a white box in Figure 8 shows this problem distinctively. Mask R-CNN(ResNet-50) identified a cluster of peaches as a single peach with a high confidence score of 0.72, whereas Mask R-CNN(Swin-T) individually identified each peach in a cluster. YOLACT showed predictions that trespass neighboring peaches. This separation capability can lead to improved performance in various farming tasks in which the identification of each fruit is crucial, such as fruit counting, robot harvesting, and yield estimation.

4.3. Speed of Two-Stage Detector and Agricultural Robot

In this study, we applied Swin-T to Mask R-CNN, a type of two-stage detector. The two-stage detector is divided into a region proposal network and a detection stage. In Mask R-CNN, the second stage includes pixel-level mask generation to represent precise shapes. Two-stage architectures have advantages such as high accuracy, precise object detection, and flexibility. Flexibility facilitates easy integration with various backbone networks and enables effective application in object segmentation. For these reasons, Mask R-CNN is utilized to verify the Swin Transformer.

However, two-stage detectors are generally slower than one-stage detectors because additional computations are required for the region proposal. Two-stage detectors also require more computational resources, resulting in limitations in applications requiring large datasets or real-time processing (≥ 30 fps). For these reasons, Mask R-CNN(Swin-T) cannot be appropriate for fast tasks but can be applied to agricultural engineering because agricultural robots usually operate and move slowly [2]. During the harvesting process, it takes a robot between a minimum of 2–3 s to a maximum of 64 s to pick a single fruit, and it moves at a speed of 0.19 m/s, indicating that a transformer model, which requires an average processing time of 20.92 ms per frame, can be adequately applied at this speed. However, improvements in the speed of transformer-based models should be studied because the necessity for real-time processing is increasingly prominent for automated spraying systems [2]. Recently, transformer-based models that show low latency, such as EfficientFormer [35], have been studied.

The transformer-based segmentation approach can be widely applied to some major field operations, such as plant monitoring, phenotyping, spraying, harvesting, yield estimation, and multipurpose robotic systems [1–4]. Particularly, the transformer-based approach can be applied to tasks that require detailed representation, such as phenotyping, because they demonstrate overwhelming performance in challenging cases, as described in Section 3.2. This approach can also be expanded to weeding and disease and insect detection by changing the target from fruit to other objects.

5. Conclusions

In this study, we proposed a transformer-based segmentation model Mask R-CNN(Swin-T) for fruit segmentation in a cluttered orchard. We also presented a novel dataset of peach tree images captured from a peach orchard. This dataset comprises cluttered images including significant obscured objects that reflect real scenes in the agricultural field.

Mask R-CNN(Swin-T), a two-stage detector, shows superior performance over comparative CNN-based models. In particular, Mask R-CNN(Swin-T) recorded an AP of 60.2 and had the fewest missed peaches (FN). The advantage of Mask R-CNN(Swin-T) is revealed in the detection of small and obscured objects, AP 40.4 and 59.2. The proposed model, based on the two-stage detector, is not capable of real-time processing (processing speed of 4.78 fps); however, it can be sufficiently applied because the average speed of agricultural robots is 11.9 s per fruit.

The findings of this study indicate the potential applicability and advantages of transformer-based vision models in agricultural settings, using data that reflect real field

environments. Further, transformer-based vision models can be expanded to various intelligent agricultural areas, such as spraying, harvesting, and yield estimation.

Author Contributions: Conceptualization, I.-S.O.; methodology, I.-S.O. and D.S.; software, D.S.; validation, D.S. and I.-S.O.; formal analysis, D.S.; investigation, D.S.; resources, S.K.L. and J.G.K.; data curation, J.G.K.; writing—original draft preparation, D.S.; writing—review and editing, I.-S.O.; visualization, D.S.; supervision, I.-S.O.; project administration, S.K.L.; funding acquisition, S.K.L. and I.-S.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Cooperative Research Program for Agriculture Science and Technology Development (Project No. RS-2021-RD009224), Rural Development Administration, the Republic of Korea.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in the GitHub repository at <https://github.com/ssomda21/Dataset-for-Instance-Segmentation-of-Fruits-in-Peach-Tree-Image> (accessed on 23 February 2023).

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

References

1. Droukas, L.; Doulergi, Z.; Tsakiridis, N.L.; Triantafyllou, D.; Kleitsiotis, I.; Mariolis, I.; Giakoumis, D.; Tzovaras, D.; Kateris, D.; Bochtis, D. A Survey of Robotic Harvesting Systems and Enabling Technologies. *J. Intell. Robot. Syst.* **2023**, *107*, 1–29. [CrossRef] [PubMed]
2. Fountas, S.; Mylonas, N.; Malounas, I.; Rodias, E.; Santos, C.H.; Pekkeriet, E. Agricultural Robotics for Field Operations. *Sensors* **2020**, *20*, 2672. [CrossRef]
3. 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]
4. Yang, Y.Y.; Han, Y.X.; Li, S.; Yang, Y.D.; Zhang, M.; Li, H. Vision based fruit recognition and positioning technology for harvesting robots. *Comput. Electron. Agric.* **2023**, *213*, 108258. [CrossRef]
5. Lu, Y.Z.; Young, S. A survey of public datasets for computer vision tasks in precision agriculture. *Comput. Electron. Agric.* **2020**, *178*, 105760. [CrossRef]
6. Yu, Y.; Zhang, K.L.; Yang, L.; Zhang, D.X. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
7. Ge, Y.Y.; Xiong, Y.; From, P.J. Instance Segmentation and Localization of Strawberries in Farm Conditions for Automatic Fruit Harvesting. *IFAC-PapersOnLine* **2019**, *52*, 294–299. [CrossRef]
8. Wang, J.H.; Zhang, Z.Y.; Luo, L.F.; Wei, H.L.; Wang, W.; Chen, M.Y.; Luo, S.M. DualSeg: Fusing transformer and CNN structure for image segmentation in complex vineyard environment. *Comput. Electron. Agric.* **2023**, *206*, 107682. [CrossRef]
9. Wang, D.D.; He, D.J. Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Comput. Electron. Agric.* **2022**, *196*, 106864. [CrossRef]
10. Hāni, N.; Roy, P.; Isler, V. MinneApple: A Benchmark Dataset for Apple Detection and Segmentation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 852–858. [CrossRef]
11. Zhao, Z.; Hicks, Y.; Sun, X.F.; Luo, C.X. Peach ripeness classification based on a new one-stage instance segmentation model. *Comput. Electron. Agric.* **2023**, *214*, 108369. [CrossRef]
12. Assunção, E.T.; Gaspar, P.D.; Mesquita, R.J.M.; Simoes, M.P.; Ramos, A.; Proença, H.; Inacio, P.R.M. Peaches Detection Using a Deep Learning Technique-A Contribution to Yield Estimation, Resources Management, and Circular Economy. *Climate* **2022**, *10*, 11. [CrossRef]
13. Liu, P.Z.; Yin, H. YOLOv7-Peach: An Algorithm for Immature Small Yellow Peaches Detection in Complex Natural Environments. *Sensors* **2023**, *23*, 5096. [CrossRef] [PubMed]
14. Xiao, F.; Wang, H.B.; Xu, Y.Q.; Zhang, R.Q. Fruit Detection and Recognition Based on Deep Learning for Automatic Harvesting: An Overview and Review. *Agronomy* **2023**, *13*, 1625. [CrossRef]
15. Espinoza, S.; Aguilera, C.; Rojas, L.; Campos, P.G. Analysis of Fruit Images with Deep Learning: A Systematic Literature Review and Future Directions. *IEEE Access* **2024**, *12*, 3837–3859. [CrossRef]
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
17. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 7–12 December 2015; pp. 91–99.

18. He, K.M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [\[CrossRef\]](#)
19. Bolya, D.; Zhou, C.; Xiao, F.Y.; Lee, Y.J. YOLACT Real-time Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9156–9165. [\[CrossRef\]](#)
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Pt I, pp. 21–37.
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
23. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.X.; Zhang, Z.; Lin, S.; Guo, B.N. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [\[CrossRef\]](#)
24. Wang, J.H.; Zhang, Z.Y.; Luo, L.F.; Zhu, W.B.; Chen, J.W.; Wang, W. SwinGD: A Robust Grape Bunch Detection Model Based on Swin Transformer in Complex Vineyard Environment. *Horticulturae* **2021**, *7*, 492. [\[CrossRef\]](#)
25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.M.; 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. 936–944. [\[CrossRef\]](#)
26. Wada, K. Labelme: Image Polygonal Annotation with Python. Available online: <https://www.mdpi.com/authors/references> (accessed on 1 June 2022).
27. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
28. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
29. 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; Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755. [\[CrossRef\]](#)
30. Follmann, P.; König, R.; Härtinger, P.; Klostermann, M.; Böttger, T. Learning to See the Invisible: End-to-End Trainable Amodal Instance Segmentation. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1328–1336. [\[CrossRef\]](#)
31. Yao, N.; Ni, F.C.; Wu, M.H.; Wang, H.Y.; Li, G.L.; Sung, W.K. Deep Learning-Based Segmentation of Peach Diseases Using Convolutional Neural Network. *Front. Plant Sci.* **2022**, *13*, 876357. [\[CrossRef\]](#)
32. Tang, X. The role of artificial intelligence in medical imaging research. *BJR Open* **2020**, *2*, 20190031. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Blok, P.M.; van Henten, E.J.; van Evert, F.K.; Kootstra, G. Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosyst. Eng.* **2021**, *208*, 213–233. [\[CrossRef\]](#)
34. Gené-Mola, J.; Ferrer-Ferrer, M.; Gregorio, E.; Blok, P.M.; Hemming, J.; Morros, J.R.; Rosell-Polo, J.R.; Vilaplana, V.; Ruiz-Hidalgo, J. Looking behind occlusions: A study on amodal segmentation for robust on-tree apple fruit size estimation. *Comput. Electron. Agric.* **2023**, *209*, 107854. [\[CrossRef\]](#)
35. Li, Y.; Yuan, G.; Wen, Y.; Hu, E.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. EfficientFormer: Vision Transformers at MobileNet Speed. *arXiv* **2022**, arXiv:2206.01191.

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.