

Learn from Foundation Model: Fruit Detection Model without Manual Annotation

Yanan Wang¹ Zhenghao Fei^{1,2} Ruichen Li³ Yibin Ying^{1,2}

mmwang@zju.edu.cn zfei@zju.edu.cn answer0319@zju.edu.cn ybying@zju.edu.cn

¹College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, 310058, China

²ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China

³Faculty of Science, National University of Singapore, 119077, Singapore



Figure 1: SDM-D can simultaneously detect and segment input images based on the prompts, and enable distillation of knowledge from foundation models to faster, smaller models.

Abstract

Recent breakthroughs in large foundation models have enabled the possibility of transferring knowledge pre-trained on vast datasets to domains with limited data availability. Agriculture is one of the domains that lacks sufficient data. This study proposes a framework to train effective, domain-specific, small models from foundation models without manual annotation. Our approach begins with SDM (Segmentation-Description-Matching), a stage that leverages two foundation models: SAM2 (Segment Anything in Images and Videos) for segmentation and OpenCLIP (Open Contrastive Language-Image Pretraining) for zero-shot open-vocabulary classification. In the second stage, a novel knowledge distillation mechanism is utilized to distill compact, edge-deployable models from SDM, enhancing both inference speed and perception accuracy. The complete method, termed **SDM-D** (Segmentation-Description-Matching-Distilling), demonstrates strong performance across various fruit detection tasks (object detection, semantic segmentation, and instance segmentation) without manual annotation. It nearly matches the performance of models trained with

abundant labels. Notably, SDM-D outperforms open-set detection methods such as Grounding SAM and YOLO-World on all tested fruit detection datasets. Additionally, we introduce MegaFruits, a comprehensive fruit segmentation dataset encompassing over 25,000 images, and all code and datasets are made publicly available at <https://github.com/AgRoboticsResearch/SDM-D.git>.

Keywords: Fruit Detection, Foundation Models, Knowledge Distillation, Zero-Shot Learning, Agriculture

1. Introduction

Labor challenges have become a pressing issue in the agricultural sector. According to a report by the Economic Research Service of the U.S. Department of Agriculture, the sector’s labor costs reached \$42.57 billion in 2022, with nearly 10% of production expenses allocated to labor alone (Giri et al., 2023). Increasing labor shortages, especially during peak seasons, highlight the need for agricultural automation (Fei and Vougioukas, 2022), where visual perception and understanding in open-world scenarios are fundamen-

tal. For instance, accurate fruit detection and segmentation are crucial for applications like robotic fruit picking, yield estimation, and quality assessment.

In the last few years, deep learning (DL) methods have emerged as a mainstream approach in agricultural visual perception (Vougioukas, 2019), showing success in various tasks such as weed detection (Pai et al., 2024), disease identification (Hassan and Maji, 2022), and plant stress detection (Gozzovelli et al., 2021). DL approaches are especially effective in fruit detection (Zhou et al., 2024; Villacrés et al., 2023). However, training high-performance convolutional networks frequently requires vast datasets, often comprising thousands of images, and even state-of-the-art (SOTA) DL detection methods falter under limited data (Gupta et al., 2019). The computer vision community gathered a vast array of general open-source datasets, such as COCO (Lin et al., 2014) and ImageNet (Deng et al., 2009), which have become invaluable for advancing research and development in the field. Similarly, specialized open-source datasets have been developed for autonomous driving (Xu et al., 2023), and face identification (Huang et al., 2023; Lin et al., 2021). In addition to these, commercial entities also establish their own private datasets for proprietary model training. However, in the domain of agriculture, factors including environmental variability, crop seasonality, and concerns about data ownership pose challenges to the creation of large, diverse datasets. Fruit instance segmentation introduces additional difficulties, as the labor-intensive process of pixel-level labeling complicates data preparation. Although some segmentation datasets, such as MangoNet (Kestur et al.), StrawDI_Db1 (Pérez-Borrero et al., 2020), and MinneApple (Hani et al., 2020), are available (Table 1 summarizes publicly available, representative fruit datasets), they tend to contain a limited number of images. Additionally, the datasets are often highly specific to particular plant varieties, horticultural practices, lighting conditions, seasons, and even camera types, making it exceedingly difficult to train a general fruit segmentation model that can effectively generalize across different contexts (Fei et al., 2021). Therefore, developing an efficient fruit detection model generation method without the need for manual annotation is an open problem at present.

Recently, foundation models (FMs) have revolutionized the fields of natural language processing and computer vision. Typically trained on expansive datasets often through large-scale self-supervision, these models demonstrate strong generalization across various downstream tasks (Bommasani et al., 2022). For example, the Segment Anything Model (SAM) (Kirillov et al., 2023) demonstrates zero-shot segmentation that generalizes exceptionally well to unfamiliar objects and images. This is achieved by training the model on the largest segmentation dataset available at the time, which includes over 1 billion masks on 11 million

images. CLIP (Radford et al., 2021; Ilharco et al., 2021) enables the model to understand both visual and textual data through large-scale pretraining. It excels in tasks such as image captioning, visual question answering, and image-text retrieval. Despite their advantages, training FMs is extremely resource-intensive, requiring significant computational power, large datasets, and extensive energy consumption. For instance, GPT-3, an “outdated” large language model, has 175 billion parameters (Brown et al., 2020). According to Lambda Labs, the estimated cost of training GPT-3 is approximately 4.6 million US dollars and would take 355 years on a single GPU as of 2020 (Li, 2020). Similarly, CLIP requires a dataset of 400 million image-text pairs, and its largest ResNet model requires 18 days to train on 592 V100 GPUs (Radford et al., 2021). Hence, efficient utilization of FMs in appropriate downstream tasks is essential to optimize their cost-effectiveness over their lifecycle and also benefit domains with limited data availability (Hernandez and Brown, 2020). To achieve this, we aim to transfer knowledge from FMs to smaller, more efficient models, thereby enabling the use of pre-trained FMs’ knowledge in the agricultural domain. We also take into account the practical aspects of real-time edge deployment.

In this study, we introduce SDM-D, a framework designed to distill knowledge from FMs and realize the panoramic perception of complex agricultural scenes without any manual annotation. In addition, we contribute MegaFruits, a high-quality segmentation dataset aimed at advancing agricultural robotics and precision farming. The key contributions of this paper are as follows:

- We propose a novel Segmentation-Description-Matching-Distilling framework that efficiently distills agricultural-specific domain knowledge from FMs and transfers it to a small student model without the need for manual annotation.
- Comprehensive experimental evaluations demonstrate that the models distilled using our method outperform existing open-set detection methods in both speed and accuracy while achieving performance comparable to models trained on extensive manual annotations.
- We introduce a high-quality, comprehensive fruit instance segmentation dataset to advance agricultural perception, which includes 20,242 images of strawberries with 569,382 pseudo masks, 2,400 manually labeled images of yellow peaches with 10,169 masks, and 2,540 manually labeled images of blueberries with 20,656 masks. Utilizing the capabilities of our method, we are able to generate such a large scale of pseudo-segmentation labels. To our knowledge, this is the largest open dataset currently available for fruit segmentation.

-
- We have open-sourced the code used in this paper, enabling anyone to generate fruit segmentation models as needed. This code supports further research and development, facilitating the application of FMs in agricultural robotics.

2. Related Work

2.1. Open-vocabulary Detection

Traditional fruit detection methods primarily involve training closed-set detection models on specifically collected and labeled datasets (Chalapathy and Chawla, 2019). Consequently, these models can only respond to objects within a fixed set of labeled categories. Their performance is constrained by the size and quality of manually annotated datasets, and they typically cannot generalize well to unfamiliar domains. The rise of FMs has shifted the focus toward open-vocabulary object detection (OVD), which can detect objects in categories not explicitly labeled during training and generalize to unfamiliar images (Tseng et al., 2024). This shift is particularly beneficial in the field of agriculture, where labeled data are often scarce. GLIP (Li et al., 2022) and OWL-ViT (Minderer et al., 2022) leveraged CLIP’s capability to understand both visual and textual features in images, thereby expanding detection to open-vocabulary tasks. Grounding DINO (Liu et al., 2024) advanced OVD by incorporating referring expression comprehension (REC) (Qiao et al., 2021), which is crucial for scenarios where objects are described based on their properties. This advancement aids in distinguishing between objects of the same category. While these models excel in OVD, they can not handle pixel-level segmentation tasks. LLaVA-Grounding (Zhang et al., 2024) connected a large multimodal model LLaVA (Liu et al., 2023) with a grounding model to realize grounded visual chat, supporting both object and pixel-level grounding. Grounded SAM (Ren et al., 2023) presented an innovative combination of open-set detector Grounding DINO (Liu et al., 2024) with the foundation segmentation model SAM (Kirillov et al., 2023). This approach effectively addresses open-set segmentation tasks by initially conducting object detection based on the input text prompt, and then performing segmentation using the detection outputs. Similarly, YOLO-World (Cheng et al., 2024) employs CLIP (Radford et al., 2021) for text encoding within a YOLO structure (Varghese and M., 2024), achieving high inference speeds for open-set detection. However, both Grounded SAM and YOLO-World adopt a prompt-then-segment paradigm, which enhances efficiency but may reduce precision, especially in dense scenarios such as those found in agriculture. Additionally, to encode general visual-text knowledge, these models tend to be large and resource-consuming, making them difficult to deploy in real-time edge applications such as robotics.

2.2. Application of FMs in Agriculture

There are some initial studies focused on deploying FMs in agriculture. This study (Yang et al., 2024) evaluated SAM’s zero-shot segmentation on chickens using part-based and infrared thermal images, finding it outperformed SegFormer (Xie et al., 2021) and SETR (Zheng et al., 2021) in both whole and part-based chicken segmentation. This work (Williams et al., 2023) introduced ”Leaf Only SAM,” a zero-shot segmentation pipeline for potato leaves. They highlight the potential of FMs to perform effectively with minimal labeled data. This work (Li et al., 2024) fine-tuned Grounding DINO on MetaFruit for open fruit object detection, demonstrating its impressive adaptability in learning. Nevertheless, this system lacks the capability for fruit segmentation and the performance of inference speed is still limited. To the best of our knowledge, in fruit segmentation, an effective framework for training a well-performed model without manual annotation is still lacking.

2.3. Knowledge Distillation

Another related field to this study is knowledge distillation. Although FMs generalize well to unfamiliar domains and tasks, they often need substantial computational resources, making them challenging to deploy efficiently on edge devices such as robots (Ishtiaq et al., 2021). Knowledge distillation has been explored to address these issues (Kozlov et al., 2021). In knowledge distillation, a ”teacher” model transfers its knowledge to a smaller ”student” model, enabling the student to achieve comparable performance while being more resource-efficient (Hinton et al., 2015). In a typical knowledge distillation process, the student model being trained to mimic the output probabilities (or logits) of the teacher model, and a loss function is used to measure the gap between the student’s and teacher’s predictions. Additionally, Xie et al. (2020) demonstrated that distillation could be achieved by propagating pseudo-labels to unlabeled data in a self-supervised pipeline, linking knowledge distillation to pseudo-labeling without relying on output matching. This establishes an important connection between knowledge distillation and pseudo-labeling. Our work builds on this relation and extends knowledge distillation to scenarios where no manual labels are available.

3. Methodology

To efficiently extract the agricultural-specific domain knowledge for fruit segmentation from FMs and address challenges related to duplicate detections and insufficient detections in dense fruit scenes, we propose a segment-then-prompt approach named SDM. Unlike prompt-then-segment referenced in Grounded SAM (Qiao et al., 2021) and YOLO-World (Liu et al., 2023), this paradigm shifts segmentation to occur before prompting. This approach fully unleashes

Table 1: List of publicly available fruit detection datasets and our MegaFruit dataset.

Datasets	Annotation categories	Images	Instances	Labels	Task
MangoYOLO (Koirala et al., 2019)	Mango	1,730	9,067	Bounding-box	Object detection
DeepBlueberry (Gonzalez et al., 2019)	Blueberry	7	228	Mask	Instance segmentation
KFuji RGB-DS (Gené-Mola et al., 2019)	Apple	293	10,161	Bounding-box	Object detection
MetaFruit (Li et al., 2024)	Apple, orange, lemon, tangerine, grapefruit	967	12,839	Bounding-box	Object detection
MangoNet (Kestur et al.)	Mango	4,248	248,015	Bounding-box	Object detection
MinneApple (Hani et al., 2020)	Apple	49	6,799	Mask	Instance segmentation
StrawDL_Db1 (Pérez-Borrero et al., 2020)	Apple	1,001	41,325	Mask	Instance segmentation
	Strawberry	3,100	17,938	Mask	Instance segmentation
	Ripe strawberry, unripe strawberry, leaf, stem, others	20,242	569,382	Pseudo-mask	Instance segmentation
MegaFruits (ours)	Blueberry	2,540	20,656	Mask	Instance segmentation
	Peach	2,400	10,129	Mask	Instance segmentation

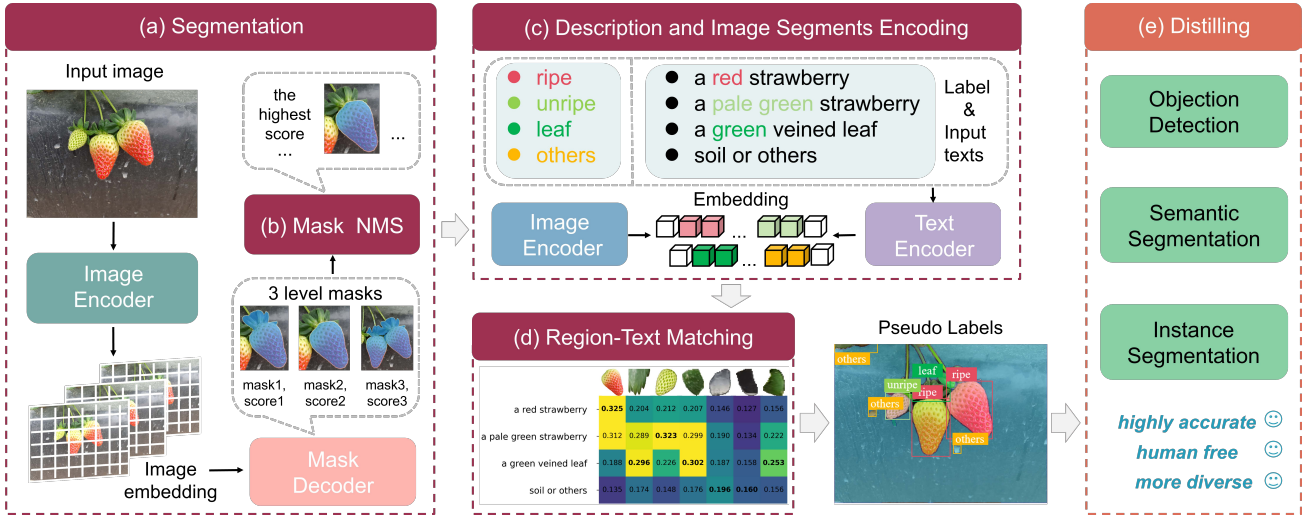


Figure 2: Overall framework of SDM-D. (a) Segmentation: is based on SAM2, utilizing the sam2_hiera_large weight with a 32×32 grid of points and no additional prompts. (b) Mask NMS: was proposed to reserve the optimal mask. (c) Description and Image Segments Encoding: OpenCLIP is used to encode the description and image segments. (d) Region-Text Matching: this is used to image regions with corresponding textual descriptions. (e) Distilling: involves transferring knowledge to smaller models that are faster and perform better. The SDM consists of sections (a), (b), (c), and (d) with pink headings, and SDM-D is a combination of SDM and section (e).

the power of the image segmentation foundation model and aligns visual and textual inputs more effectively. Additionally, we augment this method with a model distillation pipeline to further enhance both segmentation performance and runtime efficiency, particularly on resource-limited edge devices. The overall framework, as illustrated in Figure 2, is referred to as SDM-D.

(a) Segmentation. SAM2 (Ravi et al., 2024) is a prompt-driven segmentor that comprises an image encoder, a prompt encoder, and a lightweight mask decoder. The image encoder is an MAE [51] pre-trained ViT [52], encoding an input image with resolution $H \times W$ into a $\frac{H}{16} \times \frac{W}{16}$ image embedding. Given the complexity of agricultural scenes and the need for method generalization, we opt to use a 32×32 regular grid of points as prompt, rather than specific

prompts, to achieve fully automated mask generation. Each point in the grid is mapped to a 256-dimensional vectorial embedding. The two-layer mask decoder then maps the image embedding and prompt embeddings to a set of masks that correspond to potential valid objects. Since SAM2 is a model with ambiguity-awareness, for a grid that lies on a part or sub-part, it will return the top three mask outputs based on loss ranking. We keep all the segmentation results and remove the ambiguity in post-processing.

(b) Mask NMS. The segmentation step often generates multiple masks with significant overlap or redundancy, where an object can appear in multiple masks, or a single mask may cover multiple objects. This issue is particularly common in fruit images, such as strawberries with calyxes (see Figure 2(a)). In robotic operations, such inaccuracies can lead

to either damaging the fruit or failing to grasp it entirely, reducing efficiency and increasing waste. To address this, we propose a Mask NMS mechanism designed to retain the optimal mask that covers only a single fruit instance (e.g., mask2 in Figure 2(a) and eliminate ambiguity. The decision criterion is shown in Formula 1. Unlike traditional NMS, which relies on bounding box IoU, our approach calculates the overlap area between each pair of masks. If the overlap ratio, based on the smaller mask, surpasses a confidence threshold, we retain the mask with the higher score. The pseudo-code implementation of Mask NMS is provided in Algorithm 1.

$$M_1 = \begin{cases} M_2, & \text{if } \frac{|M_1 \cap M_2|}{|M_1|} > C \text{ and } S(M_2) > S(M_1) \\ M_1, & \text{else} \end{cases} \quad (1)$$

where M_i represents the area of the mask. C is the predefined confidence threshold, which was set at 0.9 in our experiment. S_i is the stability_score of the i^{th} mask output by SAM2.

Algorithm 1 Mask Non-Maximum Suppression

```

1: # masks: list of segmentation masks from SAM2
2: # areas: list of the area of each mask
3: # scores: list of the stability confidence of each mask
4: # keep: a list initialized to all True
5: def Mask_NMS(masks, threshold):
6:   for i in range(length(masks)):
7:     if not keep[i]:
8:       continue
9:     for j in range(i+1, len(masks)):
10:      if not keep[j]:
11:        continue
12:      inter = Intersection(masks[i], masks[j])
13:      smaller_area = Min(areas[i], areas[j])
14:      if inter > threshold * smaller_area:
15:        if scores[i] < scores[j]:
16:          keep[i] = False
17:        else:
18:          keep[j] = False
19:   filtered_masks = [mask for i, mask in enumerate(masks) if keep[i]]
20:   return filtered_masks

```

(c) Description and Image Segments Encoding. To facilitate seamless image-text comparisons, allowing for more accurate object recognition and classification in agricultural contexts, we utilize the open-source OpenCLIP (Ilharco et al., 2021) and emphasize the importance of REC (Qiao et al., 2021). As illustrated in Figure 2(b), instead of only inputting the labels of the instances within the image, we also include their corresponding descriptions to generate feature-rich embeddings. With its dual-encoder setup and extensive pre-training on image-text pairs, OpenCLIP adeptly handles a large vocabulary, including out-of-vocabulary words. For

instance, in open-vocabulary detection scenarios, the text embedding $t.c$ for the c^{th} object category is generated by inputting the c^{th} input description text into the text encoder. Simultaneously, the segmentation masks are transformed into image embeddings using the image encoder. This capability not only enhances the model’s flexibility but also significantly improves its performance in identifying and classifying diverse objects.

(d) Region-Text Matching. To align image regions with textual descriptions, OpenCLIP (Ilharco et al., 2021) calculates the cosine similarities between the normalized image and text embeddings. Given a batch of N (image, text) pairs, OpenCLIP can determine which of the $N \times N$ possible (image, text) pairings across a batch actually match. By jointly training an image encoder and text encoder to learn a multi-modal embedding space, OpenCLIP can maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairs. As shown in Figure 2(d), we also inherit the matrix representation from CLIP (Radford et al., 2021), providing an intuitive interpretation of the match. The label of each mask is then returned by the index of maximum similarity. This method enhances the understanding of image-text relationships, supporting advanced applications like image captioning, visual question answering, and semantic segmentation in complex agricultural environments.

(e) Distilling. To facilitate efficient deployment on edge devices, we implement distillation. We let small, edge-deployable models (students) learn from the pseudo labels generated by SDM, bypassing the need for costly manual annotation. Unlike traditional distillation, which typically operates at the feature or logit level using manually labeled data, our approach performs distillation at the label level via pseudo labels, significantly reducing labor costs and alleviating data scarcity. Interestingly, the inherent noise in pseudo-labels encourages students to acquire broader knowledge, enhancing their learning. Moreover, our method is versatile, with no restrictions on the type of student model, any compact model suited to the downstream task can undergo distillation. By transferring domain knowledge from large models to smaller ones, we create models more lightweight and adaptable to specific agricultural tasks, bridging the gap between powerful foundation models and practical small models. This approach also enables seamless deployment on edge devices, a necessity for real-time agricultural applications, while ensuring high accuracy and reducing computational demands.

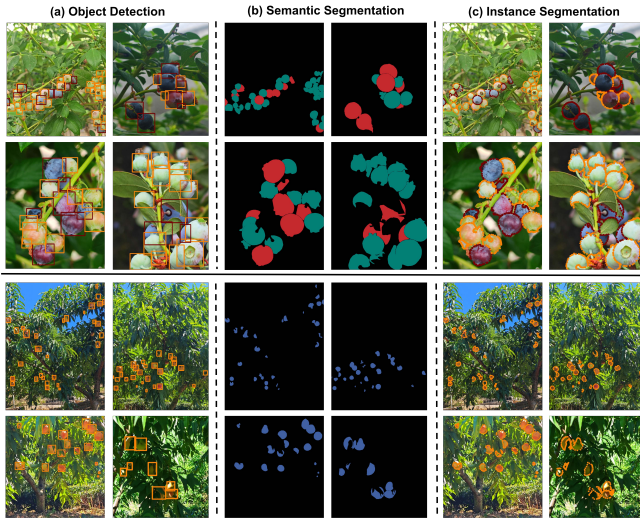


Figure 3: Representative examples of MegaFruits: (a) Object detection task: the label is the smallest rectangular box to enclose each fruit. (b) Semantic segmentation task: all the instances in an image are divided into masks and stored together in a single-channel mask image. (c) Instance segmentation task: the label is a set of polygon points around individual fruit instances.

4. Experiments and Results

4.1. Datasets and Metrics

MegaFruits Dataset. To rigorously assess our method’s performance, a comprehensive fruit segmentation dataset is essential. Such a dataset should encompass various fruit types and provide abundant segmentation masks. However, to the best of our knowledge, no publicly available dataset currently meets these criteria. To address this gap and facilitate future research, we propose the creation of the MegaFruits dataset, a large-scale, annotated segmentation dataset for fruits. This dataset was collected and annotated between October 20, 2023, and July 6, 2024, in Hangzhou, Zhejiang, China. Images were captured using an Honor Magic5 Pro smartphone and a GoPro Hero4 Black camera. The dataset presents several challenges for fruit segmentation, including varying lighting conditions, shadows, occlusions, and the presence of branches, veins, and leaves. The MegaFruits dataset comprises three subsets: Mega_Strawberry, Mega_Blueberry, and Mega_Peach. The Mega_Strawberry includes 20,242 images and 569,382 pseudo-masks generated using our SDM method. The Mega_Blueberry consists 2,540 images with 20,656 masks, and the Peach subset contains 2,400 images with 10,129 instances. To ensure unbiased evaluation, all test sets were independently collected from distinct orchard regions, separate from those used for training data. Three trained personnel carefully labeled the Mega_Blueberry and Mega_Peach using the Labelme tool (Russell et al., 2008). Annota-

tion criteria required creating precise polygonal boundaries around visible fruit, ensuring bare fruit skin was labeled while occluded parts were excluded. Instances such as background blueberries that were indiscernible due to distance, occlusion, position at the image edge, or immaturity were left unlabeled. The blueberry subset is categorized into two classes—ripe and unripe—while the peach subset includes only one class. This dataset supports object detection, semantic segmentation, and instance segmentation tasks. Example images from each subset are displayed in Figure 3. It is worth noting that the datasets used in all following experiments are detailed in Table 2, including Mega_Blueberry and Mega_Peach from our MegaFruits, as well as the StrawDI_Db1 dataset from (Pérez-Borrero et al., 2020). In the StrawDI_Db1, we unified the ripe and unripe strawberry classes into a single class. All datasets use manually annotated labels as evaluation ground truth. The Mega_Strawberry is not involved in the following experiments.

Evaluation Metrics. For object detection and instance segmentation, we adhere to the standard evaluation metrics established by COCO (Lin et al., 2014), focusing on three key metrics: $mAP_{50:95}$, mAP_{50} , and $mAR_{50:95}$. For semantic segmentation, we employ the VOC (Everingham et al., 2010) evaluation metrics, focusing on class accuracy, mIOU, and FWIOU. These evaluation criteria enable a comprehensive and rigorous assessment of model performance across different tasks.

4.2. Zero-shot Open-vocabulary Perception

In this section of the experiments, we evaluated the zero-shot open-vocabulary perception performance of the SDM method across three tasks: object detection, semantic segmentation, and instance segmentation. The SDM method does not require the training of any models; it can work directly in unfamiliar fruit domains using only descriptions of each target class. Two SOTA OVD methods Grounded SAM (Qiao et al., 2021) and YOLO-World (Liu et al., 2023), were used as comparison methods, using the same prompts as SDM. The experiments involved comparing the direct prediction output by these methods to the ground truth human annotation using the evaluation metrics mentioned in Section 4.1.

4.2.1. OBJECT DETECTION

The detection of the fruit bounding box is one of the essential tasks in agriculture, as it facilitates robotic fruit picking and allows for precise monitoring of fruit yield and maturity. Table 3 shows the object detection results of three methods across three datasets, where SDM demonstrates superior performance by a large margin across all three datasets in all metrics. This result indicates that SDM has a strong

Table 2: Distribution table for datasets used in the experiment.

Subset	StrawDI.Db1				Mega Blueberry				Mega Peach			
	Train	Val	Test	Total	Train	Val	Test	Total	Train	Val	Test	Total
Images	2,800	100	200	3,100	1,778	254	508	2,540	1,680	240	480	2,400
Instances	16,234	572	1,132	17,938	12,898	1,802	5,956	20,656	6,349	980	2,800	10,129



Figure 4: Comparison of object detection results on a strawberry image.

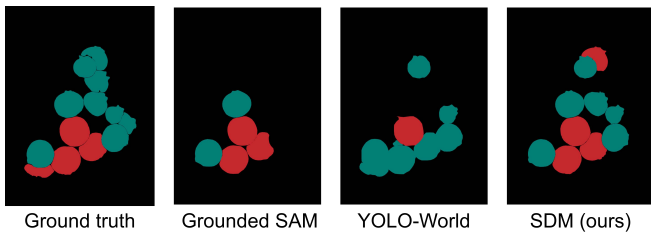


Figure 5: Comparison of semantic segmentation results.

understanding of the general concept of fruits, allowing it to effectively detect fruits in a zero-shot setting, even though the models within SDM have never been trained on these specific domains. A visual comparison of object detection results using the three methods is shown in Figure 4.

4.2.2. SEMANTIC SEGMENTATION

Semantic segmentation provides pixel-level classification which has a wide range of applications in agricultural perceptions. Table 4 presents three key metrics for semantic segmentation. Our algorithm achieves the best performance across all three datasets in all metrics. A visual comparison of semantic segmentation results is shown in Figure 5.

4.2.3. INSTANCE SEGMENTATION

Instance segmentation offers higher precision and fine-grained details for complex agricultural tasks by accurately predicting the boundaries of each fruit, which is essential for downstream tasks. Table 5 presents all evaluation metrics of instance segmentation of all methods, the results indicate our proposed SDM method consistently outperforms other methods in all metrics by a large margin. In terms of $mAP_{50:95}$, SDM achieving 2.21, 1.59, and 2.59 times the performance of the second-best algorithms on strawberry, blueberry, and peach datasets. This substantial improvement across all metrics underscores SDM’s capability complex general fruit scenes at the level of individual fruits and to



Figure 6: Comparison of instance segmentation results.

capture fine-grained boundary details. A visual comparison of instance segmentation results is shown in Figure 6.

4.2.4. PANORAMIC PERCEPTION

Unlike traditional segmentation models, which are restricted to predicting a limited set of specific categories predefined during training, SDM’s capability for general understanding of objects theoretically allows it to segment any object based on given prompts. In this section, we prompt the models to segment all potentially interested objects in three types of fruit scenes, including ripe and unripe fruits, leaves, stems, flowers, and the background. Since the dataset does not include ground truth labels for panoramic segmentation, no quantitative results are included in this section. Fig.7 presents the original images and zero-shot segmentation results of in strawberry, blueberry, and peach domains. For clarity, the background class mask is not shown. As illustrated in the figure, SDM demonstrates robust panoramic segmentation capabilities, accurately identifying multiple instances with minimal overlap. This advantage is evident compared to Grounded SAM and YOLO-World, which, due to their prompt-then-segment approach, often assign multiple labels to the same region or miss detection of some obvious fruits. This is particularly highlighted in Figure 7(c) and 7(d), where a significant number of the instances are redundantly segmented while some are missed. However, the SDM method also has limitations. As observed in Figure 7(b), SDM struggles to detect objects with unclear boundaries, such as leaves, in the peach images.

To further explore the generalization ability of SDM, we conducted panoramic segmentation experiments on various fruit types beyond strawberry, blueberry, and peach. As shown in FFigure 8, SDM consistently shows striking zero-shot segmentation performance, indicating it has strong general-

Table 3: Results of object detection. The **best** and second-best results for each evaluation metric are highlighted in **bold** and underlined, respectively.

Methods	StrawDI_Db1			Mega Blueberry			Mega Peach		
	mAP ₅₀	mAP _{50:95}	mAR _{50:95}	mAP ₅₀	mAP _{50:95}	mAR _{50:95}	mAP ₅₀	mAP _{50:95}	mAR _{50:95}
Grounded SAM	<u>0.216</u>	<u>0.322</u>	0.401	0.232	0.251	0.356	0.190	0.228	0.220
YOLO-World	0.173	0.232	<u>0.429</u>	<u>0.233</u>	<u>0.314</u>	<u>0.517</u>	<u>0.287</u>	<u>0.410</u>	<u>0.335</u>
SDM (Ours)	0.540	0.635	0.639	0.411	0.462	0.633	0.524	0.596	0.686

Table 4: Results of semantic segmentation. The **best** and second-best results for each evaluation metric are highlighted in **bold** and underlined, respectively.

Method	StrawDI_Db1			Mega Blueberry			Mega Peach		
	ClassAcc.	mIOU	FWIOU	ClassAcc.	mIOU	FWIOU	ClassAcc.	mIOU	FWIOU
Grounded-SAM	<u>0.936</u>	<u>0.832</u>	<u>0.959</u>	0.614	<u>0.553</u>	<u>0.829</u>	0.683	0.644	0.858
YOLO-World	0.863	0.768	0.944	<u>0.621</u>	0.536	0.820	<u>0.901</u>	<u>0.875</u>	0.948
SDM (Ours)	0.959	0.917	0.981	0.813	0.760	0.901	0.914	0.882	0.951

ization ability across diverse fruit categories. This suggests that our model can play a critical role in zero-shot agricultural fruit perception at the pixel level, potentially benefiting downstream tasks such as robotic harvesting, yield monitoring, and orchard management. These tasks, which require fruit identification, can be accomplished without the need for human annotation or additional model training.

4.3. Distilled Edge-deployable Models

One of the key innovations of this work is the distillation of domain knowledge from large, computationally intense foundation models to smaller, edge-deployable student models. The student models were trained using pseudo labels generated by SDM, Grounded SAM, and YOLO-World, while reference models utilized the same architecture as the student models but were trained using manually annotated labels. The following experiments compare the distilled student, trained from these three zero-shot perception methods, with the reference models in object detection, semantic segmentation and instance segmentation across three fruit datasets.

4.3.1. OBJECT DETECTION

For the object detection task, we selected YOLOv8s and EfficientDet-D2 as student model architectures. As shown in Table 6, the models distilled from SDM outperform those from Grounded SAM and YOLO-World by a significant margin, regardless of the student model architecture. Moreover, the performance of the student models, which do not utilize any manual annotations, is comparable to that of the reference models trained with manual labels. For example, the mAP_{50:95} of the best student models reached 84.8%, 86.9%, and 90.6% of the corresponding reference models in

the strawberry, blueberry, and peach domains, respectively.

4.3.2. SEMANTIC SEGMENTATION

For the semantic segmentation task, we selected DeepLabv3+ as the student model architecture, and the performance of the distilled models is reported in Table 7. SDM outperformed two comparison algorithms across all metrics on three datasets, achieving mIOU rates of 98.6%, 87.5%, and 96.7% in the strawberry, blueberry, and peach datasets, respectively, compared to the manually labeled baselines.

4.3.3. INSTANCE SEGMENTATION

For the instance segmentation task, YOLOv8s was selected as the student model architecture. From Table 8, the mAP_{50:95} of the distilled model from SDM reached 85.8%, 88.7%, and 66.6% of the corresponding baselines in the strawberry, blueberry, and peach datasets, respectively. Notably, on the blueberry dataset, which contains two classes, the mAP_{50:95} of our algorithm surpasses Grounded SAM by over 2.21 times and exceeds YOLO-World by over 1.47 times. This highlights our algorithm’s superior performance, particularly in handling tasks with similar object descriptions, further proving its robustness and effectiveness in instance segmentation.

4.3.4. FOUNDATION MODEL VERSUS DISTILLED MODEL

In this section, we compare the performance of the directly using foundation models with the distilled student model to emphasize the impact of distillation. There are two key aspects we are interested in: inference efficiency and perception accuracy.

Table 5: Results of instance segmentation. The **best** and second-best results for each evaluation metric are highlighted in **bold** and underlined, respectively.

Methods	StrawDI_Db1			Mega Blueberry			Mega Peach		
	mAP ₅₀	mAP _{50:95}	mAR _{50:95}	mAP ₅₀	mAP _{50:95}	mAR _{50:95}	mAP ₅₀	mAP _{50:95}	mAR _{50:95}
Grounded SAM	0.247	0.376	0.438	0.234	0.255	0.355	0.175	0.212	0.212
YOLO-World	0.119	0.222	0.239	<u>0.258</u>	<u>0.319</u>	<u>0.511</u>	0.163	<u>0.234</u>	<u>0.622</u>
SDM (Ours)	0.548	0.632	0.666	0.411	0.461	0.633	0.454	0.565	0.634

Table 6: Results of object detection. The manual labels trained results (*), serving as the baselines, are marked with an asterisk. The **best** and second-best results except the baseline for each evaluation metric are highlighted in **bold** and underlined, respectively. During training, the original hyper-parameters of YOLOv8s and EfficientDet were used without modification to maintain consistency.

Teacher	Student model	StrawDI_Db1			Mega Blueberry			Mega Peach		
		mAP _{50:95}	mAP ₅₀	mAR _{50:95}	mAP _{50:95}	mAP ₅₀	mAR _{50:95}	mAP _{50:95}	mAP ₅₀	mAR _{50:95}
Manual*		0.826	0.937	0.846	0.781	0.880	0.844	0.781	0.921	0.843
Grounded SAM	YOLOv8s	<u>0.369</u>	<u>0.542</u>	<u>0.620</u>	0.395	0.441	0.544	<u>0.542</u>	<u>0.716</u>	<u>0.641</u>
YOLO-World		0.352	0.469	0.618	<u>0.397</u>	<u>0.534</u>	<u>0.616</u>	0.421	0.645	0.471
SDM (Ours)		0.701	0.836	0.743	0.679	0.785	0.817	0.708	0.840	0.801
Manual*		0.738	0.879	0.778	0.731	0.865	0.846	0.668	0.822	0.779
Grounded SAM	Efficient-Det-D2	<u>0.306</u>	<u>0.536</u>	0.541	<u>0.361</u>	0.432	<u>0.678</u>	<u>0.480</u>	<u>0.607</u>	<u>0.603</u>
YOLO-World		0.272	0.453	0.554	<u>0.357</u>	<u>0.535</u>	0.594	0.365	0.584	0.442
SDM (Ours)		0.640	0.776	0.699	0.551	0.658	0.799	0.643	0.794	0.741

Table 7: Results of semantic segmentation. The manual labels trained results (*), serving as the baselines, are marked with an asterisk (*). The **best** and second-best results except the baselines for each evaluation metric are highlighted in **bold** and underlined, respectively. During training, the original hyper-parameters of DeepLabv3+ were used without modification to maintain consistency.

Teacher	StrawDI_Db1			Mega Blueberry			Mega Peach		
	Class Acc.	mIOU	FWIOU	Class Acc.	mIOU	FWIOU	Class Acc.	mIOU	FWIOU
Manual*	0.980	0.959	0.989	0.927	0.865	0.912	0.963	0.929	0.973
Grounded SAM	<u>0.945</u>	<u>0.835</u>	<u>0.952</u>	0.603	0.533	0.711	0.866	0.838	0.940
YOLO-World	0.875	0.786	0.938	0.666	<u>0.583</u>	0.751	0.942	<u>0.897</u>	0.960
SDM (Ours)	0.966	0.946	0.986	0.830	0.757	0.848	0.949	0.898	0.961

Table 8: Results of instance segmentation of the distilled models. The manual labels trained results (*), serving as the baselines, are marked with an asterisk (*). The **best** and second-best results except the baseline for each evaluation metric are highlighted in **bold** and underlined, respectively. During training, the original hyper-parameters of YOLOv8 were used without modification to maintain consistency.

Teacher	StrawDI_Db1			Mega Blueberry			Mega Peach		
	mAP50	mAP50:95	mAR50:95	mAP50	mAP50:95	mAR50:95	mAP50	mAP50:95	mAR50:95
Manual*	0.796	0.944	0.950	0.767	0.905	0.833	0.901	0.919	0.853
Grounded SAM	<u>0.453</u>	<u>0.657</u>	0.627	0.307	0.42	0.551	0.505	0.709	0.63
YOLO-World	0.221	0.413	<u>0.634</u>	<u>0.461</u>	<u>0.564</u>	<u>0.572</u>	<u>0.548</u>	<u>0.739</u>	<u>0.651</u>
SDM (Ours)	0.684	0.848	0.949	0.68	0.802	0.725	0.600	0.809	0.745

Table 9: Comparison of inference time and GPU memory allocation for each method. In bold is the optimal results.

Method	Grounded SAM	YOLO-World	SDM	SDM-D (YOLOv8s)
Inference Time (ms)	8,090.81	99.32	7,615.08	18.96
GPU-Memory Allocation (MiB)	7,602	2,268	6,650	878



Figure 7: Comparison of zero-shot open-vocabulary segmentation results. (a) The original images. (b) The prediction results of SDM. (c) The prediction results of Grounded SAM. (d) The prediction results of YOLO-World. On the far left of each row is a diagram of categories and colors. The mask and bounding box of the same category have the same color.

Inference Efficiency. Consideration of inference speed and the runtime resource requirement are the main reasons for distilling knowledge from foundation models to smaller student models. We compared the inference time and runtime GPU memory allocation of the best SDM-D student models, utilizing the YOLOv8s architecture, against methods that directly employ foundation models for the task of instance segmentation. This evaluation was conducted on StrawDLdb1 images using an NVIDIA RTX 3090 GPU. We performed 1,000 inferences and calculated the average results, as shown in Table 9. Unsurprisingly, the student model exhibited the fastest inference speed, achieving a reduction of over 99.7% in inference time compared to the direct using Grounded SAM and SDM. It also demonstrated an 80.9% reduction in inference time compared to YOLO-World, an inference-time optimized method. Regarding runtime GPU memory allocation, the distilled SDM-D student

model also had the lowest GPU memory usage, consuming only 11.5%, 13.2%, and 38.7% relative to the Grounded SAM, SDM, and YOLO-World respectively. The result indicates that the SDM-D student model is significantly more efficient in terms of both inference speed and GPU memory footprint compared to directly using foundation models, facilitating its deployment on edge devices and enabling real-time inference.

Accuracy. To evaluate the impact of distillation on perception performance, we compared the distilled student models to their teacher models across all the tasks. The results are shown in Figure 9. Surprisingly, the process of distilling foundation models into smaller student models not only accelerates inference speed but also improves the perception performance in the target domain, which is also distinct from the traditional knowledge distillation method with a

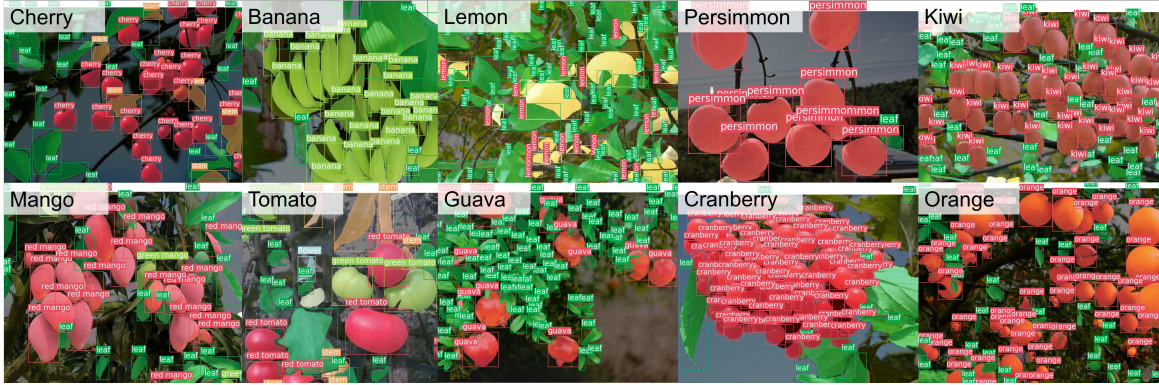


Figure 8: The results of SDM’s zero-shot open-vocabulary segmentation on various fruit images.

distillation loss. Figure 9 illustrates that all distilled models outperform their foundation model teachers in three evaluated tasks. Interestingly, this phenomenon is observed not only in the SDM method proposed in this paper but also in Grounded SAM (Ren et al., 2023) and YOLO-World (Cheng et al., 2024). We refer to this improvement as a distillation improvement. While the distillation improvement is minor in some tasks, such as semantic segmentation (as shown in Figure 9b), it is quite significant in object detection and instance segmentation, as shown in Figure 9a and 9c. One possible explanation for this distillation improvement is that the distillation process introduces information from target domain images (albeit without any human labeling) into the models, thereby narrowing down the overall prediction space of the foundation models. Another possible explanation is that the distillation model training process helps to average out the noisy, incorrect pseudo labels, allowing the student model to develop a robust perception capability. Similar improvements have also been observed by Xie (2020).

4.4. Fine-tuning with Few Manual Labels

Although the SDM-D method can generate fruit perception models without any manual annotations, the performance of the generated model still lags behind reference models trained with extensive manual annotation, as we presented in Table 6, 7, 8. Despite the imperfection of the SDM method, a class label can be interpreted in various ways, resulting in different labeling criteria. For example, "strawberry" can refer to the fruit with or without the calyx. To align the model’s intent with the ground truth labels and correct other sources of error, fine-tuning the model is a viable approach.

The models distilled without any manual label are already well-trained base models capable of extracting features from images effectively. Their performance can be further enhanced when limited labels are available. Few-shot learning involves refining the model using a small number of labeled samples. Figure 10 presents the experimental results com-

paring the best-distilled student model, fine-tuned using a few manually labeled images, to the models trained from scratch using manually labeled images. The model architectures used here are the YOLOv8s object detection model and the YOLOv8s instance segmentation model. The tasks involved are object detection and instance segmentation, and the dataset used for evaluation is StrawDI.Db1. We used the $mAP_{50:90}$ metric for both tasks.

For the distilled student models, we fine-tuned them with 1, 50, and 100 labeled training images, randomly selected from the training set. Results were averaged over 10 trials to mitigate the dependency on training data selection. The models trained from scratch used between 1 and 2,900 labeled images, with the first 50 images labeled in increments of 5 and the remaining images in increments of 50. Images were randomly selected from the full training dataset to ensure a comprehensive comparison. The experimental results are presented in Figure 10. Figure 10 shows that the performance of the initial distilled student models using SDM-D method were already comparable to the same models trained from scratch with 200 manually labeled samples for both tasks. One-shot fine-tuning allowed the distilled model to reach 91.6% and 91.8% of the performance of a model trained with 2,900 labels in object detection and instance segmentation, respectively. To achieve comparable performance of our 1-shot and 50-shot models in instance segmentation, a purely manually trained model would require an additional 250 and 1,050 labeled images, respectively. This underscores the remarkable data efficiency of our approach, where the rich features embedded in the pseudo-labels significantly enhance performance with minimal manual labeling, offering substantial support for reducing human labor.

4.5. Mask NMS

Mask NMS is one of the innovations introduced in this work, designed to eliminate the mask ambiguity and overlapping produced by SAM2. We conducted ablation experiments on mask NMS for object detection and instance segmen-

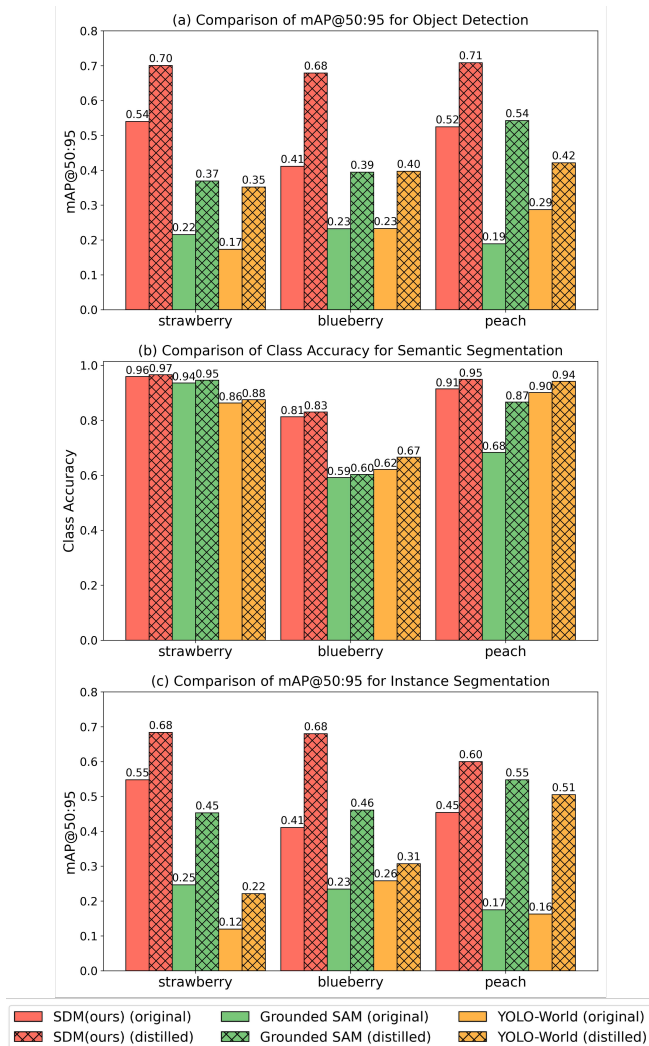


Figure 9: Comparison of foundation models and distilled models. (a) Comparison of mAP@50:95 for object detection. (b) Comparison of Class Accuracy for semantic segmentation. (c) Comparison of mAP@50:95 for instance segmentation.

tation across three datasets listed in Table 2. As shown in Figure 11, applying Mask NMS consistently improved the mAP_{50:95} scores for all three datasets when predicted by SDM, demonstrating the effectiveness of Mask NMS within the SDM pipeline.

4.6. Prompts Design

According to the experiment, we find that the design of prompts greatly affects the model performance. According to our statistics, about 80% of labeling errors come from the region-text matching step. We summarize an effective prompt template: a/an color shape object with feature. Among them, the color description is the most crucial. It can be seen from Figure 12(a) and 12(b) that the wrong clas-

sification of backgrounds similar to fruits can be avoided through reasonable design of prompt words. Figure 12(c) shows the all labels of segmentation, highlighting some errors indicated by the orange arrows. Although this error can be avoided by adding a new description (e.g., "black background"), considering the generality of the entire dataset, we didn't do that. Regarding the design of the number of prompt texts, we recommend that readers consider the characteristics of objects within the entire scene. While an excessive number of prompts may lead to higher accuracy, it can adversely affect the model's generalization ability, rendering it less suitable for large-scale datasets and requiring a lot of time and effort.

5. Conclusion

This paper presents an innovative framework, SDM-D. The primary contribution of this work is the establishment of a comprehensive framework that leverages the knowledge within pre-trained foundation models for fruit perception and distills this knowledge into edge deployable models that excel in both speed and accuracy. Experimental results demonstrate that our method performs remarkably well in multiple perception tasks across various fruit scenes, surpassing SOTA OVD methods. The distilled student model achieves satisfying perception performance without any manual annotation, reaching over 86.6% of the performance compared to the reference model trained using extensive labeled images in the instance segmentation task at the strawberry scene. And further reference over 91.8% of the reference model's performance with 1-shot fine-tuning.

This approach has broad implications for agricultural perception. By utilizing our framework, significant annotations can be saved, thereby reducing both the cost and time required to develop high-performance fruit perception models. This advancement accelerates the development and deployment of agricultural robots, enhancing efficiency and scalability in tasks such as fruit monitoring and harvesting. Furthermore, we believe this approach holds potential applications beyond agriculture, extending to fields such as healthcare, autonomous driving, and robotics, where open vocabulary segmentation is required. We also present a high-quality dataset, MegaFruits, which includes over 25,000 annotated images of strawberries, peaches, and blueberries, making it the largest open fruit segmentation dataset. We hope this resource can advance research and applications in agricultural perception.

There are still some limitations in SDM-D. While the distilled models for all experimental tasks are competent, they do not yet fully match the accuracy of the reference models trained on extensive human-labeled datasets. And the distilled student models with high inference speed, show limited adaptability compared to their foundation model teach-

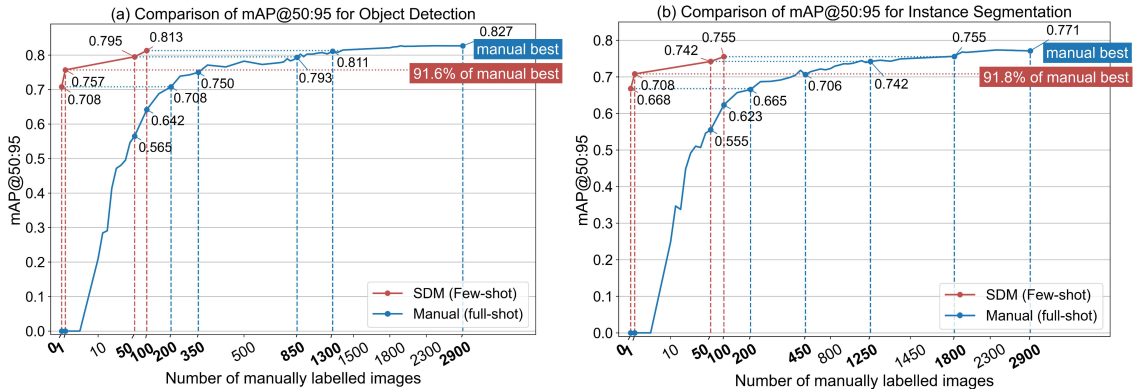


Figure 10: Student model training results with the amount of training data image on StrawDI dataset. (a) Comparison of mAP@50:95 for object detection. (b) Comparison of mAP@50:95 for instance segmentation. The red dot lines represent the few-shot learning results of the model generated by SDM-D and it starts with the zero-shot result of training on the pseudo labels generated by SDM. The blue dot lines represent the training with purely manually labeled data. During the training we incremented by 5 images per step for the first 50, and by 50 images per step from 50 to 2,900.

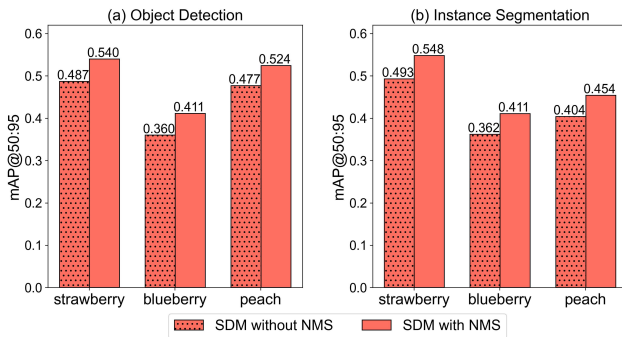


Figure 11: The ablation results of mask NMS. (a) mAP@50:95 of the generated labels for object detection. (b) mAP@50:95 of the generated labels for instance segmentation. The bars filled with spots represent the foundation models, and the bars filled with grid lines represent the distilled models.

ers, requiring re-distillation as environmental conditions or task requirements change. While the overall zero-shot performance of SDM is satisfying, it occasionally misses detail structures or introduce small, disconnected components that should belong to a single object. SDM is designed for generality and does not require any training, which contributes to its convenience but may impact accuracy when handling domains requiring specialized annotations. Future research can focus on further improving perception accuracy, enhancing the adaptability of the student small models to various fruit domains and automatic prompt design.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bern-

stein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL <https://arxiv.org/abs/2108.07258>.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie

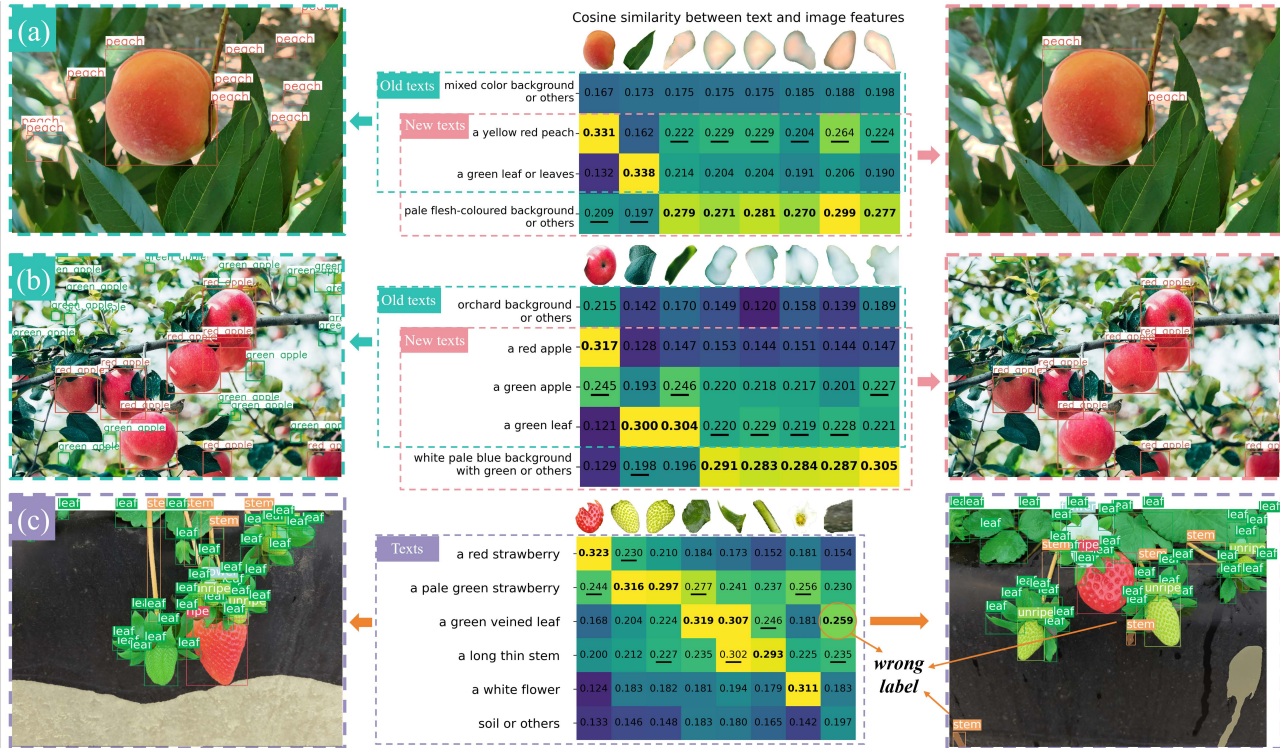


Figure 12: Influence of prompts on label assignment. (a) and (b) show the influence of color prompt words on label classification. (c) Examples of matching between segmentation masks and prompts of strawberry. The numbers in bold represent the highest similarity. The numbers with an underline is the second-highest similarity. The figures on the left are the results of old prompts, and on the right are the results of new prompts.

Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 2020-December, 2020. URL <https://api.semanticscholar.org/CorpusID:218971783>.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *ArXiv*, abs/1901.03407, 2019. URL <https://api.semanticscholar.org/CorpusID:57825713>.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xingang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024.

doi: 10.48550/arXiv.2401.17270. URL <https://doi.org/10.48550/arXiv.2401.17270>.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL <https://ieeexplore.ieee.org/document/5206848>.

Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88, 2010. ISSN 09205691. doi: 10.1007/s11263-009-0275-4. URL <https://doi.org/10.1007/s11263-009-0275-4>.

Zhengkao Fei and Stavros Vougioukas. Row-sensing templates: A generic 3d sensor-based approach to robot localization with respect to orchard row centerlines. *Journal of Field Robotics*, 39, 2022. ISSN 15564967. doi: 10.1002/rob.22072. URL <https://onlinelibrary.wiley.com/doi/full/10.1002/rob.22072>.

Zhengkao Fei, Alex Olenskyj, Brian N. Bailey, and Mason Earles. Enlisting 3d crop models and GANs for more

- data efficient and generalizable fruit detection. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1269–1277. IEEE, 2021. ISBN 978-1-66540-191-3. doi: 10.1109/ICCVW54120.2021.00147. URL <https://ieeexplore.ieee.org/document/9607528/>.
- Jordi Gené-Mola, Verónica Vilaplana, Joan R Rosell-Polo, Josep-Ramon Morros, Javier Ruiz-Hidalgo, and Eduard Gregorio. Kfuji rgb-ds database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected ir data. *Data in brief*, 25:104289, August 2019. ISSN 2352-3409. doi: 10.1016/j.dib.2019.104289. URL <https://europepmc.org/articles/PMC6685673>.
- Anil K. Giri, Dipak Subedi, Jonathan Law, Tatiana Borisova, , and Marcelo Castillo. Agricultural labor expenses forecast to increase by almost 2 percent in 2023, 2023. URL <http://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=107320>.
- Sebastian Gonzalez, Claudia Arellano, and Juan E. Tapia. Deepblueberry: Quantification of blueberries in the wild using instance segmentation. *IEEE Access*, 7:105776–105788, 2019. doi: 10.1109/ACCESS.2019.2933062. URL <https://ieeexplore.ieee.org/document/8787818>.
- Riccardo Gozzovelli, Benjamin Franchetti, Malik Bekmurat, and Fiora Pirri. Tip-burn stress detection of lettuce canopy grown in plant factories. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1259–1268, 2021. doi: 10.1109/ICCVW54120.2021.00146. URL <https://ieeexplore.ieee.org/document/9607809>.
- Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, 2019. doi: 10.1109/CVPR.2019.00550. URL <https://ieeexplore.ieee.org/document/8954457>.
- Nicolai Hani, Pravakar Roy, and Volkan Isler. Minneapolis: A benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters*, 5(2): 852–858, April 2020. ISSN 2377-3774. doi: 10.1109/lra.2020.2965061. URL <http://dx.doi.org/10.1109/LRA.2020.2965061>.
- Sk Mahmudul Hassan and Arnab Kumar Maji. Plant disease identification using a novel convolutional neural network. *IEEE Access*, 10:5390–5401, 2022. doi: 10.1109/ACCESS.2022.3141371. URL <https://ieeexplore.ieee.org/document/9674894>.
- Danny Hernandez and Tom B. Brown. Measuring the algorithmic efficiency of neural networks, 2020. URL <https://arxiv.org/abs/2005.04305>.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL <https://api.semanticscholar.org/CorpusID:7200347>.
- Baojin Huang, Zhongyuan Wang, Kui Jiang, Qin Zou, Xin Tian, Tao Lu, and Zhen Han. Joint segmentation and identification feature learning for occlusion face recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10875–10888, 2023. doi: 10.1109/TNNLS.2022.3171604.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. URL https://github.com/mlfoundations/open_clip.
- Arhum Ishtiaq, Sara Mahmood, Maheen Anees, and Neha Mumtaz. Model compression, 2021. URL <http://arxiv.org/abs/2105.10059>. version: 1.
- Ramesh Kestur, Avadesh Meduri, and Omkar Narasipura. Mangonet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. 77:59–69. ISSN 09521976. doi: 10.1016/j.engappai.2018.09.011. URL <https://linkinghub.elsevier.com/retrieve/pii/S0952197618301970>.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.00371. URL <https://ieeexplore.ieee.org/document/10378323>.
- Anand Koirala, Kerry Walsh, Zhenglin Wang, and C. McCarthy. Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of ‘mangoyolo’. *Precision Agriculture*, 20, 12 2019. doi: 10.1007/s11119-019-09642-0. URL <https://doi.org/10.1007/s11119-019-09642-0>.
- Alexander Kozlov, Ivan Lazarevich, Vasily Shamporov, Nikolay Lyalyushkin, and Yury Gorbachev. Neural network compression framework for fast model inference. In *Lecture Notes in Networks and*

-
- Systems*, volume 285, 2021. doi: 10.1007/978-3-030-80129-8_17. URL https://doi.org/10.1007/978-3-030-80129-8_17.
- Chuan Li. Openai’s gpt-3 language model: A technical overview, 2020. URL <https://lambdalabs.com/blog/demystifying-gpt-3>.
- Jiajia Li, Kyle Lammers, Xunyuan Yin, Xiang Yin, Long He, Renfu Lu, and Zhaojian Li. Metafruit meets foundation models: Leveraging a comprehensive multi-fruit dataset for advancing agricultural foundation models. *ArXiv*, abs/2407.04711, 2024. URL <https://api.semanticscholar.org/CorpusID:271051066>.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022. doi: 10.1109/CVPR52688.2022.01069.
- Hong Lin, Rita Tse, Su-Kit Tang, Yanbing Chen, Wei Ke, and Giovanni Pau. Near-realtime face mask wearing recognition based on deep learning. In *2021 IEEE 18th Annual Consumer Communications and Networking Conference (CCNC)*, pages 1–7, 2021. doi: 10.1109/CCNC49032.2021.9369493.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL <https://api.semanticscholar.org/CorpusID:14113767>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2023. ISSN 10495258. URL <https://dl.acm.org/doi/10.5555/3666122.3667638>.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*, 2024. URL <https://eccv.ecva.net/virtual/2024/poster/395>.
- Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13670 LNCS, 2022. doi: 10.1007/978-3-031-20080-9_42.
- Deepthi G Pai, Radhika Kamath, and Mamatha Balachandra. Deep learning techniques for weed detection in agricultural environments: A comprehensive review. *IEEE Access*, 12:113193–113214, 2024. doi: 10.1109/ACCESS.2024.3418454.
- Isaac Pérez-Borrero, Diego Marín-Santos, Manuel E. Gegúndez-Arias, and Estefanía Cortés-Ancos. A fast and accurate deep learning method for strawberry instance segmentation. 178:105736, 2020. ISSN 01681699. doi: 10.1016/j.compag.2020.105736. URL <https://linkinghub.elsevier.com/retrieve/pii/S0168169920300624>.
- Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2021. doi: 10.1109/TMM.2020.3042066.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL <https://arxiv.org/abs/2408.00714>.
- Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, He Cao, Kunchang Li, Jiayu Chem, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded-sam: Detect, segment and generate anything. In *Proceedings of the 2023 International Conference on Computer Vision (ICCV 2023)*, Paris, France, October 2-6 2023. URL <https://hdl.handle.net/1783.1/135998>.
- Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-

-
- based tool for image annotation. *International Journal of Computer Vision*, 77, 2008. ISSN 09205691. doi: 10.1007/s11263-007-0090-8.
- Yu-Wen Tseng, Hong-Han Shuai, Ching-Chun Huang, Yung-Hui Li, and Wen-Huang Cheng. Language-guided negative sample mining for open-vocabulary object detection. In *2024 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4, 2024. doi: 10.1109/ICEIC61013.2024.10457133.
- Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. doi: 10.1109/ADICS58448.2024.10533619.
- Juan Villacrés, Michelle Viscaïno, José Delpiano, Stavros Vougioukas, and Fernando Auat Cheein. Apple orchard production estimation using deep learning strategies: A comparison of tracking-by-detection algorithms. 204:107513, 2023. ISSN 0168-1699. doi: 10.1016/j.compag.2022.107513. URL <https://www.sciencedirect.com/science/article/pii/S0168169922008213>.
- Stavros G. Vougioukas. Agricultural robotics. *Annu. Rev. Control. Robotics Auton. Syst.*, 2:365–392, 2019. URL <https://api.semanticscholar.org/CorpusID:219600694>.
- Dominic Williams, Fraser Macfarlane, and Avril Britten. Leaf only sam: A segment anything pipeline for zero-shot automated leaf segmentation. *ArXiv*, abs/2305.09418, 2023. URL <https://api.semanticscholar.org/CorpusID:258714696>.
- Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems*, 2021. URL <https://api.semanticscholar.org/CorpusID:235254713>.
- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020. doi: 10.1109/CVPR42600.2020.01070.
- Runsheng Xu, Hao Xiang, Xu Han, Xin Xia, Zonglin Meng, Chia Ju Chen, Camila Correa-Jullian, and Jiaqi Ma. The openca open-source ecosystem for cooperative driving automation research. *IEEE Transactions on Intelligent Vehicles*, 8, 2023. ISSN 23798858. doi: 10.1109/TIV.2023.3244948.
- Xiao Yang, Haixing Dai, Zihao Wu, Ramesh Bahadur Bist, Sachin Subedi, Jin Sun, Guoyu Lu, Changying Li, Tianming Liu, and Lilong Chai. An innovative segment anything model for precision poultry monitoring. *Computers and Electronics in Agriculture*, 222:109045, 2024. ISSN 0168-1699. doi: 10.1016/j.compag.2024.109045. URL <https://www.sciencedirect.com/science/article/pii/S0168169924004368>.
- Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang Chunyuan Li, and Jainwei Yang. Llava-grounding: Grounded visual chat with large multimodal models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLIII*, page 19–35, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72774-0. doi: 10.1007/978-3-031-72775-7_2. URL https://doi.org/10.1007/978-3-031-72775-7_2.
- Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886, 2021. doi: 10.1109/CVPR46437.2021.00681.
- Wei Zhou, Yifan Cui, Hongpu Huang, Haitian Huang, and Chen Wang. A fast and data-efficient deep learning framework for multi-class fruit blossom detection. 217:108592, 2024. ISSN 0168-1699. doi: 10.1016/j.compag.2023.108592. URL <https://www.sciencedirect.com/science/article/pii/S0168169923009808>.