American Society of Agricultural and Biological Engineers
2950 Niles Road | St. Joseph MI 49085-9659 | USA
269.429.0300 | fax 269.429.3852 | hq@asabe.org | www.asabe.org

An ASABE Meeting Presentation
DOI: https://doi.org/10.13031/aim.202401398
Paper Number: 2401398

Advancing Orchard Fruit Detection: An Innovative Agricultural Foundation Model Approach

Jiajia Li[1], Kyle Lammers[1], Xunyuan Yin[2], Xiang Yin[3], Long He[4], Renfu Lu[5], Zhaojian Li[6]

[1]Department of Electrical and Computer Engineering, Michigan State University, MI, USA

[2]School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore

[3]Department of Automation and Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Shanghai, China

[4]Department of Agricultural and Biological Engineering, Pennsylvania State University, USA

[5]United States Department of Agriculture Agricultural Research Service, East Lansing, MI, USA

[6]Department of Mechanical Engineering, Michigan State University, East Lansing, MI, USA

Written for presentation at the
2024 ASABE Annual International Meeting
Sponsored by ASABE
Anaheim, CA
July 28-31, 2024

ABSTRACT. *Fruit harvesting poses a significant labor and financial burden on the fruit industry, which underscore the urgent need for advancements in robotic harvesting solutions. Despite considerable progress in leveraging deep learning and machine learning techniques for fruit detection, a common shortfall is the inability to swiftly extend the developed models across different orchards and/or various fruit species. Additionally, the limited availability of pertinent data further compounds these challenges. In this work, we introduce MetaFruit, the largest publicly available multi-class fruit dataset, comprising 4,248 images and 248,015 manually labeled instances across diverse U.S. orchards. Furthermore, this study proposes an innovative open-set fruit detection system leveraging advanced Vision Foundation Models (VFMs) for fruit detection that can adeptly identify a wide array of fruit types under varying orchard conditions. This system not only demonstrates remarkable adaptability in learning from minimal data through few-shot learning but also shows the ability to interpret human instructions for subtle detection tasks. The performance of the developed foundation model is comprehensively evaluated across several metrics, outperforming existing state-of-the-art algorithms in our MetaFruit, thereby setting a new benchmark in the field of agricultural technology and robotic harvesting. The MetaFruit dataset (https://www.kaggle.com/datasets/jiajiali/metafruit) and detection framework (https://github.com/JiajiaLi04/FMFruit) are open-sourced to foster future research in vision-based fruit harvesting, marking a significant stride toward addressing the urgent needs of the agricultural sector.*

Keywords. *Fruit harvesting, fruit detection, foundation model*

Introduction

Fruit harvesting is a significant labor and financial burden in modern orchards, accounting for over 10 million worker hours each year and about 15% of the overall production costs in the U.S. (Bergerman et al., 2015; Chu et al.,2021, 2023). Moreover, fruit growers are struggling with an increasing labor shortage due to a declining interest in agricultural work (Zhou et al., 2022). This situation is further aggravated by recent international travel restrictions caused by public health crises, such as the COVID-19 pandemic, and geopolitical tensions, like the Russia-Ukraine conflict, significantly reducing agricultural productivity by limiting the availability of migrant labor (Ben Hassen and El Bilali, 2022; Laborde et al., 2020). Therefore, there is a critical need for the innovation of robotic harvesting technologies to mitigate labor shortages, minimize human injury risks, and boost the efficiency and economic viability of the fruit industry (Chu et al., 2021; Zhou et al., 2022).

The perception system is essential in harvesting robots, as it enables the identification of fruits within the target area and guides the robot in executing subsequent tasks (Chu et al., 2021; Zhao et al., 2016). More recently, deep learning (DL) based approaches have rapidly evolved and attracted significant attention in various agricultural sectors, such as plant disease identification (Xu et al., 2022), weed detection (Li et al., 2024a; Rai and Sun, 2024), and plant breeding (Li et al., 2024b). These DL methods have also proven effective in fruit detection (Koirala et al., 2019; Ukwuoma et al., 2022; Xiao et al., 2023). For instance, Faster-RCNN (Girshick, 2015) has been successfully applied for apple (Fu et al., 2020; Gao et al., 2020), Kiwifruit (Fu et al., 2018), and multiple fruits detection (mangoes, almonds and apples) (Bargoti and Underwood, 2017). In addition, YOLO models (Terven and Cordova-Esparza, 2023) are also applied for fruit detection and recognition such as apple (Tian et al., 2019), mango (Shi et al., 2020), orange (Mirhaji et al., 2021), and cherry (Gai et al., 2023). In our previous research, state-of-theart DL techniques based on Mask-RCNN (He et al., 2017) and Faster RCNN (Girshick, 2015) are also developed for accurate apple detection for dense orchard settings (Chu et al., 2021, 2023). Despite the aforementioned successes, developing DL models from scratch faces several challenges. Firstly, it relies heavily on large, accurately annotated image datasets, which are generally costly to obtain (Chen et al., 2022). Secondly, the training phase is remarkably time intensive and demands significant computational resources (LeCun et al., 2015). Moreover, while these specialized models excel in their designated tasks, they often encounter difficulties when applied to novel scenarios, such as different orchard conditions or fruit species, demonstrating limited capabilities in generalization (Kamilaris and PrenafetaBoldú, 2018).

It is widely acknowledged that a comprehensive set of annotated images are essential for the development of highperforming DL models in visual fruit detection tasks (Sun et al., 2017). In Lu and Young (2020), the authors provide an overview of various publicly accessible fruit image datasets aimed at robotic harvesting. For instance, mango-related datasets, such as MangoNet (Kestur et al., 2019) and MangoYOLO (Koirala et al., 2019) contain 49 and 1730 images for mango segmentation and detection, respectively. There are specialized apple datasets for apple detection, including KFuji RGB-DS (Gené-Mola et al., 2019) WSUApple (Bhusal et al., 2019), LFuji-air dataset (Gené-Mola et al., 2020a), and MinneApple (Häni et al., 2020), along with two apple datasets from our previous studies (Chu et al., 2021, 2023). Additionally, DeepBlueberry (Gonzalez et al., 2019) is a dataset including 294 images for blueberry detection. However, most of these datasets are species-specific and not transferable to different fruit types. Recently, there's been an increasing interest in multi-fruit datasets. For instance, FruitNet (Meshram and Patil, 2022) and Fruit3601 feature 19,500 and 41,322 images across 5 and 80 fruit species, respectively, catering to fruit classification tasks. In terms of fruit detection, OrchardFruit (Bargoti and Underwood, 2017) and DeepFruits (Sa et al., 2016) provide open-source access to 3,232 and 587 images for 3 and 7 fruit species, respectively. Yet, these datasets are typically designed for specific orchard environments with less dense fruit clusters. Table 1 summarizes these datasets, providing an overview of the resources available for advancing research in fruit detection.

Lately, the rise of large pre-trained models, commonly known as foundation models (FMs), such as ChatGPT-4 (Achiam et al., 2023), Segment Anything Model (SAM) (Kirillov et al., 2023), have demonstrated outstanding performance in both language and vision tasks across diverse domains (Bommasani et al., 2021; Li et al., 2023). These models undergo extensive training on diverse datasets spanning multiple domains and modalities. Once fully trained, they exhibit the capability to perform a range of tasks requiring minimal fine-tuning and without extensive reliance on task-specific labeled data. There has been growing interest in applying FMs within the field of agriculture, offering innovative solutions and insights. As an example, Yang et al. (2023) employs SAM for chicken segmentation tasks in a zero-shot manner, integrating part-based segmentation and the use of infrared thermal imagery. The experimental findings reveal that SAM outperforms other vision foundation models (VFMs) like SegFormer and SETR in accuracy for both whole and partial chicken segmentation. Williams et al. (2023) introduce "Leaf Only SAM", an automatic leaf segmentation pipeline designed for zero-shot segmentation of potato leaves. Compared to a fine-tuned Mask R-CNN model tailored for annotated potato leaf datasets, this innovative approach demonstrates superior effectiveness. These developments underscore the potential of FMs in various agricultural applications. However, to the best of our knowledge, FMs have not yet been applied to fruit harvesting tasks involving multiple fruit classes.

In this study, we introduce a comprehensive multiclass fruit dataset (also named MetaFruit), gathered from commercial orchards across various U.S. states during the growth seasons of 2022 and 2023. Building on this, we develop an innovative open-set fruit detection system, leveraging the power of advanced vision FMs (VFMs) to identify a wide range of fruits. This paper's contributions are significant and can be summarized as follows:

1. We introduce a uniquely comprehensive and diverse fruit dataset, including 4,248 images with 248,015 manually labeled fruit instances, meticulously collected from commercial orchard fields across multiple U.S. states.
2. We propose a novel FM-based open-set fruit detection framework designed for multi-class fruit detection, which is not only capable of identifying various and novel types of fruit but also integrates the ability to process human language inputs.
3. Comprehensive experiments have been conducted to rigorously assess the performance of our proposed framework FMFruit on our dataset MetaFruit.
4. Both curated dataset and developed software are open-sourced, making them accessible for further research and engineering integration in vision-based fruit harvesting and related fields.

Table 1: List of publicly available fruit datasets and our new MetaFruit dataset.

Datasets	Fruit Variety	Modality	# Imgs	# Instances	Tasks
MangoNet (Kestur et al., 2019)	Mango	RGB	49	-	Fruit segmentation
MangoYOLO (Koirala et al., 2019)	Mango	RGB	1,730	9,067	Fruit detection
DeepBlueberry (Gonzalez et al., 2019)	Blueberry	RGB	293	10,161	Fruit detection
StrawDI_Db1 (Pérez-Borrero et al., 2020)	Strawberry	RGB	3,100	17,938	Fruit instance segmentation
KFuji RGB-DS (Gené-Mola et al., 2019)	Apple	RGB-D	967	12,839	Fruit detection
WSUApple (Bhusal et al., 2019)	Apple	RGB	2,298	-	Fruit detection
Fuji-SfM (Gené-Mola et al., 2020b)	Apple	RGB	288	1,455	Fruit detection
LFuji-air dataset (Gené-Mola et al., 2020)	Apple	LiDAR	-	-	Fruit detection
MinneApple (Häni et al., 2020)	Apple	RGB	1,001	41,325	Fruit detection and segmentation
OrchardFruit (Bargoti and Underwood, 2017)	Apple, mango, almond	RGB	3,232	-	Fruit detection
DeepFruits (Sa et al., 2016)	Strawberry, rockmelon, otrange, mango, capsicum, avocado, apple	RGB	587	-	Fruit detection
FruitNet (Meshram and Patil, 2022)	Apple, banana, guava, lime, orange, and pomegranate	RGB	>19,500	-	Fruit quality classification
Fruit360 (https://www.kaggle.com/datasets/moltean/fruits)	80 classes of fruits	RGB	41,322	-	Fruit classification
MSUAppleDataset (Ours) (Chu et al., 2021)	Apple	RGB	1,500	19,528	Fruit detection
MSUAppleDatasetv2 (Ours) (Chu et al., 2023)	Apple	RGB	1,246	14,518	Fruit detection
Ours	Apple, orange, lemon, tangerine, grapefruit	RGB	4,248	248,015	Fruit detection

Materials and Methods

In this section, we first present our collected dataset, MetaFruit, and the VFMs used for multi-class fruit detection. We then detail the few-shot learning, evaluation metrics, and experimental setups employed in our study.

MetaFruit dataset

The multi-class fruit dataset, MetaFruit, introduced in this study is collected utilizing advanced imaging technology, comprising both a high-definition camera and a sophisticated LiDAR system (with a resolution of 1920 × 1080), from commercial orchards in North Michigan and California, USA. To guarantee a diverse and varied collection of images that enhance model robustness (Lu and Young, 2020), the dataset includes images taken under natural field lighting conditions

across various weather scenarios (e.g., sunny, cloudy, and overcast) during the peak harvest season of the fruit growth stage. The dataset contains 4,247 images, featuring five distinct fruit types: apples, oranges, lemons, grapefruits, and tangerines. **Figure 1** shows representative samples for each fruit category. Unlike existing datasets, MetaFruit is characterized by more realistic/complex orchard environments with fruits frequently appearing in clusters, presenting a challenging yet realistic scenario for model training and evaluation. Notably, the dataset also includes multiple varieties within each fruit category. For example, the apple class includes both red and green species, adding another layer of diversity and complexity to the dataset.

The images acquired for the MetaFruit dataset are meticulously labeled by trained personnel. These annotators utilized the Labelme (Wada, 2011) tool to accurately draw bounding boxes around individual fruit instances in the images. This meticulous process results in the acquisition of 248,015 manually labeled bounding boxes. The distribution of the MetaFruit dataset is detailed in Table 2. Overall, the dataset exhibits an even distribution among apples, oranges, lemons, and tangerines, each with a similar number of images, whereas grapefruits are represented with slightly fewer images, totaling 490. Tangerines are particularly well represented in the dataset with 1,063 images and 85,785 labeled instances, averaging 81 bounding boxes per image. The average number of bounding boxes per image sheds light on the density of fruits captured in the images, whereas the average size of these instances provides insight into the physical size of the objects. Notably, the smaller the size of the instances, the greater the challenge in detecting them accurately. Interestingly, while the lemon class does not have the highest average number of bounding boxes per image, it features the smallest average size of instances (823 pixels per instance), indicating lemons' smaller physical presence within the images, which presents its unique detection challenges.

The MetaFruit dataset, in terms of instance numbers, significantly surpasses previous collections, being more than 10 times larger than the dataset for multi-class fruit species featured in OrchardFruit (Bargoti and Underwood, 2017) (as shown in Table 1). To the best of our knowledge, it represents the most extensive publicly available dataset for fruit detection specifically designed for commercial orchard systems, establishing a new benchmark for research and development in agricultural technology and robotic harvesting.

Table 2: Statistics of MetaFruit dataset.

	# Imgs	# Bboxes	# Avg. bboxes/images	# Avg. size/instances	Region
Apple	812	62,040	76	1,193	Michigan & California
Orange	926	45,834	49	1,176	California
Lemon	958	42,238	44	823	California
Grapefruit	490	12,118	25	2,232	California
Tangerine	1,062	85,785	81	1,068	California
Total	4,248	248,015	58	1,133	Michigan & California

VFMs for fruit detection

In recent years, DL approaches have made significant strides in advancing fruit detection models. Prominent among the object detectors employed are FCOS (Tian et al., 2020), Faster-RCNN (Girshick, 2015), and YOLO series (Terven and Cordova-Esparza, 2023), all of which are designed as closed-set detectors. Such models operate under the assumption that the categories of objects to be detected are predefined and known during both the training and testing phases, thereby limiting their capacity to recognize previously unseen categories. Furthermore, these approaches depend on extensive, meticulously labeled image datasets—a process that is both labor-intensive and demands significant resources. In contrast, recent focus has shifted towards openset object detection (Geng et al., 2020) and the exploration of LLMs and FMs (Bommasani et al., 2021; Li et al., 2023). These open-set detectors are capable of not only precisely detecting the known classes but also efficiently handling the unknown ones. Therefore, the language data needs to be added for model training to solve the situation that a testing sample comes from some unknown classes. Similarly, LLMs and FMs, which are trained on extensive datasets covering a wide range of domains and modalities, demonstrate a remarkable ability to perform a variety of open-set tasks after training, which is achieved with minimal fine-tuning and reduced reliance on extensive, task-specific labeled data.

Figure 1: Representative examples of MetaFruit dataset, including five fruit classes: (a) apple, (b) orange, (c) lemon, (d) grapefruit, and (e) tangerine.

To facilitate open-set fruit detection across a diverse array of fruit categories, this study employs a vision foundation model (VFM), specifically the Grounding DINO (Liu et al., 2023) model, for the detection task. Grounding DINO is an open-set detector predicated on the DETR-like architecture, DINO (Zhang et al., 2022), which integrates endto-end Transformer-based detection mechanisms. A pivotal aspect of enabling open-set detection capabilities is the integration of linguistic elements for the generalization of unseen objects. This approach involves training the model on existing bounding box annotations, augmented through language generalization, to facilitate the identification of a broader array of objects beyond those seen during training.

The overall workflow and architectural design are illustrated in **Figure 2**. Initially, the process involves extracting fundamental features from both images and text through respective image and text backbones, i.e., the Swin Transformer (Liu et al., 2021) module. These foundational features serve as inputs to a feature enhancer network dedicated to the fusion of cross-modality features, facilitating a comprehensive integration of image and textual information. Following the acquisition of enriched cross-modality text and image features, the system employs a language-guided query selection module (Liu et al., 2023) to meticulously select cross-modality queries based on the image features, thereby harnessing the synergistic potential of linguistic cues and visual data. This selection process mirrors the transformative approach of integrating diverse modalities to enhance detection precision and contextual understanding through the strategic alignment of textual and visual elements. Subsequently, these cross-modal queries are introduced into a cross-modal decoder to extract and refine the desired features from the combined bimodal information, continually updating its parameters to reflect the insights gained from the cross-modal analysis. Ultimately, the decoder's output queries are used to predict object bounding boxes and identify relevant textual phrases, culminating in a sophisticated system capable of precise object detection and association with appropriate linguistic descriptors. The loss function is defined as

$$L = L_1 + L_{GIOU} + L_{Cons} \quad (1)$$

where L_1 and GIOU (Rezatofighi et al., 2019) are utilized for the regression of bounding boxes. The contrastive loss, L_{Cons}, is also incorporated as in GLIP (Li et al., 2022), to fine-tune the classification of predicted objects and language tokens (Liu et al., 2023; Zhang et al., 2022).

The Grounding DINO model (Liu et al., 2023) leverages the foundational DINO architecture (Zhang et al., 2022),

and to save computational resources and training time, the Grounding DINO model is transferred from DINO weights instead of training from scratch (Zhuang et al., 2020). The DINO model is trained on the O365 data (Shao et al., 2019), which is a large-scale object detection dataset containing 365 categories and 2 million images. Based on the pre-trained DINO weights, the grounding DINO with swin-transformer tiny backbone is trained on a combined data set including O365, GoldG, and Cap4M, where GoldG contains images in Flickr30k entities (Plummer et al., 2015) and Visual Genome (Krishna et al., 2017), and Cap4M is from (Li et al., 2022) but not publicly available. Similarly, the grounding DINO with Swin-transformer large backbone is also transferred from DINO, but with more data (e.g., O365, GoldG, Cap4M, OI (Krasin et al., 2017), RefCOCO/+/g (Kazemzadeh et al., 2014), and COCO). To tailor the Grounding DINO model for the specific task of detecting a wide array of fruits in open-set conditions, we conducted fine-tuning using our MeteFruit dataset based on the pre-trained Grounding DINO weights.

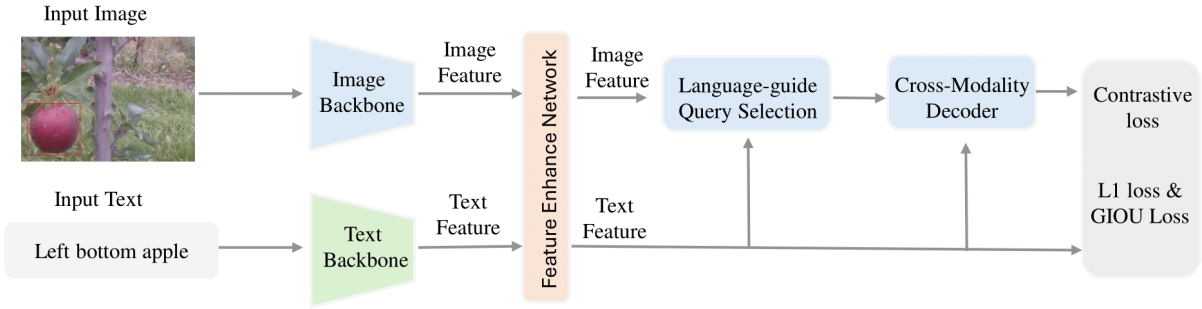

Figure 2: The framework of the VFM for fruit detection based on the Grounding DINO (Liu et al., 2023) model

Few-shot learning

Contemporary fruit detection algorithms, while yielding promising results, often struggle to generalize across varying data distributions, such as different fruit classes and orchard settings, especially when faced with a lack of extensive data (Wang et al., 2020). The scarcity of data can be attributed not only to the inherent challenges of the task or privacy issues but also to the significant costs associated with data preparation, including collection, preprocessing, and labeling. In response to these challenges, few-shot learning has gained recognition as a promising learning method, demonstrating the significant potential for quickly learning underlying patterns from merely a few or even zero samples (Song et al., 2023). Zero-shot transfer learning refers to scenarios where no training samples are utilized, and models are directly deployed on testing images, aiming to make accurate predictions based solely on their pre-existing knowledge and capabilities. On the other hand, few-shot learning involves using a minimal number of samples to refine and adjust the models. For example, in 5-shot learning, precisely five samples are employed for model fine-tuning. It is important to note that while few-shot learning allows models to adapt to new tasks with limited data, the performance of such models, when only a few samples are used for fine-tuning, can sometimes be constrained. The effectiveness of the fine-tuning process is heavily dependent on the quality and representativeness of the selected samples, their alignment with the task at hand, and the model's inherent ability to generalize from minimal information (Song et al., 2023; Wang et al., 2020). This delicate balance between sample selection and model adaptability is critical for maximizing the potential of few shot learning approaches in diverse application scenarios, including those within the domain of fruit detection where variability across classes and environments is high.

In this study, we employ few-shot learning frameworks to evaluate the generalizability of the FMFruit model across various fruit categories. Specifically, the zero-shot learning scenario is utilized by deploying the FMFruit model on new fruit classes without any model fine-tuning. Concurrently, for the few-shot learning experiments, a minimal number of samples are randomly selected from these new fruit categories to slightly adjust the model.

Evaluation metrics

The performance of DL models in fruit detection tasks is rigorously evaluated using key detection accuracy metrics, such as Average Precision (AP), mean Average Recall (mAR), and mean Average Precision (mAP) (Dang et al., 2023). These metrics collectively offer a detailed assessment of a model's proficiency in both identifying and precisely locating fruits within images. AP, with a specific focus on precision at a 50% overlap threshold (AP50), and mAP, which calculates the average precision across a range of overlap thresholds (from 0.5 to 0.95, in increments of 0.05), together provide insights into the precision aspects of model performance. Meanwhile, mAR evaluates the model's recall capabilities over a spectrum of Intersection over Union (IoU) ranging from 0.5 to 1.0, thereby gauging the model's effectiveness in capturing the true positive detections across various conditions.

Experimental setups

Extensive experiments are conducted based on the following four settings:

- Zero-shot transfer, few-shot learning, and fine-tuning on our MetaFruits.
- Cross-class generalization ability evaluation by finetuning with four kinds of fruits and evaluating on the remaining novel one.
- Case study of language-referring object detection.

We have Swin-T (Liu et al., 2021) as the image backbone of FMFruit model. Following BERT-base (Devlin et al., 2018), Hugging Face (Wolf et al., 2019) is used as the text backbone. All the models are trained for 100 epochs with the AdamW optimizer. The learning rate is set to be 1e-4 with the weight decay as 0.0001, but the learning rate for the image and text backbone is set to be 1e-5. To expedite the model training process, we leverage transfer learning based on pre-trained DINO and pre-trained Grounding DINO (Zhuang et al., 2020). The fine-tuning procedure involves using a batch size of 4 over 100 epochs, and we utilize the PyTorch framework (version 1.10.1) (Paszke et al., 2019). Both the training and testing phases of the models take place on a server running Ubuntu 20.04. This server is equipped with two GeForce RTX 2080Ti GPUs, each offering 12GB of GDDR6X memory.

Results

In this section, we first evaluate the zero-shot and few-shot transfer learning performance of FMFruit in comparison with leading-edge fruit detection algorithms on our MetaFruit dataset. Then, we examine its ability of crossclass generalization and evaluate its effectiveness on other publicly available fruit datasets. Lastly, we present initial findings on its capability to integrate text inputs and comprehend referring expressions.

Few-shot fruit detection performance

In this subsection, we examine the zero-shot and fewshot transfer learning capabilities of our proposed model across five distinct fruit types from our MetaFruits data. We compare our model's performance with that of leading object detection models, including Fully Convolutional OneStage (FCOS) object detector (Tian et al., 2020), FasterRCNN (Girshick, 2015), RetinaNet (Lin et al., 2017), and RTMDet (Lyu et al., 2022). The performance comparison is presented in **Table 3**. It's noteworthy that traditional CNNbased models such as FCOS, despite being trained on the comprehensive COCO dataset (Lin et al., 2014), which encompasses 80 categories including apples and oranges, fail to achieve any positive mAP and mAR scores in fruit detection tasks across all fruit classes. This highlights a critical limitation of conventional object detection algorithms, which struggle with generalization across diverse datasets and are typically fine-tuned for narrow, specific detection scenarios. Among the baseline models, RTMDet emerges as one of the best-performing models following comprehensive training across all evaluated fruit types in terms of mAP and mAR metrics, while RetinaNet is observed to lag behind the rest of the baseline models in performance.

Conversely, our foundation model-based fruit detection model, FMFruit, demonstrates exceptional zero-shot transfer performance across all evaluated fruit classes. Notably, for FMFruit, two out of the five fruit classes achieve a mAP score exceeding 30, alongside a mAR score surpassing 46 across all types of fruits. Cumulatively, FMFruit yields an overall 28.7 mAP and a mAR of 46.9 across all fruit classes, underlining its impressive capability to accurately detect and identify a wide range of fruits without specific prior training in those classes. FMFruit's performance on apples shows a specific challenge, achieving a zero-shot 24.1 mAP score. This performance can be attributed to the presence of densely clustered fruits, with an average of 76 apples per image, as detailed in **Table 2**. Similarly, the model's detection capability for lemons, which achieves a 29.4 mAP score, highlights the difficulty in accurately identifying fruits that occupy very small areas within images, with the average size being only 823 pixels per lemon, as also indicated in **Table 2**.

In few-shot learning scenarios, FMFruit exhibits promising performance across all fruit classes. It achieves impressive overall mAP and mAR scores of 46.7 and 52.7 in the 1-shot setting, using just a single image per fruit class for fine-tuning. Expanding to a 5-shot scenario, where five images per class are used for retraining, FMFruit maintains excellent performance, akin to the full-shot setting where all available images are used for fine-tuning. FMFruit also demonstrates excellent few-shot performance in individual fruit classes. Specifically, FMFruit yields a significant improvement in performance on the apple class, with a remarkable 78.99 improvement, increasing the AP50 from 45.5 to 81.8 under the 1-shot setting. Moreover, with the 10-shot setting, it achieves a 96.5 improvement in performance, further illustrating the model's impressive ability to rapidly adapt and excel with minimal training data.

Figure 3 presents examples of detection outputs achieved by the FMFruit model under various few-shot configurations. These visualizations underscore FMFruit's robust open-set detection capabilities, particularly highlighting its impressive performance in zero-shot settings where the model undergoes no fine-tuning. This illustrates FMFruit's inherent ability to generalize and accurately identify fruits even without direct prior exposure to specific fruit class data. It is noted that in certain cases, such as with lemons, the model under the zero-shot setting may miss some fruit instances due to occlusion and the small size of the fruits, as illustrated in **Figure 3** (c). However, the model's detection capabilities are significantly enhanced through fine-tuning with just one single image (i.e., 1-shot), by effectively adapting to address challenges

associated with fruit occlusion and small size.

Table 3: Zero-shot and few-shot performance on our MetaFruits dataset.

		Apple			Orange			Lemon			Grapefruit			Tangerine		
		mAP	AP50	mAR	mAP	AP50	mAR	mAP	AP50	mAR	mAP	AP50	mAR	mAP	AP50	mAR
Retinanet	Full-shot	38.0	67.4	42.9	41.2	70.6	46.0	37.8	68.3	43.6	43.8	81.3	51.0	37.1	62.1	40.6
Faster-rcnn		48.4	78.4	53.3	50.9	83.1	55.6	46.7	78.6	52.7	51.3	86.8	56.5	43.9	70.3	47.4
FCOS		49.8	80.9	55.3	52.5	85.1	58.4	47.5	80.1	54.2	54.5	90.3	61.5	45.8	71.4	49.7
RTMDet		52.4	81.5	58.0	53.5	83.5	59.4	49.0	79.8	55.7	60.3	91.2	66.9	46.9	71.0	50.4
FMFruit	Zero-shot	24.1	45.7	**46.1**	36.6	68.9	**52.2**	29.4	52.1	**47.3**	37.2	64.4	**52.7**	29.6	60.3	**43.8**
	1-shot	**45.5**	**81.8**	**55.5**	**45.9**	**81.3**	**54.7**	**37.5**	**72.3**	**49.3**	**48.6**	**83.1**	**59.7**	**42.0**	**83.6**	**47.3**
	5-shot	48.0	85.2	56.5	48.4	82.7	56.8	42.7	76.3	52.6	47.0	81.9	58.8	38.6	79.2	46.2
	10-shot	53.2	89.8	59.3	52.4	85.9	59.9	48.4	81.2	56.0	54.6	88.8	62.9	44.9	87.8	49.3
	20-shot	55.3	91.3	60.9	54.4	87.4	61.7	49.8	82.9	57.4	57.8	91.0	65.6	46.7	90.3	50.9
	Full-shot	59.4	**94.1**	64.7	60.1	**92.0**	66.5	56.0	**88.0**	62.6	64.0	**94.7**	70.9	50.4	**93.7**	54.4

Figure 3: Zero-shot and few-shot fruit detection visualization examples for (a) apple, (b) orange, (c) lemon, (d) grapefruit, and (e) tangerine. The bounding box confidence threshold is set as 0.2 and 0.3 for zero-shot and few-shot, respectively. Best view via zoom in.

Performance of cross-class generalization

In this subsection, we evaluate the cross-class generalization capability of FMFruit to assess the impact of training on existing fruit classes on the detection performance of an unseen fruit class. Specifically, in this evaluation, the model is first trained on four fruit classes and subsequently tested on the fifth, unseen class. For instance, to test the model's generalization

capability to detect lemons with cross-variety training data of other fruits, the model is first fine-tuned using data from oranges, apples, grapefruits, and tangerines, and then tested for its ability to detect lemons, class not seen during training. This assessment helps us understand FMFruit's adaptability and effectiveness in recognizing new fruit types based on learned features from other fruit classes.

Table 4 summarizes the performance of FMFruit across three distinct training settings: zero-shot, where the model receives no training on any of the five fruit classes; crossclass, where the model is trained on four fruit classes and evaluated on the fifth, unseen class; and full-shot, where the model undergoes training on the specific fruit classes. The data clearly demonstrate the efficacy of cross-class training in enhancing fruit detection capabilities. Specifically, crossclass training significantly boosts detection performance by 98.9, with an AP50 improvement from 45.7 to 90.9, nearly matching the performance in the full-shot setting, which achieves an AP50 of 92.7. This outcome underscores the potential of cross-class training to effectively prepare models for recognizing new fruit types.

Table 4: Cross-class generalization performance

	Apple			Orange			Lemon			Grapefruit			Tangerine		
	mAP	AP50	mAR	mAP	AP50	mAR	mAP	AP50	mAR	mAP	AP50	mAR	mAP	AP50	mAR
Zero-shot	24.1	45.7	46.1	36.6	68.9	52.2	29.4	52.1	47.3	37.2	64.4	52.7	29.6	60.3	43.8
Cross-class	53.0	90.2	59.4	58.0	89.4	64.3	52.0	83.8	59.6	60.8	90.1	71.1	47.8	92.2	52.0
Full-shot	59.4	94.1	64.7	60.1	92.0	66.5	56.0	88.0	62.6	64.0	94.7	70.9	50.4	93.7	54.4

Performance of Referring Expression Comprehension (REC)

In this subsection, we present an initial evaluation of our FMFruit model's ability to REC. The model is tasked with processing human instructions provided in natural language, identifying the critical elements of these instructions, and selecting features that accurately correspond to the described text.

Figure 5 shows the REC results. The first illustrative set involves the model detecting apples with minimal occlusion, guided by the specific instruction "apple with less occlusion". FMFruit demonstrates proficiency in accurately isolating and excluding apples that are heavily occluded by leaves, adhering closely to the given instructions. The second example demonstrates the model's ability to filter out apples occluded by branches, following the instruction "apple without occlusion by branch". Unsurprisingly, FMFruit exhibits exceptional adaptability by focusing detection on apples without branch occlusion. These scenarios highlight FMFruit's precise interpretation and execution based on specific linguistic instructions, underscoring its sophisticated ability to utilize referring expressions for enhanced fruit detection accuracy.

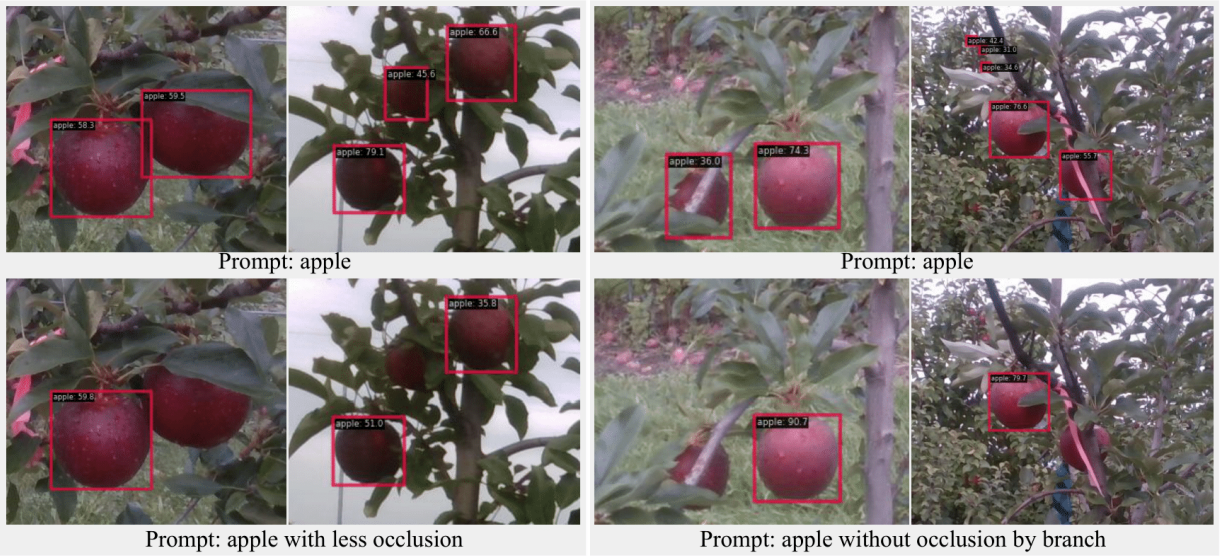

Figure 5: Visualization examples of referring object detection.

Discussion

Fruit detection is both a widely studied research topic and a practical challenge. Traditional DL methods have shown considerable success, yet they tend to be specialized for certain fruit types and specific scenarios, limiting their applicability to new orchard environments and different fruit classes. In response to this limitation, our study delves into the potential of

VFMs to tackle a wider range of fruit detection challenges. Additionally, we introduce the MetaFruit dataset, encompassing 248,015 labeled instances across five fruit classes, to support and enhance the development and evaluation of advanced fruit detection models. Despite its contributions, this study acknowledges certain limitations, paving the way for future enhancements as elaborated below.

Challenges in real-world deployment

Implementing FMs in agricultural applications introduces some challenges, particularly regarding inference speed and model size which often require significant computing resources (Bommasani et al., 2021). As shown in **Table 6**, our proposed FMFruits has the largest inference time, which limits the deployment of FM in many on field agricultural settings, as the downstream tasks often require immediate action based on the model's outputs. For example, after outputting the fruit location, the fruit-picking system needs to implement other actions immediately, such as decision-making and path planning. In addition, the complexity and size of FMs demand large computing resources and memory bandwidth, which is not practical for real-world deployment.

Recent researchers have tackled challenges through model optimization, notably model compression techniques. These methods, including quantization, knowledge distillation, and pruning, significantly shrink model size and speed up inference without sacrificing performance. For instance, SqueezeLLM's post-training quantization framework achieves lossless compression, enhancing quantization performance under memory constraints (Kim et al., 2023). Edge computing strategies further accelerate inference by processing data near its source, exemplified in agriculture by drones or field sensors enabling on-site decision-making, as seen in MobileSAM (Kirillov et al., 2023). This model distills knowledge from heavy encoders to lightweight ones, achieving a 12ms inference time and 9.66M parameters, a remarkable improvement over the original SAM's 456ms inference time and 615M parameters.

Table 6: Model inference time

Model	FPS (imgs/s)	Inference time per image (ms
Retinanet	21.9	45.7
Faster RCNN	20.2	49.5
FCOS	18.9	52.9
RTMDet	53.2	18.8
FMFruit	5.5	181.3

Integration of LLMs

The realm of Large Language Models (LLMs) and Foundation Models (FMs) has advanced remarkably, finding applications in various fields like ChatGPT, robotics, and agriculture. Preliminary investigations suggest promising integration of LLMs and FMs into farming technologies, enhancing agricultural practices. Section 3.4 explores Referring Expression Comprehension, getting detection outcomes through human instructions. Using a language-guided query selection method (Liu et al., 2023), it aligns features with input text, promising more precise detection. However, it requires well-organized and labeled pairs for training, which is time-intensive and complex. Integrating mature LLM and FM developer APIs, like OpenAI's ChatGPT API, presents exciting possibilities. Human-robot interaction (HRI) offers another opportunity, integrating LLMs and FMs into fruit-harvesting robots for enhanced comprehension of natural language instructions (Wang et al., 2024), improving adaptability in orchard environments.

Conclusions

Fruit detection is a pivotal component in the development of robotic fruit harvesting systems. Central to successful fruit detection is the assembly of a substantial, accurately labeled fruit dataset and the subsequent development of robust DL models. This paper introduces, to date, the most extensive fruit detection dataset pertinent to U.S. commercial orchards, encompassing 4,248 images across 5 fruit classes, annotated with a total of 248,015 bounding boxes, gathered under diverse natural field lighting conditions. Additionally, we have developed an innovative open-set fruit detection system that utilizes the advanced capabilities of VFMs to identify a wide range of fruits. This model exhibits outstanding detection performance in both zero-shot and few-shot learning scenarios, consistently surpassing the FOCS network. Furthermore, the model effectively demonstrates cross-class generalization capabilities by being trained on known fruit classes and then tested on unseen classes, showcasing its exceptional open-set detection ability. Lastly, we explore the potential of a human-robot-interaction (HRI) framework within our developed system, further enhancing its applicability and versatility in real-world agricultural scenarios. The fruit detection dataset and source codes for model development and evaluation are now publicly accessible to the research community.

References

Bergerman, M., Maeta, S.M., Zhang, J., Freitas, G.M., Hamner, B., Singh, S. and Kantor, G., 2015. Robot farmers: Autonomous orchard vehicles help tree fruit production. *IEEE Robotics & Automation Magazine*, *22*(1), pp.54-63.

Chu, P., Li, Z., Lammers, K., Lu, R. and Liu, X., 2021. Deep learning-based apple detection using a suppression mask R-CNN. *Pattern Recognition Letters*, *147*, pp.206-211.

Chu, P., Li, Z., Zhang, K., Chen, D., Lammers, K. and Lu, R., 2023. O2rnet: Occluder-occludee relational network for robust apple detection in clustered orchard environments. Smart Agricultural Technology, 5, p.100284.

Zhou, H., Wang, X., Au, W., Kang, H. and Chen, C., 2022. Intelligent robots for fruit harvesting: Recent developments and future challenges. Precision Agriculture, 23(5), pp.1856-1907.

Ben Hassen, T. and El Bilali, H., 2022. Impacts of the Russia-Ukraine war on global food security: towards more sustainable and resilient food systems?. Foods, 11(15), p.2301.

Laborde, D., Martin, W., Swinnen, J. and Vos, R., 2020. COVID-19 risks to global food security. Science, 369(6503), pp.500-502.

Zhao, Y., Gong, L., Huang, Y. and Liu, C., 2016. A review of key techniques of vision-based control for harvesting robot. *Computers and Electronics in Agriculture*, *127*, pp.311-323.

Gongal, A., Amatya, S., Karkee, M., Zhang, Q. and Lewis, K., 2015. Sensors and systems for fruit detection and localization: A review. *Computers and Electronics in Agriculture*, *116*, pp.8-19.

Zhang, K., Lammers, K., Chu, P., Li, Z. and Lu, R., 2021. System design and control of an apple harvesting robot. *Mechatronics*, *79*, p.102644.

Syal, A., Garg, D. and Sharma, S., 2014, December. Apple fruit detection and counting using computer vision techniques. In *2014 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-6). IEEE.

Chaivivatrakul, S. and Dailey, M.N., 2014. Texture-based fruit detection. *Precision Agriculture*, *15*, pp.662-683.

Xu, M., Yoon, S., Fuentes, A., Yang, J. and Park, D.S., 2022. Style-consistent image translation: A novel data augmentation paradigm to improve plant disease recognition. *Frontiers in Plant Science*, *12*, p.773142.

Li, J., Chen, D., Yin, X. and Li, Z., 2024. Performance Evaluation of Semi-supervised Learning Frameworks for Multi-Class Weed Detection. *arXiv preprint arXiv:2403.03390*.

Rai, N. and Sun, X., 2024. WeedVision: A single-stage deep learning architecture to perform weed detection and segmentation using drone-acquired images. *Computers and Electronics in Agriculture*, *219*, p.108792.

Li, Q., Ma, W., Li, H., Zhang, X., Zhang, R. and Zhou, W., 2024. Cotton-YOLO: Improved YOLOV7 for rapid detection of foreign fibers in seed cotton. *Computers and Electronics in Agriculture*, *219*, p.108752.

Koirala, A., Walsh, K.B., Wang, Z. and McCarthy, C., 2019. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precision Agriculture*, *20*(6), pp.1107-1135.

Ukwuoma, C.C., Zhiguang, Q., Bin Heyat, M.B., Ali, L., Almaspoor, Z. and Monday, H.N., 2022. Recent advancements in fruit detection and classification using deep learning techniques. *Mathematical Problems in Engineering*, *2022*, pp.1-29.

Xiao, F., Wang, H., Xu, Y. and Zhang, R., 2023. Fruit detection and recognition based on deep learning for automatic harvesting: an overview and review. *Agronomy*, *13*(6), p.1625.

Girshick, R., 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

Fu, L., Majeed, Y., Zhang, X., Karkee, M. and Zhang, Q., 2020. Faster R–CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. *Biosystems Engineering*, *197*, pp.245-256.

Gao, F., Fu, L., Zhang, X., Majeed, Y., Li, R., Karkee, M. and Zhang, Q., 2020. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. *Computers and Electronics in Agriculture*, *176*, p.105634.

Fu, L., Feng, Y., Majeed, Y., Zhang, X., Zhang, J., Karkee, M. and Zhang, Q., 2018. Kiwifruit detection in field images using Faster R-CNN with ZFNet. *IFAC-PapersOnLine*, *51*(17), pp.45-50.

Bargoti, S. and Underwood, J., 2017, May. Deep fruit detection in orchards. In *2017 IEEE international conference on robotics and automation (ICRA)* (pp. 3626-3633). IEEE.

Terven, J. and Cordova-Esparza, D., 2023. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *arXiv preprint arXiv:2304.00501*.

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E. and Liang, Z., 2019. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Computers and electronics in agriculture*, *157*, pp.417-426.

Shi, R., Li, T. and Yamaguchi, Y., 2020. An attribution-based pruning method for real-time mango detection with YOLO network. *Computers and electronics in agriculture*, *169*, p.105214.

Mirhaji, H., Soleymani, M., Asakereh, A. and Mehdizadeh, S.A., 2021. Fruit detection and load estimation of an orange orchard using the YOLO models through simple approaches in different imaging and illumination conditions. *Computers and Electronics in Agriculture*, *191*, p.106533.

Gai, R., Chen, N. and Yuan, H., 2023. A detection algorithm for cherry fruits based on the improved YOLO-v4 model. *Neural Computing and Applications*, *35*(19), pp.13895-13906.

Chen, D., Lu, Y., Li, Z. and Young, S., 2022. Performance evaluation of deep transfer learning on multi-class identification of common weed species in cotton production systems. *Computers and Electronics in Agriculture*, *198*, p.107091.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *nature*, *521*(7553), pp.436-444.

Kamilaris, A. and Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, *147*, pp.70-90.

Sun, C., Shrivastava, A., Singh, S. and Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (pp. 843-852).

Lu, Y. and Young, S., 2020. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, *178*, p.105760.

Kestur, R., Meduri, A. and Narasipura, O., 2019. MangoNet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. *Engineering Applications of Artificial Intelligence*, *77*, pp.59-69.

Koirala, A., Walsh, K.B., Wang, Z. and McCarthy, C., 2019. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precision Agriculture*, *20*(6), pp.1107-1135.

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.R., Ruiz-Hidalgo, J. and Gregorio, E., 2019. KFuji RGB-DS database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected IR data. *Data in brief*, *25*, p.104289.

Bhusal, S., Karkee, M. and Zhang, Q., 2019. Apple dataset benchmark from orchard environment in modern fruiting wall.

Gené-Mola, J., Gregorio, E., Cheein, F.A., Guevara, J., Llorens, J., Sanz-Cortiella, R., Escola, A. and Rosell-Polo, J.R., 2020. LFuji-air dataset: Annotated 3D LiDAR point clouds of Fuji apple trees for fruit detection scanned under different forced air flow conditions. *Data in brief*, *29*, p.105248.

Häni, N., Roy, P. and Isler, V., 2020. MinneApple: a benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters*, *5*(2), pp.852-858.

Gonzalez, S., Arellano, C. and Tapia, J.E., 2019. Deepblueberry: Quantification of blueberries in the wild using instance segmentation. *Ieee Access*, *7*, pp.105776-105788.

Meshram, V. and Patil, K., 2022. FruitNet: Indian fruits image dataset with quality for machine learning applications. *Data in Brief*, *40*, p.107686.

Bargoti, S. and Underwood, J., 2017, May. Deep fruit detection in orchards. In *2017 IEEE international conference on robotics and automation (ICRA)* (pp. 3626-3633). IEEE.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T. and McCool, C., 2016. Deepfruits: A fruit detection system using deep neural networks. *sensors*, *16*(8), p.1222.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P., 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015-4026).

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Li, J., Xu, M., Xiang, L., Chen, D., Zhuang, W., Yin, X. and Li, Z., 2023. Foundation models in smart agriculture: Basics, opportunities, and challenges. *arXiv preprint arXiv:2308.06668*.

Yang, X., Dai, H., Wu, Z., Bist, R., Subedi, S., Sun, J., Lu, G., Li, C., Liu, T. and Chai, L., 2023. Sam for poultry science. *arXiv preprint arXiv:2305.10254*.

Williams, D., MacFarlane, F. and Britten, A., 2023. Leaf only SAM: a segment anything pipeline for zero-shot automated leaf segmentation. *arXiv preprint arXiv:2305.09418*.

K. Wada. Labelme: Image Polygonal Annotation with Python, 2011. URL https://github.com/wkentaro/labelme.

Terven, J. and Cordova-Esparza, D., 2023. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *arXiv preprint arXiv:2304.00501*.

Geng, C., Huang, S.J. and Chen, S., 2020. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), pp.3614-3631.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. and Zhang, L., 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M. and Shum, H.Y., 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. and Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 658-666).

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N. and Chang, K.W., 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10965-10975).

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q., 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), pp.43-76.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J. and Sun, J., 2019. Objects365: A large-scale, high-quality dataset for object

detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8430-8439).

Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J. and Lazebnik, S., 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision* (pp. 2641-2649).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A. and Bernstein, M.S., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, *123*, pp.32-73.

I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github. com/openimages, 2(3):18, 2017.

Kazemzadeh, S., Ordonez, V., Matten, M. and Berg, T., 2014, October. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 787-798).

Wang, Y., Yao, Q., Kwok, J.T. and Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, *53*(3), pp.1-34.

Song, Y., Wang, T., Cai, P., Mondal, S.K. and Sahoo, J.P., 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, *55*(13s), pp.1-40.

Wang, Y., Yao, Q., Kwok, J.T. and Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, *53*(3), pp.1-34.

Dang, F., Chen, D., Lu, Y. and Li, Z., 2023. YOLOWeeds: a novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems. *Computers and Electronics in Agriculture*, *205*, p.107655.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q., 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), pp.43-76.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S. and Chen, K., 2022. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

Pérez-Borrero, I., Marín-Santos, D., Gegundez-Arias, M.E. and Cortés-Ancos, E., 2020. A fast and accurate deep learning method for strawberry instance segmentation. *Computers and Electronics in Agriculture*, *178*, p.105736.

Kim, S., Hooper, C., Gholami, A., Dong, Z., Li, X., Shen, S., Mahoney, M.W. and Keutzer, K., 2023. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*.

Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S. and Hong, C.S., 2023. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Wang, C., Hasler, S., Tanneberg, D., Ocker, F., Joublin, F., Ceravola, A., Deigmoeller, J. and Gienger, M., 2024. Large language models for multi-modal human-robot interaction. *arXiv preprint arXiv:2401.15174*.