A Dataset for Semantic and Instance Segmentation of Modern Fruit Orchards

Tieqiao Wang, Abhinav Jain, Liqiang He, Cindy Grimm, Sinisa Todorovic
Oregon State University
{wangtie, jainab, heli, grimmc, sinisa}@oregonstate.edu
https://github.com/tqosu/MFO

Abstract

Automating orchard tasks, such as pruning tree branches, requires tree-structure understanding – a significant challenge for computer vision. This paper introduces the first large-scale dataset for semantic and instance segmentation of modern fruit orchards. It consists of videos showing Cherry and Apple trees in modern-orchard scenes, and includes both labeled synthetic and real data, along with synthetic tree meshes. To address prohibitive costs of annotating numerous tree branches, we study unsupervised domain adaptation from synthetic to real data. For this setting, we propose a new Semantically-Guided Depth Refinement (SGDR) that leverages zero-shot depth estimation and semantic-aware smoothing. SGDR outperforms strong baselines and state of the art. Furthermore, we also benchmark the dataset in the supervised setting, where the initial annotations from the first frame are automatically propagated throughout the video using the foundation Segment Anything Model (SAM). The resulting pseudo labels are then manually corrected to generate the ground truth. For the supervised setting, we introduce SAM-Mask2Former (SAM-M2F) aimed at instance segmentation. By providing this dataset and benchmarking for both settings, we aim to enable new research for precision agriculture.

1. Introduction

Labor shortages and increasing global food demand are driving the need for agricultural automation, particularly in labor-intensive tasks within modern fruit orchards with trees bearing apples, cherries or other fruits [4, 17, 19]. In a modern orchard, trees are arranged in rows and trained into specific shapes (e.g., flat, wall-like structures or V-shaped trellises). Pruning – a critical operation for maintaining tree health, spacing of fruiting locations, and adequate sunlight penetration – is the second most labor-intensive task in such orchards [21, 36, 50]. The decision-making process for pruning is complex, requiring detailed knowledge of tree structure, from large trunks and support branches down to the small, tertiary branches where fruiting buds reside.

Figure 1. Left: Real Envy apple tree. Right: Procedurally generated Envy apple tree generated with custom LPy[9] software.

Recent advances in semantic and instance segmentation, combined with the accessibility of low-cost RGB-D cameras, have paved the way for developing effective robotic systems to automate this process [77]. However, achieving precise robotic manipulation requires a highly accurate segment of the overall plant architecture. This is challenging due to the highly-variable imaging conditions in the outdoor environment, the vast range of scales – from large trunks to millimeter-thin branches – the presence of long slender structures, large variations in tree structure within and across orchards of different fruit types, and significant background clutter from surrounding orchard trees.

While recent supervised object segmentation methods, such as SegFormer [71] and Mask2Former [14], have shown promise, they require extensive training data. Unfortunately, no such large-scale datasets are publicly available for modern fruit orchards. This paper introduces a new, large-scale database encompassing both images and videos of modern fruit orchards, and including both synthetic and real-world data. We believe this dataset will be a critical resource for developing and benchmarking advanced scene understanding algorithms for vision in agriculture.

The proposed dataset is benchmarked in two settings – namely, the supervised setting with limited annotations and the unsupervised domain adaptation (UDA) setting from synthetic to real data. For the former, we mitigate the substantial labeling effort by first automatically propagating the annotations provided for the initial frame throughout the

video, and then by manually correcting any errors in the resulting pseudo labels to generate the ground truth for training. In the context of robotic pruning, structural errors (e.g., "floating" branches, incorrectly merged branches) are far more detrimental than minor boundary inaccuracies. Therefore, in the supervised setting, we investigate both semantic and instance segmentation.

In the UDA setting, we rely solely on synthetic data annotations for training, and use real-data labels only for evaluation. We focus solely on UDA semantic segmentation, as prior work has reported that UDA instance segmentation remains too challenging, even in simpler domain contexts.

Most existing instance segmentation methods are fully supervised, requiring fully annotated big datasets [14, 41]. Unfortunately, manual labeling of these images is very challenging and prone to error. Other approaches perform label transfer based on unrelated images [7, 68], which may not include object classes of our interest, such as tertiary spur branches or trellis wires. Furthermore, as input, existing UDA segmentation methods [30, 32] either use RGB images or require complex multi-modal fusion to incorporate depth information [74].

To address the limitations of existing approaches, this paper makes the following key contributions:

- **Large-Scale Orchard Dataset**. We introduce the first large-scale, labeled synthetic dataset and partially labeled real-world dataset designed for semantic and instance segmentation of fruit trees in modern orchards. The synthetic and real datasets focus on two distinct tree structures: (1) Cherry trees grown in an Upright Fruiting Offshoot (UFO) system and (2) Apple trees grown in a V-Trellis system. The synthetic dataset consists of 5,000 images per tree type, encompassing nine semantic classes: foreground trunk (leader), foreground secondary branches (branch), foreground tertiary branches (spur), wires, ground, sky, background (other trees), foreground/background posts. This dataset can be easily expanded if needed. The real-world dataset comprises 132 videos (68,760 frames) of UFO cherry trees and 91 videos (91,470 frames) of Envy apple trees. The real dataset is partially labeled for evaluation, where 387 cherry images are manually labeled with the spur, branch, and leader classes for semantic segmentation, and 24 apple and 15 cherry images are labeled for instance segmentation. We are also releasing 500 synthetically generated tree meshes, representing both UFO and V-Trellis tree structures, to provide a valuable resource for future research in 3D reconstruction, simulation, and synthetic dataset generation.
- **Semantic Depth-Fusion**. We benchmark state-of-the-art UDA methods (e.g., MIC: Masked Image Consistency) and propose a novel approach called Semantically-Guided Depth Refinement (SGDR). SGDR incorporates zero-shot depth estimation to address noise in depth estimates. SGDR leverages ground-truth labels available in the synthetic dataset to smooth the estimated depth, leading to more stable and accurate domain adaptation.
- **Efficient Video Label Extension**. We introduce SAM-Mask2Former (SAM-M2F) that uses a foundation segmentation model, SAM 2, for efficient zero-shot video segmentation and label propagation.

2. Related Work

2.1. Segmentation Datasets for Dormant Trees

Prior research on dormant tree segmentation has primarily focused on small datasets and limited label categories. For instance, researchers have used datasets of around 450 images to train models like mask2former for truck segmentation [66] or YOLO for detecting trunks and branches [57]. Similarly, [48] used a dataset containing 356 images to segment various tree parts. [76] introduced synthetic data generated in Blender [1] to segment foreground trees from orchard backgrounds. Similarly, accurate synthetic images can be generated using 3D-reconstructed trees [5], though tree reconstruction remains an open challenge. [8] created a high-fidelity synthetic dataset by compositing images of cherry trees onto various backgrounds to train a Mask R-CNN for segmenting leaders. However, the omission of key labels like trellis wires and spurs – due to limited data, long annotation times, or inaccurate meshes – reduces the practical utility of these models. The development of tools to procedurally generate 3D meshes of trees [2, 3, 16] has enabled the generation of high-fidelity synthetic environments, primarily utilizing Blender. For example, [10] leveraged such environments to generate synthetic point cloud data for forest tree segmentation. They addressed the domain gap between the real and synthetic point clouds through domain adaptation techniques. Existing tree mesh generation tools lack support for modeling orchard-specific practices such as pruning and tying. To overcome this limitation, inspired by [16], we extend LPy [9]–an open-source tree modeling tool–to incorporate the tying and pruning processes. This allows us to produce high-fidelity synthetic datasets tailored for fruit tree segmentation.

2.2. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) transfers knowledge from a labeled source domain to an unlabeled target domain, commonly using adversarial training [22, 26–28, 56, 61, 63] or self-training [6, 11, 23, 30, 31, 39, 49, 51, 54, 59, 60, 70, 73, 79, 80, 83, 84]. Because semantics and geometry are related [62, 72, 81], many UDA methods use depth. Some treat depth estimation as an auxiliary task [13, 29, 40, 55, 62, 64, 65, 69], while others jointly optimize depth and segmentation in a multi-task framework

[24, 33, 37, 47, 65, 72, 82]. RGB-D segmentation methods [12, 35] also use depth to enhance RGB features, often with Squeeze-and-Excitation (SE) blocks [34] or attention [12, 35, 47, 67, 72, 78]. In contrast, we use *pre-computed* depth as an *additional input*, similar to RGB-D segmentation but within a UDA context. This setup is like MICDrop [74], but crucially, we avoid their complex attention-based fusion and masking. Our method directly integrates depth for a simpler, more efficient UDA solution.

2.3. Semi-supervised Instance Segmentation

Semi-supervised instance segmentation improves performance by using both labeled and unlabeled data. Early methods like Mask-RCNN [25] and ViTDet [44] used detection followed by segmentation, while transformers like MaskFormer [15] and Mask2Former [14] adopted mask classification. Pseudo-labeling, where a teacher model labels unlabeled data for a student model, is common [20, 68], often relying on techniques like EMA teacher updates and differential data augmentation. Noisy Boundaries [68] used a fixed teacher, and Polite Teacher [20] used EMA [59]. Guided Distillation (GD) [7] pre-trained the student with pseudo-labels from a fixed teacher. Other knowledge distillation (KD) methods exist [18, 42, 43, 46], but primarily complement pseudo-labeling. However, existing methods often require complex teacher-student architectures and EMA updates [20, 68], are sensitive to pseudo-label accuracy [7], and heavily rely on image augmentation consistency [58]. In contrast, our approach targets practical agricultural applications. We incorporate minimal human-in-the-loop feedback to refine easily-revisable pseudo-labels, ensuring high-quality annotations. Furthermore, we move beyond image-based augmentation by exploring video-based pseudo-label generation through tracking and segmentation, a novel approach not explored in prior work [7, 42, 58, 68].

3. The Tree Segmentation Dataset

Modern fruit orchard trees are trained into specific structures optimized for increasing yield and reducing time to harvest. These structures are typically formed by tying trunks and branches to wires supported by wooden posts and are maintained through annual cycles of pruning and tying. Figure 1 (Left) shows an example of a tree in a V-Trellis architecture. A modern orchard can be specified by the tree structure and the layout of the orchard. The layout commonly features trees growing in rows with specific distances between adjacent trees and adjacent rows of trees. These characteristics give the orchard a unique appearance, different from typical trees we see in nature.

The image segmentation dataset introduced in this paper includes images of trees in an orchard trained in two specific structures: Upright Fruiting Offshoots (UFO) and the V-trellis system. The V-trellis system features a trunk that grows at an incline to the ground, with secondary branches on either side tied to wires that run perpendicular to the trunk, forming a planar structure. In contrast, the UFO system consists of a horizontal trunk with secondary branches growing upward, supported by wires to keep them vertical and in the same plane. These secondary branches further grow tertiary branches – the fruiting sites of the tree.

Figure 2. Sample images from the real-world datasets, illustrating a range of conditions including varying illumination, image resolutions, times of day (e.g., morning, afternoon), and weather conditions (e.g., sunny, cloudy, rainy). Top row: UFO subsets (UFO-1, 2, 3, 4). Bottom row: Envy subsets (Envy-1, 2, 3).

For vision-enabled robotic pruning of a modern orchard, a segmentation algorithm must be capable of accurately phenotyping the relevant parts of the foreground tree – the tree closest to the camera – such as the trunk, secondary and tertiary branches. It must also be able to identify orchard elements such as posts, wires, ground, sky, and background trees.

To support learning of such a segmentation algorithm we build a dataset consisting of images from two domains: (i) synthetically generated orchard images with their corresponding semantic annotations, and (ii) real orchard images. Annotations of the synthetic images comprise three *tree-related* and six *environment-related* categories. The tree-related categories are: 1) Foreground Trunk, 2) Foreground Secondary Branches, and 3) Foreground Tertiary Branches. The environmental categories are: 1) Wires, 2) Ground, 3) Sky, 4) Background trees, 5) Background Posts, and 6) Foreground Posts. A subset of real images is manually labeled with the same categories for evaluation.

3.1. Real-World Data

We introduce a new real-world dataset collected from orchards in Prosser, Washington, U.S.A. between 2021 and 2024. The videos were captured under unconstrained conditions, resulting in significant variability across several factors: (1) orchard location, (2) time of day and year, (3) camera model, (4) video resolution, (5) camera motion patterns, (6) weather conditions, and (7) availability of depth information. Figure 2 illustrates the diversity of the collected data. The specifications of the real-world dataset are outlined below.

UFO Cherries – 132 videos collected in 4 settings.

1. **Cherry_UFO_1 (December 2021):** Recorded with a handheld cellphone (no depth sensing) at 1080×1920 resolution under overcast conditions. The camera moved in a rectangular pattern close to the tree. This subset comprises 3 videos, totaling 50 seconds or 1500 frames.
2. **Cherry_UFO_2 (January 2022):** Recorded with a handheld Intel RealSense D435 (depth sensing) at 1920×1080 under mixed sunny/cloudy conditions. Camera moved vertically close to the tree. 95 videos, 348 seconds, 10,440 frames.
3. **Cherry_UFO_3 (March 2022):** Recorded with a handheld Intel RealSense D415 (depth sensing) at 1280×720 under sunny conditions. Camera moved randomly close to the tree. 20 videos, 107 seconds, 3210 frames.
4. **Cherry_UFO_4 (January 2023):** Recorded with a handheld Azure Kinect DK (depth sensing) at 1920×1080 under overcast conditions. Camera moved horizontally farther from the tree. 14 videos, 1787 seconds, 53,610 frames.

Envy Apples – 91 videos collected in 3 settings.

1. **Envy_V_Trellis_1 (January 2022):** Recorded with a handheld Intel RealSense D435 (depth sensing) at 480×640 under overcast conditions. Camera moved randomly close to the tree. 12 videos, 55 seconds, 1,650 frames.
2. **Envy_V_Trellis_2 (January 2023):** Recorded with a robot-mounted Azure Kinect DK (depth sensing) at 1920×1080 under cloudy/overcast conditions. Camera moved horizontally and vertically at a far distance from the tree. 71 videos, 2,511 seconds, 75,330 frames.
3. **Envy_V_Trellis_3 (January 2024):** Recorded with a robot-mounted Intel RealSense D435 (depth sensing) at 424×240 under cloudy/overcast conditions. Camera moved in an S-shaped pattern, both close to and far from the tree. 8 videos, 483 seconds, 14,490 frames.

3.2. Synthetic Data

The primary goal of our synthetic dataset generation is to enable unsupervised domain adaptation for segmentation methods by: (i) replicating the geometries and visual characteristics of trees growing in an orchard and (ii) generating images of these trees with variability similar to that of real-world data along with their annotations.

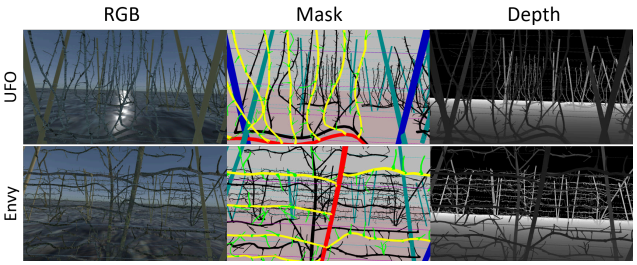

Figure 3. Synthetic dataset: RGB images, labels, and depths.

To achieve this, we built a 3D virtual orchard that allows us to place a camera and take pictures. The generation of this virtual orchard involves three key steps: a) Synthesis of trees with realistic geometries; b) Creating a virtual orchard with a desired layout and environmental elements such as posts and wires; c) Generating artificial images that exhibit the same variability as real-world ones. These three steps are explained below.

Generating the trees. We use L-Systems to simulate tree growth, making modifications to LPy[9] that allow for continuous tying and pruning to generate tree meshes similar to modern fruit trees. These meshes have been validated by horticultural experts to be visually consistent with the target modern fruit trees. We also encode the segmentation label of each mesh triangle as a color.

Generating a virtual orchard. We import the tree meshes, along with meshes of posts and wires in Blender [1], and place them according to the given specifications of an orchard. For example, for V-Trellis we place the wires to be running along with the tree branches and the trees at an angle to form a V when viewed from the end of the row.

Generating images. We randomize the camera distance from the tree, lighting, and tree textures. The sun is uniformly placed in different positions in the sky, whereas we use 7 different bark, and 3 different ground textures and randomly apply them. We use sky textures provided by Blender. For segmentation ground truth, we re-render with just the colored mesh vertices (no lighting).

Using this method, we generate **UFO-synthetic** and **Envy-synthetic** datasets. Each dataset consists of 5000 independent RGB images, their corresponding labels, and depth images. The datasets can be arbitrarily increased as they are synthetically generated. Figure 3 shows examples of the RGB, labeled annotations, and depth from the UFO-Synthetic and Envy-Synthetic data. We also release the tree meshes that were used to generate these datasets. An example mesh is shown in Figure 1 (Right).

4. Unsupervised Domain Adaptation (UDA)

This section presents our multimodal unsupervised domain adaptation (UDA) approach to semantic segmentation. Our goal is to train a model using labeled source data (X_s, Y_s) and unlabeled target data (X_t), minimizing the domain gap to enhance performance on the target domain. As input, we use both RGB and depth images, and produce a semantic segmentation map as output. Standard supervised learning is used for the source domain, while we leverage the unlabeled target data via a student-teacher framework, drawing inspiration from related methods [30–32, 60].

The teacher network, updated using an Exponential Moving Average (EMA) [59] of the student's weights, generates pseudo-labels $(X_t, \hat{Y}_t)$ for the target data. These

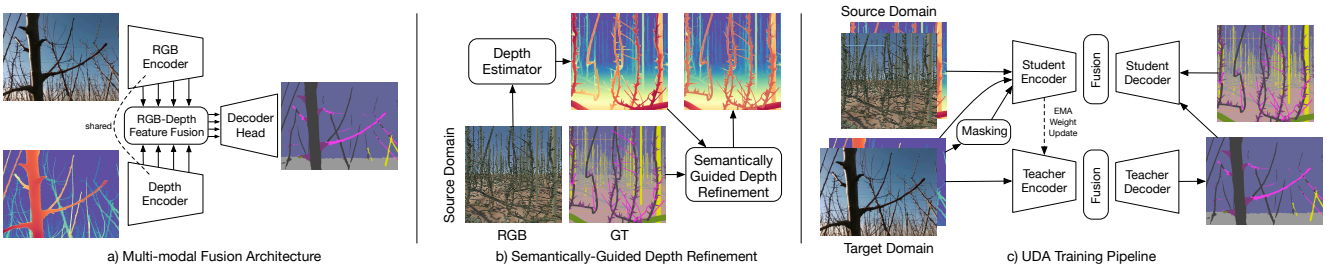

Figure 4. Overview of our UDA approach. a) Multi-modal Fusion Architecture: Shared RGB and Depth encoders with multi-scale feature fusion. b) Semantics-Aware Depth Smoothing: Improved synthetic depth estimation via semantic category injection for better boundaries and consistency. c) UDA Training Pipeline: a student-teacher framework. Both student and teacher use multimodal feature fusion. The student is trained on the source-domain ground truth (GT) and target-domain pseudo-labels generated by the EMA-updated teacher.

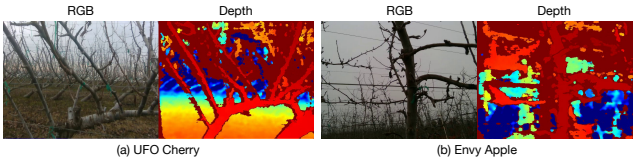

Figure 5. RGB images and their depth from the real-world dataset.

pseudo-labels provide supervision for the student's training on the target domain. To promote robustness, strong and weak augmentations are provided for the student and teacher inputs [60], respectively. Masked image consistency (MIC) [32] is incorporated to enforce consistent predictions between masked views of the target images and the pseudo-labels generated from the unmasked images.

Building upon hierarchical encoders [30, 31], which provide multi-resolution feature maps, our approach integrates depth estimation. Depth is estimated for both source and target data using the Depth Anything V2 model [75], allowing our method to generalize to datasets with missing, noisy (Fig. 5), or domain-shifted (Fig. 3) depth.

Our contributions are twofold: (1) improving accuracy of depth estimation; and (2) fusing the improved depth estimates with RGB features to achieve more robust UDA.

Datasets. For RGB-based unsupervised domain adaptation (UDA), we use a subset of the UFO cherry orchard dataset, specified in Section 3.2. Our UDA training set consists of 4,550 unlabeled real-world images, randomly selected from the dataset, and 9,009 synthetic images, generated using the algorithm detailed in Section 3.2. For evaluation, we use a separate, held-out set of 387 labeled real-world images from the same dataset.

Models. Our baseline, MIC [32], is a domain adaptation method that uses spatial context by enforcing consistency between masked and unmasked views of target images. We propose two extensions of this baseline: MIC with Depth (MIC-D) and Semantically-Guided Depth Refinement (SGDR). MIC-D integrates zero-shot depth features [75] to provide geometric guidance, thereby improving segmentation accuracy. These depth features are not fine-tuned. SGDR, in contrast, focuses on improving accuracy of the zero-shot depth estimation using a lightweight adapter. Leveraging labeled data from the source domain, SGDR refines depth estimates, significantly improving stability and sharpness, especially at object boundaries and for small, thin, or narrow structures, as appropriate for our orchard domain. This depth refinement is achieved by generating a semantically-aware depth map: depth values are averaged within regions defined by each semantic mask from the source domain. We did not compare our method with MICDrop [52] – another UDA method that uses depth as a parallel input to RGB data – because its source code is not publicly available and its architecture is considerably more complex. MICDrop employs duplicate, heavyweight encoders for RGB and depth data, a complex masking strategy to balance RGB and depth features, and an additional attention mechanism for depth-based semantic reasoning. Instead, our approach uses a single shared encoder for RGB and depth. The RGB and depth features are concatenated and projected to a lower-dimensional combined feature. This design is substantially more efficient in terms of both parameter count and computational cost.

Metrics. Following [32], we evaluate our semantic segmentation on the target domain using the following metrics: per-class Intersection-over-Union (IoU) and accuracy (Acc), mean IoU (mIoU) and mean accuracy (mAcc) calculated across three classes (leader, branch, and spur), and overall accuracy (aAcc).

Comparison with SOTA. The proposed MIC-D and SGDR are compared with the MIC baseline [32] in Table 1. Our SGDR method significantly outperforms both MIC-D and MIC. Specifically, SGDR improves upon MIC by +5.2 in aAcc, +8.6 in mIoU, and +10.9 in mAcc.

Analysis. Figure 6 illustrates the effectiveness of our SGDR method compared to the state-of-the-art MIC. Qualitatively, SGDR's segmentation aligns more closely with the ground truth across both synthetic (9 classes: spur, branch, leader, sky, ground, other trees, wires, post, trunk) and real-world images (3 classes: spur, branch, leader).

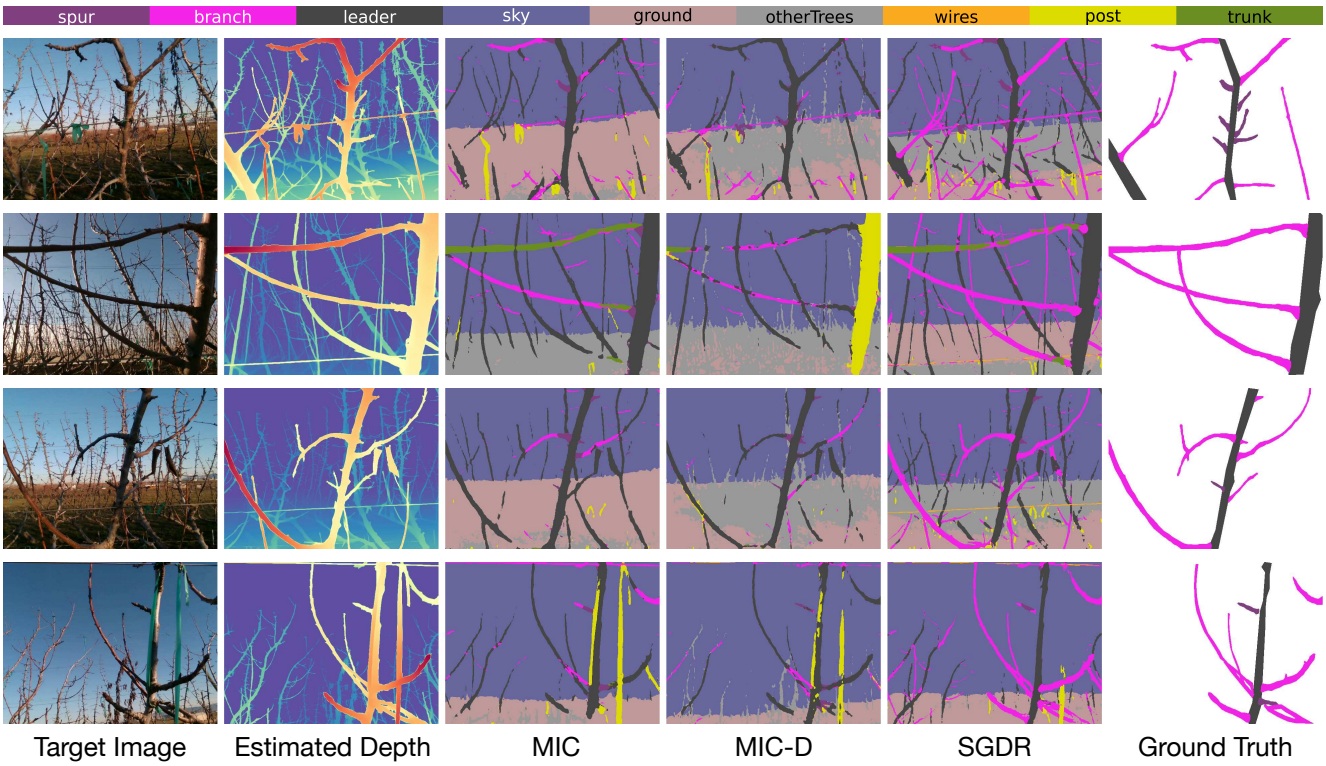

Figure 6. Qualitative comparison of MIC, MIC-D and SGDR. SGDR demonstrates superior performance in segmenting branches (rows 2-4) and accurately predicting leaders and branches, particularly those located at the image boundary (row 1).

	IoU			Acc			Overall		
Method	spur	branch	leader	spur	branch	leader	aAcc	mIoU	mAcc
MIC [32]	28.7	37.7	**76.5**	45.3	39.4	**87.1**	69.7	47.6	57.3
MIC-D	27.0	29.2	64.3	39.4	30.0	76.4	59.7	40.2	48.6
SGDR	**33.6**	**58.6**	76.4	**55.6**	**69.5**	79.5	**74.9**	**56.2**	**68.2**

Table 1. Performance comparison of the baseline MIC [32], MIC with Depth (MIC-D) and Semantically-Guided Depth Refinement (SGDR). Metrics include overall accuracy (aAcc), mean Intersection-over-Union (mIoU), and mean class accuracy (mAcc).

While our domain adaptation strategy yields promising results for semantic segmentation, achieving good performance in instance segmentation – required for many agricultural applications – remains too challenging. Rather than relying solely on domain adaptation, we further explore maximizing the utility of limited labeled data. To enable instance segmentation, the following sections present our investigations into: (1) efficient data extension in the sparse and few-shot annotation settings; and (2) multi-modal foundation models.

5. Supervised Setting

Creating a large-scale, high-quality dataset with instance-level annotations for our orchard imagery presents significant challenges. The diverse, elongated, and curved shapes of tree branches, combined with the frequent intermingling of different instances, make manual annotation extremely laborious and time-consuming. Therefore, in this work, we pursue efficient annotation extension techniques, leveraging the capabilities of state-of-the-art (SOTA) multi-modal foundation models. This section specifies our exploration of two distinct approaches: (1) image-based data extension, where a single annotated sample serves as a visual prompt for one-shot novel object detection by a foundation model; and (2) zero-shot video object segmentation (ZVOS), where annotation of a single frame is propagated to the remaining frames of a video sequence. The results and analysis of both approaches are presented in the subsequent subsections. In the following, we show that the image-based extension approach encountered significant difficulties, while the ZVOS method demonstrated promising performance.

T-Rex2 Image Visual Prompt. T-Rex2 [38], a recently in-

(a) Visual Prompt Box (b) Intelligent Annotate

Figure 7. One-shot detection using T-Rex2 [38]. (a) Input image with a single bounding box annotation provided as a visual prompt, specifying a tree branch. (b) T-Rex2's generated detections based on the visual prompt in (a).

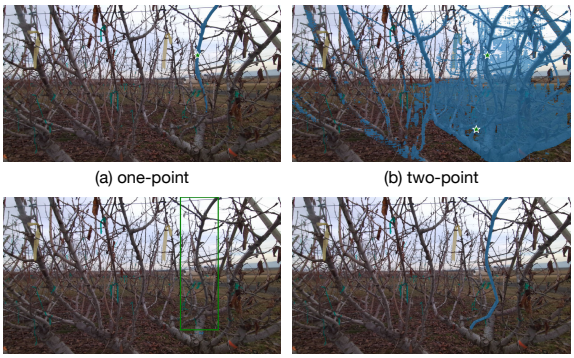

Figure 8. Evaluation of SAM2 [53] segmentation performance with different visual prompt types. (a) Single-point prompt. (b) Two-point prompt. (c) Bounding box prompt. (d) Mask prompt. Input prompts (points and boxes) are shown in green; the resulting SAM-generated segmentation masks are shown in blue.

troduced object detection model, integrates both textual and visual prompts via contrastive learning to achieve flexible zero-shot detection across a wide range of object categories and scenarios. While T-Rex2 demonstrates impressive performance on common object classes such as animals and fruits, our experiments revealed limitations in its ability to accurately detect tree branches in our orchard dataset. As illustrated in Figure 7, T-Rex2 exhibits a high rate of missed detections and frequently detects only partial segments of the branches, rather than the complete structure. We hypothesize that this limitation stems from the inherent complexity of our orchard imagery. The intricate and overlapping nature of the tree branches, coupled with variations in lighting and perspective, presents a significant challenge even for state-of-the-art zero-shot detection models like T-Rex2, hindering their generalization capabilities in this specific context.

SAM 2 Frame Visual Prompt. Segment Anything Model 2 (SAM 2) [53] is a powerful foundation model for promptable image segmentation. A critical question for our agricultural application is determining the optimal prompting strategy to achieve accurate instance segmentation of tree parts (e.g., leaders and branches) within a complex orchard environment. Our goal is to achieve effective segmentation

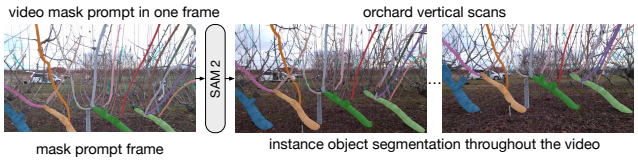

Figure 9. Zero-shot video instance segmentation: We leverage moving cameras and initial frame annotations to generate a large labeled dataset, expanding the initial labels hundreds of times.

with minimal annotation effort. Initial experiments evaluating SAM with single-point, two-point, and bounding box prompts yielded unsatisfactory results (Fig.8). A single positive point prompt resulted in incomplete segmentation of the target branch (Fig.8 (a)). A two-point prompt generated numerous false positives, incorrectly segmenting other branches and failing to distinguish between foreground and background objects (Figure 8 (b)). A bounding box prompt was similarly ineffective (Fig.8 (c)), highlighting the challenges posed by the intricate and noisy nature of our orchard imagery. Consequently, we adopted a mask-based prompting strategy (Fig.8 (d)) for our experiments.

Given the promising results with mask prompts, we aim to minimize the number of manually labeled frames while still generating a diverse and comprehensive training dataset. To this end, we propose a per-tree vertical scanning strategy for efficient label extension. The vertically oriented growth pattern of trees ensures that portions of the leader and branches remain visible as the camera moves upwards or downwards along the trunk. The adopted vertical camera scan facilitates reliable tracking and segmentation of tree branches using a zero-shot tracker, initialized with the initial mask prompt from SAM. By systematically performing vertical scans of each tree, we can leverage the foundation model's capabilities to generate diverse high-quality segmentation masks, effectively extending our initial labeling.

Experimental Setup: Datasets and Models. To evaluate the effectiveness of our proposed data extension approach, we conducted experiments using per-tree vertical scans of Envy apple and UFO cherry trees. The dataset comprised 14 Envy apple scans and 5 UFO cherry scans, containing a total of 56 leader instances, 187 branches, and 10 posts across all scans. For initial training, a single representative frame from each tree scan was manually annotated. These initial annotations served as the training data for the baseline Mask2Former model [14]. Our proposed method, SAM-M2F, uses the Segment Anything Model 2 (SAM 2) [53] for label extension. The initial frame annotations were propagated through their respective video sequences using SAM 2. Manual corrections were made to the SAM-generated labels as needed to ensure accuracy. This expanded the training set, comprising both the initial manual annotations and the corrected ones. The expanded

Method	AP	AP50	AP75	APm	APl	leader	branch	post
Mask2Former [14]	54.289	81.997	54.246	24.341	56.321	57.290	31.146	**74.431**
SAM-M2F	**64.095**	**91.145**	**79.060**	**53.384**	**66.157**	**74.922**	**43.086**	74.277

Table 2. Instance segmentation results on 20 evaluation scenes. SAM-M2F outperforms the strong Mask2Former [14] baseline by a significant margin. Mask2Former was trained without label extension from SAM 2 [53].

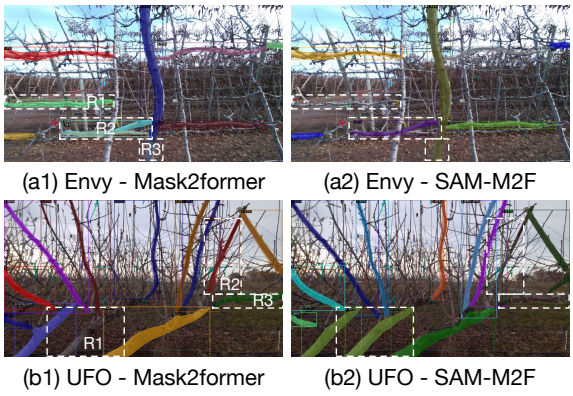

Figure 10. Qualitative comparison of Mask2Former and our proposed SAM-M2F for instance segmentation on Envy Apple (a) and UFO Cherry (b) trees. (a1) and (b1) depict Mask2Former results; (a2) and (b2) depict SAM-M2F results.

dataset was used to train the Mask2Former instance segmentation model. Both models used COCO [45] pre-trained weights [14] and were trained for 500 iterations with a batch size of 6 and a learning rate of 5e-5. For evaluation, we randomly selected 20 scenes not used in the training set. The evaluation set included 10 scenes from each tree type (Envy apple and UFO cherry), containing a total of 49 leader instances, 143 branches, and 6 posts.

Metrics. For instance segmentation, we use the standard average precision (AP) metrics, including per-category AP and overall AP across all categories [14, 45].

Comparison with SOTA. Table 2 demonstrates that SAM-M2F, using our proposed label extension, achieves significantly better performance compared to Mask2Former. We observe a 10% increase in overall AP, a 12% increase in branch segmentation AP, and a 17% increase in leader segmentation AP. This substantial improvement highlights the effectiveness of our proposed label extension.

Analysis. Figure 10 presents a qualitative comparison between Mask2Former and our proposed SAM-M2F. Each instance box displays the predicted category (0: leader, 1: branch) and a confidence score, with different colors distinguishing individual instances. Dashed white boxes (R1, R2, R3) in each image pair highlight regions of improvement. We observe several key advantages of SAM-M2F. In the Envy Apple comparison ((a1) vs. (a2)), R1 and R2 demonstrate that Mask2Former incorrectly segments some background tree branches, leading to false positives. SAM-M2F, in contrast, avoids these errors. Furthermore, R3 shows that Mask2Former fails to completely segment a leader at the bottom of the image, whereas SAM-M2F provides a more complete and accurate segmentation. In the UFO Cherry comparison ((b1) vs. (b2)), R1 illustrates a case where Mask2Former misses an obvious leader, while SAM-M2F successfully detects it. Finally, in R2 and R3, Mask2Former misclassifies a supporting post and part of a leader as branches, while SAM-M2F successfully avoids both false positives and misclassification errors. These examples clearly demonstrate SAM-M2F's improved ability to handle complex orchard scenes and its adherence to our specific segmentation goals.

6. Conclusion

This work is aimed at enabling vision-based automation of labor-intensive tasks in modern orchards. We have introduced the first large-scale dataset for semantic and instance segmentation of fruit trees in modern orchards. The dataset encompasses both synthetic images and real-world videos, as well as a collection of synthetic 3D meshes of trees. The provided manual annotations of the dataset are aimed at supporting semantic and instance segmentation. The dataset is benchmarked in the supervised setting with limited initial annotations and the unsupervised domain adaptation setting. For instance segmentation in the supervised setting with limited labeled data, we have proposed a new SAM-M2F model, which leverages the SAM 2 foundation model to provide an effective data extension for a supervised training of the Mask2Former. For the unsupervised domain adaptation setting, we have proposed the SGDR model, which effectively refines and integrates depth estimation for enhanced semantic segmentation. The new dataset and proposed SAM-M2F and SGDR models fill a critical gap in the field toward advancing vision research for agriculture applications.

7. Acknowledgement

We thank Martin Churuvija, Alexander You, Josyula Krishna, Ellie Camp, and Utsav Bhandari for their help in data collection. This work has been supported by USDA NIFA award No.2021-67021-35344 (AgAID AI Institute) and the Washington Tree Fruit Research Commission.

References

[1] Blender. https://www.blender.org/. 2, 4

[2] The Grove 3D. https://www.thegrove3d.com/. 2

[3] The Plant Factory. https://www.bentley.com/software/e-on-software-downloads/. 2

[4] Ag America Lending. (infographic) the u.s. labor shortage. 6 2022. 1

[5] Shayan A Akbar, Somrita Chattopadhyay, Noha M Elfiky, and Avinash Kak. A novel benchmark rgbd dataset for dormant apple trees and its application to automatic pruning. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 81–88, 2016. 2

[6] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, 2021. 2

[7] Tariq Berrada, Camille Couprie, Karteek Alahari, and Jakob Verbeek. Guided distillation for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 475–483, 2024. 2, 3

[8] Daniel Borrenpohl and Manoj Karkee. Automated pruning decisions in dormant sweet cherry canopies using instance segmentation. *Computers and Electronics in Agriculture*, 207:107716, 2023. 2

[9] Frédéric Boudon, Christophe Pradal, Thomas Cokelaer, Przemyslaw Prusinkiewicz, and Christophe Godin. L-py: an l-system simulation framework for modeling plant architecture development based on a dynamic language. *Frontiers in plant science*, 3:76, 2012. 1, 2, 4

[10] Mitch Bryson, Ahalya Ravendran, Celine Mercier, Tancred Frickey, Sadeepa Jayathunga, Grant Pearse, and Robin JL Hartley. Domain adaptation of deep neural networks for tree part segmentation using synthetic forest trees. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 14:100078, 2024. 2

[11] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. *ACM Multimedia*, 2023. 2

[12] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, 2020. 3

[13] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 2

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 2, 3, 7, 8

[15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 3

[16] Evelyne Costes, Colin Smith, Michael Renton, Yann Guédon, Przemyslaw Prusinkiewicz, and Christophe Godin. Mapplet: simulation of apple tree development using mixed stochastic and biomechanical models. *Functional Plant Biology*, 35(10):936–950, 2008. 2

[17] Jeff Daniels. From strawberries to apples, a wave of agriculture robotics may ease the farm labor crunch. CNBC, 3 2018. 1

[18] Peijie Dong, Lujun Li, and Zimian Wei. DisWOT: Student architecture search for distillation without training. In *CVPR*, 2023. 3

[19] New American Economy. Immigrants and american agriculture. 2019. 1

[20] Dominik Filipiak, Andrzej Zapała, Piotr Tempczyk, Anna Fensel, and Marek Cygan. Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. *IEEE Access*, 2024. 3

[21] Suzette P Galinato, R Karina Gallardo, and Carol A Miles. Cost estimation of establishing a cider apple orchard in western washington. *Washington State Univ. Ext. Publ. FS141E*, 2013. 1

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 2

[23] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 17, 2004. 2

[24] Vitor Guizilini, Jie Li, Rareș Ambruș, and Adrien Gaidon. Geometric unsupervised domain adaptation for semantic segmentation. In *ICCV*, 2021. 3

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *iccv*, 2017. 3

[26] Liqiang He and Sinisa Todorovic. Attention decomposition for cross-domain semantic segmentation. In *European Conference on Computer Vision*, pages 414–431. Springer, 2024. 2

[27] Liqiang He, Wei Wang, Albert Chen, Min Sun, Cheng-Hao Kuo, and Sinisa Todorovic. Bidirectional alignment for domain adaptive detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18775–18785, 2023.

[28] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[29] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *CVPR*, pages 11130–11140, 2021. 2

[30] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 2, 4, 5

[31] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*, 2022. 2, 5

[32] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 2, 4, 5, 6

[33] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *IJCV*, 2023. 3

[34] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation net-

works. In *CVPR*, 2018. 3

[35] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *ICIP*, 2019. 3

[36] INFACO. How farm labor crisis impacting pruning. Accessed August 8, 2022 [Online]. 1

[37] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *CVPR*, 2020. 3

[38] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2024. 6, 7

[39] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 2

[40] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. Spigan: Privileged adversarial learning from simulation. *arXiv preprint arXiv:1810.03756*, 2018. 2

[41] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 2

[42] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *ECCV*, 2022. 3

[43] Lujun Li and Jin Zhe. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *NeurIPS*, 2022. 3

[44] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *eccv*, 2022. 3

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 8

[46] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. NORM: Knowledge distillation via N-to-one representation matching. In *ICLR*, 2023. 3

[47] Ivan Lopes, Tuan-Hung Vu, and Raoul de Charette. Cross-task attention mechanism for dense multi-task learning. In *WACV*, 2023. 3

[48] Yaqoob Majeed, Jing Zhang, Xin Zhang, Longsheng Fu, Manoj Karkee, Qin Zhang, and Matthew D Whiting. Deep learning based segmentation for automated training of apple trees on trellis wires. *Computers and Electronics in Agriculture*, 170:105277, 2020. 2

[49] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, 2020. 2

[50] Gary Moulton, Jacky King, and Washington State University Mount Vernon Research and Extension Unit. Pruning tree fruit - the basics. Washington State University College of Agricultural, Human, and Natural Resource Sciences. 1

[51] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, 2019. 2

[52] Saba Parvez, Chelsea Herdman, Manu Beerens, Korak Chakraborti, Zachary P Harmer, Jing-Ruey J Yeh, Calum A MacRae, H Joseph Yost, and Randall T Peterson. Mic-drop: A platform for large-scale in vivo crispr screens. *Science*, 373(6559):1146–1151, 2021. 5

[53] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 7, 8

[54] Suman Saha, Lukas Hoyer, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Edaps: Enhanced domain-adaptive panoptic segmentation. In *ICCV*, 2023. 2

[55] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. In *CVPR*, 2021. 2

[56] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 2

[57] Ranjan Sapkota, Dawood Ahmed, and Manoj Karkee. Comparing yolov8 and mask r-cnn for instance segmentation in complex orchard environments. *Artificial Intelligence in Agriculture*, 13:84–99, 2024. 2

[58] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 2020. 3

[59] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2, 3, 4

[60] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 2, 4, 5

[61] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 2

[62] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 2

[63] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[64] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *CVPR*, 2019. 2

[65] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *ICCV*, 2021. 2, 3

[66] T Wang, P Sankari, J Brown, A Paudel, L He, M Karkee, A Thompson, C Grimm, JR Davidson, and S Todorovic. Automatic estimation of trunk cross sectional area using deep

learning. In *Precision agriculture'23*, pages 491–498. Wageningen Academic, 2023. 2

[67] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, 2022. 3

[68] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *CVPR*, 2022. 2, 3

[69] Quanliang Wu and Huajun Liu. Unsupervised domain adaptation for semantic segmentation using depth distribution. *Advances in Neural Information Processing Systems*, 35:14374–14387, 2022. 2

[70] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE TPAMI*, 2023. 2

[71] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 1

[72] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018. 2, 3

[73] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 2

[74] Linyan Yang, Lukas Hoyer, Mark Weber, Tobias Fischer, Dengxin Dai, Laura Leal-Taixé, Marc Pollefeys, Daniel Cremers, and Luc Van Gool. Micdrop: masking image and depth features via complementary dropout for domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 329–346. Springer, 2024. 2, 3

[75] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 5

[76] Alexander You, Cindy Grimm, and Joseph R. Davidson. Optical flow-based branch segmentation for complex orchard environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9180–9186, 2022. 2

[77] Alexander You, Nidhi Parayil, Josyula Gopala Krishna, Uddhav Bhattarai, Ranjan Sapkota, Dawood Ahmed, Matthew Whiting, Manoj Karkee, Cindy M Grimm, and Joseph R Davidson. Semiautonomous precision pruning of upright fruiting offshoot orchard systems: An integrated approach. *IEEE Robotics & Automation Magazine*, 2023. 1

[78] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 3

[79] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 2021. 2

[80] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *NeurIPS*, 2019. 2

[81] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *ECCV*, 2018. 2

[82] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 3

[83] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2

[84] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 2