

PhenoLeaf-TS: A Time-Series Benchmark for Leaf Instance Segmentation, Tracking, and Growth Stage Classification

Rijad Sarić^{1,*}, Basim Azam², Sarmad Khan³, and Edhem Čustović⁴

¹ ARC Research Hub for Protected Cropping (PC Hub), La Trobe University, Melbourne, Australia

`*r.saric@latrobe.edu.au`

² School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia

³ Department of Computer Science, Technological University Dublin, Dublin, Ireland

⁴ Photon Systems Instruments (PSI), Drásov, Czech Republic

Abstract. Image-based high-throughput plant phenotyping leverages instance-level understanding of individual leaves growth over experimental time, yet existing datasets lack temporal depth and annotation consistency to jointly benchmark segmentation, tracking, and growth stage classification. We introduce *PhenoLeaf-TS*, a time-series dataset of 17,082 top-down RGB images spanning 21 *Arabidopsis thaliana* genotypes, totalling 318 plant replicates, each annotated with colour-coded leaf instance masks that maintain consistent identity throughout the growth sequence. We define three benchmark computer vision tasks with standardised protocols and evaluate 21 distinct models: 9 instance segmentation architectures, 6 multi-object trackers, and 6 classification architectures. Mask R-CNN with ResNet-50 achieves 73.2 mAP for leaf segmentation, ByteTrack reaches 84.1% MOTA for tracking individual leaves in sequences, and Swin-T achieves 91.7% accuracy for growth stage classification. Cross-dataset experiments on CVPPP and Komatsuna show that fine-tuning from PhenoLeaf-TS weights yields up to +51 mAP over zero-shot transfer. The dataset, the evaluation code, and the trained models are publicly accessible at <https://pisyntor.github.io/PhenoLeaf-TS>.

Keywords: Plant phenotyping · Leaf instance segmentation · Leaf tracking · Growth stage classification · Benchmark dataset

1 Introduction

Non-invasive measurement of leaf-level traits such as, projected area, growth rate, length, width, and shape is central to high-throughput plant phenotyping [21, 29]. These measurements require instance-level understanding, where each leaf must be individually segmented, tracked through time, and associated with a developmental stage. This is challenging because the leaves overlap and occlude each other, and their shapes change as the plant grows [32].

Existing benchmarks phenotyping RGB datasets have advanced leaf segmentation but remain limited in temporal scope. The CVPPP Leaf Segmentation Challenge [20, 28] provides *Arabidopsis thaliana* and tobacco images with instance masks, but each image is an independent snapshot with no temporal ordering. PhenoBench [35] covers field-grown crops at the semantic and panoptic level but does not provide per-leaf instance labels across time. Komatsuna [33] offers 15 plant sequences with tracking annotations, but is limited in scale (900 images) and genetic diversity. No current dataset simultaneously provides (i) temporally ordered images of the same plant, (ii) consistent per-leaf instance labels across the image sequences, and (iii) enough scale to benchmark segmentation, tracking, and growth stage classification jointly.

We address this gap with **PhenoLeaf-TS**, a dataset of 17,082 RGB images captured throughout the full growth trajectory of 21 *Arabidopsis* genotypes (Fig. 1). Each leaf instance is assigned a unique color from a fixed 31-color palette, and this assignment persists across all frames in which the leaf appears. The colour-coded scheme enables direct evaluation of temporal consistency without requiring a post-hoc label matching step. Our key contributions are:

1. A publicly available time-series dataset of 17,082 RGB images across 21 genotypes and 318 plant replicates, with temporally consistent, colour-coded leaf instance annotations (Sec. 3).
2. Three benchmark CV tasks: instance segmentation, leaf tracking, and growth stage classification with standardized evaluation protocols (Sec. 4).
3. Baselines spanning nine instance segmentation architectures, six trackers, and six classifiers, together with cross-dataset transfer experiments on CVPPP and Komatsuna that demonstrate the dataset’s value as a pre-training source for plant phenotyping research (Sec. 5).

In conducted experiments, Mask R-CNN (R50-FPN) achieves 73.2 mAP for segmentation, ByteTrack achieves 84.1% MOTA for leaf instances, and Swin-T achieves 91.7% accuracy for growth stage classification. Fine-tuning of the PhenoLeaf-TS weights improves mAP segmentation by up to 51 points on an external dataset of different plant species.

2 Related Work

Plant Phenotyping Datasets. The CVPPP Leaf Segmentation Challenge (LCC) [20, 28] generated *Arabidopsis* and tobacco datasets (Ara2012, Ara2013) with per-leaf instance masks, becoming the standard benchmark for leaf segmentation in plant phenotyping. Later releases expanded the collection [2]. Komatsuna [33] provides 15 plant sequences (900 images) with tracking annotations, but with restricted genetic diversity MSU-PID [9] offers multi-modal plant imagery with temporal sequences but only partial instance labels. PhenoBench [35] targets UAV field-level sugat beet segmentation with panoptic annotations, but does not provide per-leaf instance labels throughout time. However, our dataset provides temporally consistent instance annotations across 21 genotypes at a substantially larger scale, enabling joint evaluation of segmentation, tracking, and growth stage classification.

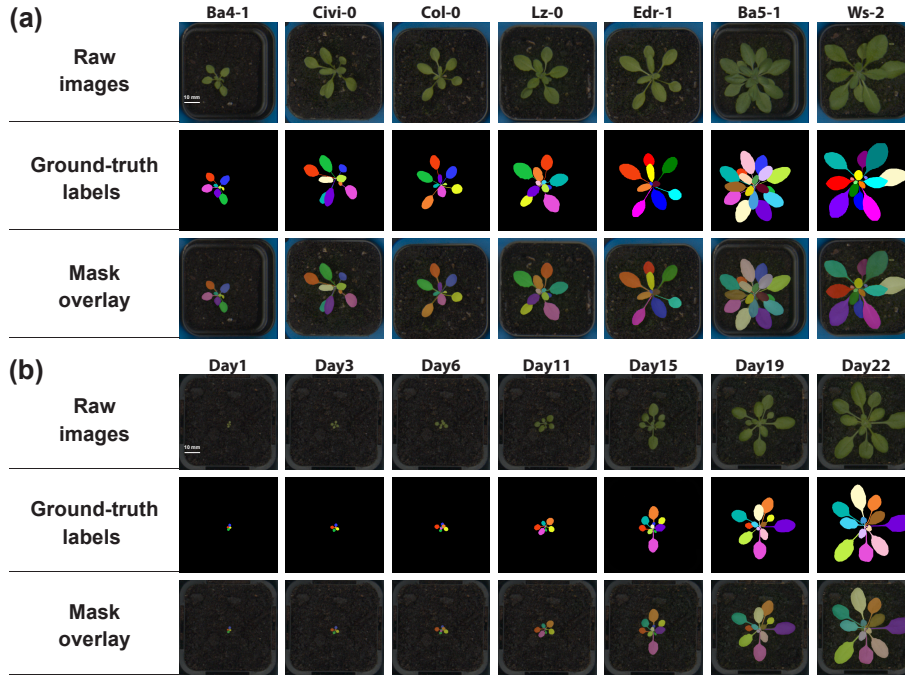


Fig. 1: Overview of the PhenoLeaf-TS dataset. (a) Representative samples across seven Arabidopsis genotypes, showing raw RGB images (top), colour-coded leaf instance labels (middle), and mask overlays (bottom). Each leaf receives a unique colour from a fixed 31-colour palette. (b) Temporal sequence of a single plant from experimental Day 1 to Day 22. Leaf colours persist across frames, enabling direct evaluation of tracking consistency.

Leaf Instance Segmentation. Early approaches used watershed transforms [23] and recurrent mask decoders [25, 26]. Modern detection-based frameworks such as Mask R-CNN [13] and YOLO variants [15] achieve strong results when fine-tuned with plant data [34]. Transformer-based Mask2Former [7] and foundation models such as SAM2 [24] offer zero-shot capabilities. In this study, we evaluated both paradigms using internally created large-scale leaf dataset.

Multi-Object Tracking. The dominant paradigm is tracking-by-detection [19]. DeepSORT [36] combines Kalman filtering with a learned appearance descriptor; ByteTrack [38] retains low-confidence detections for secondary association; BoTSORT [1] adds camera-motion compensation; Deep-OC-SORT [5] handles occlusion via observation-centric updates; and StrongSORT [12] improves DeepSORT with stronger features. These trackers were primarily designed for pedestrians and vehicles, where objects frequently enter and exit the scene. However, leaf tracking presents a different regime: objects grow continuously, new leaves emerge between existing ones, and disappearances are not that common. We adapt six trackers to this setting and report standard multi-object tracking (MOT) metrics.

Growth Stage Classification. Classifying plant developmental stages from images has been addressed with Convolutional Neural Networks (CNNs) [39] and Vision Transformers (ViTs) [11]. Self-supervised features of DINOv2 [22] provide strong transfer representations. Our benchmark defines a three-class growth stage problem derived from the leaf count and evaluates six architectures.

3 The PhenoLeaf-TS Dataset

3.1 Data Collection

Phenotyping experiments were conducted using 21 Arabidopsis genotypes grown in an environmentally controlled growth chamber and screened from above at regular intervals (2 or 4 images per day) with a commercial RGB camera (4.92 MP) mounted on automated gantry system. Each of the 318 plant replicates was placed on the adjustable-height table and recorded as a separate image sequence, capturing the monitored growth period from approximately 14 days after sowing to a mature rosette. Images have a resolution of approximately 530×525 pixels, with the plant centred in the frame against a soil background [27].

3.2 Annotation Protocol

Leaf instance masks were manually annotated using the CVAT labelling tool with a fixed palette of 31 perceptually distinct colours. The first leaf to appear is assigned colour 1 (red), the second colour 2 (blue), and so on following a deterministic ordering (Fig. 2d). When a new leaf emerges in a later frame, it receives the next colour in the predefined sequence. This scheme ensures temporal consistency, i.e., the same physical leaf retains the same colour label across all frames in which it appears.

The deterministic colour-to-identity mapping simplifies evaluation because predicted instance labels can be compared directly with the colour-coded ground truth, without requiring Hungarian matching. Boundary pixels followed leaf contours and occluded regions were marked conservatively. Annotation was conducted by 12 trained annotators and checked by 4 reviewers using frame-to-frame bootstrapping with fixed leaf IDs. Temporal consistency was reviewed for growth, occlusion, emergence, interpolation, and identity swaps, while inter-annotator agreement was not reported as a separate quantitative statistic.

3.3 Dataset Statistics

Tab. 1 summarises the PhenoLeaf-TS dataset. The 17,082 images are distributed across 21 genotypes with a mean of 53.7 frames per replicate and 9.7 leaves per image (Fig. 2a). Genotype representation ranges from 4 replicates (Ei-6) to 38 (Kz-9), reflecting natural variation in availability (Fig. 2b).

Table 1: Dataset summary.

Statistic	Value
Genotypes	21
Plant replicates	318
Total images	17,082
Resolution	$\sim 530 \times 525$
Frames / replicate	53.7
Leaves / image	9.7
Min / max leaves	4 / 26
Growth stages	3
Leaf unique colours	31

Growth stages were defined from leaf count as **Early** (4–6 leaves; 22.6%), **Intermediate** (7–10 leaves; 39.8%), and **Mature** (11+ leaves; 37.6%). Labels were obtained from instance-mask leaf counts (Fig. 2c), with the early stage moderately underrepresented.

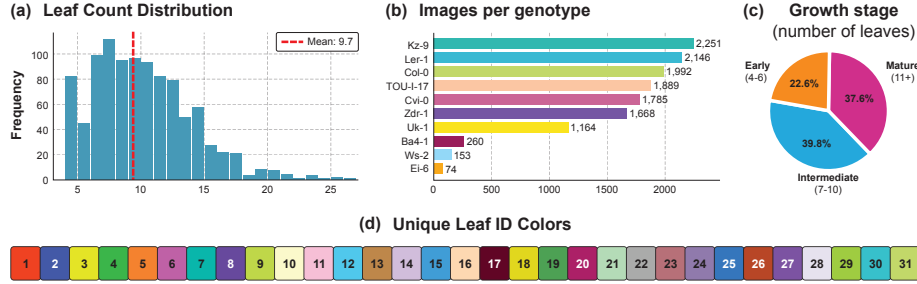


Fig. 2: Dataset statistics. (a) Leaf count distribution (mean 9.7, range 4–26). (b) Number of images per genotype. (c) Growth stage distribution. (d) The 31-colour palette used for leaf instance annotation.

Table 2: Comparison with existing public top-view RGB phenotyping datasets. PhenoLeaf-TS is the only dataset providing temporally consistent instance labels that enable joint evaluation of segmentation, tracking, and growth stage classification. ✓ = supported, – = not supported.

Dataset	Species	Images	Instance	Time-series	Tracking	Growth	Public
Ara2012 [20]	<i>A. thaliana</i>	120	✓	–	–	–	✓
Ara2013 [20]	<i>A. thaliana</i>	27	✓	–	–	–	✓
Ara2014 [2]	<i>A. thaliana</i>	168	✓	–	–	–	✓
Tobacco [2, 20]	Tobacco	89	✓	–	–	–	✓
MSU-PID [9]	<i>A. thaliana</i> /Bean	2,000+	Partial	✓	–	–	✓
Komatsuna [33]	Komatsuna	900	✓	✓	✓	–	✓
PhenoBench [35]	Sugar beet	1,500+	Panoptic	–	–	–	✓
PhenoLeaf-TS (Ours)	<i>A. thaliana</i> (21)	17,082	✓	✓	✓	✓	✓

3.4 Comparison with Existing Datasets

Tab. 2 positions PhenoLeaf-TS relative to existing benchmarks. CVPPP datasets (Ara2012, Ara2013, Ara2014) provide instance masks but are single-time point collections without temporal ordering. Komatsuna is the closest in scope, providing time-series data with tracking annotations, but covers only one species and 900 images. MSU-PID includes temporal sequences but lacks full instance-level labels. PhenoBench provides panoptic segmentation of field crops but at the semantic level without per-leaf identities. PhenoLeaf-TS is the first dataset to combine large-scale temporally ordered imagery, consistent per-leaf instance labels, and annotations for all three tasks at a scale of over 17,000 images across 21 *Arabidopsis* genotypes.

4 Benchmark Tasks and Evaluation Protocols

We define three benchmark tasks. All experiments use a 70/15/15 random split at the frame level within each plant replicate. For tracking, each plant replicate is treated as a temporal sequence. Frames are split within the same replicate, and the held-out test frames are evaluated in chronological order.

4.1 Task 1: Leaf Instance Segmentation

Given a single RGB image, the goal is to produce a set of binary masks, one per leaf instance. We report mask-level mean Average Precision (mAP) at IoU thresholds of 0.50 and 0.75, following COCO conventions, alongside the Symmetric Best Dice (SBD) [28] used in the CVPPP challenge, the Dice coefficient, and per-class IoU.

4.2 Task 2: Leaf Tracking

Given a temporal sequence of images of the same plant, the goal is to maintain consistent assigned identity for each leaf across frames. We formulate this as multi-object tracking (MOT), where detections from Task 1 models serve as input. We adopt the standard MOT metrics [3, 18]: MOTA, MOTP, HOTA, IDF1, and identity switches (IDSW), along with precision and recall.

4.3 Task 3: Growth Stage Classification

Given a single RGB image, the goal is to classify the plant into one of three growth stages: Early (4–6 leaves), Intermediate (7–10 leaves), or Mature (11+ leaves). Ground-truth labels are derived from the instance masks. We report accuracy, macro F1, precision, and Matthews Correlation Coefficient (MCC) [8].

5 Experiments

5.1 Implementation Details

Instance segmentation. YOLOv11-seg and YOLO26-seg are trained using Ultralytics [15] for 100 epochs (batch 16, lr 0.01, input 640×640). Mask R-CNN variants and Cascade Mask R-CNN use Detectron2 [37] for 100 epochs (batch 8, lr 0.001, SGD with momentum). Mask2Former uses a Swin-L backbone [7] for 25 epochs (batch 4, lr 10^{-4} , AdamW with cosine schedule). SAM2 [24] (IS-7) uses automatic mask generation with no fine-tuning; SAM3 (IS-9) combines Grounding DINO [16] text prompts with SAM3 mask generation.

Object tracking. All six trackers use IS-1 (YOLOv11-seg) as the detector, following the tracking-by-detection paradigm. We use the BoxMOT library with default hyperparameters for each tracker.

Table 3: Instance segmentation results on the PhenoLeaf-TS test set. Best in **bold**, second-best underlined. Models grouped by architecture family.

ID	Model	Backbone	mAP	mAP ₅₀	mAP ₇₅
<i>Single-stage detectors</i>					
IS-1	YOLOv11-seg [15]	CSPDarknet-X	68.2	84.7	–
IS-2	YOLO26-seg [15]	CSPDarknet-L	67.3	84.3	–
<i>Two-stage detectors</i>					
IS-3	Mask R-CNN [13]	R50-FPN	73.2	<u>94.3</u>	82.9
IS-4	Mask R-CNN [13]	R101-FPN	<u>72.4</u>	93.4	81.6
IS-5	Mask R-CNN [13]	X101-FPN	71.3	91.4	80.4
IS-6	Cascade Mask R-CNN [4]	R50-FPN	71.9	91.5	81.8
<i>Transformer-based</i>					
IS-8	Mask2Former [7]	Swin-L	71.9	94.6	<u>82.7</u>
<i>Foundation models</i>					
IS-7	SAM2 (AMG) [24]	Hiera-L	55.0	66.7	63.2
IS-9	SAM3 [6, 16, 24]	Hiera-L	71.1	87.9	81.2

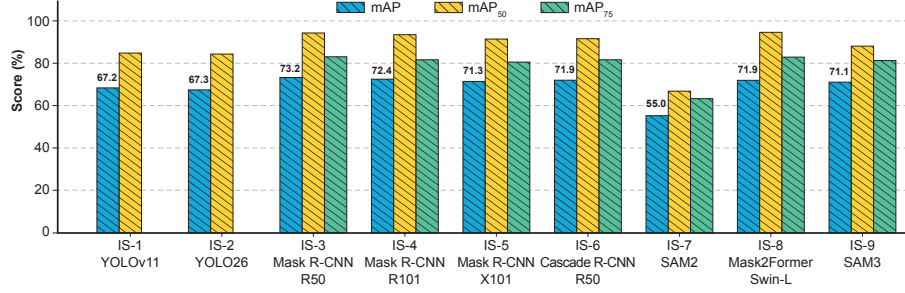
Classification. All classifiers are fine-tuned from ImageNet [10] pre-trained weights (DINOv2 [22] for CL-5) for 100 epochs with early stopping (patience 10), batch 32, input 224×224 , AdamW with cosine annealing. All experiments run on 4x NVIDIA A100 GPUs.

5.2 Instance Segmentation Results

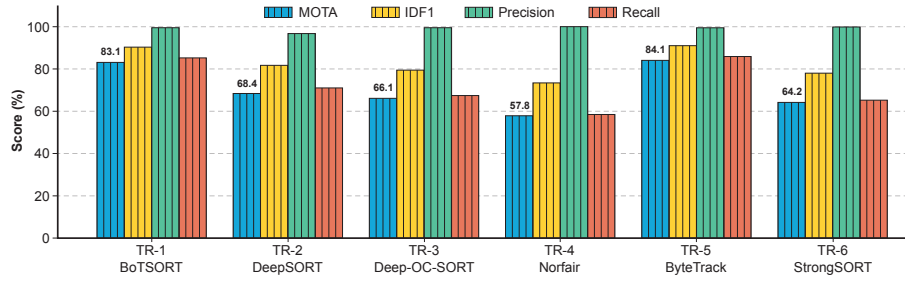
Tab. 3 compares nine architectures spanning four families. Fig. 3a visualises the achieved results.

Two-stage detectors lead overall. Mask R-CNN with R50-FPN (IS-3) achieves the highest mAP of 73.2, followed by the R101 (IS-4, 72.4) and X101 (IS-5, 71.3) variants. Increasing backbone capacity consistently *decreases* performance: R50 > R101 > X101. With over 17,000 training images, the bottleneck appears to be data diversity rather than model capacity, as larger backbones overfit more readily. The cascade refinement stage (IS-6) does not improve over single-stage Mask R-CNN, suggesting that leaf geometry does not benefit from multi-threshold bounding box regression.

Transformer and foundation models. Mask2Former (IS-8) achieves the highest mAP₅₀ (94.6) and matches Cascade Mask R-CNN in mAP (71.9), indicating that attention-based decoders capture leaf boundaries effectively at coarser IoU thresholds. Zero-shot SAM2 with automatic mask generation (IS-7) reaches only 55.0 mAP an 18.2-point gap behind IS-3. SAM3 (IS-9), guided by text prompts via Grounding DINO (IS-9), performance rises to 71.1 mAP. This 16.1-point gap between prompted and unprompted SAM2 underscores the importance of domain-specific guidance for foundation models in plant phenotyping.

(a) Leaf instance segmentation results: mAP, mAP₅₀, mAP₇₅ across nine models.

(b) Multi-object tracking results: MOTA, IDF1, Precision, and Recall across six trackers.

**Fig. 3: Benchmark results for instance segmentation and leaf tracking.**

(a) Two-stage detectors (IS-3 to IS-6) dominate overall; Mask2Former (IS-8) leads at mAP₅₀. Zero-shot SAM2 (IS-7) shows a large gap that text prompting (IS-9) partially closes. (b) ByteTrack (TR-5) and BoTSORT (TR-1) achieve the highest MOTA and IDF1.

Single-stage detectors. YOLOv11-seg (IS-1) and YOLO26-seg (IS-2) achieve mAP of 68.2 and 67.3, trailing the best two-stage model by ~ 5 points but offering real-time inference suitable for automated phenotyping pipelines.

5.3 Tracking Results

Tab. 4 reports tracking performance across the 318 test sequences; Fig. 3b provides a visual comparison.

ByteTrack (TR-5) achieves the highest MOTA (84.1%) and IDF1 (91.0%), narrowly surpassing BoTSORT (TR-1, 83.1% MOTA). ByteTrack’s two-threshold association retains every detection, including low-confidence proposals, for secondary matching. This is well-suited for leaf tracking, where newly emerged leaves often produce weak detector responses that single-threshold methods discard. BoTSORT ranks second because its camera-motion compensation also accommodates the gradual spatial shifts of growing leaves.

DeepSORT (TR-2) produces the fewest identity switches (43 vs. ByteTrack’s 590), owing to its learned appearance descriptor that reliably re-identifies leaves despite morphological changes. However, its conservative matching threshold

Table 4: Leaf tracking results on PhenoLeaf-TS test sequences. All trackers use IS-1 (YOLOv11-seg) detections. Best in **bold**. $\uparrow$: higher is better; $\downarrow$: lower is better.

ID	Tracker	MOTA $\uparrow$	MOTP $\downarrow$	IDF1 $\uparrow$	IDSW $\downarrow$	Prec $\uparrow$	Recall $\uparrow$
TR-5	ByteTrack [38]	84.1	29.8	91.0	590	99.5	85.9
TR-1	BoTSORT [1]	83.1	27.1	90.3	692	99.5	85.1
TR-2	DeepSORT [36]	68.4	33.7	81.7	43	96.7	70.9
TR-3	Deep-OC-SORT [5]	66.1	23.4	79.5	417	99.6	67.4
TR-6	StrongSORT [12]	64.2	24.2	78.0	345	99.8	65.2
TR-4	Norfair [31]	57.8	–	73.4	71	99.3	58.4

Table 5: Growth stage classification results on the PhenoLeaf-TS test set. Best in **bold**, second-best underlined.

ID	Model	Pre-train	Acc	F1 _{mac}	Prec	MCC
CL-6	Swin-T [17]	ImageNet	91.7	91.7	92.4	0.873
CL-3	EfficientNetV2-S [30]	ImageNet	<u>91.4</u>	<u>91.4</u>	<u>91.5</u>	<u>0.869</u>
CL-2	ResNet-101 [14]	ImageNet	89.7	89.6	90.1	0.843
CL-1	ResNet-50 [14]	ImageNet	89.4	89.4	89.9	0.838
CL-5	DINOv2-B [22]	DINOv2	87.1	87.1	88.2	0.803
CL-4	ViT-B-16 [11]	ImageNet	85.1	85.3	86.1	0.771

limits recall to 70.9%. Deep-OC-SORT (TR-3) achieves the best MOTP (23.4) but suffers from frequent identity switches (417), as motion-only cues are insufficient when leaves grow at varying rates. Norfair (TR-4) has the lowest MOTA (57.8%), indicating that the lightweight distance-based association is insufficient for this domain.

The results reveal a fundamental difference from pedestrian tracking: identity maintenance during growth, rather than entry/exit handling, is the primary challenge. ByteTrack’s success without explicit appearance features demonstrates that inclusive detection handling matters more than appearance-based re-identification in this growth-dominated setting.

5.4 Growth Stage Classification Results

Tab. 5 evaluates six classifiers in the three-class growth stage task. Swin-T (CL-6) and EfficientNetV2-S (CL-3) perform within 0.3 percentage points of each other (91.7% vs. 91.4% accuracy), with MCC scores above 0.86 confirming balanced performance across classes. ResNet-50 and ResNet-101 (CL-1, CL-2) achieve 89.4–89.7%, with minimal gain from the deeper variant.

ViT-B-16 (CL-4) underperforms at 85.1%, consistent with observations that ViTs require larger datasets or stronger regularisation to match CNNs [11]. DINOv2-B (CL-5) reaches 87.1% with a frozen backbone and only a linear head fine-tuned, showing that self-supervised features transfer to plant phenotyping without full fine-tuning.

5.5 Qualitative Analysis

Fig. 4 illustrates qualitative segmentation predictions from representative models on two test images: a mature rosette with 15+ leaves (Fig. 4, row a) and an intermediate-stage rosette (Fig. 4, row b). This comparison reveals architecture-specific failure modes that complement the quantitative results in Tab. 3.

Mature rosettes (row a). In dense rosettes, two-stage detectors (Mask R-CNN R50, R101; Cascade R50) produce masks that closely match the ground truth, correctly separating overlapping leaf pairs and preserving petiole structure. The Cascade variant with Swin-L backbone achieves comparable segmentation quality, with slightly smoother mask boundaries owing to the transformer’s global context. YOLOv11 and YOLO26 capture the overall structure but produce noisier leaf boundaries, consistent with their ~ 5 -point mAP deficit. Mask2Former (IS-8) generates clean masks that closely match the ground truth, but occasionally merges tightly overlapping leaf tips. SAM2 (IS-7) over-segments the scene producing spurious mask regions from soil texture and pot edges while SAM3 (IS-9) constrains predictions to leaf-like regions, recovering most instances.

Intermediate rosettes (row b). In plants of younger age, the difficulty shifts from overlap resolution to small-object detection. All fine-tuned models correctly segment the 9–10 visible leaves, but quality differences emerge at petioles and leaf primordia. Mask R-CNN R50 (IS-3) produces the tightest petiole boundaries, while R101 (IS-4) generates marginally coarser masks consistent with the inverse backbone scaling trend observed quantitatively. SAM2 (IS-7) again over-segments the background, producing spurious coloured regions across the soil, while SAM3 constrains segmentation to the plant but still misses the smallest emerging leaves. The contrast between rows (a) and (b) illustrates that model ranking is consistent across growth stages: Mask R-CNN R50 and Mask2Former lead regardless of plant maturity, while foundation models require domain guidance at all stages.

Tracking analysis. The tracking results in Tab. 4 reveal distinct failure modes across tracker families. Among the genotypes in our dataset, we observed distinct rosette architectures posing varied challenges: Ba4-1 develops compact rosettes with overlapping leaves, Col-0 exhibits rapid radial expansion, and Cvi-0 produces elongated, well-separated leaves. Temporal annotation sequences illustrating these morphological differences and the corresponding tracking behaviour are provided in the supplementary material.

Tracking failure modes. ByteTrack’s identity switches (590 IDSW) predominantly occur when newly emerged leaves receive temporary identities during their first 2–3 frames before merging with correct tracks an artefact of two-threshold association where small leaf detections lack spatial overlap with existing tracks. BoTSORT (692 IDSW) shows similar behaviour from over-correcting growth-induced spatial shifts. DeepSORT (43 IDSW) maintains identity via its

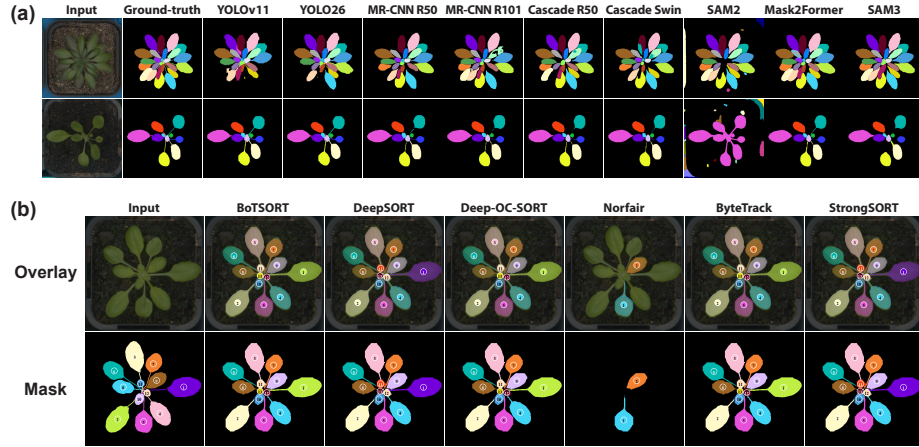


Fig. 4: Qualitative leaf instance segmentation and tracking results. (a) Instance segmentation predictions for a mature rosette (top row) and an intermediate-stage rosette (bottom row). Columns (left to right) show the input RGB image, ground truth, YOLOv11 (IS-1), YOLO26 (IS-2), Mask R-CNN R50 (IS-3), Mask R-CNN R101 (IS-4), Cascade Mask R-CNN R50 (IS-6), Cascade Mask R-CNN Swin-L (IS-7), SAM2, Mask2Former (IS-8), and SAM3 (IS-9). (b) Tracking-by-detection results using IS-1 detections with BoT-SORT, DeepSORT, Deep-OC-SORT, Norfair, ByteTrack, and StrongSORT; rows show the ID overlay and the corresponding tracked instance masks. appearance descriptor but limits recall to 70.9% due to conservative matching; a domain-adapted appearance model would likely improve recall. Deep-OC-SORT and StrongSORT achieve high precision ($>99.5\%$), but low recall (65–67%), while Norfair breaks down in dense rosettes (57.8% MOTA).

Classification confusion patterns. Fig. 5 presents confusion matrices for six chosen classifiers. The dominant error across all models is misclassification between adjacent stages: Early→Intermediate confusion accounts for 5–17% of Early-class samples, while Intermediate→Mature confusion affects 6–13% of Intermediate-class samples, concentrated at leaf counts of 6–7 and 10–11 where the visual difference is subtle. Swin-T (CL-6) and EfficientNetV2-S (CL-3) show the most balanced confusion matrices with near-symmetric error distributions. ViT-B-16 (CL-4) disproportionately confuses Early plants as Intermediate (17% of Early samples), while DINOv2-B (CL-5) achieves a cleaner boundary despite a frozen backbone, indicating that self-supervised features capture growth-stage patterns without task-specific fine-tuning. No classifier confuses Early with Mature (0% across all models), confirming that the three-class definition captures visually distinct phenotypic states at the extremes.

5.6 Cross-Dataset Transfer

A key measure of a benchmark dataset’s utility is whether models trained on it transfer to new domains. We evaluate transferability in two settings: (i) *zero-shot*, where models trained on PhenoLeaf-TS are directly applied to external

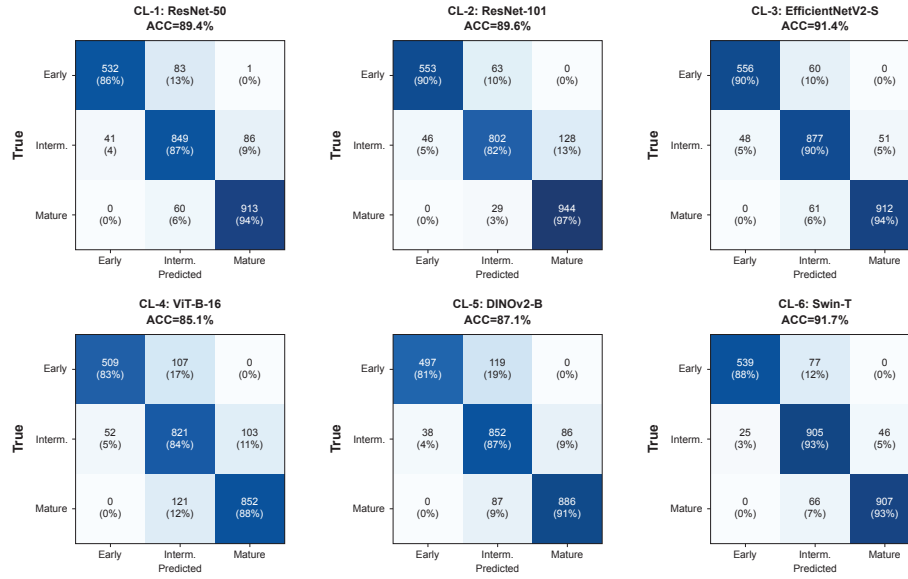


Fig. 5: Classification confusion matrices. All six classifiers show errors concentrated at adjacent stage boundaries (Early/Intermediate at 6–7 leaves, Intermediate/Mature at 10–11 leaves). No model confuses Early with Mature. Swin-T (CL-6) achieves the most balanced error distribution with 91.7% overall accuracy.

datasets without further training, and (ii) *fine-tuned*, where PhenoLeaf-TS pre-trained weights initialise training on the same 70/15/15 split of each external dataset. Tab. 6 and Fig. 6 present the transfer results.

Zero-shot segmentation. Within the same species family (Arabidopsis), both models transfer at 51–60 mAP on Ara2012 and Ara2013. Mask2Former generalises better on Ara2012 (59.9 vs. 55.3), likely because attention captures more transferable boundary features. Cross-species transfer to Komatsuna is harder: Mask R-CNN drops to 34.7 mAP.

Fine-tuning gains. Fine-tuning from PhenoLeaf-TS weights consistently improves performance. The largest gain occurs in Komatsuna, where Mask R-CNN jumps from 34.7 to 86.1 mAP ($\Delta = +51.4$), confirming that our annotations capture leaf morphology features that transfer across different species. On Ara2012 and Ara2013, fine-tuning yields +11–22 mAP for leaf instance segmentation.

Classification transfer. Ara2012 classification serves mainly as a sanity check because all 120 images depict mature plants, so the 100% zero-shot result collapses to a single-class setting. Ara2013 is more diagnostic: zero-shot accuracy is only 21%, while fine-tuning improves performance to 67–78%, with Swin-T outperforming EfficientNetV2 by 11 points. Thus, the strongest transfer evidence comes from Ara2013 classification together with the segmentation gains observed on Ara2012, Ara2013, and Komatsuna.

Table 6: Cross-dataset transfer results. Zero-shot: models trained only on PhenoLeaf-TS. FT: fine-tuned from PhenoLeaf-TS weights. Δ : absolute improvement from fine-tuning. Segmentation metric: mAP; classification metric: accuracy (%).

Dataset	Task Model	Zero-shot	FT	Δ
Ara2012 [20]	Seg. Mask R-CNN (IS-3)	55.3	73.9	+18.6
	Seg. Mask2Former (IS-8)	59.9	71.1	+11.2
	Cls. EfficientNetV2 (CL-3)	100.0	100.0	0.0
	Cls. Swin-T (CL-6)	100.0	100.0	0.0
Ara2013 [20]	Seg. Mask R-CNN (IS-3)	53.5	75.2	+21.7
	Seg. Mask2Former (IS-8)	51.3	66.3	+15.0
	Cls. EfficientNetV2 (CL-3)	21.0	67.0	+46.0
	Cls. Swin-T (CL-6)	21.0	78.0	+57.0
Komatsuna [33]	Seg. Mask R-CNN (IS-3)	34.7	86.1	+51.4
	Seg. Mask2Former (IS-8)	47.1	76.7	+29.6

6 Discussion and Limitations

Practical guidance. For segmentation, Mask R-CNN R50-FPN offers the best accuracy; YOLO variants sacrifice ~ 5 mAP for real-time throughput. For tracking, ByteTrack is the best overall choice (84.1% MOTA), while DeepSORT is preferable when minimising identity switches is critical. For classification, Swin-T and EfficientNetV2 perform nearly identically; the latter is more efficient for deployment. In combined pipelines, the detector choice is consequential: switching from IS-1 to IS-3 would likely improve tracking metrics, as higher-quality detections reduce the association burden.

Backbone scaling. Across both segmentation (R50 > R101 > X101) and classification (ResNet-50 $\approx$ ResNet-101), larger backbones do not improve overall results in this dataset, suggesting that data diversity rather than model capacity is the limiting factor with $\sim 17,000$ images. This has practical implications for edge deployment in indoor settings, where smaller backbones deliver equivalent accuracy at lower cost. Performance gains from larger architectures would likely emerge with additional genetic diversity or field-condition imagery.

Foundation models. The 16.1-point mAP gap between unprompted SAM2 (IS-7, 55.0) and text-prompted SAM3 (IS-9, 71.1) highlights that zero-shot generalisation to plant imagery remains unreliable without domain guidance. Text prompting constrains generation to leaf-like objects, but the remaining 2.1-point gap to IS-3 (73.2) indicates prompt engineering alone does not close the domain gap relevant to the plant phenotyping where foundation models are increasingly adopted without fine-tuning.

Growth tracking Leaf tracking departs fundamentally from pedestrian tracking: objects rarely leave the scene, and the principal challenges are continuous

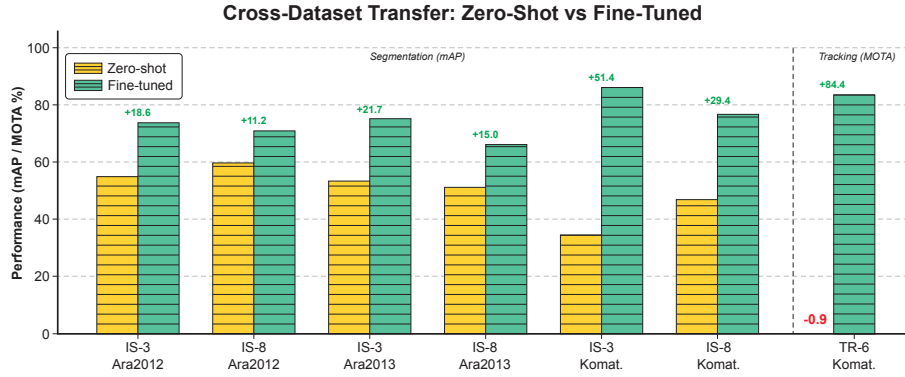


Fig. 6: Cross-dataset transfer. Zero-shot vs. fine-tuned performance on Ara2012, Ara2013, and Komatsuna. Fine-tuning from PhenoLeaf-TS weights demonstrates consistent improvements, with the largest gain (+51.4 mAP) occurring for cross-species transfer to Komatsuna.

morphological deformation, gradual spatial displacement, and emergence of new leaves between existing ones. ByteTrack’s 84.1% MOTA demonstrates that inclusive detection handling is more valuable than appearance-based re-identification in this setting.

Pre-training value. The cross-dataset experiments demonstrate that PhenoLeaf-TS serves as an effective pre-training source, with fine-tuning gains of +18–51 mAP on external datasets. The largest improvements occur for cross-species transfer (Komatsuna), confirming that the dataset captures general leaf morphology features that transfer across different plant species. The Ara2012/Ara2013 classification asymmetry is instructive: Ara2012’s 100% zero-shot accuracy reflects its homogeneous maturity distribution, while Ara2013’s recovery from 21% to 67–78% after fine-tuning confirms that PhenoLeaf-TS provides developmental representations requiring only light adaptation.

Limitations. (i) The dataset covers a controlled indoor environment with top-view imaging; field environments introduce variable illumination and weed occlusion. (ii) The fixed 31-colour palette simplifies evaluation but requires re-mapping for external datasets with different labeling conventions. (iii) Although the annotations have been carefully prepared, minor labelling errors may occur, such as imprecise leaf boundary delineation, misjudged leaf elevation order, or swapped colour IDs for small, closely spaced leaves at low resolution. These infrequent cases do not significantly affect overall benchmark reliability. (iv) Growth stage boundaries at leaf counts of 6 and 10 may be ambiguous even for human annotators. (v) Standard MOT metrics, designed for pedestrian tracking, may not fully capture plant-specific challenges such as leaf emergence detection and maintaining temporally consistent growth trajectories. (vi) Extension to field conditions, continuous growth regression, and self-supervised pre-training on the temporal sequences are directions for future work.

7 Conclusion

We introduced PhenoLeaf-TS, a time-series dataset of 17,082 RGB images spanning 21 Arabidopsis genotypes with temporally consistent, colour-coded leaf instance annotations, and established baselines across three tasks: instance segmentation, leaf tracking, and growth stage classification. The benchmark results show strong performance across all tasks, with Mask R-CNN R50 achieving 73.2 mAP for leaf instance segmentation, ByteTrack reaching 84.1% MOTA for tracking individual leaves across image sequences, and Swin-T achieving 91.7% accuracy for growth stage classification. Cross-dataset transfer experiments further demonstrate the dataset’s value as a pre-training source, with fine-tuning gains of up to +51 mAP across species boundaries. By combining temporal depth, consistent leaf identities, and standardised evaluation protocols, PhenoLeaf-TS provides a practical benchmark for developing and comparing computer-vision methods for leaf-level plant phenotyping. Future work will extend this benchmark towards outdoor conditions, major crop species, continuous growth-regression tasks, and self-supervised learning from temporal plant image sequences.

Acknowledgments

The authors gratefully acknowledge the valuable support and collaboration of Prof. James Whelan at Zhejiang University. We also sincerely thank Prof. Mathew G. Lewsey from the Australian Plant Phenomics Network (APPN) node at La Trobe University, as well as Prof. Tony Bacic and Dr. Veronica Borret from the La Trobe Institute for Sustainable Agriculture and Food (LISAF) and the ARC Research Hub for Protected Cropping (PC Hub), La Trobe University, for their guidance, expertise, and continued support throughout this work. We extend our appreciation to Martin Trtilek, Klara Panzarova, Ivan Kashkan, and Sajid Ullah from Photon Systems Instruments (PSI) for their technical support, collaboration, and assistance with plant phenotyping systems and data-related activities. Their contributions, expertise, and collaborative input have been highly valuable in supporting the development and completion of this research.

This work was supported by the Australian Research Council Industrial Transformation Research Hub for Medicinal Agriculture (ARC MedAg Hub) through Grant Number IH180100006 and the ARC Industrial Transformation Research Hub for Protected Cropping through Grant Number IH240100024.

References

1. Aharon, N., Orfaig, R., Bobrovsky, B.Z.: BoT-SORT: Robust associations multi-pedestrian tracking. In: arXiv preprint arXiv:2206.14651 (2022)
2. Bell, J., Dee, H.M.: An open image dataset for plant segmentation and counting. Proceedings of the European Workshop on Visual Information Processing (EUVIP) (2016)

3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. In: *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10 (2008). <https://doi.org/10.1155/2008/246309>
4. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: *CVPR*. pp. 6154–6162 (2018)
5. Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In: *CVPR*. pp. 9686–9696 (2023)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), <https://arxiv.org/abs/2511.16719>
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *CVPR*. pp. 1290–1299 (2022)
8. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**(1), 6 (2020). <https://doi.org/10.1186/s12864-019-6413-7>
9. Cruz, J.A., Yin, X., Liu, X., Imber, S.M., Sozzani, R., Kramer, D.M., Chen, J.: Multi-modality imagery database for plant phenotyping. *Machine Vision and Applications* **27**(5), 735–749 (2016). <https://doi.org/10.1007/s00138-015-0734-6>
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *CVPR*. pp. 248–255 (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR* (2021)
12. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: StrongSORT: Make DeepSORT great again. *IEEE TPAMI* **45**(9), 11394–11404 (2023). <https://doi.org/10.1109/TPAMI.2023.3270921>
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: *ICCV*. pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778 (2016)
15. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO. <https://github.com/ultralytics/ultralytics> (2023)
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: *ECCV* (2024)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *ICCV*. pp. 10012–10022 (2021)
18. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: A higher order metric for evaluating multi-object tracking. In: *IJCV*, vol. 129, pp. 548–578 (2021). <https://doi.org/10.1007/s11263-020-01375-2>
19. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Kim, T.K.: Multiple object tracking: A literature review. *Artificial Intelligence* **293**, 103448 (2021). <https://doi.org/10.1016/j.artint.2020.103448>

20. Minervini, M., Abdelsamea, M.M., Tsaftaris, S.A.: Image-based plant phenotyping with incremental learning and active contours. *Ecological Informatics* **23**, 35–48 (2014). <https://doi.org/10.1016/j.ecoinf.2013.07.004>
21. Minervini, M., Fischbach, A., Tsaftaris, S.A., Scharr, H.: Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognition Letters* **81**, 80–89 (2016). <https://doi.org/10.1016/j.patrec.2015.10.013>
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024)
23. Pape, J.M., Klukas, C.: Utilizing machine learning approaches to improve the prediction of leaf counts and individual leaf segmentation of rosette plant images. In: *Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP)* (2015)
24. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024)
25. Ren, M., Zemel, R.S.: End-to-end instance segmentation with recurrent attention. In: *CVPR*. pp. 6656–6664 (2017)
26. Romera-Paredes, B., Torr, P.H.S.: Recurrent instance segmentation. In: *ECCV*. pp. 312–329 (2016)
27. Sarić, R., Čustović, E., Akagić, A., Trtílek, M., Lewsey, M.G., Whelan, J.: Computer vision-based recognition and distinction of arabidopsis thaliana ecotypes using supervised deep learning models. *The Plant Phenome Journal* **8**(1), e70041 (2025). <https://doi.org/https://doi.org/10.1002/ppj2.70041>
28. Scharr, H., Minervini, M., French, A.P., Klukas, C., Kramer, D.M., Liu, X., Luengo, I., Pape, J.M., Polder, G., Vukadinovic, D., et al.: Leaf segmentation in plant phenotyping: a collation study. In: *Machine Vision and Applications*. vol. 27, pp. 585–606 (2016). <https://doi.org/10.1007/s00138-015-0737-3>
29. Singh, A., Ganapathysubramanian, B., Singh, A.K., Sarkar, S.: Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science* **21**(2), 110–124 (2016). <https://doi.org/10.1016/j.tplants.2015.10.015>
30. Tan, M., Le, Q.V.: EfficientNetV2: Smaller models and faster training. In: *ICML*. pp. 10096–10106 (2021)
31. Tryolabs: Norfair: Customizable lightweight Python library for real-time multi-object tracking. <https://github.com/tryolabs/norfair> (2022)
32. Tsaftaris, S.A., Minervini, M., Scharr, H.: Machine learning for plant phenotyping needs image processing. *Trends in Plant Science* **21**(12), 989–991 (2016). <https://doi.org/10.1016/j.tplants.2016.10.002>
33. Uchiyama, H., Sakurai, S., Mishima, M., Arita, D., Okada, T., Takamatsu, T., Taniguchi, R.i.: An easy-to-setup 3D phenotyping platform for Komatsuna dataset. *ICCV Workshops* (2017)
34. Ward, D., Moghadam, P.: Scalable learning for bridging the species gap in image-based plant phenotyping. *Computer Vision and Image Understanding* **197**, 103009 (2020). <https://doi.org/10.1016/j.cviu.2020.103009>
35. Weyler, J., Magistri, F., Marks, E., Chong, Y.L., Sodano, M., Roggiolani, G., Behley, J., Stachniss, C.: PhenoBench: A large multi-sensor dataset for crop-weed segmentation in agricultural fields. *IEEE Robotics and Automation Letters* (2024). <https://doi.org/10.1109/LRA.2024.3382524>

36. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: IEEE International Conference on Image Processing (ICIP). pp. 3645–3649 (2017)
37. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
38. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: ByteTrack: Multi-object tracking by associating every detection box. In: ECCV. pp. 1–21 (2022)
39. Zheng, Y.J., Kong, J.L., Jin, X.B., Wang, X.Y., Su, T.L., Zuo, M.: CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors* **19**(5), 1058 (2019). <https://doi.org/10.3390/s19051058>