LeafTrackNet: A Deep Learning Framework for Robust Leaf Tracking in Top-Down Plant Phenotyping

Shanghua Liu
Leibniz Institute for Agricultural Engineering and Bioeconomy (ATB)
14469 Potsdam, Germany
sliu@atb-potsdam.de

Majharulislam Babor
ATB
14469 Potsdam, Germany
MBabor@atb-potsdam.de

Christoph Verduyn
BASF Belgium Coordination Center CommV
2040 Antwerpen, Belgium
christoph.verduyn@basf.com

Breght Vandenberghe
BASF Belgium Coordination Center CommV
2040 Antwerpen, Belgium
breght.vandenberghe@basf.com

Bruno Betoni Parodi
BASF Metabolome Solutions GmbH
10589 Berlin, Germany
bruno.betoni@basf.com

Cornelia Weltzien
ATB, Technical University Berlin
10623 Berlin, Germany
CWeltzien@atb-potsdam.de

Marina M.-C. Höhne
ATB, University of Potsdam
14469 Potsdam, Germany
mhoehne@atb-potsdam.de

Abstract

Leaf-level phenotyping can reveal early signals of plant growth and stress, making it a key step toward understanding crop development. However, tracking individual leaves over time is still challenging, especially for structurally complex crops such as canola. This difficulty stems both from the scarcity of realistic, publicly available benchmark datasets and from the limitations of current methods: existing plant-specific tracking methods often rely on Intersection-over-Union (IoU) thresholds to associate leaves between frames, which can break down when leaves overlap, grow, or change shape. Generic multi-object tracking (MOT) methods, on the other hand, are designed for approximately rigid objects like cars or pedestrians and struggle with the continuous deformation and complex motion patterns of leaves. Therefore, the contribution of our work is two-folded - First, we introduce **CanolaTrack**, a high-resolution dataset of 5,704 top-down RGB images with 31,840 annotated leaf instances spanning the early growth stages of 184 canola plants. Second, we propose **LeafTrackNet**, an efficient lightweight framework for long-term leaf tracking. It combines a YOLOv10 detector with a MobileNetV3 embedding head and links identities via cosine similarity and Hungarian assignment, without geometric motion priors. On CanolaTrack, LeafTrackNet outperforms both plant-specific tracking methods and state-of-the-art MOT baselines, improving HOTA by 9.73%. Our work provides a realistic benchmark dataset and a simple, effective framework for long-term leaf tracking, contributing to AI-driven plant phenotyping. Code and dataset are available at https://github.com/shl-shawn/LeafTrackNet.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Automated plant phenotyping enables high-throughput measurements essential for data-driven precision agriculture. However, whole-plant phenotyping can obscure intra-plant dynamics that anticipate physiological stress and genotype-specific traits [1, 2, 3]. Here, leaves are particularly informative: they drive photosynthesis, respond locally to biotic and abiotic stress, and exhibit measurable traits such as emergence timing, growth rate, and morphological change [4, 5, 6]. Leaf-level phenotypes thus provide temporally resolved, function-valued insights that aggregate metrics cannot capture [7, 8].

However, tracking individual leaves over weeks is a challenging, long-term, non-rigid multi-object tracking (MOT) problem. Leaves emerge, occlude, senesce, or reappear; their appearance evolves with growth and lighting; and pot rotation can induce large global shifts. These dynamics violate common MOT assumptions—near-rigid motion, consistent appearance, and smooth trajectories [9, 10].

In this work, we focus on *Brassica napus* (canola), an economically and ecologically important crop whose vegetative growth stage introduces domain-specific challenges for identity preservation [11, 12]. The rosette structure leads to low inter-leaf visual variance, as leaves often exhibit similar textures, shapes, and reflectance. As the plant develops, leaves continuously emerge, expand, and deform, producing dynamic occlusions and appearance drift across time. At the top of the plant, overlapping leaves and stems frequently obscure one another, causing repeated occlusions hindering tracking of individual leaves over time. Additionally, posture changes and experimental pot rotations disrupt frame-to-frame spatial alignment, making motion-based heuristics unreliable.

Our contributions are as follows:

- **CanolaTrack.** A high-resolution, top-down benchmark dataset for long-term leaf tracking: 184 canola plants imaged daily for 31 days, totaling 5,704 RGB frames and 31,840 leaf instances. It captures realistic biological leaf events, such as birth, death, occlusion, reoccurrence, non-uniform growth, as well as pot rotation.
- **LeafTrackNet.** An efficient tracking framework combining a fine-tuned YOLOv10 detector [13] with a MobileNetV3 embedding head [14] with triplet margin loss. Leaf identities are associated using cosine similarity and Hungarian assignment, without reliance on motion prediction. LeafTrackNet outperforms both plant-specific and general-purpose MOT baselines on CanolaTrack.

2 Related Work

2.1 Leaf-level Tracking Datasets

Compared to leaf classification, segmentation, and counting [15, 16, 17], the task of leaf-level tracking is underexplored, and only a few leaf tracking datasets are available, which are summarized in Table 1. **LeTra** [18] provides 513 chlorophyll fluorescence images from nine *Arabidopsis thaliana* plants across 57 time points with 204 annotated leaves. **KOMATSUNA** [19] contains $\sim$300 RGB-D images of five *Komatsuna* plants recorded every four hours over ten days. **MSU-PID** [20] includes fluorescence, infrared, RGB, and depth for *Arabidopsis* and been plants, with subsets (576 and 172 images respectively) annotated for tracking. Beyond top-down view, **PhenoTrack3D** [21] captures side-view, multi-camera images of maize for 3D reconstruction, but requires complex calibration and alignment, limiting scalability for long-term studies.

2.2 Plant-Specific Leaf Tracking Methods

In the following we provide an overview of the few methods explicitly designed for leaf-level tracking. LeTra [18] segments leaves region with Mask R-CNN and associates leaf instances using IoU-based mask matching, which degrades under heavy overlap and frequent emergence/disappearance/recurrence. PlantDoctor [22] pairs YOLOv8 with DeepSORT to add Re-Identification (ReID) features, but embeddings are not tailored to morphology and growth-stage variation in dense rosettes. In practice, most plant-specific approaches combine generic detectors and tracking heuristics, rather than addressing long-term identity maintenance in plant-specific scenes.

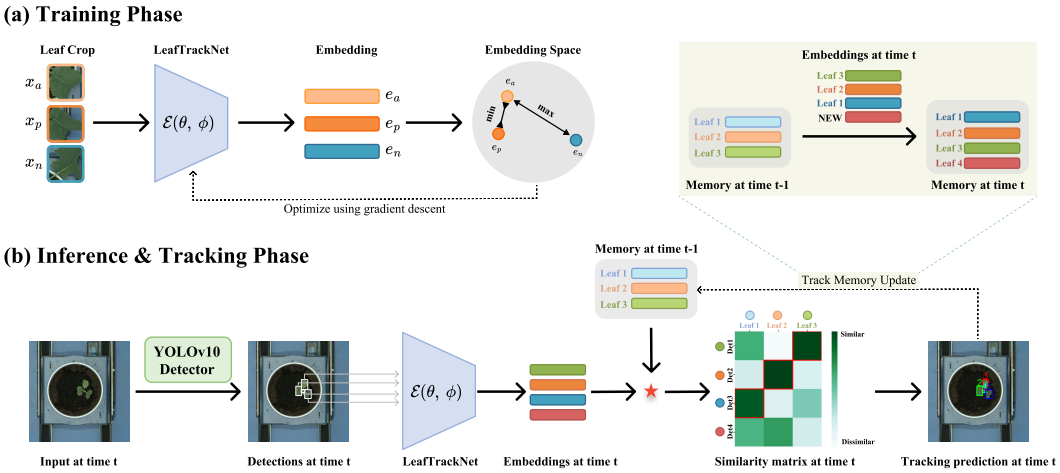

Figure 1: Two-phase framework for leaf tracking. **(a) Training Phase:** Anchor–positive–negative leaf crops are passed through LeafTrackNet and trained via triplet-margin loss to learn a discriminative, temporally consistent embedding space. **(b) Inference & Tracking Phase:** An input RGB image is fed into a fine-tuned YOLOv10 to detect leaves. Detected regions are embedded and compared to stored embeddings in the memory bank using cosine similarity ($\star$) to compute a similarity matrix. Hungarian matching then updates matched tracks' embedding, initializes new tracks and prunes inactive tracks.

2.3 General Multi-Object Tracking Methods

Recent MOT methods have achieved rapid progress in pedestrian and vehicle domains [9, 10, 23, 24]. ByteTrack [10] improves IoU-based association by incorporating low-confidence detections but still relies on box geometry. BoT-SORT [9] adds appearance cues and Kalman filtering, yet assumes smooth, near-rigid motion. End-to-end approaches like MOTRv2 [23] use transformer-based query propagation to jointly detect and track objects and achieves state-of-the-art results on DanceTrack [24] and BDD100K [25]. However, these methods are developed for high-frame-rate videos with strong frame-to-frame continuity. Leaf tracking, by contrast, requires identity persistence across large temporal gaps (e.g., daily frames), with substantial appearance shifts.

3 Method

3.1 Motivation

The design of LeafTrackNet is guided by failure cases observed when applying MOT trackers to biologically complex plant growth sequences. First, geometric association methods, such as IoU or Kalman-based tracking, fail under occlusions and pose changes, which are frequent due to overlapping leaves and rotational artifacts. Second, generic embedding extractors often lack the discriminative capacity to distinguish visually similar leaf instances within a single plant. Finally, end-to-end transformers designed for high-frame-rate pedestrian tracking fail to generalize in temporally sparse, biologically dynamic sequences. These observations motivate a framework that decouples spatial localization from identity matching (Figure 1). A leaf embedding network is trained with triplet margin loss to enforce temporal consistency and intra-plant discriminability. During inference, cosine similarity and memory-based matching support identity propagation without relying on geometric continuity.

3.2 Training Phase

Triplet Sampling. Let I_k^t denote the RGB image of plant k at time t. The annotated leaves are $\mathcal{G}_k^t = \{b_{k,i}^t\}$, where $b_{k,i}^t = (u, v, w, h)$ is the i^{th} leaf bounding box with top-left coordinates (u, v), width w, and height h. Leaf crops are extracted via a crop-and-resize operator ψ, yielding $x_{k,i}^t = \psi(I_k^t, b_{k,i}^t)$. The triplets (x_a, x_p, x_n) are formed as follows: for an anchor $x_a = x_{k,i}^{t_a}$, the

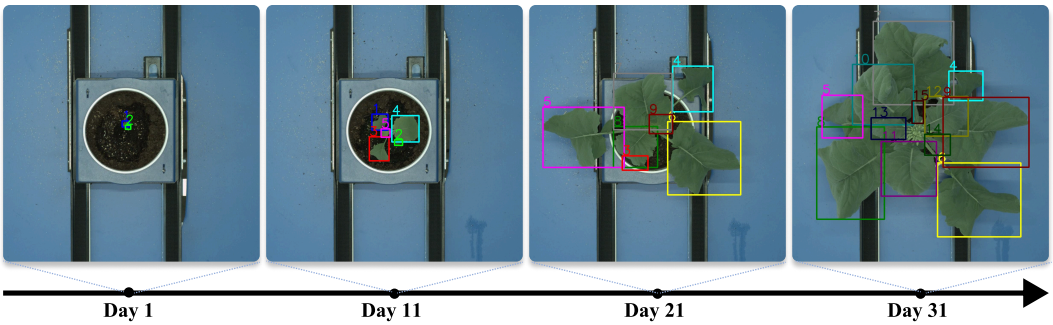

Figure 2: Example images of the plant sample Plant-003 from days 1, 11, 21, and 31 with color-coded bounding boxes indicating individual leaves over time.

positive sample is the same leaf at a different time, $x_p = x_{k,i}^{t_p}, t_p \neq t_a$; the negative is a different leaf from the same plant, $x_n = x_{k,j}^{t_n}, j \neq i$. Thus $(x_a, x_p, x_n) = \left(x_{k,i}^{t_a},\, x_{k,i}^{t_p},\, x_{k,j}^{t_n}\right)$.

Model Architecture. The embedding network $\mathcal{E} = \mathcal{F}_\phi\big(\mathcal{N}_\theta(x)\big) : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^D$ maps each cropped leaf region x to a fixed-dimensional descriptor. It consists of a MobileNetV3 backbone $\mathcal{N}_\theta : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^F$, where F denotes the dimension of the intermediate feature vector, followed by a linear projection head $\mathcal{F}_\phi : \mathbb{R}^F \to \mathbb{R}^D$ that produces the final embedding. The parameters θ and ϕ are jointly optimized during training.

Loss Function. Given a triplet (x_a, x_p, x_n), the corresponding embeddings are: $e_a = \mathcal{E}(x_a), e_p = \mathcal{E}(x_p), e_n = \mathcal{E}(x_n)$. We use the triplet margin loss [26] to enforce that the anchor–positive distance is smaller than the anchor–negative distance by a margin m:

$$\mathcal{L}(x_a, x_p, x_n) = \max\big\{0,\ \|e_a - e_p\|_2^2 - \|e_a - e_n\|_2^2 + m\big\}. \tag{1}$$

3.3 Inference and Tracking Phase

In contrast to the training phase, where triplets are sampled randomly, the inference phase proceeds images sequentially. Inspired by MOTRv2 [23], we use YOLOv10 [13] as our leaf detector, fined-tuned on the CanolaTrack training set. We filter out leaf detections with confidence score below 0.5. For plant k at time t, let $\mathcal{D}^t = \big\{b_i^t\big\}$ be the detected leaf boxes and define their embeddings as $\big\{e_i^t\big\} = \big\{\mathcal{E}\big(\psi(I_k^t,\, b_i^t)\big)\big\}$.

Tracking Memory Bank. We maintain a tracking memory bank $\mathcal{T}^t = \big\{(p_\ell^t, a_\ell^t)\big\}$, where ℓ indexes each active tracks at time t, $p_\ell \in \mathbb{R}^D$ is the prototype embedding for that track, and $a_\ell \in \mathbb{N}$ is the number of consecutive images in which the corresponding identity has not been observed.

Initialization at $t = 1$. Since no tracks exist initially, all detections within the first image are treated as new tracks. Let $\{e_j^1\}_{j=1}^{N_1}$ denote the embeddings extracted from the detections at time $t = 1$. The memory bank is initialized as: $\mathcal{T}^1 = \big\{(p_j^1 = e_j^1,\, a_j^1 = 0)\big\}_{j=1}^{N_1}$, where N_1 is the number of detected leaf in the first image.

Sequential update for $t > 1$. At each subsequent time t, given the active tracks from the memory bank $\{p_\ell^{t-1}\}_{\ell=1}^{N_{t-1}}$ and the current detection embeddings $\{e_j^t\}_{j=1}^{N_t}$, we compute a similarity matrix $S_{\ell j} = \big(p_\ell^{t-1}\big)^\top e_j^t$. A one-to-one assignment $\mathcal{A}$ is obtained by solving a Hungarian match on the cost matrix $C = 1 - S \in \mathbb{R}^{N_{t-1}\times N_t}$. After matching, we apply the similarity threshold τ_s: any pair $(\ell, j) \in \mathcal{A}$ with $S_{\ell j} < \tau_s$ is rejected and treated as unmatched. The resulting sets of matched pairs, unmatched detections, and unmatched tracks are then used to update the memory bank.

Table 1: Comparison of publicly available top-down view leaf tracking datasets.

Dataset	Species	#A. Images	Resolution	#Plants	#Leaves	Δt	Modality	Rot.
LeTra	Arabidopsis	513	266×266	9	204	8	F	✗
KOMATSUNA	Komatsuna	300	$\sim 480 \times 480$	5	–	4	R, D	✗
MSU-PID	Arabidopsis	576	$\sim 120 \times 120$	16	–	1.6	F, I, R, D	✗
MSU-PID	Bean	172	380×720	5	–	1.8	F, I, R, D	✗
CanolaTrack(Ours)	Canola	5,704	1200×1200	184	31,840	24	R	✓

"#A. Images" = number of annotated images; Δt = hours between successive images per plant; "Rot." = pot rotation included; "F", "I", "R", "D" = Fluorescence, Infrared, RGB, Depth; "–" = not reported in original publication.

(i) *Matched Detections (Persistent or Reappearing Leaves):* For each matched pair (ℓ, j), the track's prototype is updated by an exponential moving average (EMA):

$$p_\ell^t = \alpha \cdot e_j^t + (1-\alpha) \cdot p_l^{t-1}, \quad a_\ell^t = 0,$$

where $\alpha \in [0,1]$ control temporal smoothing.

(ii) *Unmatched Detections (New Leaves):* Each unmatched detection j is initialized as a new track in the memory bank:

$$p_{\ell'}^t = e_j^t, \quad a_{\ell'}^t = 0,$$

where $\ell' = N_{t-1} + 1,\ N_{t-1} + 2,\ \cdots$.

(iii) *Unmatched Tracks (Disappearing Leaves):* Any existing tracked leaves in the memory bank that are not matched to the current set of leafs are considered disappeared temporally or forever. For any unmatched leaf ℓ, we keep the embedding prototype unchanged and increment its age:

$$p_\ell^t = p_\ell^{t-1}, \quad a_\ell^t = a_\ell^{t-1} + 1.$$

Any tracked leaf older than age threshold ($a_\ell^t > \tau_a$) is removed from the memory bank.

4 Experiments

4.1 Dataset

CanolaTrack[1] compromises daily top-down RGB images of 184 *Brassica napus* (canola) plants over 31 days, including multiple genotypes (distinct genetic lines) and nutrient regimes (different fertilization levels) to elicit diverse growth patterns. Image acquisition begins at first leaf emergence and continues until floral buds formation, capturing the full vegetative growth phrase with rapid, non-uniform leaf expansion (Figure 2). In total, the dataset contains 5,704 high-resolution images and 31,840 leaf instances annotated with tight bounding boxes. A comparison to existing top-down tracking datasets is provided in Table 1.

We randomly split the data 80/20 by plant: the training set has 147 plants (4,557 images, 25,485 leaves), and test set has 37 unseen plants (1,147 images, 6,355 leaves). All models are trained on the training set and are evaluated on the test set.

4.2 Results

We evaluate LeafTrackNet on CanolaTrack using TrackEval[2], benchmarking against general MOT methods (BoT-SORT, ByteTrack, MOTRv2) and plant-specific trackers (LeTra, Plant-Doctor). All models use the same fine-tuned YOLOv10 detector for per-frame proposals, replacing original detectors (e.g., YOLOX in MOTRv2). Unless stated otherwise, we use official implementations with default hyperparameters and train on the CanolaTrack training split. Implementation details are in Appendix A.

[1] © BASF SE

[2] `https://github.com/JonathonLuiten/TrackEval`

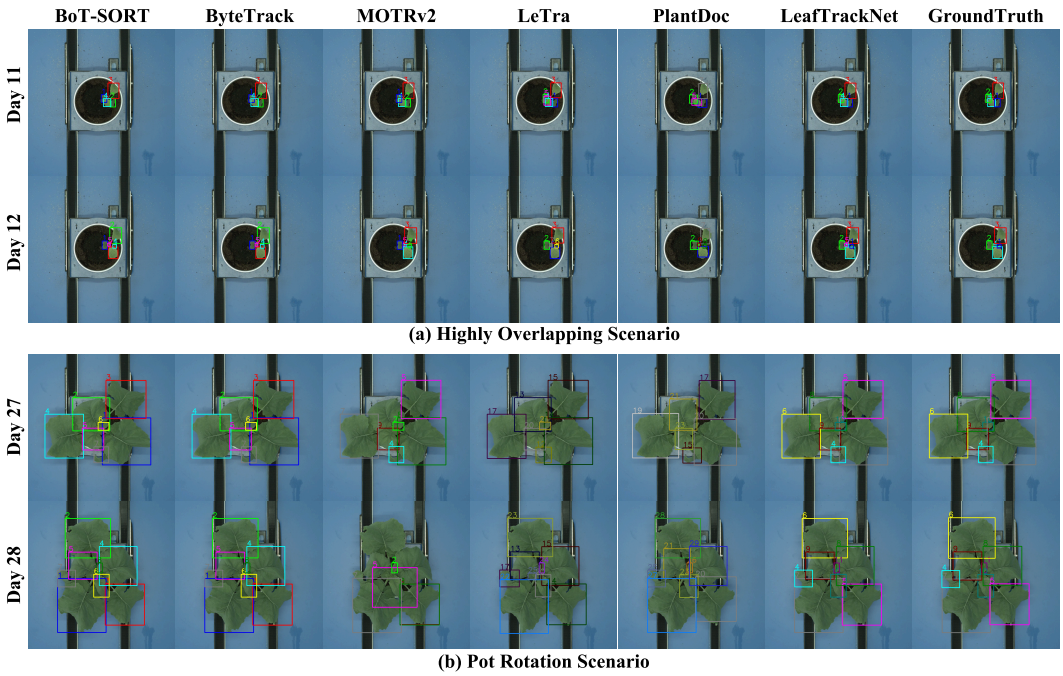

Figure 3: Qualitative tracking results on the plant sample Plant-158 from the CanolaTrack dataset. (a) High-overlap scenario between Day 11 and Day 12. (b) Pot rotation scenario between Day 27 and Day 28. Best shown in the GroundTruth column.

Quantitative Benchmarking. Table 2 reports performance on standard MOT metrics (see Appendix B). Among general trackers, MOTRv2 achieves the best association (e.g., AssA and IDF1), benefiting from transformer query propagation. BoT-SORT and ByteTrack show strong detection performance (DetA: 91.30/91.94) but weak association (AssA: 12.18/12.29). Plant-specific baselines, LeTra and Plant-Doctor, better align with leaf-level structure and achieve improved overall performance (e.g., LeTra HOTA: 67.02, MOTA: 83.09), yet still struggle to maintain identities through occlusion and emergence. LeafTrackNet achieves state-of-the-art performance across all five MOT metrics, outperforming the best competing method by a 9.73 HOTA margin.

Table 2: Tracking performance on the CanolaTrack dataset. Best results are in **bold**; second best are underlined. Results are reported as mean ± standard deviation across three runs. "Improvement" denotes the margin of LeafTrackNet over the strongest competing method for each metric.

Domain	Method	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
General	BoT-SORT	33.32±0.91	91.30±0.21	12.18±0.65	40.35±1.91	26.13±0.85
	ByteTrack	33.58±0.90	91.94±0.15	12.29±0.69	41.88±1.79	26.20±0.83
	MOTRv2	78.30±1.85	77.33±2.72	79.36±1.07	79.68±3.05	83.78±1.94
Plant	LeTra*	67.02±0.04	82.03±0.14	54.98±0.16	82.09±0.19	69.06±0.10
	Plant-Doctor	59.74±0.04	74.42±0.03	48.20±0.09	79.71±0.06	69.56±0.03
	LeafTrackNet	**88.03±0.24**	**92.25±0.03**	**84.07±0.49**	**93.64±0.18**	**92.90±0.35**
	Improvement	+9.73	+0.31	+4.71	+11.55	+9.12

*LeTra originally matches leaves using segmentation masks; here we adapt it to bounding boxes due to the annotation format in CanolaTrack.

Qualitative Analysis. Early growth (Figure 3(a), Day 11 to 12) exhibits frequent occlusions among small, low-contrast leaves. Generic MOT methods and plant-specific baselines tend to misassociate or lose identities, whereas LeafTrackNet maintains identities (e.g., Leaves 1, 2, 4) by combining occlusion/scale-tolerant embeddings with a prototype memory that smooths associations across frames. A ∼90° clockwise pot rotation (Figure 3(b), Day 27 to 28) breaks spatial continuity and

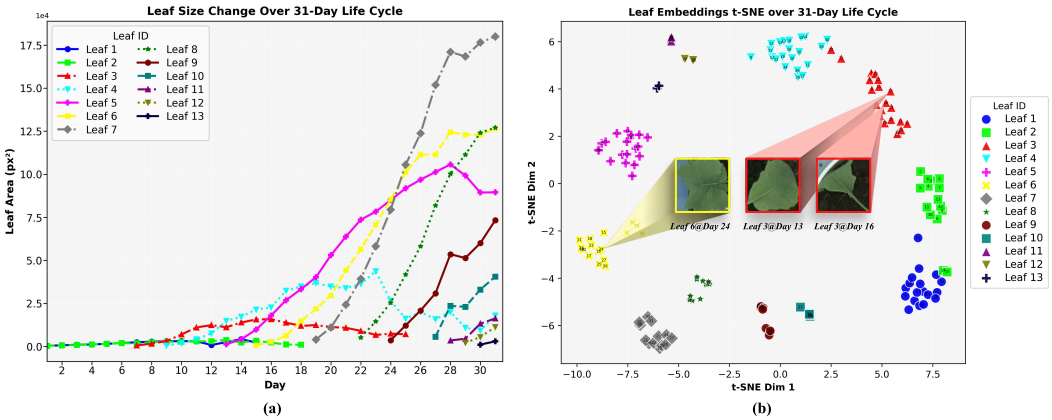

Figure 4: Analysis for the plant sample Plant-158 over 31 days. (a) Per-leaf area trajectories illustrating emergence, disappearance and growth dynamics. (b) t-SNE projection of learned leaf embeddings with day indices. Same-leaf instances cluster tightly despite appearance changes (e.g., Leaf 3 on Days 13 and 16), while different leaves remain well separated (e.g., Leaf 6 on Day 24 vs. Leaf 3).

violates smooth-motion priors. Geometry–Kalman trackers (BoT-SORT, ByteTrack) and mask–IoU association (LeTra) fail under the induced rearrangement. Plant-Doctor's ReID features are sensitive to orientation changes, and MOTRv2's position-encoded query propagation inherits misaligned anchors after rotation. By performing association in feature space with cosine–Hungarian matching over a compact per-leaf memory, LeafTrackNet remains stable under both occlusion and global reorientation.

Temporal Consistency of the Learned Embedding. We analyze Plant-158, a challenging plant sequence with overlap, emergence, deformation, and rotation. Here, *overlap* means occlusion between leaves, *emergence* is the first appearance of new leaves in the frame, *deformation* refers to growth-induced changes in in blade shape, size, and pose, and *rotation* denotes tray or pot movement independent of leaf motion. These plant-specific events induce sparse, non-rigid temporal dynamics that differ from common pedestrian/vehicle MOT. Figure 4(a) shows leaf area trajectories of 13 leaves across 31 days, revealing diverse growth dynamics. In Figure 4(b), a t-SNE [27] visualization of all detected leaves over times shows clusters that are compact and well-separated by identity despite morphological, scale, and orientation changes, indicating a temporally stable and discriminative feature space. Additional comparisons across methods appear in Appendix Figure 6.

4.3 Ablation Study

Backbone. We used the MobileNetV3 as our backbone and compare it against other backbone models, i.e., ResNet variants (18/34/50/101) and ViT-B/16 in Table 3. Under identical training settings and detectors, MobileNetV3 achieves the strongest identify metrics (HOTA/AssA/IDF1) while using only $\sim$3M parameters. DetA is effectively flat across backbones, as expected with a shared detector. Higher model capacity does not directly translate to better identity maintenance for structured, non-rigid motion leaf trackers. Deeper ResNets and ViT-B/16 increase computation by 4–30$\times$ without improving tracking. This suggests that, at our data scale and with a triplet objective in biological sceen, a compact CNN is sufficient to learn discriminative, temporally stable embeddings. We therefore adopt MobileNetV3 for the main results.

Triplet sampling. We study how negatives are chosen while anchors and positives are the same leaf on different days. We compare three strategies: *(i) cross-plant negatives*, sampled from any plant; *(ii) intra-plant full-cycle negatives*, a different leaf from the same plant without temporal constraints; *(iii) intra-plant windowed negatives*, restricted to a ΔT-day neighborhood around the anchor. As shown in Table 4, unconstrained negatives—cross-plant or full-cycle—deliver the strongest and statistically similar performance, whereas windowed sampling underperforms, with the largest drop at small ΔT and again at large ΔT. Small windows yield easy negatives that provide little discriminative pressure

Table 3: Backbone ablation. Metrics are reported as mean ± standard deviation over three runs. Best values are **bold**; second best are underlined.

	Parms(M)	MACs(G)	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
MobileNetV3	2.97	0.23	**88.03±0.24**	92.25±0.03	**84.07±0.49**	93.64±0.18	**92.90±0.35**
ResNet18	11.18	1.82	87.67±0.62	92.22±0.01	83.41±1.18	**93.65±0.13**	92.55±0.72
ResNet34	21.28	3.68	87.31±1.29	92.28±0.06	82.69±2.40	93.53±0.80	91.98±1.21
ResNet50	23.51	4.13	87.70±0.04	92.27±0.03	83.44±0.10	93.60±0.16	92.45±0.13
ResNet101	42.50	7.86	87.10±0.47	**92.29±0.07**	82.28±0.82	93.49±0.43	91.79±0.55
ViT_B16	86.57	17.61	86.79±0.75	92.24±0.03	81.75±1.43	92.97±0.42	91.61±0.88

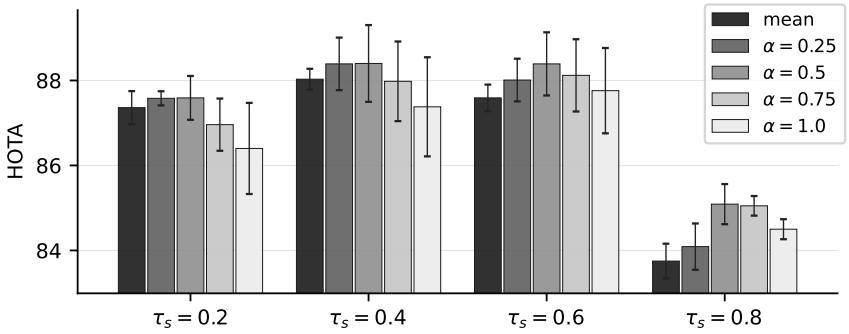

Figure 5: Inference ablation of the similarity threshold τ_s and temporal smoothing coefficient α. Error bars indicate ± one standard deviation over three trainings.

within a rosette, while very large windows bias training toward trivially separable pairs. We adopt intra-plant full-cycle sampling as the default.

Table 4: Ablation on triplet sampling strategies and temporal window size (ΔT).

Strategy	ΔT	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
(i) cross-plant flexible	–	**88.30±0.24**	**92.25±0.02**	**84.59±0.46**	**94.13±0.23**	**93.15±0.14**
(ii) intra-plant full-cycle	–	88.03±0.24	92.25±0.03	84.07±0.49	93.64±0.18	92.90±0.35
(iii) intra-plant temporal windows	1	59.29±0.67	92.23±0.06	38.18±0.88	83.59±0.56	59.79±1.00
	2	64.83±0.58	92.14±0.05	45.69±0.79	86.29±0.12	65.09±0.37
	5	72.91±8.25	92.22±0.03	58.19±13.42	88.09±2.86	74.76±9.92
	10	64.96±1.48	92.16±0.08	45.84±2.09	84.02±0.80	65.43±2.13
	20	59.10±4.13	92.14±0.07	38.07±5.22	80.17±2.91	58.02±5.75

Inference hyperparameters. We ablate the similarity threshold τ_s and the EMA coefficient α in Figure 5. HOTA forms a plateau at $\tau_s \in [0.4, 0.6]$ and drops at $\tau_s = 0.8$ (over-pruning) and $\tau_s = 0.2$ (noisy associations). Within each τ_s, performance improves with moderate smoothing and declines for $\alpha = 0.25$ (history-dominated, slow to adapt) and $\alpha = 1.0$ (one-frame memory that overwrites history). The history *mean* baseline (uniform average of past embeddings) is consistently below EMA with $\alpha = 0.5$ for all τ_s, indicating that equal weighting underemphasizes recent morphology. Error bars are small (std ≤ 1 HOTA) across seeds.

5 Conclusion and Discussion

We propose **LeafTrackNet**, an effective framework for long-term leaf tracking from top-down RGB sequences of canola. By combining a high-accuracy leaf detector with an embedding-based association strategy, LeafTrackNet handles key biological and environmental-induced challenges such as leaf emergence, occlusion, deformation, and rotation. We also released **CanolaTrack**, a

high-resolution dataset comprising 184 plants tracked over 31 days. LeafTrackNet achieves state-of-the-art performance on this benchmark, outperforming both general MOT methods and plant-specific baselines across standard metrics, and enabling accurate, scalable, temporally consistent leaf identity tracking. While this study focuses on a single crop species (canola) in a controlled environment with bounding-box annotations, future work will extend toward cross-species generalization, field environments with greater variability (e.g., wind, clutter, and changing light), and richer annotation schemes such as instance masks.

Acknowledgment

This work was funded by the Federal Ministry of Research, Technology and Space through project DCropS4OneHealth (ref. 16LW0528K) and REFRAME (ref. 01IS24073B).

References

[1] Joaquim Miguel Costa, Jorge Marques da Silva, Carla Pinheiro, Matilde Barón, Photini Mylona, Mauro Centritto, Matthew Haworth, Francesco Loreto, Baris Uzilday, Ismail Turkan, and Maria Margarida Oliveira. Opportunities and limitations of crop phenotyping in southern european countries. *Frontiers in Plant Science*, Volume 10 - 2019, 2019.

[2] Tom Rankenberg, Batist Geldhof, Hans van Veen, Kristof Holsteens, Bram Van de Poel, and Rashmi Sasidharan. Age-dependent abiotic stress resilience in plants. *Trends in Plant Science*, 26(7):692–705, 2021.

[3] Jinhai Cai, Mamoru Okamoto, Judith Atieno, Tim Sutton, Yongle(Leo) Li, and Stanley Miklavcic. Quantifying the onset and progression of plant senescence by color image analysis for high throughput applications. *PLOS ONE*, 11:e0157102, 06 2016.

[4] Weiming Yan, Yangquanwei Zhong, and Zhouping Shangguan. A meta-analysis of leaf gas exchange and water status responses to drought. *Scientific Reports*, 6:20917, 2016.

[5] Minsoo Jeong, Sihyun Park, Sook-Min Kwon, KyeongMo Lim, Da-Ryung Jung, Hong-Seok Lee, Hei Kim, and Jae-Ho Shin. Rapid detection of soybean nutrient deficiencies using yolov8s: Advancing precision agriculture. *Scientific Reports*, page 13810, 04 2025.

[6] Yingying Zhang, Xue Li, Meiqing Wang, Tao Xu, Kai Huang, Yuanhao Sun, Quanchun Yuan, Xiaohui Lei, Yannan Qi, and Xiaolan Lv. Early detection and lesion visualization of pear leaf anthracnose based on multi-source feature fusion of hyperspectral imaging. *Frontiers in Plant Science*, Volume 15 - 2024, 2024.

[7] Alper Adak, Seth C. Murray, and Jacob D. Washburn. Deciphering temporal growth patterns in maize: integrative modeling of phenotype dynamics and underlying genomic variations. *New Phytologist*, 242(1):121–136, 2024.

[8] David Hobby, Hao Tong, Marc Heuermann, Alain Mbebi, Roosa Laitinen, Matteo Dell'Acqua, Thomas Altmann, and Zoran Nikoloski. Predicting plant trait dynamics from genetic markers. *Nature Plants*, 11:1018–1027, 04 2025.

[9] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[10] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision*, pages 1–17. Springer, 2022.

[11] Yanina S. Correndo, Ana J.P. Carcedo, Mario A. Secchi, Michael J. Stamm, P.V. Vara Prasad, Sara Lira, Carlos D. Messina, and Ignacio A. Ciampitti. Identifying environments for canola oil production under diverse seasonal crop water stress levels. *Agricultural Water Management*, 302:108996, 2024.

[12] Feryel Lassoued, Peter Slade, and Ashly Dyck. Crop rotations and canola yields: Evidence from field-level data in western canada. *Agronomy Journal*, 117(1):e21739, 2025.

[13] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.

[14] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019.

[15] David. P. Hughes and Marcel Salathe. An open access repository of images on plant health to enable the development of mobile disease diagnostics, 2016.

[16] Tianqi Wei, Zhi Chen, Xin Yu, Scott Chapman, Paul Melloy, and Zi Huang. Plantseg: A large-scale in-the-wild dataset for plant disease segmentation, 2024.

[17] Nima Teimouri, Mads Dyrmann, Per Rydahl Nielsen, Solvejg Kopp Mathiassen, Gayle J. Somerville, and Rasmus Nyholm Jørgensen. Weed growth stage estimator using deep convolutional neural networks. *Sensors*, 18(5), 2018.

[18] Federico Jurado-Ruiz, Thu-Phuong Nguyen, Gerrit Polder, and Mark GM Aarts. Letra: a leaf tracking workflow based on convolutional neural networks and intersection over union. *Plant Methods*, 20(1):1–12, 2024.

[19] Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rin-ichiro Taniguchi. An easy-to-setup 3d phenotyping platform for komatsuna dataset. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[20] Jeffrey A. Cruz, Xi Yin, Xiaoming Liu, Saif M. Imran, Daniel D. Morris, David M. Kramer, and Jin Chen. Multi-modality imagery database for plant phenotyping. *Machine Vision and Applications*, 27(5):735–749, July 2016.

[21] Lionel Daviet et al. Phenotrack3d: tracking the development of maize organs using 3d reconstructions. *Plant Phenomics*, 2022:1–12, 2022.

[22] Marc Josep Montagut-Marquès et al. Plant doctor: A hybrid machine learning and image segmentation software to quantify plant damage in video footage. *Measurement*, 215:112345, 2025.

[23] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023.

[24] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20993–21002, June 2022.

[25] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[26] Daniel Ponsa Vassileios Balntas, Edgar Riba and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11. BMVA Press, September 2016.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[28] Xiangyu Zhu, Zhenbo Luo, Pei Fu, and Xiang Ji. Voc-reld: Vehicle re-identification based on vehicle-orientation-camera. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2566–2573, 2020.

[29] Jonathon Luiten, Aljossa Ossep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vision*, 129(2):548–578, February 2021.

A Implementation Details

Training Details. The embedding network parameters (θ, ϕ) are optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. Following the setup in [28], the triplet loss margin m is set to 0.3. All leaf crops are resized to $W = H = 224$, as adopted in [14]. Training is conducted for up to 80 epochs with a batch size of 48 on four NVIDIA Tesla V100S GPUs, using an early stopping training strategy.

Inference Details. We use a similarity threshold $\tau_s = 0.4$, age threshold $\tau_a = 5$, and smoothing coefficient $\alpha = 0.5$ during inference.

B Evaluation Metrics

Following the common practice in multi-object tracking evaluation [29], we report the standard MOT metrics, including Higher Order Tracking Accuracy (HOTA), Detection Accuracy (DetA), Association Accuracy (AssA), Multi-Object Tracking Accuracy (MOTA), and Identification F1 Score (IDF1).

HOTA measures the joint performance of detection DetA and association AssA, providing a balanced evaluation of tracking quality:

$$\mathrm{HOTA} = \sqrt{\mathrm{DetA} \times \mathrm{AssA}}$$

DetA quantifies how well the tracker detects objects across images. Let TP, FP, and FN represent true positives, false positives, and false negatives respectively, DetA is computed as:

$$\mathrm{DetA} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

AssA measures the correctness of identity preservation over time. Set T is the set of time steps, and $\mathrm{TP}^t_{\mathrm{assoc}}$ is the number of correctly associated detections at time t. AssA is defined as the average fraction of correctly associated objects given that a detection is matched:

$$\mathrm{AssA} = \frac{1}{|\mathsf{T}|} \sum_{t \in \mathsf{T}} \frac{\mathrm{TP}^t_{\mathrm{assoc}}}{\mathrm{TP}^t_{\mathrm{assoc}} + \mathrm{FP}^t_{\mathrm{assoc}} + \mathrm{FN}^t_{\mathrm{assoc}}}.$$

MOTA considers missed detections, false positives, and identity switches. Let GT_t is the number of ground-truth objects at time t, and IDSW_t is the number of identity switches. MOTA is defined as:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t}.$$

IDF1 computes the F1 score of correctly identified detections, where IDTP, IDFP, and IDFN denote identity-level true positives, false positives, and false negatives:

$$\mathrm{IDF1} = \frac{2 \cdot \mathrm{IDTP}}{2 \cdot \mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}}.$$

C Accuracy Comparison

To enable a direct visual comparison of the long-term tracking performance across methods, we computed binary accuracy heatmaps that highlight the tacking performance for a single plant (Figure 6 a) and for each individual leaf of the same plant (Figure 6b).

Figure 6(a) demonstrates the impact of the embedding stability of our method on long-term tracking performance, showing the high average accuracy across 31 days. The accuracy is defined as the

proportion of leaves correctly detected and consistently tracked on a given day. While state-of-the-art methods exhibit significant performance degradation—especially after Day 9—our method maintains consistently high and stable accuracies throughout the plant's growth cycle. Figure 6(b) provides a fine-grained view by visualizing the tracking accuracy of the individual leaves, where each cell reflects the success of identifying a specific leaf on a specific day. Yellow indicates a correct leaf association (IoU ≥ 0.75 and correct ID), purple denotes failure, and blank cells indicate the absence of leaves due to complete occlusion, senescence, or not yet sprouted leaves. These visualizations clearly demonstrate that LeafTrackNet preserves long-term tracking more reliably, both in terms of average daily accuracy and individual leaf trajectories.

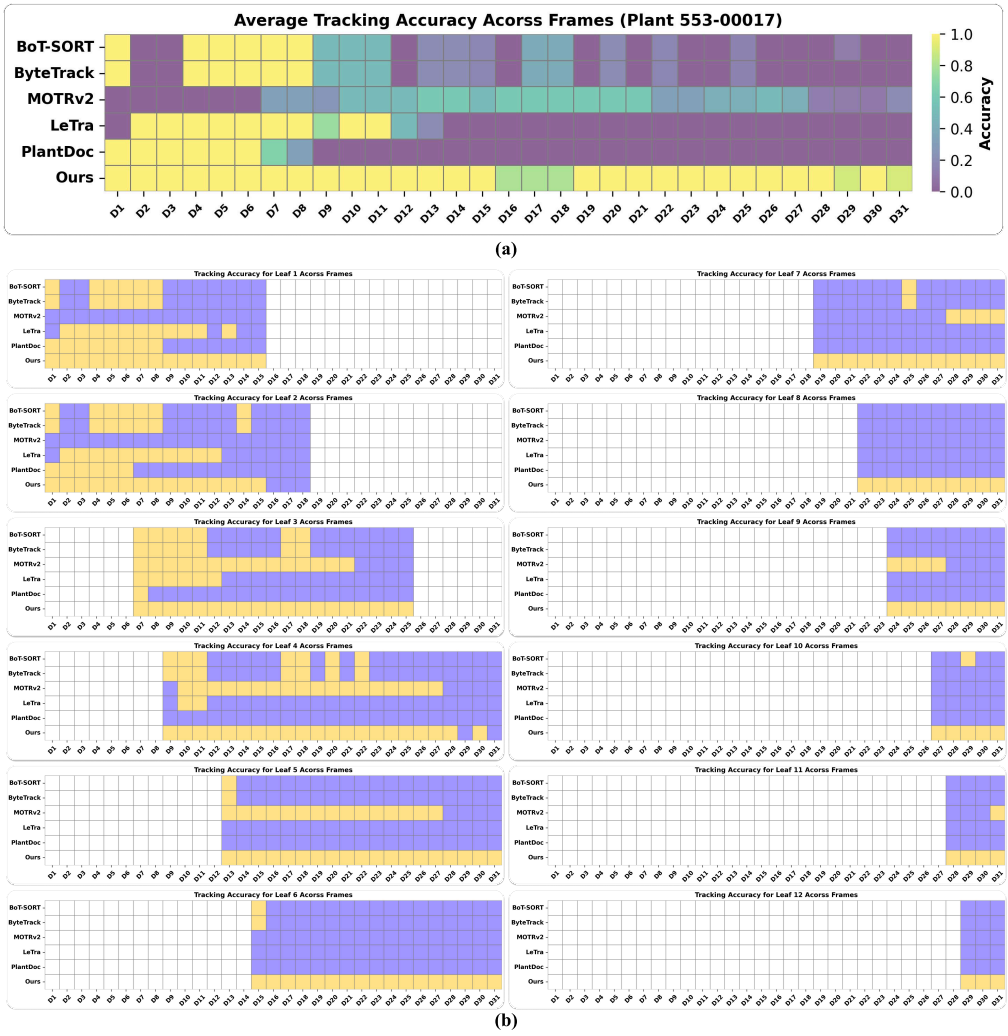

Figure 6: Tracking accuracy visualization for Plant-158. (a) Average frame-level accuracy per method per day. (b) Per-leaf binary tracking matrix: yellow = correct, purple = failure, blank = leaf absent.