Plant Phenotype

Zaeem Yousaf *
23030021@lums.edu.pk

Supervisor: Dr. Murtaza Taj
CS LUMS University, Pakistan

***Abstract*—Plant phenotyping has greatly benefited from recent advancements in deep learning, enabling precise trait analysis for precision agriculture. Open-World Semantic Segmentation [1] allows the identification of unseen plant species, addressing challenges in diverse agricultural environments. Vision Transformers (ViTs) [2] and Graph Neural Networks (GNNs) [3] enhance hierarchical plant structure analysis, improving phenotype trait prediction. Object detection models like Faster R-CNN, YOLOv8, and DETR have been widely used for plant organ detection, such as wheat spikes. DeepLabV3+ and U-Net provide state-of-the-art semantic segmentation for weed-crop differentiation. Temporal analysis techniques, including ConvLSTMs and 3D CNNs, enable the modeling of plant growth patterns over time. Self-Supervised Learning improves feature extraction with minimal labeled data, while Multi-Modal Learning integrates RGB and multi-spectral imagery for enhanced phenotypic analysis. These advancements contribute to improved growth monitoring, age estimation, and plant health assessment, making plant phenotyping more automated and scalable.**

I. State of The Art Literature Review

The GroMo [4] introduces the GroMo Challenge, focusing on two primary tasks: (1) plant age prediction and (2) leaf count estimation, both crucial for crop monitoring and precision agriculture. To support this challenge, the authors present GroMo25, a dataset [5] comprising multiview images of four crops—radish, okra, wheat, and mustard—captured from 24 different angles over multiple days. They propose a Multiview Vision Transformer (MVVT) model to address these tasks, achieving a mean absolute error (MAE) of 7.74 for age prediction and 5.52 for leaf count estimation. The study aims to advance plant phenotyping research by encouraging innovative solutions for tracking and predicting plant growth.

The paper [6] introduces The CropAndWeed Dataset [7], a large-scale dataset with multi-modal annotations (bounding boxes, semantic masks, stem positions) covering 74 crop and weed species for automated agricultural robotics. It evaluates state-of-the-art deep learning models for detection (YOLOv5, Faster R-CNN), segmentation (DeepLabV3+, Mask R-CNN), and classification (ResNet, ViTs) to benchmark weed-crop differentiation. The dataset integrates RGB and multi-spectral imaging, enabling domain adaptation and self-supervised learning techniques for generalized plant identification. Experiments show that multi-modal learning improves segmentation accuracy, and incorporating rare weed species enhances model robustness for real-world weed control applications.

In this paper [8], the authors address the issues of the rarity and cost of highly accurate, multi-temporal 3D point cloud datasets of plants needed for advanced plant analysis and machine learning. To solve this, they present Pheno4D, a new dataset featuring high-resolution registered point clouds of maize and tomato plants captured daily over several weeks, with manual labels for computer vision tasks. They acquired the data using a high-accuracy 3D laser scanning system and provided temporally consistent labels for plant organs. To demonstrate the dataset's usability, they showed baseline results in tasks like point cloud segmentation, non-rigid registration, and surface reconstruction, and also derived time series of phenotypic traits. The dataset [9] contains approximately 260 million labeled 3D points across 126 labeled point clouds and is freely accessible which would absolve researchers from generating their own dataset.

The paper [2] addresses the species gap in image-based plant phenotyping, where models trained on one species fail to generalize due to morphological and environmental variations. It proposes a scalable learning framework using domain adaptation, self-supervised learning, and meta-learning to improve cross-species generalization. The method leverages contrastive feature alignment for domain adaptation, self-supervised pretraining on unlabeled plant data, and prototypical networks with model-agnostic meta-learning (MAML) for few-shot adaptation. It integrates Vision Transformers (ViTs) and CNNs for hierarchical feature extraction. The approach significantly reduces labeled data requirements and enhances transferability across plant species, improving phenotyping robustness. Using these techniques on the Dataset [10] it outperforms the accuracy.

II. First Improvement

A. Robust Preprocessing with HSV Segmentation and Canny Edge Detection

In our controlled container-based experiments, plant images are captured against a soil background that exhibits significant chromatic variation over time as moisture levels change and the substrate dries. Even minor fluctuations in soil color or ambient lighting can introduce substantial noise in raw RGB inputs, degrading our leaf-counting Vision Transformer (ViT) model's performance. To mitigate this, we propose a two-stage preprocessing strategy:

1) **Background Removal via HSV Thresholding.** By converting each frame to the HSV color space and

applying a fixed hue–saturation–value range tuned for canopy green, non-leaf pixels (soil, container walls) are effectively masked out. This removes background artifacts that would otherwise confound the network's attention mechanisms.

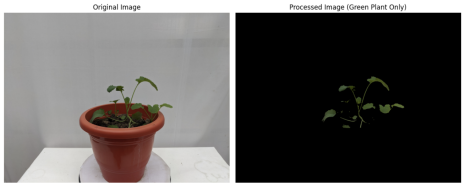

Fig. 1. HSV Filter

2) **Motivation behind Background Removal** I conducted an attention-map analysis on the original ViT outputs (see Fig.2). Although the model predominantly focuses on leaf regions, the attention overlays reveal persistent activations over container edges and soil patches, signaling that background artifacts continue to influence the network's attention. This residual noise underscores the necessity of our two-stage approach: HSV-based background removal removes the confounding substrate, and Canny edge extraction further suppresses non-leaf activations, guiding the model to learn exclusively from the botanical structures of interest.

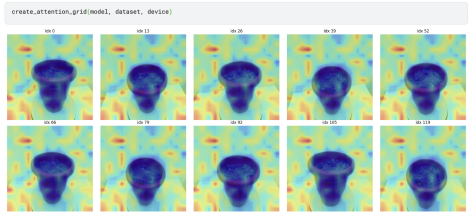

Fig. 2. Attention map

3) **Canny Edge Detection on Isolated Plant Regions.** Following background removal, I apply the Canny operator to extract clear leaf contours and vein patterns. Edge-based features emphasize morphology rather than color intensity, making our model robust to variations in leaf pigmentation, shape, or residual shading.

By focusing the ViT exclusively on high-contrast, geometry-driven features, we achieve several benefits:

- *Noise Resilience:* Eliminating soil color drift and uneven lighting prevents spurious gradient updates and reduces overfitting to background variations.
- *Shape-Centric Learning:* Edge maps encode leaf boundaries and vein structures, enabling the model to generalize across different species and developmental stages with diverse pigmentation.
- *Computational Efficiency:* Masking and edge extraction reduce irrelevant pixel information, accelerating both training convergence and inference throughput.

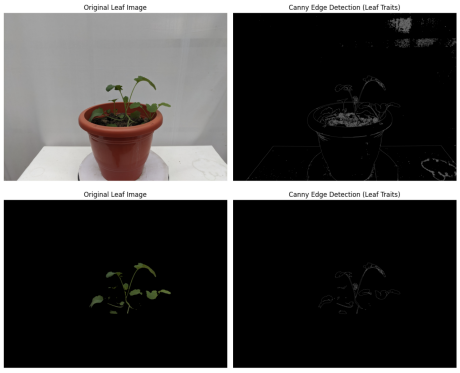

Fig. 3. outline of HSV using Canny Edge

This preprocessing enhancement lays a solid foundation for more accurate and robust leaf-count predictions, irrespective of container conditions or plant coloration, thereby strengthening the overall reliability of our agricultural phenotyping pipeline.

III. Second Improvement

A. Using Weighted Contrastive Learning Approach

In recent developments, Multiview Vision Transformers (ViTs) have shown significant promise in plant age and leaf count prediction by leveraging multiple images captured from different angles [11]. However, these methods require a complete set of images from various views of the plant, which may not be feasible in real-world scenarios, especially for mobile applications. Capturing a 360-degree view of a plant can be cumbersome and impractical in many cases, especially when the goal is to make accurate predictions using limited input. ViTs have been successfully applied in multiview learning tasks, but require more computationally expensive and data-intensive inputs [12], [13].

On the other hand, contrastive learning offers a powerful alternative by allowing the model to learn representations from a single view image, making it more suitable for real-time applications where collecting multiple images of a plant is not feasible. Contrastive learning has proven to be effective in visual representation learning by maximizing the similarity of similar samples while minimizing the similarity of dissimilar ones [14]. However, the challenge with standard contrastive learning is that it treats all positive and negative pairs equally, even though some pairs—such as images with similar leaf counts—may carry more semantic similarity than others. To address this, we propose incorporating a weighted contrastive loss, where pairs of images with more similar characteristics (e.g., similar leaf counts or age) are assigned higher weights [15]. This approach enables the model to focus more on the finer distinctions between images with differing leaf counts, thus improving prediction accuracy. By combining weighted contrastive learning with the power of multiview 360-degree data for training, we aim to develop a model that can predict plant age and leaf count from a single image, making it more practical for mobile-based applications. This approach

takes advantage of a large pool of 360-degree data during training, while only requiring a single image for inference, thus making it an ideal solution for resource-constrained mobile applications.

I also aim to benchmark three distinct contrastive learning models, comparing their performance with Vision Transformers (ViTs) in the context of plant age and leaf count prediction. While previous works have primarily explored standard contrastive learning frameworks, the focus here will be on evaluating the impact of weighted contrastive loss functions. By selecting three models with varying strategies for weight assignment in contrastive learning, we will be able to assess their effectiveness and practicality in capturing subtle visual differences between plant images. This benchmarking exercise will provide valuable insights into which model is best suited for real-world applications, especially when considering the challenges posed by limited-view imaging and the need for robust, efficient predictions on mobile platforms.

I also aim to benchmark three distinct contrastive learning models—namely, Deep InfoMax (DIM), SimCLR, and MoCo—comparing their performance with Vision Transformers (ViTs) in the context of plant age and leaf count prediction. While previous works have primarily explored standard contrastive learning frameworks, the focus here will be on evaluating the impact of weighted contrastive loss functions. By selecting three models with varying strategies for weight assignment in contrastive learning, we will be able to assess their effectiveness and practicality in capturing subtle visual differences between plant images. Specifically, DIM focuses on maximizing mutual information, SimCLR leverages large-batch contrastive learning, and MoCo uses a memory bank to efficiently store negative samples, each offering distinct advantages for the given task. This benchmarking exercise will provide valuable insights into which model is best suited for real-world applications, especially when considering the challenges posed by limited-view imaging and the need for robust, efficient predictions on mobile platforms.

IV. Mathematical Formulation of Contrastive Loss

The contrastive loss function used in this work can be expressed as follows:

$$\mathcal{L}_{contrastive} = \frac{1}{2N} \sum_{i=1}^{N} \left[y_i \cdot \mathrm{d}(z_i, z_i^+)^2 + (1 - y_i) \right. \quad (1)$$

$$\left. \cdot \max\left(0, m - \mathrm{d}(z_i, z_i^-)^2\right)^2 \right. \quad (2)$$

Where:

- $\mathcal{L}_{contrastive}$ is the contrastive loss.
- N is the total number of image pairs.
- z_i and z_i^+ are the feature representations (embeddings) of positive pairs (similar images).
- z_i^- is the feature representation of the negative pair (dissimilar image).
- $y_i \in \{0, 1\}$ is the binary label: $y_i = 1$ for positive pairs, and $y_i = 0$ for negative pairs.
- $\mathrm{d}(z_i, z_j)$ is the Euclidean distance (or any other distance metric) between the feature embeddings of images i and j.
- m is the margin that separates positive and negative pairs.

The first term in the loss function minimizes the distance between similar pairs, while the second term ensures that dissimilar pairs are sufficiently far apart by enforcing a margin m. This formulation is typically used for contrastive learning, but we plan to extend this approach with weighted loss functions to focus on more difficult-to-differentiate pairs based on the plant's leaf count.

V. Results from 1st Improvements

The performance comparison between the 3-layer and 6-layer models clearly highlights the importance of model complexity in relation to input data characteristics. Canny edge data, by its very nature, provides a simplified representation of the image, focusing primarily on boundary information while discarding most of the texture, color, and other nuanced features present in original RGB images. This reduced feature space limits the amount of meaningful information that can be extracted by deeper networks. As a result, using a 6-layer model, which introduces higher capacity and more parameters, can lead to overfitting — where the model begins to memorize the training data rather than generalize from it. This is evident in the drastic drop in validation performance for the 6-layer model, particularly for age prediction, where the validation R^2 drops to a negative value, indicating poor generalization.

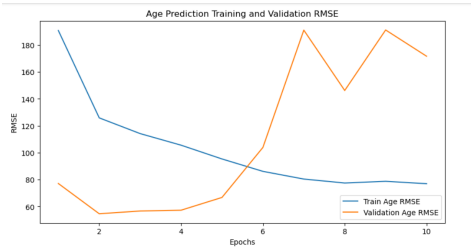

Fig. 4. Overfitting when using Baseline model with Canny Edge

In contrast, the 3-layer model demonstrates better generalization with more balanced performance across both training and validation datasets. This suggests that a shallower architecture is more suitable for scenarios involving sparse or minimalistic input features like those derived from edge detection. By limiting the number of layers, the model maintains a controlled capacity that aligns well with the simplicity of the input, thus reducing the risk of overfitting while still achieving competitive results. Therefore, when working with simplified representations such as Canny edge images, it is advisable to use smaller, less complex architectures to achieve optimal balance between performance and generalization.

TABLE I
COMPARISON OF 3-LAYER VS. 6-LAYER MODELS (CANNY EDGE DATA)

Metric	3-Layer Model	6-Layer Model	Observation
Train MAE (Leaf)	3.5169	3.2941	Slightly better in 6-layer
Train MAE (Age)	3.1567	4.2150	Worse in 6-layer
Train R^2 (Leaf)	0.6548	0.6884	Slightly better in 6-layer
Train R^2 (Age)	0.8618	0.7601	Better in 3-layer
Validation MAE (Leaf)	3.3747	4.1471	Better in 3-layer
Validation MAE (Age)	3.5723	15.2131	Much better in 3-layer
Validation R^2 (Leaf)	0.6078	0.4843	Better in 3-layer
Validation R^2 (Age)	0.8471	-3.4531	Severe overfitting in 6-layer
Train Loss Trend (Leaf)	Decreasing	Decreasing	Stable for both
Val Loss Trend (Leaf)	Slight fluctuations	Ends higher	Slight overfitting in 6-layer
Train Loss Trend (Age)	Consistent decrease	Flattens at end	Stagnation in 6-layer
Val Loss Trend (Age)	Mostly decreasing	Severe spikes after epoch 5	Overfitting in 6-layer

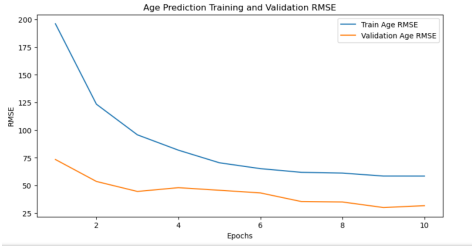

Fig. 5. Overfitting resolved with 3 layers

TABLE II
PERFORMANCE COMPARISON OF CONTRASTIVE LEARNING MODELS ON LEAF COUNT PREDICTION

Metric	Deep InfoMax (DIM)	SimCLR	MoCo
MAE (↓)	8.43	10.86	**7.50**
RMSE (↓)	10.80	13.00	**11.20**
R^2 (↑)	-0.657	-0.475	**-0.600**

VI. SECOND IMPROVEMENT RESULTS

Commentary Paragraphs: Why First and Second Improvements Differ The first improvement focuses on enhancing input quality through preprocessing techniques such as HSV filtering and Canny edge detection. These methods isolate leaf structures by removing noisy background and emphasizing shape-based features. This is especially valuable in tightly controlled environments where visual distractions such as soil, shadows, or container edges can misguide the model's attention. The approach is effective in scenarios with consistent backgrounds and relies on architectural simplification (like 3-layer models) to mitigate overfitting due to the minimalistic feature set.

On the other hand, the second improvement tackles the learning methodology itself, proposing a weighted contrastive learning framework that improves the model's ability to differentiate between subtle visual differences even with limited input (e.g., a single image). Unlike the first improvement, which manipulates the raw input space, this method enhances internal representation learning through loss function engineering and multi-view pretraining. It's particularly advantageous for real-world, mobile scenarios where image capture may be constrained, and extensive preprocessing might not be feasible. The model becomes semantically aware of leaf and age similarity through intelligent weighting rather than being dependent on clean visual isolation.

REFERENCES

[1] M. Sodano, F. Magistri, L. Nunes, J. Behley, and C. Stachniss, "Open-world semantic segmentation including class similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3184–3194.

[2] D. Ward and P. Moghadam, "Scalable learning for bridging the species gap in image-based plant phenotyping," *Computer Vision and Image Understanding*, vol. 197, p. 103009, 2020.

[3] H.-C. Yang, J.-P. Zhou, C. Zheng, Z. Wu, Y. Li, and L.-G. Li, "Phenologynet: A fine-grained approach for crop-phenology classification fusing convolutional neural network and phenotypic similarity," *Computers and Electronics in Agriculture*, vol. 229, p. 109728, 2025.

[4] R. Bhatt, S. Bansal, A. Chander, R. Kaur, M. Singh, M. Kankanhalli, A. E. Saddik, and M. K. Saini, "Gromo: Plant growth modeling with multiview images," *arXiv preprint arXiv:2503.06608*, 2025.

[5] M. Lab, "Gromo: Plant growth modeling with multiview images," https://github.com/mriglab/GroMo-Plant-Growth-Modeling-with-Multiview-Images, 2024, accessed: March 25, 2025.

[6] D. Steininger, A. Trondl, G. Croonen, J. Simon, and V. Widhalm, "The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3729–3738.

[7] Crop and W. D. Contributors, "Crop and weed dataset," https://github.com/cropandweed/cropandweed-dataset/tree/main, 2024, accessed: March 25, 2025.

[8] D. Schunck, F. Magistri, R. A. Rosu, A. Cornelißen, N. Chebrolu, S. Paulus, J. Léon, S. Behnke, C. Stachniss, H. Kuhlmann *et al.*, "Pheno4d: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis," *Plos one*, vol. 16, no. 8, p. e0256340, 2021.

[9] I. of Photogrammetry and U. o. B. Robotics, "Pheno4d dataset," https://www.ipb.uni-bonn.de/data/pheno4d/, 2024, accessed: March 25, 2025.

[10] X. Shuai, "Gwfss competition dataset," https://huggingface.co/datasets/XIANG-Shuai/GWFSS-competition, 2024, accessed: March 25, 2025.

[11] A. Dosovitskiy, M. Berman, and J. Mairal, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1734–1746, 2015.

[12] M. Caron, P. Bojanowski, M. Douze, H. Jégou, and J. Ponce, "Deep clustering for unsupervised learning of visual features," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 132–149.

[13] H. Touvron, M. Cord, and M. Douze, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning (ICML)*, 2021, pp. 2001–2012.

[14] X. Chen, L. Xie, and K. He, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[15] Z. Wu, L. Xie, and R. Girshick, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.