

Multibranch ensemble head in vision foundation models for efficient cucumber phenological stage classification*

Kostiantyn Khabarлак^{1,*}

¹ Dnipro University of Technology, D. Yavornytskoho Av., 19, Dnipro, 49005, Ukraine

Abstract

Vision Foundation Models (VFM) pretrained with self-supervised learning, such as DINOv2, provide highly transferable features, yet adapting them to small, domain-shifted datasets is still challenging. We propose a regularizing multibranch ensemble head: a training and inference scheme that pairs low- and high-capacity branches, samples a single branch stochastically at each training step, and averages both branches in probability space at inference. The design adds negligible parameters and essentially no inference overhead over a single heavy head, while improving the accuracy-efficiency trade-off under data-scarce conditions. Accurate classification of plant growth stages is critical for precision agriculture, enabling timely decisions on fertilization, irrigation, and harvesting. Therefore, as a challenging practical case we use cucumber phenological stage classification, in which flowers and fruits are frequently occluded by foliage, and stage transitions manifest primarily as subtle changes in leaf morphology. The NALEF Food Safety dataset contains images of five cucumber growth stages captured under diverse real-world conditions. On it we evaluate a broad range of backbones, including DINOv2 and DINOv3 ViT and ConvNeXt variants, against strongly supervised counterparts (ResNet, EfficientNet, ViT). The proposed DINOv2 ViT-S with multibranch head achieves a Macro F1 of 0.978 and Cohen's Kappa of 0.971, matching the performance of models more than ten times larger in parameters and GFlops, while requiring only 12.2 GFlops and 22.3M parameters, making it deployable on smartphones and edge devices for phenological stage classification.

Keywords

plant growth stage classification, vision foundation models, self-supervised learning, DINOv2, DINOv3, regularization, transfer learning, cucumber phenology

1. Introduction

The timing and progression of plant developmental phases are highly dependent on environmental factors, particularly temperature, precipitation, and photoperiod [1]. Accurate classification of plant growth stages is critical for optimizing agricultural management, including decisions on fertilization, herbicide application, and harvesting. Automating this process reduces reliance on labor-intensive and subjective manual analysis, ultimately supporting higher and more consistent crop yields.

Monitoring of plant growth can be performed using ground-level cameras, unmanned aerial vehicles (UAVs), or remote-sensing satellite imagery.

Remote-sensing satellite imagery can be used to estimate growth stages of large farmlands. For instance, Sentinel-2 satellites provide public access to multi-spectral imagery, including red-edge bands specifically designed for vegetation analysis [2]. However, this approach is unsuitable for use in cloudy conditions or in greenhouses, and its low spatial resolution leads to poor performance at early growth stages when the canopy does not fully cover the field, or for small fields.

UAVs and stationary cameras offer much greater flexibility and precision. They can be used regardless of weather conditions and inside greenhouses. Additionally, smartphones enable rapid,

* ADT-ISMDDC 2026: Advanced Digital Technologies: Intelligent Systems, Modeling, and Data-Driven Control, September 22–24, 2026, Odesa, Ukraine

^{1*} Corresponding author.



Khabarлак.K.S@nmu.one (K. Khabarлак)



0000-0003-4263-0871 (K. Khabarлак)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

on-site plant growth evaluation by non-experts. Camera-based plant growth monitoring is often formulated as a classification task [3], [4].

In particular, automated classification of cucumber growth stages from visual data remains an open problem despite the crop’s considerable agronomic significance. While computer vision methods have been applied to cucumbers across several related tasks including disease detection [5], fruit localization for robotic harvesting [6], quality grading [7], and morphological phenotyping [8], the problem of growth stage classification has received comparatively little attention in the literature. This task poses unique challenges: flowers and fruits are frequently occluded by foliage, and cucumber fruit shares near-identical coloration with the surrounding canopy. As a result, transitions between intermediate stages manifest primarily as subtle changes in leaf morphology.

Vision-based plant monitoring is typically performed by fine-tuning a backbone trained with strong supervision on a large labeled dataset. However, recent works have shown that backbones trained with self-supervised objectives offer far more universal features that can be adapted to diverse downstream tasks using a simple linear probe [9], [10]. These Vision Foundation Models (VFM) have demonstrated applicability across object classification, detection, tracking, scene depth estimation, and many other tasks; yet their use for plant growth stage classification remains comparatively underexplored.

In this work we study how to adapt self-supervised VFMs to small, domain-shifted image datasets through the design of the classification head. We introduce a regularizing multibranch ensemble head: a general training and inference scheme in which a low-capacity linear branch and a higher-capacity MLP branch are attached to the backbone. A single branch is sampled stochastically at each training step, and the two branches are averaged in probability space at inference. This yields an in-model ensemble that regularizes head training under data scarcity while adding negligible parameters and essentially no inference cost. We use cucumber phenological stage classification as a challenging fine-grained case study: on the recently introduced NALEF Food Safety [11] dataset, the proposed head substantially improves the accuracy-efficiency trade-off over the DINOv2 ViT-L fine-tuning approach previously reported for this dataset [12].

Our main contributions are as follows:

1. We propose the regularizing multibranch ensemble head, a lightweight, backbone-agnostic scheme for training classification heads on top of Vision Models under data-scarce conditions, combining stochastic per-step branch sampling with probability-space aggregation at inference.
2. Using a cucumber phenological stage classification dataset, we systematically compare a broad range of self-supervised (DINOv2, DINOv3) and strongly supervised backbones, and show that self-supervised pretraining is particularly beneficial in the small-model regime despite the large domain gap between its pre-training data and plant imagery.
3. The resulting DINOv2 ViT-S model substantially improves over existing solutions on the NALEF Food Safety dataset at a small fraction of its computational cost, requiring few GFlops and parameters and remaining deployable on smartphones and edge devices.

2. Literature Overview

Fine-tuning from a strongly pretrained backbone is the standard approach for training deep neural networks on downstream tasks such as plant growth stage classification. The authors of [3] collect a strawberry growth dataset and report that a fine-tuned EfficientNet-B7 achieves 83.7% classification accuracy. In [13], the authors highlight the importance of camera-based plant growth

monitoring as (1) such systems are low-cost and (2) the collected data is crucial for nutrient management and harvesting scheduling; they develop a YOLOv8-based detection system for multi-plant basil growth tracking. In [4], a ResNet-50-based butterhead lettuce growth stage classification system is presented.

Cucumber-specific research is an active area. [14] investigate the relationship between cucumber root traits and crop yield to support the breeding of high-yielding, stress-tolerant cultivars. Cucumber production further benefits from automated harvesting systems. Accurate detection of harvest-ready cucumbers remains challenging due to the visual similarity between fruits and surrounding foliage, as well as frequent occlusions. To address this problem, [6] propose a lightweight YOLO-based detection network tailored for cucumber harvesting scenarios. Quality assessment has also been explored using deep learning, with [7] introducing a CNN-based framework for classifying cucumber fruits into defective, curved, and straight/sound categories. For phenotyping applications, [8] present a lightweight keypoint detection model that enables automated extraction of fruit phenotypic traits. [15] propose a pest classification method applicable to different plant species, including cucumber. Closest to our task, [12] fine-tune a single DINOv2 ViT-L backbone with progressive layer unfreezing for cucumber growth stage classification.

In contrast to strongly supervised approaches, large language models (LLMs) have demonstrated remarkable progress through self-supervised learning, typically by training on large text corpora with a next-token prediction objective. This approach enables the network to learn strong linguistic representations from unlabeled data.

Vision Transformers have successfully adapted the concept of tokens to images, often surpassing convolutional neural networks. Initially, however, they were still trained with strong supervision, as the design of effective self-supervised objectives for vision is more challenging, and the notion of a “next image token” is less self-evident.

To pretrain a vision backbone with self-supervision, several pretext tasks have been proposed: image re-colorization, prediction of image transformations, and others. DINOv2 uses masked image modeling combined with a feature-consistency loss between teacher and student networks observing different crops of the same image. A large Vision Transformer (1.1B parameters [9]) is trained and then distilled into several smaller ViTs. This approach was later extended by DINOv3 [10], which identified limitations in DINOv2’s architecture scaling under extended training and proposed a Gram loss, among other improvements. The resulting 7B model achieves state-of-the-art performance on many downstream tasks, and the authors distill it into both ViT and ConvNeXt variants. Models trained in this fashion are commonly referred to as Vision Foundation Models, as the rich features they produce transfer broadly across diverse downstream tasks.

The self-supervised training objective allows the collection of large unlabeled datasets for training. For instance, the authors of DINOv2 assembled the LVD-142M dataset comprising 142 million images without any labels, including ImageNet-22K and many other sources. Most of these are not plant-focused; while the authors include the small Flowers dataset, the larger nature-focused iNaturalist dataset is absent from the training set [9]. The DINOv3 authors introduce the LVD-1689M dataset (1,689 million Instagram photos). While substantially larger, this dataset likely entails a greater domain gap than DINOv2’s web-sourced collection. Despite the domain gap, recent works have shown that self-supervised methods can be successfully adapted to plant-related tasks [16], [17].

Despite these advances, most existing approaches to plant growth stage classification still rely on strongly supervised backbones, limiting their transferability to specialized agricultural domains with scarce labeled data. Furthermore, despite its agronomic importance, cucumber growth stage classification has received only limited attention in the literature, with VFM-based solutions for this specific task explored only in a single large-backbone configuration [12]. Finally, the task stands to

benefit from additional in-model regularization, as classification must rely primarily on subtle leaf morphology cues due to frequent occlusion of flowers and fruits; the potential of lightweight ensemble head designs to improve the accuracy-efficiency trade-off in this setting remains unexplored.

3. Materials and Methods

To address plant growth stage classification, we develop an approach based on Vision Foundation Models paired with a regularizing multibranch ensemble head. We compare the proposed method against the architectures listed in Table 1, spanning both convolutional neural networks (CNNs) and Transformer architectures.

Table 1
Backbones considered

Model	Type	Sizes
Strongly Supervised Backbones		
ResNet [18]	CNN	50, 101
ConvNeXt [19]	CNN	Tiny (T)
EfficientNet [20]	CNN	B3, B7
EfficientNetV2 [21]	CNN	Small (S), Large (L)
ViT [22]	Transformer	Base (B), Large (L)
Self-Supervised Backbones		
DINOv2 ViT [9]	Transformer	Small (S), Large (L)
DINOv3 ViT [10]	Transformer	Small (S), Large (L)
DINOv3 ConvNeXt [10]	CNN	Tiny (T), Small (S), Large (L)

The classification head is an important component of the proposed architecture. Given the rich features extracted by the backbone, the head must adapt them to the downstream task. We develop three head variants. Figure 1a shows a lightweight head, also known as a linear probe, often used with VFMs as recommended in [9], [10], and frequently sufficient given the descriptive power of VFM features. To address the domain shift between VFM pretraining and downstream agricultural data, a head with a larger number of parameters may be required. Figure 1b is a heavy head comprising two MLP layers with GELU activation followed by Dropout regularization (probability 0.3). Figure 1c shows the proposed multibranch ensemble head, described in detail below. To improve the performance of VFMs, an ensemble of models can be used; however, this increases inference time by a factor of N . Several works instead build more robust representations within a single network, effectively forming an in-model ensemble. The prototypical example is Dropout [23]: randomly masking activations during training samples a different weight-sharing subnetwork at each step, and test-time weight scaling approximates averaging the predictions of exponentially many such subnetworks. The authors of PTA [24] make this idea explicit, building an in-model ensemble from multiple parallel branches of different depth within the backbone, improving classification quality.

StochasticDepth [25] demonstrates that dynamically varying network depth during training acts as a regularizer; this technique has been adopted by EfficientNet V1 and V2, DINOv2, and DINOv3. [26] proposes collaborative networks where multiple branches are trained simultaneously, each supervised by a combination of ground-truth labels and soft targets from the other heads. [27] proposes a multi-head multi-loss classifier in which each head is assigned higher weight to a subset of classes, inducing specialization, with the final prediction obtained by averaging across heads.

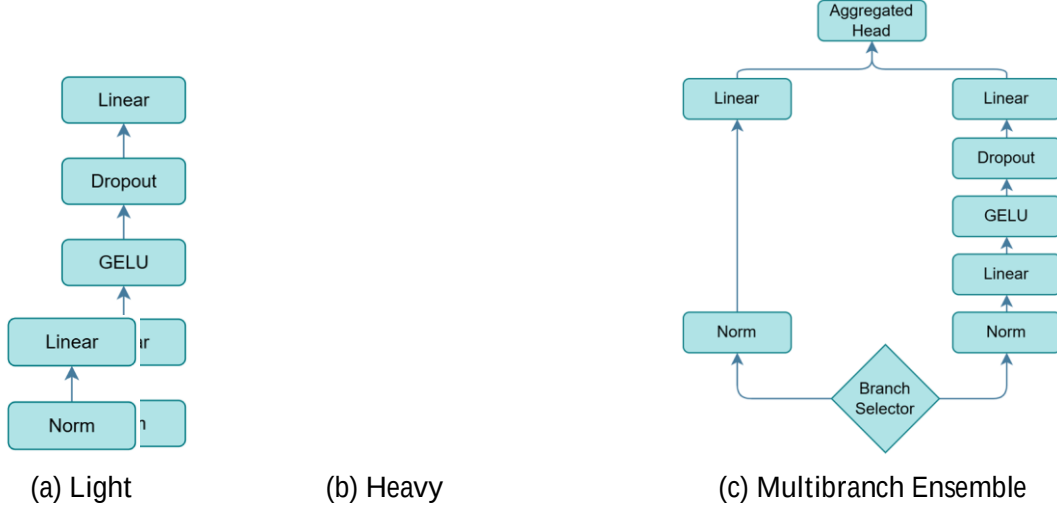


Figure 1: Neural network heads proposed for fine-tuning.

In this work we propose a multibranch ensemble head (Figure 1c). The two branches differ in depth and capacity: the left branch is a simple linear probe, while the right branch consists of several MLP layers with Dropout, equivalent to the heavy head described above (Figure 1b). During training, the active branch is randomly selected for each batch, so each branch learns to solve the task independently. The difference in capacity allows the light and heavy branches to capture simpler and more complex feature transformations, respectively. At inference, an in-model ensemble is formed by aggregating the outputs of both branches:

$$\hat{y} = \log\left(\frac{1}{2}(p^{(\text{light})} + p^{(\text{heavy})}) + \epsilon\right), \quad (1)$$

where the branch output $p^{(k)}, k \in \{\text{light}, \text{heavy}\}$ is defined as:

$$p^{(k)} = \text{softmax}(f^{(k)}(z)). \quad (2)$$

Here $f^{(k)}$ is the function computed by branch k , and $z = g(X)$ are the features produced by the backbone g for the input batch X . The vector $\hat{y} \in \mathbb{R}^C$ holds the aggregated log-probabilities, and ϵ is a small positive constant that ensures numerical stability of the logarithm. Aggregation is performed in probability space rather than logit space (Eqs. 1–2), since logits from branches of different capacity may have substantially different scales. Probability-space aggregation is consistent with standard mixture-of-experts averaging and corresponds to a uniform Bayesian model average treating each head as an equally plausible hypothesis about the class distribution. The branch selector shown in Figure 1c activates exactly one branch per training step. At each step a branch is drawn independently from a Bernoulli distribution, $k \sim \text{Bernoulli}(\pi)$, with $\pi = 0.5$ assigning equal probability to the two branches. Only the selected branch produces the prediction for that step, and only its parameters, together with those of the backbone, receive gradients from the cross-entropy loss; the other branch is left unchanged. No knowledge-distillation or cross-branch consistency term is applied, so each branch learns to solve the task independently. Algorithm 1 summarizes the

procedure. Models are trained using the AdamW optimizer with a batch size of 32. The head is randomly initialized while the backbone uses pretrained weights. To mitigate catastrophic forgetting, training proceeds in two stages: 15 epochs of head-only fine-tuning at a learning rate of 3×10^{-3} , followed by 15 epochs of full-network fine-tuning at a reduced learning rate of 5×10^{-6} .

Algorithm 1

Training step with the multibranch ensemble head

	Input: mini-batch (X, Y) ; backbone g ; branches $f^{(\text{light})}, f^{(\text{heavy})}$; branch probability π
1	$z \leftarrow g(X)$ (<i>backbone features</i>)
2	draw $k \sim \text{Bernoulli}(\pi), k \in \{\text{light}, \text{heavy}\}$
3	$\ell \leftarrow \text{CrossEntropy}(f^{(k)}(z), Y)$
4	Update parameters of g and $f^{(k)}$ using $\nabla \ell$ (<i>other branch weights are not updated</i>)

4. Experiments

To evaluate the proposed approach we use the NALEF Food Safety dataset [11], published in 2026. The dataset contains images of cucumber crops captured under real-world conditions, annotated with five growth stages: plantula, early vegetative, vegetative growth, flowering, and fruit setting. Overall there are 787 training and 239 test images. As the public test set does not include ground-truth annotations, we perform a stratified train/validation split with an 80/20 ratio for all reported experiments. Images are captured using different devices (professional and smartphone cameras) under varying environmental conditions. The dataset is strongly imbalanced across growth stages, as summarized in Table 2. The smallest class (plantula) contributes an order of magnitude fewer samples than the largest (vegetative growth), leaving only about ten validation images for the minority stages. Accuracy is unsuitable as the primary metric for highly imbalanced datasets. We therefore evaluate all models using Macro F1 score and Cohen’s Kappa.

Table 2

Per-class image counts for the stratified 80/20 train/validation split

Growth stage	Train	Val	Total
Plantula	40	10	50
Early vegetative	42	11	53
Vegetative growth	307	77	384
Flowering	70	17	87
Fruit setting	170	43	213
Total	629	158	787

To improve model generalization, the following augmentations are applied: randomly resized crop, random horizontal flip, and color jitter. All images are resized to 224×224 .

5. Results

Experiment results are presented in Table 3. For each model we report the head configuration that achieved the best result, the type of backbone supervision, evaluation metrics, and computational complexity. For strongly supervised models we report the model name and size; for self-supervised backbones we report the DINO version, backbone architecture, and size. Most strongly supervised backbones are outperformed by self-supervised ones of comparable size. The exception is ViT-L, which is on par with the best DINO-pretrained models and attains the highest Macro F1 overall (0.979). Therefore, the advantage of self-supervised pretraining is concentrated in the small- and mid-sized models (Figure 2).

Table 3

Evaluation results of the considered models and heads. DINOv2 ViT-S with the multibranch head (highlighted in red) achieves the best quality-efficiency trade-off

Model	Head	B. Supervision	F1	Kappa	GFlops	Params (M)
ViT-L	Heavy	Strong	0.979	0.972	123.108	303.831
DINOv2 ViT-L	Heavy	Self	0.978	0.971	162.025	304.898
DINOv2 ViT-S	Multibranch	Self	0.978	0.971	12.248	22.260
DINOv3 ConvNeXt-L	Light	Self	0.970	0.962	68.720	196.238
DINOv3 ViT-S	Heavy	Self	0.968	0.953	9.397	21.797
DINOv3 ConvNeXt-T	Light	Self	0.968	0.962	8.910	27.824
DINOv3 ViT-L	Heavy	Self	0.963	0.952	125.681	303.659
ConvNeXt-T	Heavy	Strong	0.959	0.952	8.910	28.218
DINOv3 ConvNeXt-S	Light	Self	0.955	0.943	17.366	49.459
ViT-B	Heavy	Strong	0.934	0.914	35.127	86.196
EfficientNet V2-L	Heavy	Strong	0.931	0.894	24.461	117.893
ResNet101	Multibranch	Strong	0.913	0.875	15.601	43.562
ResNet50	Heavy	Strong	0.897	0.874	8.176	24.560
EfficientNet-B7	Heavy	Strong	0.883	0.874	10.337	65.101
EfficientNet V2-S	Heavy	Strong	0.875	0.845	5.699	20.836
EfficientNet-B3	Light	Strong	0.871	0.810	1.922	10.704

Figure 2 shows Macro F1 score as a function of computational complexity (Gflops). Strongly supervised models are shown in red and self-supervised models in blue. The advantage of DINO-pretrained backbones is most pronounced in the small-model regime. The best result for a small model (Macro F1 of 0.978) is achieved by DINOv2 ViT-S with the proposed multibranch ensemble

head (highlighted in red in Table 3), matching the strongly supervised ViT-L model, which is over 10x larger. A similarly sized ConvNeXt-Tiny model achieves a substantially lower Macro F1 of 0.959; DINOv3-style pretraining of ConvNeXt-Tiny raises this to 0.968.

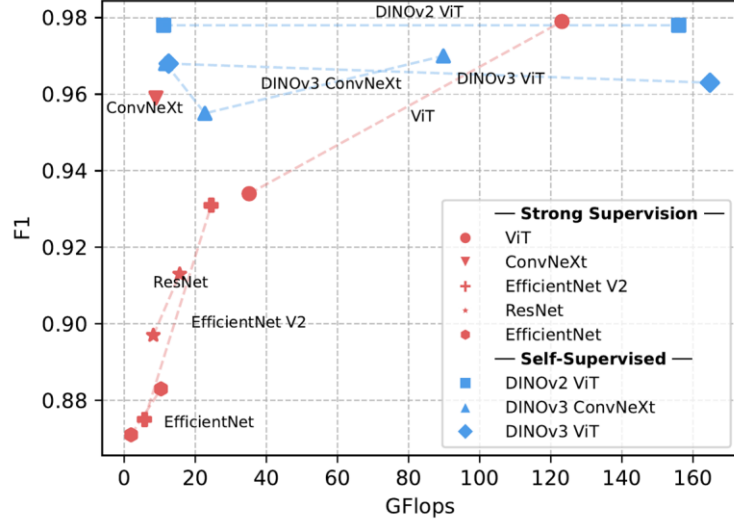


Figure 2: Macro F1 score vs. model computational complexity (GFlops). Self-supervised backbones (blue) hold a clear advantage in the small- and mid-compute regime.

DINOv2 models also outperform their DINOv3 ViT and ConvNeXt counterparts. As discussed above, DINOv3 models are trained on a larger dataset sourced primarily from Instagram, and the greater domain shift is the likely cause of their slightly inferior performance.

6. Discussion

Table 4 details the incremental improvements to DINOv2 ViT-S proposed in this work. As a baseline we train DINOv2 ViT-S with a light head (Figure 1a) for 15 epochs.

Replacing the linear probe with a heavy head improves F1 to 0.955. Switching to the multibranch head further raises F1 to 0.960. Subsequently fine-tuning the full backbone yields the final F1 of 0.978. Importantly, the multibranch head adds negligible parameters and essentially no additional inference overhead compared to the single-branch heavy head, since backbone computation dominates.

Table 4

Ablation study of the proposed improvements over DINOv2 ViT-S

Modification	F1	Kappa	GFlops	Params (M)
DINOv2 ViT-S (baseline)	0.915	0.895	12.247	22.059
+ multibranch head	0.960	0.944	12.248	22.260
+ backbone fine-tuning	0.978	0.971	12.248	22.260

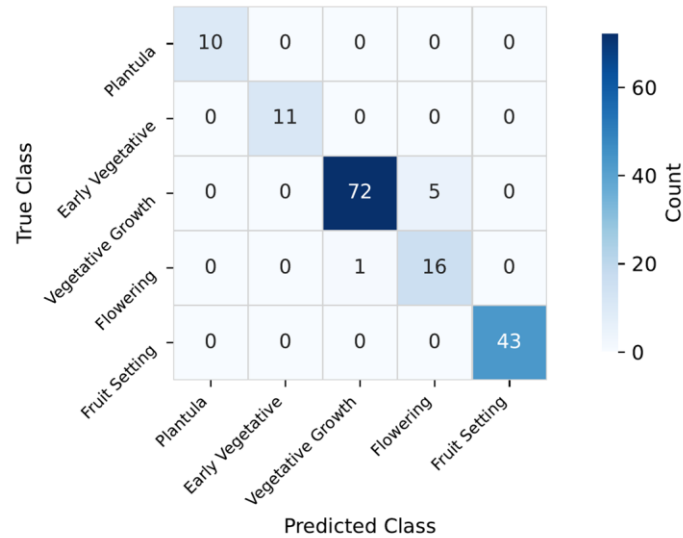


Figure 3: Confusion matrix of DINOv2 ViT-S with multibranch head.

As shown in Figure 3, the proposed DINOv2 ViT-S with multibranch head is a robust classifier: it makes no errors between non-adjacent growth stages, and despite significant class imbalance it remains accurate on minority classes.

The only misclassifications occur between the vegetative growth and flowering stages. At first glance these stages should be visually distinct, separable by the presence of flowers. However, closer inspection of the NALEF Food Safety dataset reveals that images primarily show cucumber leaves, with flowers or fruits often occluded by foliage. The model must therefore discriminate stages based on subtle differences in leaf and vine appearance alone, which is a considerably harder task.

Limitations. The dataset is captured from a leaf-centric viewpoint in which flowers and fruits are frequently occluded, which is also the main source of the residual vegetative-growth vs. flowering confusion. Typically, field examination is performed during daytime and under good environmental conditions. We have not evaluated robustness to adverse conditions such as heavy shading, rain, night-time capture, or lens contamination on smartphone and edge devices. In this work we target the NALEF Food Safety dataset. Therefore, retraining will be needed to adapt the model to other cultivars or environmental conditions. Also, the validation split is highly imbalanced, so small differences in Macro F1 between the top models should be interpreted with this in mind.

The multibranch ensemble head targets any fine-grained classification task where a lightweight head on top of frozen or fine-tuned features must be trained on limited data and is at risk of overfitting. Applying and evaluating it on other fine-grained benchmarks, for instance, species recognition or disease severity grading, is a natural next step that would establish how far the regularization benefit observed here carries over.

7. Conclusions

In this work we proposed a regularizing multibranch ensemble head for adapting Vision Foundation Models to small, domain-shifted datasets, and studied it on cucumber growth stage classification as a fine-grained case study. Experiments across a broad set of strongly supervised and self-supervised backbones on the NALEF Food Safety dataset indicate that self-supervised pretraining yields stronger feature representations for this task in the small- and mid-sized model regime. DINOv2 and DINOv3 backbones outperform ResNet, EfficientNet, and strongly supervised ViT variants of comparable size, while at the largest scale the strongly supervised ViT-L remains competitive; overall, VFMs transfer well to agricultural imagery despite substantial pre-training domain gaps, with the clearest benefit

when model capacity is limited. Notably, DINOv2 ViT-S paired with the proposed multibranch ensemble head achieves a Macro F1 of 0.978 and Cohen's Kappa of 0.971, matching models with over 300M parameters while requiring only 12.2 GFlops and 22.3M parameters, making it suitable for deployment on smartphones and edge devices.

The introduced multibranch ensemble head trains a lightweight linear branch and a heavier MLP branch simultaneously via stochastic branch sampling. At inference, predictions are aggregated in probability space as a uniform Bayesian model average. Ablation analysis confirms that the multibranch head improves over both the linear probe and the heavy head, and that subsequent full-network fine-tuning yields further substantial gains. The method adds negligible parameters and essentially zero inference overhead compared to a single-branch heavy head.

Error analysis reveals that misclassifications are confined to adjacent phenological stages, specifically between vegetative growth and flowering, a consequence of the dataset's leaf-centric imaging perspective in which flowers and fruits are frequently occluded. The absence of errors between non-adjacent stages indicates well-structured internal representations.

Future work includes extending the approach to additional crop species and growth stage taxonomies, as well as incorporating explainability by integrating large language models.

Acknowledgements

This research was carried out as part of the scientific project "Software and algorithmic solutions for agricultural information and analytical systems based on artificial intelligence in variable agroclimatic conditions" (state registration number 0126U001122) funded by the Ministry of Education and Science of Ukraine at the expense of the state budget.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-5.6 in order to: Grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- [1] N. Katal, M. Rzanny, P. Mäder, and J. Wäldchen, Deep learning in plant phenological research: a systematic literature review, *Frontiers in Plant Science*, vol. Volume 13-2022, 2022, doi: 10.3389/fpls.2022.805738.
- [2] Sentinel-2 | Copernicus Data Space Ecosystem. Accessed: May 14, 2026. [Online]. Available: <https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-missions/sentinel-2>
- [3] D. Darlan et al., SmartBerry for AI-based growth stage classification and precision nutrition management in strawberry cultivation, *Sci Rep*, vol. 15, no. 1, p. 14019, Apr. 2025, doi: 10.1038/s41598-025-97168-z.
- [4] J.-S. G. Kim, S. Chung, M. Ko, J. Song, and S. H. Shin, Comparison of Image Preprocessing Strategies for Convolutional Neural Network-Based Growth Stage Classification of Butterhead Lettuce in Industrial Plant Factories, *Applied Sciences*, vol. 15, no. 11, p. 6278, Jun. 2025, doi: 10.3390/app15116278.
- [5] J. Liu, X. Wang, Q. Chen, P. Yan, and X. Liu, Deep learning method for cucumber disease detection in complex environments for new agricultural productivity, *BMC Plant Biol*, vol. 25, no. 1, p. 888, Jul. 2025, doi: 10.1186/s12870-025-06841-y.

- [6] B. Li, G. Zhong, and W. Ke, LMS-Res-YOLO: Lightweight and Multi-Scale Cucumber Detection Model with Residual Blocks, *Sensors*, vol. 25, no. 23, p. 7305, Dec. 2025, doi: 10.3390/s25237305.
- [7] S. Haseli Golzar, H. Bagherpour, and J. Amiri Parian, A New Method to Optimize Deep CNN Model for Classification of Regular Cucumber Based on Global Average Pooling, *Journal of Food Processing and Preservation*, vol. 2024, no. 1, p. 5818803, Jan. 2024, doi: 10.1155/2024/5818803.
- [8] H. Yu, X. Zhang, J. Yan, and X. Meng, YOLO11m-SCFPose: An Improved Detection Framework for Keypoint Extraction in Cucumber Fruit Phenotyping, *Horticulturae*, vol. 11, no. 7, p. 858, Jul. 2025, doi: 10.3390/horticulturae11070858.
- [9] M. Oquab et al., DINOv2: Learning robust visual features without supervision, *Trans. Mach. Learn. Res.*, vol. 2024, 2024, [Online]. Available: <https://openreview.net/forum?id=a68Sut6zFt>
- [10] O. Siméoni et al., DINOv3, *CoRR*, vol. abs/2508.10104, 2025, doi: 10.48550/ARXIV.2508.10104.
- [11] A. C. Menco Tovar, E. Puertas, J. C. Martinez-Santos, and J. Serrano Castañeda, NALEF food safety 2026. Zenodo, Mar. 2026. doi: 10.5281/zenodo.19192941.
- [12] K. Khabarlak, Fine-tuned DINOv2 ViT-L for cucumber growth stage classification: NALEF food safety challenge solution, in *Proceedings of the NALEF 2026 workshop, Cartagena, Colombia, in CEUR workshop proceedings*. 2026.
- [13] J.-S. G. Kim, S. H. Shin, and S. Chung, Instance-level phenotype-based growth stage classification of basil in multi-plant environments, *Front. Plant Sci.*, vol. 16, p. 1707985, Nov. 2025, doi: 10.3389/fpls.2025.1707985.
- [14] C. Zhu, H. Yu, T. Lu, Y. Li, W. Jiang, and Q. Li, Deep learning-based association analysis of root image data and cucumber yield, *The Plant Journal*, vol. 118, no. 3, pp. 696–716, May 2024, doi: 10.1111/tpj.16627.
- [15] K. Khabarlak, I. Laktionov, and G. Diachenko, Feature knowledge distillation using group convolutions for efficient plant pest recognition, in *Proceedings of the 12th International Conference Information Control Systems & Technologies (ICST 2024)*, Odesa, Ukraine, September 23–25, in *CEUR workshop proceedings*, vol. 3790. 2024, pp. 377–387. doi: <https://doi.org/10.55056/ceur-ws.org/Vol-3790/paper33.pdf>.
- [16] R. Singh, R. Bidese-Puhl, K. Dhakal, and S. Sornapudi, Few-shot adaptation of grounding DINO for agricultural domain, in *IEEE/CVF conference on computer vision and pattern recognition workshops, CVPR workshops 2025, nashville, TN, USA, june 11-15, 2025*, Computer Vision Foundation / IEEE, 2025, pp. 5332–5342. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025W/V4A/html/Singh_Few-Shot_Adaptation_of_Grounding_DINO_for_Agricultural_Domain_CVPRW_2025_paper.html
- [17] A. Picon et al., Mitigating domain drift in multi species segmentation with dinov2: a cross-domain evaluation in herbicide research trials. 2026. [Online]. Available: <https://arxiv.org/abs/2508.07514>
- [18] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

- [19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, A ConvNet for the 2020s, in IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, LA, USA, june 18-24, 2022, IEEE, 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.
- [20] M. Tan and Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, in Proceedings of machine learning research, vol. 97. PMLR, 2019, pp. 6105–6114.
- [21] M. Tan and Q. V. Le, EfficientNetV2: Smaller models and faster training, in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, PMLR, 2021, pp. 10096–10106.
- [22] A. Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recognition at scale, in 9th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021, OpenReview.net, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [23] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, CoRR, vol. abs/1207.0580, 2012, [Online]. Available: <http://arxiv.org/abs/1207.0580>
- [24] K. Khabarлак, Post-train adaptive MobileNet for fast anti-spoofing, in Proceedings of the 3rd International Workshop on Intelligent Information Technologies & Systems of Information Security, Khmelnytskyi, Ukraine, March 23-25, 2022, in CEUR workshop proceedings, vol. 3156. CEUR-WS.org, 2022, pp. 44–53. doi: 10.55056/ceur-ws.org/Vol-3156/keynote5.pdf.
- [25] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, Deep networks with stochastic depth, in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part IV, in Lecture notes in computer science, vol. 9908. Springer, 2016, pp. 646–661. doi: 10.1007/978-3-319-46493-0_39.
- [26] G. Song and W. Chai, Collaborative learning for deep neural networks, in Proceedings of the 32nd international conference on neural information processing systems, in NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, pp. 1837–1846.
- [27] A. Galdran, J. W. Verjans, G. Carneiro, and M. A. González Ballester, Multi-head multi-loss model calibration, in Medical image computing and computer assisted intervention – MICCAI 2023, Cham: Springer Nature Switzerland, 2023, pp. 108–117.