

Agri-FM+: A Self-Supervised Foundation Model for Agricultural Vision

Md Jaber Al Nahian, Tapotosh Ghosh, Farnaz Sheikhi, Farhad Maleki

Department of Computer Science, University of Calgary, Calgary, AB, Canada

{mdjaberal.nahian, tapotosh.ghosh, farnaz.sheikhi, farhad.maleki1}@ucalgary.ca

Abstract

Foundation models have revolutionized computer vision, yet their adoption in precision agriculture remains limited due to significant domain shifts from natural images. Existing agricultural foundation models focus primarily on remote sensing applications; to date, no dedicated foundation model exists for close-field agricultural vision. In this paper, we propose Agri-FM+, a self-supervised foundation model specifically tailored for agricultural vision, trained via a two-stage continual learning pipeline. Starting from publicly available unsupervised ImageNet weights from the SlotCon, Agri-FM+ is continually adapted on a curated 147K-image agricultural dataset using SlotCon. Evaluated across eight diverse benchmarks—covering object detection, semantic segmentation, and instance segmentation tasks—Agri-FM+ consistently outperforms both ImageNet-pretrained and randomly initialized models. Under full supervision, it achieves average gains of +1.27% over supervised ImageNet-pretrained and +8.25% over random initialization. Even when trained with only 10% of the annotated data, Agri-FM+ maintains robust performance, achieving gains of +1.02% and +4.54% over supervised ImageNet pretraining and random initialization, respectively. These results demonstrate the ability of Agri-FM+ to provide domain-adapted, label-efficient representations that scale effectively across real-world agricultural vision tasks. The code, weights, and more details will be made available at: <https://github.com/FarhadMaleki/AgriFMPlus>.

1. Introduction

Foundation Models (FMs) pretrained on large-scale datasets have recently demonstrated success across various computer vision tasks, including classification [78], detection [99], and segmentation [50]. However, the direct application of general-purpose foundation models to domain-specific applications, such as precision agriculture, remains a significant challenge. Agricultural images significantly differ from conventional natural images, such as

those in ImageNet [32] and COCO [53] datasets, in several critical aspects. Agricultural objects of interest typically exhibit smaller sizes, higher density, and more repetitive spatial distributions compared to objects commonly depicted in natural scenes. Additionally, these objects are often visually similar and grouped closely together, presenting unique challenges for precise identification, segmentation, and counting tasks. Moreover, agricultural scenes frequently require capturing subtle, fine-grained visual differences critical for tasks like disease detection, pest identification, and crop phenotyping. Agricultural images also show significant heterogeneity due to substantial variations in lighting, growth stages, occlusions, and environmental conditions [55, 91, 100]. Such distinctive characteristics significantly limit the effectiveness and generalizability of traditional foundation models in precision agriculture, underscoring the necessity for specialized, domain-specific foundation models tailored explicitly to agricultural vision tasks.

Some foundation models are developed for agricultural applications [23, 45, 54, 80, 83]. However, they primarily focus on remote sensing imagery from satellites and drones. Despite being effective for large-scale field monitoring and crop classification, these models fail to capture the fine-grained details required for close-field agricultural vision tasks, such as plant disease detection, pest identification, and crop phenotyping. To bridge this gap, a dedicated self-supervised foundation model, tailored specifically for close-field agricultural vision is essential.

To address these challenges, we introduce *Agri-FM+*, a self-supervised foundation model tailored for agricultural vision tasks in close-field settings. We first develop *Agri-FM*, which is pretrained solely on our curated *Agri-147K* dataset—collected and refined from 35 publicly available agricultural datasets. *Agri-FM* achieves performance comparable to supervised ImageNet pretraining, highlighting the effectiveness of domain-specific self-supervised learning (SSL). Building upon this, we propose *Agri-FM+*, which adopts a continual pretraining strategy, first learning from unsupervised ImageNet weights and then further pretraining on *Agri-147K*. This approach enables *Agri-FM+*

to retain general feature representations while adapting to agricultural-specific challenges, mitigating catastrophic forgetting, and enhancing domain adaptation. Our key contributions to this study are as follows:

- **Curated Agri-147K dataset:** We introduce *Agri-147K*, a high-quality agricultural dataset curated from 35 publicly available datasets, ensuring diverse and representative training data for agricultural vision tasks.
- **Agri-FM+ as the first agricultural vision foundation model for close-field tasks:** We propose *Agri-FM+*, leveraging continual self-supervised pretraining to bridge the gap between general and domain-specific feature representations.
- **Comprehensive evaluation:** We benchmark *Agri-FM+* on three core downstream vision tasks—i.e., object detection, semantic segmentation, and instance segmentation—using eight diverse agricultural datasets.

The rest of the paper is organized as follows: Section 2 reviews related work. While Section 3 details our methodology, describes downstream tasks, and evaluation protocol, Section 4 presents the results. Section 5 provides a discussion, and finally, Section 6 concludes the paper.

2. Related Work

Vision foundation models. Trained on large-scale datasets, vision foundation models have significantly advanced computer vision by improving model generalization across diverse tasks. Segment Anything Model (SAM) from Meta AI [49, 68] facilitates zero-shot segmentation with broad applicability. DINOv2 [61] leverages SSL for robust feature representation. SEEM [102] integrates multi-modal information for open-set segmentation. OwlViT [64] enhances open-world object detection using vision-language alignment. CLIP from OpenAI [66] enables text-guided image recognition through contrastive pretraining. BLIP-2 [51] improves vision-language tasks, including image captioning and visual question answering. InternImage [84] achieves state-of-the-art performance in detection and segmentation through hierarchical vision transformers. Florence from Microsoft [98] serves as a unified visual representation model for diverse vision tasks. I-JEPA from Meta AI [16] applies self-supervised learning for robust scene understanding. Stable Diffusion [34] and DALL·E 3 [19] generate high-fidelity images from text prompts, expanding creative applications. These models establish new benchmarks for scalable and generalizable vision systems.

Self-supervised pretraining for dense prediction. Foundation models can be trained using supervised learning (e.g., CLIP [66]), weakly supervised learning (e.g., SAM [49, 68]), or self-supervised learning (e.g., DINO [22, 61]). Self-supervised approaches are particularly valuable in precision agriculture, where obtaining labeled data

is costly and time-intensive. Self-supervised learning has emerged as a powerful technique for learning visual representations without labeled data, making it particularly effective for tasks such as object detection and segmentation. By leveraging intrinsic patterns in raw data, SSL eliminates the need for manual annotations and improves generalization between tasks. Early SSL methods relied on pretext tasks, such as rotation prediction [36], context prediction [33], solving jigsaw puzzle [59], and colorization [101], to extract meaningful representations. However, these approaches often lacked scalability and failed to generalize beyond specific datasets, leading to the widespread adoption of contrastive self-supervised learning as the dominant paradigm for dense prediction tasks.

Contrastive learning enforces feature consistency across augmented views by treating transformations of the same image as positive pairs while considering different images as negative pairs. Early frameworks, such as SimCLR [25] and MoCo [26, 43], demonstrated that contrastive learning significantly improves representation learning by maximizing agreement between similar views of images while simultaneously minimizing agreement with negative samples, i.e., augmented views from different images. MoCo-V3 [27] later enhanced the training stability by eliminating the need for a memory queue, while DINO [22, 61] introduced self-distillation to learn feature representations without explicit negative pairs. However, these methods primarily focus on instance-level discrimination, which is insufficient for dense prediction tasks that require spatial consistency across pixel-level features.

For segmentation and object detection, pixel-level contrastive learning approaches [44, 63, 85, 87, 92] have been more effective than instance-based methods. DenseCL [85] and PixCon [63] work by pixel-wise correspondence between augmented views of images, enhancing the robustness of segmentation and detection models by maintaining spatial alignment. SlotCon [87] further improves it by introducing a group-wise contrastive learning framework that clusters pixel embeddings into learnable slot prototypes, ensuring semantic consistency across spatial locations. This approach is particularly beneficial for agricultural vision tasks, where fine-grained plant structures, disease regions, and pest-affected areas require structured feature representations.

Several studies have examined the continual pretraining strategy, where pretrained encoder weights undergo further refinement using self-supervised learning on a target dataset before fine-tuning for downstream tasks [17, 30, 47, 56, 69]. Continual pretraining enhances domain adaptation by utilizing self-supervised signals prior to task-specific supervised learning. The second pretraining stage can either involve the entire model [17, 69] or focus on specific components of a model [30]. In this work, we implement continual pre-

Table 1. Summary of 35 agricultural datasets used in this study

Ref.	#Images	Ref.	#Images	Ref.	#Images	Ref.	#Images
[7]	829	[5]	2599	[10]	40457	[2]	267
[76]	4000	[75]	3000	[57]	24881	[60]	17509
[95]	3600	[48]	14870	[3]	182	[46]	12354
[6]	900	[8]	251	[29]	2355	[11]	2950
[15]	17724	[4]	392	[9]	100	[93]	20639
[65]	13878	[81]	3824	[1]	670	[37]	5539
[79]	2822	[20]	1081	[35]	21397	[82]	360
[90]	520	[73]	4402	[74]	2400	[18]	3694
[58]	15402	[88]	129000	[31]	6514		

training in the agricultural domain to further enhance the representations learned by the models.

Self-supervised learning in precision agriculture. SSL has transformed precision agriculture by reducing reliance on labeled data for plant disease classification, crop phenotyping, pest detection, and species identification [14]. Contrastive learning-based SSL methods, such as SimCLR [25], are effective in disease detection. However, they require large batch sizes and extensive augmentations [21]. Hybrid approaches, including Attentive Self-Supervised Contrastive Learning (ASCL), improve model robustness and transferability by integrating multi-branch contrastive learning and attention mechanisms [96].

Beyond plant disease classification, SSL has been applied to multi-label plant species identification using DINOv2 [39] and crop phenotyping with 3D plant simulations [52]. SSL also contributes to biodiversity monitoring, as demonstrated by BIOCLIP [78], a vision model trained on TREEOFLIFE-10M. However, FMs for high-resolution, close-up agricultural images, specifically for dense prediction tasks (object detection and segmentation), are still underdeveloped.

3. Methodology

Agri-FM+ is developed through a two-stage continual self-supervised learning process using SlotCon, where a ResNet-50 backbone is first pretrained on ImageNet and then continually adapted on the curated Agri-147K dataset. An overview of the methodology is presented in Figure 1, which also highlights the downstream evaluations across eight datasets and the ablation studies conducted on key design choices.

3.1. Large-scale Agri-147K dataset curation for pretraining

To enable effective self-supervised pretraining for agricultural vision tasks, we introduce *Agri-147K*, a large-scale, unlabeled dataset specifically curated for close-up agricultural image representation learning. *Agri-147K* consists of 147,285 high-quality images which are systematically selected from 35 publicly available datasets (see Table 1), encompassing crop fields, plant diseases, pests, and general

agricultural conditions.

Our curation process was rigorous and multi-faceted, ensuring dataset quality, domain relevance, and applicability to real-world agricultural challenges. In the initial collection phase, we collected 381,362 unlabeled images from diverse sources. However, indiscriminate aggregation of datasets introduces issues such as inconsistent quality, domain mismatch, and redundancy, necessitating a meticulous refinement pipeline. To ensure dataset integrity and relevance, we applied a systematic filtering process. Images below 224×224 pixels were discarded to maintain high visual fidelity, and manual inspection was conducted to remove blurry or noisy images. Domain relevance was enforced by removing datasets containing aerial and satellite imagery as well as lab-captured plant images, which lack real-world complexity. Additionally, licensing verification was rigorously conducted to confirm that all datasets adhered to permissive licensing policies, thereby mitigating ethical and legal concerns and ensuring unrestricted usability for downstream research.

Following this systematic refinement process, *Agri-147K* comprises 147,285 diverse, high-quality images optimized for pretraining agricultural FMs. By eliminating irrelevant, low-quality, and redundant samples, *Agri-147K* ensures that self-supervised pretraining is driven by high-resolution, domain-relevant agricultural imagery, ultimately enhancing representation learning for real-world agricultural applications.

3.2. Foundation model development

We develop two foundation models: *Agri-FM* and *Agri-FM+*. *Agri-FM* is pretrained exclusively on the *Agri-147K* dataset, while *Agri-FM+* undergoes continual pretraining, first on ImageNet and then on *Agri-147K*. Both of these models are pretrained in self-supervised manner, using SlotCon [87] framework. The rationale for choosing this framework was its superior performance compared to most of the group-level or pixel-level approaches, such as DenseCL [85], DetCon [44], and PixPro [92]. Moreover, SlotCon is also designed for learning from scene-centric data which is similar to the agriculture domain. We have pretrained a ResNet-50 network for both of the cases.

SlotCon [87] is a group-level contrastive learning framework that employs a student-teacher architecture to learn structured representations by clustering pixel embeddings into learnable slot prototypes. Given two augmented views of the same image, student and teacher encoders project features into a latent space, where pixel-to-semantic grouping assignments are computed via dot-product similarity with slot prototype and softmax operation. To handle spatial misalignment induced by augmentations, inverse transforms are applied to align assignments across views. A cross-entropy-based grouping loss enforces consistent as-

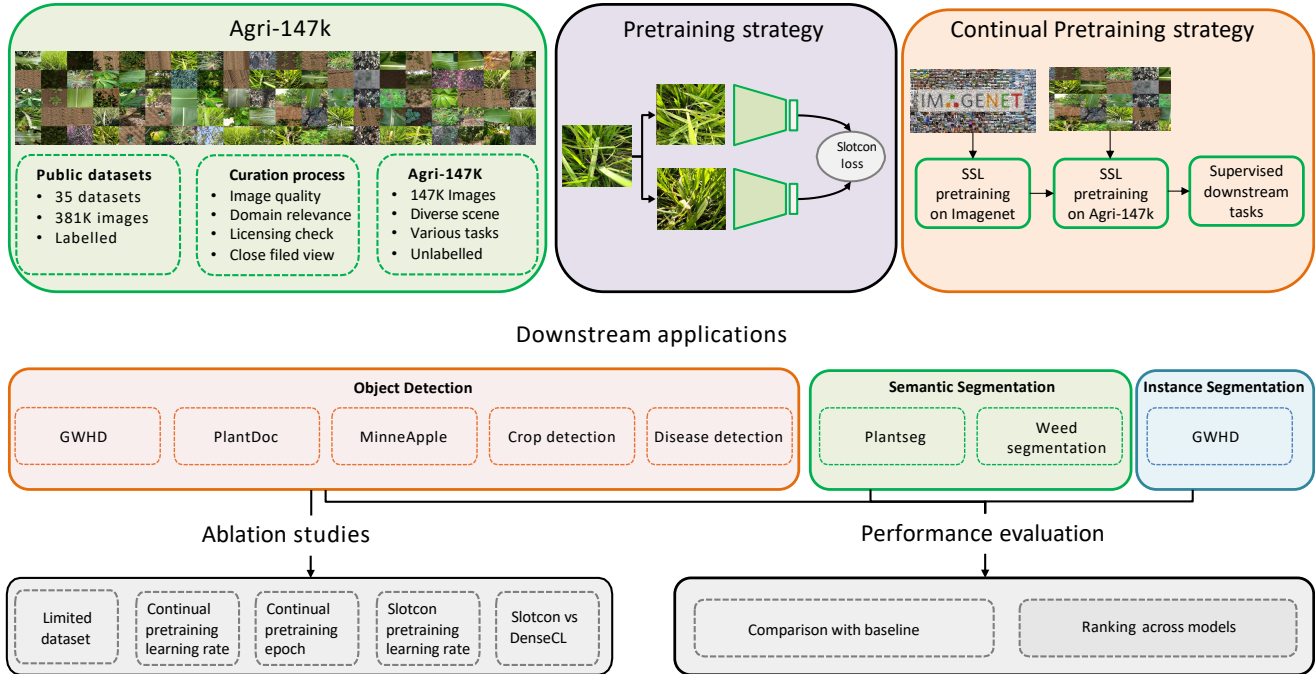


Figure 1. Overview of the study: The Agri-147K dataset was curated for SSL pretraining, leading to the development of Agri-FM and Agri-FM+ through pretraining and continual pretraining with SlotCon. Evaluation was conducted across eight datasets for detection, semantic, and instance segmentation tasks. Ablation studies on the limited dataset, learning rates, epochs, and SlotCon vs DenseCL were performed, along with comparisons with the baseline models.

signments, stabilized via a mean logit adjustment to avoid prototype collapse (i.e: all pixels become assigned to the same prototype).

To improve slot expressiveness, attentive pooling aggregates pixel features into slot embeddings using learned attention weights. Irrelevant slots are filtered, and a contrastive InfoNCE loss pulls semantically matching slots together while separating unrelated ones. A predictor head further enhances slot alignment, with the final contrastive loss applied symmetrically across views. This approach enables SlotCon to learn semantically consistent and spatially aligned representations, beneficial for dense prediction tasks such as segmentation and detection.

Agricultural images differ significantly from natural images, yet pretraining on large-scale natural image datasets like ImageNet can still provide valuable feature representations. However, to achieve optimal performance in the agricultural domain, it is crucial to adapt these pretrained weights using domain-specific data. To accomplish this, we employed a continual pretraining strategy to develop Agri-FM+. Our approach begins with pretraining a ResNet-50 backbone, along with all its auxiliary layers, on ImageNet to establish strong foundational features. Following this, we further pretrain the model on our curated agricultural dataset, leveraging the prior knowledge from ImageNet,

while reducing the domain gap. During this second pre-training phase, we use a smaller learning rate to ensure that the model retains the general features learned from ImageNet while gradually adapting to the agricultural domain. Both pretraining stages are conducted using the SlotCon contrastive learning framework.

3.3. Implementation details

Pretraining setup. The Agri-FM model is pretrained following the SlotCon framework, employing a momentum-based teacher-student contrastive learning strategy. The model consists of two encoders, where the teacher encoder is an exponential moving average of the student encoder. A prototype-based contrastive loss is used to enforce semantic consistency by grouping features into learnable clusters. The student encoder is optimized with LARS [97], while prototype updates are based on batch statistics.

Data augmentation follows BYOL [38], applying 224×224 random resized crops, horizontal flipping, color distortion, grayscale conversion, Gaussian blur, and solarization. Cropped image pairs lacking sufficient overlap are discarded to maintain feature alignment.

The network consists of a ResNet-50 backbone for both encoders, with projection and prediction heads implemented as MLPs, containing a 4096-dimensional hidden

layer and a 256-dimensional output layer. Training is performed on two NVIDIA A6000 GPUs (48 GB each), using LARS optimization with a batch size of 512. The learning rate follows a cosine decay schedule, initialized at 1.0 and scaled with the batch size as $\text{LearningRate} = 1.0 \times \frac{\text{BatchSize}}{256}$.

A weight decay of 10^{-5} is applied, with a 5-epoch warm-up period for stabilization. The model undergoes 800 epochs of pretraining on the Agri-147K dataset, using the same protocol as COCO pretraining in SlotCon. Momentum updates follow a cosine schedule, where the teacher momentum starts at 0.99 and progressively increases to 1.0. Synchronized batch normalization and automatic mixed precision training are enabled. Hyperparameters are set according to SlotCon, with student and teacher temperatures of $\tau_s = 0.1$ and $\tau_t = 0.07$, center momentum $\lambda_c = 0.9$, and 256 prototypes. The contrastive loss temperature is set to $\tau_c = 0.2$, and the balancing ratio λ_g remains at 0.5.

This pretraining setup strictly follows the SlotCon framework, ensuring consistency with self-supervised learning approaches while optimizing feature representations for agricultural vision tasks.

Continual pretraining setup. The setup for Agri-FM+ follows a two-stage SSL approach. Initially, a ResNet-50 backbone is pretrained on the ImageNet dataset to establish general-purpose feature representations. We used the unsupervised ImageNet weight, published by SlotCon [87] after pretraining for 200 epochs with a learning rate of 1.0 scaled by batch size in this case. In the second stage, the model undergoes further self-supervised pretraining on the curated Agri-147K agricultural dataset with a reduced learning rate of 0.001, ensuring that prior knowledge from the first stage is retained while adapting to domain-specific agricultural data. To maintain consistency with SlotCon’s unsupervised ImageNet pretraining, the number of prototypes is set to 2048. The continual pretraining is conducted for 300 epochs, while all other hyperparameters and training settings remain identical to the initial pretraining setup.

3.4. Downstream tasks and evaluation protocol

To assess the effectiveness of our pretrained foundation model, we evaluate it through three key vision tasks: object detection, semantic segmentation, and instance segmentation. Each task is assessed using diverse agricultural datasets, ensuring a robust evaluation of domain adaptability and performance.

3.4.1. Downstream datasets

Object detection. For this task, we utilize the following five datasets, covering a wide range of agricultural scenarios, from wheat head counting to disease and fruit detection.

- **Global Wheat Head Dataset (GWHD):** The GWHD 2021 [31] dataset contains 6,514 images, designed to

detect wheat heads in real-world field conditions. The dataset is split into 4,560 training, 1,303 validation, and 651 test images. Fine-tuning is performed on the training set, and evaluation is conducted on the test set.

- **PlantDoc:** PlantDoc [77] comprises 2,569 images, spanning 13 plant species and 30 categories (healthy and diseased). The dataset contains 8,851 labeled instances, with training and test splits of 2,328 and 239 images, respectively. The images are annotated in COCO format and resized to 416×416 pixels.
- **MinneApple:** The MinneApple [40] dataset consists of 1,001 images with 41,000 annotated apple instances. The train, validation, and test sets contain 701, 200, and 100 images, respectively. The images are resized to 720×1280 pixels.
- **Disease Detection Dataset:** This dataset [71] consists of 5,493 plant leaf images across 13 disease categories. The dataset is split into 2,904 training, 1,416 validation, and 1,163 test images. The image resolution is standardized at 416×416 pixels.
- **Crop Detection Dataset:** The crop detection dataset [72] contains 570 images of grass, maize, banana, sugarcane, and coffee. The dataset is sourced from Roboflow and is split into 392 training, 107 validation, and 71 test images, with a resolution of 360×202 pixels.

Semantic segmentation. For semantic segmentation, we employ two datasets: *PlantSeg* [86] and the *Multiclass Weeds Dataset* [94]. Fine-tuning is conducted using MM-Segmentation pipeline.

- **PlantSeg:** A large-scale dataset for plant disease segmentation, containing 11,458 images across 115 disease categories and 34 plant hosts. It includes a standardized 20% test set per disease category.
- **Multiclass Weeds Dataset:** This dataset includes two weed species, *Soliva Sessilis* (*Field Burrweed*) and *Thlaspi Arvense L.* (*Field Pennycress*), comprising 7,872 augmented images with pixel-wise annotations.

Instance segmentation. Instance segmentation is evaluated using the GWHD Instance Segmentation dataset. Fine-tuning is performed using Detectron2 pipeline.

- **GWHD Instance Segmentation:** This dataset [12] includes 582 images with wheat head instance masks sourced from Roboflow. It is split into 439 training, 102 validation, and 41 test images, all resized to 1024×1024 pixels.

3.4.2. Fine-tuning and evaluation protocol

To ensure a fair evaluation, we fine-tune our pretrained models using standardized settings across all tasks. We transfer the pretrained encoder. The fine-tuning and each pretraining experiment are conducted only once.

Object detection and instance segmentation. For object detection, we employ Faster R-CNN [70] with a ResNet-

50 FPN backbone, while Mask R-CNN [42] with ResNet-50 [41] FPN is used for instance segmentation. Both models are implemented in Detectron2 [89] and fine-tuned end-to-end. The total training schedule consists of 20K iterations, with learning rate decay steps at 12K and 16K iterations. An exception is made for the MinneApple dataset, where training is limited to 6K iterations with decay steps at 2K and 4K iterations. All object detection and instance segmentation experiments are conducted on a single NVIDIA A100 (80GB) GPU.

Semantic segmentation. For this task, we use DeepLabV3+ [24] with a ResNet-50 backbone, implemented in MMSegmentation. Training follows a longer fine-tuning schedule, with the PlantSeg dataset trained for 160K iterations and the Weed Segmentation dataset trained for 80K iterations. Experiments are conducted on a single NVIDIA A100 (40GB) GPU.

Evaluation metrics. Model performance is assessed using the standard task-specific metrics. For object detection and instance segmentation, we report mean Average Precision (mAP) [62] on test sets. Moreover, for semantic segmentation, we use mean Intersection over Union (mIoU) [67] as the primary metric.

Reproducibility. The experiments on the same dataset are conducted under identical hardware and software conditions. Object detection and instance segmentation models are fine-tuned using Detectron2 v0.6 [89], while semantic segmentation experiments are performed using MMSegmentation v1.0.0 [28].

4. Results

In this section, we present the performance evaluation of Agri-FM and Agri-FM+ across multiple agricultural vision tasks. Our experiments cover object detection, instance segmentation, and semantic segmentation, comparing different pretraining strategies.

4.1. Performance of Agri-FM and Agri-FM+

Table 2 compares the performance of Agri-FM and Agri-FM+ on various object detection, semantic segmentation, and instance segmentation tasks. Agri-FM+ outperforms the supervised ImageNet weights in all tasks except the MinneApple object detection and weed semantic segmentation. Herein, Agri-FM achieves the best performance in weed semantic segmentation.

4.2. Performance in limited data scenario

Foundation models are capable of performing well with limited annotated data during fine-tuning. We tested Agri-FM and Agri-FM+ with 5%, 10%, and 20% of annotated data during fine-tuning. Agri-FM+ outperformed supervised ImageNet, Agri-FM, and random initialized weights in all scenarios except the MinneApple 5% data scenario

(depicted in Table 3). The performance gap between random initialization and Agri-FM+ is also significantly higher compared to 100% available data (reported in Table 2).

4.3. Ablation with pretraining strategy and epochs in Agri-FM

We have ablated with a group level (SlotCon) and pixel level (DenseCL) contrastive learning approach in six different downstream tasks. In both cases, the number of epochs was fixed at 800 and followed the original implementation. We found out that in most cases, SlotCon outperformed DenseCL. After selecting SlotCon as our pretraining approach, we have ablated with the number of epochs to pre-train Agri-FM on Agri-147K dataset. We have tried out with 800 and 1600 epochs, where we did not find any significant improvement even after pretraining with $2\times$ epochs. Table 4 depicts the aforementioned performances.

4.4. Ablation with learning rate and epochs in Agri-FM+

The ablation study on the learning rate and continual pretraining epochs for Agri-FM+ (Table 5) shows that a learning rate of 0.001 and 300 continual pretraining epochs provide the best overall performance. In continual pretraining scenario, the learning rate is a crucial factor. Too high of a learning rate can lead to forgetting of previously acquired knowledge with a larger dataset. We have ablated with different learning rates where the number of epochs we train Agri-FM+ with Agri-147K dataset was kept fixed at 300 (shown in Table 5). We have found an equal share in performance where the learning rate of 0.01 was the best in three tasks, whereas in other three out of four tasks, the learning rate of 0.001 performed the best. Since the performance shows no trend, we moved further with the learning rate of 0.001. On the other hand, after finding 0.001 as our go-to learning rate, we have ablated it with the number of epochs where we did not find any continuous increment or decrement. We found that continual pretraining with 300 epochs on Agri-147K dataset yielded a better overall performance.

5. Discussion

In this study, we presented Agri-FM+, a self-supervised foundation model for agricultural vision, trained via continual pretraining on ImageNet and a curated agricultural dataset (Agri-147K). Agri-FM+ demonstrated strong generalization across object detection, semantic segmentation, and instance segmentation tasks, surpassing both supervised ImageNet and single-stage self-supervised learning baselines in its best configuration.

As shown in Table 6 and the radar chart in Figure 2, Agri-FM+ achieved the highest performance across a diverse set of tasks and datasets. It ranked first in six out of

Table 2. Performance of Agri-FM and Agri-FM+ in several object detection, semantic segmentation and instance segmentation tasks.

Pretrained ResNet-50 Weight	Object Det. (mAP)				Instance Seg. (mAP)		Semantic Seg. (mIoU)	
	GWHD	PlantDoc	MinneApple	Crop Det.	Disease Det.	GWHD	PlantSeg	Weed Seg.
Random Init.	44.27	20.99	35.61	32.96	34.57	66.26	16.15	95.13
Supervised ImageNet	47.97	38.15	40.53	31.81	47.71	72.09	28.31	95.17
Agri-FM	48.18	38.53	38.21	30.01	47.90	72.41	28.07	95.87
Agri-FM+	48.48	41.61	40.41	33.41	49.83	73.48	29.76	94.95

Table 3. Performance of Agri-FM and Agri-FM+ in GWHD, PlantDoc, MinneApple, and Disease Detection Object Detection tasks with limited annotated data during fine-tuning. The results reported here are AP. We have randomly selected 5%, 10% and 20% annotated data during the fine-tuning process.

Pretrained ResNet-50 Weight	GWHD (mAP)			PlantDoc (mAP)			MinneApple (mAP)			Disease Detection (mAP)		
	5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%
Random Init.	34.57	38.32	41.91	3.72	6.19	10.84	19.82	25.81	27.11	15.94	19.13	26.16
Supervised ImageNet	38.33	39.71	41.99	9.17	17.97	22.92	23.25	28.55	32.52	22.46	27.29	36.31
Agri-FM	37.92	39.41	41.29	9.16	15.04	19.79	21.44	28.96	32.61	20.62	25.71	34.98
Agri-FM+	39.24	40.41	42.44	10.73	18.09	24.22	22.08	30.01	33.21	23.18	29.09	37.06

Table 4. Ablation study with different pretraining approaches and epochs for Agri-FM.

Task	Dataset	Pretraining Approach		Pretraining Epochs	
		SlotCon	DenseCL	800 Ep.	1600 Ep.
Obj Det (mAP)	GWHD	48.18	48.03	48.18	48.08
	PlantDoc	38.53	35.68	38.53	38.43
	MinneApple	38.21	39.57	38.21	38.23
	Disease Det.	47.90	46.88	47.90	48.75
Sem Seg (mIoU)	Weed Seg.	95.87	94.90	95.87	95.89
	PlantSeg	28.07	25.40	28.07	28.03

Table 5. Ablation on the learning rate and continual pretraining epochs for Agri-FM+.

Task	Dataset	Learning Rate			Cont. Epochs		
		0.01	0.001	0.0001	200	300	400
Obj Det (mAP)	GWHD	48.68	48.48	48.29	48.63	48.48	48.54
	PlantDoc	42.72	41.61	41.06	42.00	41.61	41.60
	MinneApple	40.50	40.41	39.76	40.33	40.41	40.68
	Crop Det.	32.63	33.41	30.56	31.83	33.41	31.06
	Disease Det.	49.30	49.83	49.56	50.60	49.83	50.43
Sem Seg (mIoU)	PlantSeg	29.01	29.76	30.93	29.91	29.76	28.88
	Weed Seg.	95.03	94.95	95.20	95.22	94.95	95.28

eight benchmarks, confirming its strength as a unified agricultural foundation model. In the fully annotated settings, it achieved an average gain of +1.27% over supervised ImageNet weights and +8.25% over random initialization. The most notable improvements were observed in object detection datasets such as PlantDoc (+3.46% vs ImageNet, +20.62% vs random) and Disease Detection (+2.12% vs ImageNet, +15.26% vs random). Performance gains also extend to instance segmentation tasks (+1.39% vs ImageNet on GWHD) and semantic segmentation (+1.45% vs ImageNet on PlantSeg), highlighting the ability of Agri-FM+

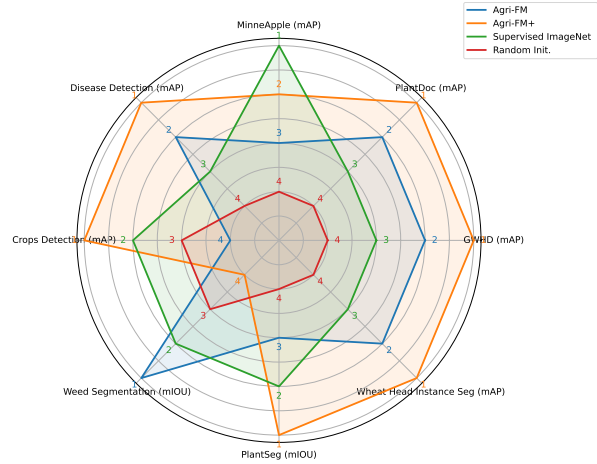


Figure 2. Model performance ranking across multiple datasets. The radar chart compares the models of Agri-FM, Agri-FM+, Supervised ImageNet, and Random Init., based on their rankings in object detection and segmentation tasks. Lower rankings (closer to the center) indicate inferior, while higher rankings (closer to the outer edge) indicate superior performance. Agri-FM+ consistently achieves the best rankings in most datasets.

to generalize across dense and sparse prediction tasks alike.

Agri-FM+ also maintained a strong performance in low-label regimes. As shown in Table 7, with only 10% of the annotated data, the model improved by +1.02% over models where the training started from supervised ImageNet initialization and by +4.54% over randomly initialized models. These results underline the data efficiency of Agri-FM+—a key advantage in agriculture, where acquiring large-scale expert-labeled datasets is often impractical. Notably, even Agri-FM, a variant trained only on the Agri-147K dataset

Table 6. Performance gains (%) of Agri-FM+ over ImageNet and Random Init under full annotation.

Dataset	Task	Δ over ImgNet	Δ over RandInit
GWHD	Obj Det	+0.51	+4.21
PlantDoc	Obj Det	+3.46	+20.62
MinneApple	Obj Det	-0.12	+4.80
Crop Det.	Obj Det	+1.60	+0.45
Disease Det.	Obj Det	+2.12	+15.26
GWHD	Inst Seg	+1.39	+7.22
PlantSeg	Sem Seg	+1.45	+13.61
Weed Seg.	Sem Seg	-0.22	-0.18
Average	—	+1.27	+8.25

Table 7. Agri-FM+ gains (%) over ImageNet and RandInit with 10% annotated data.

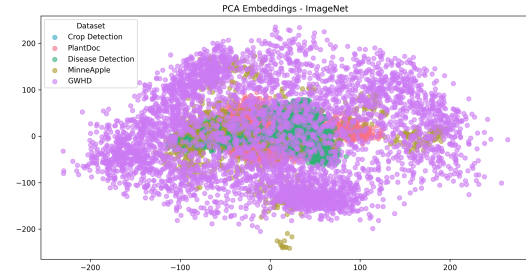
Dataset	Task	Δ ImgNet	Δ RandInit
GWHD	Obj Det	+0.70	+2.09
PlantDoc	Obj Det	+0.12	+1.90
MinneApple	Obj Det	+1.46	+4.20
Disease Det.	Obj Det	+1.80	+9.96
Average	—	+1.02	+4.54

without ImageNet pretraining, performs competitively with or better than supervised ImageNet models. This finding supports the idea that domain-specific self-supervised learning on high-quality unlabeled data can produce highly transferable representations for downstream agricultural tasks.

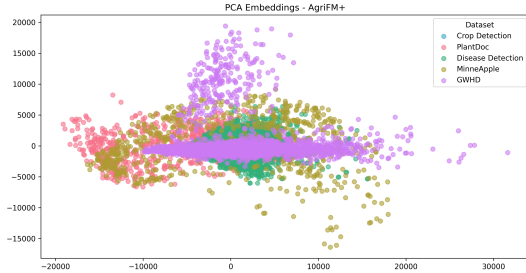
To further assess representational quality, we have conducted PCA [13] visualization (shown in Figure 3) on embeddings from five downstream datasets. Compared to ImageNet-pretrained models, Agri-FM+ embeddings form more compact, well-separated clusters, suggesting better inter-dataset discrimination. This aligns with the SlotCon framework’s objective of learning semantically consistent prototypes, which is particularly advantageous for structured agricultural imagery, involving crops, diseases, or row-level patterns.

The curation quality of the Agri-147K dataset also plays a vital role. By refining an initial pool of 381K images and removing non-agricultural or noisy samples, we ensured high domain relevance, which significantly improved pre-training outcomes. These findings reinforce the idea that, in self-supervised learning, the quality and specificity of the dataset can be more impactful than the dataset size alone.

It is important to acknowledge that our study has several limitations. Currently, Agri-FM+ is limited to close-field RGB imagery and a ResNet-50 backbone. Future work will explore larger architectures (e.g., transformers), multimodal input (e.g., hyperspectral, UAV, or temporal imagery), and zero-shot generalization using vision-language alignment.



(a) Supervised ImageNet pretrained



(b) Agri-FM+

Figure 3. PCA visualization of learned features from five datasets. (a) Supervised ImageNet (b) Agri-FM+. As depicted, Agri-FM+ shows improved feature compactness and inter-dataset separation compared to supervised ImageNet-pretrained ResNet-50, suggesting enhanced domain adaptation and generalization.

6. Conclusion

In this paper, we have proposed the first self-supervised foundation model, designed specifically for close-field agricultural vision. The proposed model, titled Agri-FM+, is built upon a continual pretraining strategy that first learns from unsupervised pretraining on ImageNet and then adapts to a curated agricultural dataset (Agri-147K). Agri-FM+ demonstrates strong generalization across object detection, semantic segmentation, and instance segmentation tasks. It consistently outperforms supervised ImageNet pretraining and random initialization, with notable gains in both fully annotated and limited annotation settings. The results confirm that domain-specific self-supervised learning, coupled with high-quality dataset curation, enables efficient and robust representation learning for agriculture. This work lays the foundation for building scalable and label-efficient vision systems, tailored to real-world agricultural challenges.

Acknowledgments

This research was generously supported by Google’s Compute Credit Scholarship, which provided the computational resources for this study. The authors extend their sincere gratitude to Dr. Ian McQuillan for his invaluable guidance.

References

- [1] Wheat spike dataset. <https://www.kaggle.com/datasets/voglinio/wheat-dataset-original>, 2020. Kaggle. Accessed: 2025-03-22. 3
- [2] Corn field mapped with ebee ag in brazil. <https://ageagle.com/data-set/corn-field-mapped-with-ebec-ag-in-brazil/>, 2021. AgEagle. Published: 2021-02-18. Accessed: 2025-03-22. 3
- [3] Mixed-use agricultural fields ebee ag. <https://ageagle.com/data-set/mixed-use-agricultural-fields-ebec-ag/>, 2021. AgEagle. Published: 2021-02-17. Accessed: 2025-03-22. 3
- [4] Perennial plants detection. <https://www.kaggle.com/datasets/benediktgeisler/perennial-plants-detection>, 2021. Kaggle. Accessed: 2025-03-22. 3
- [5] Makerere fall armyworm crop challenge dataset. <https://zindi.africa/competitions/makerere-fall-armyworm-crop-challenge/data>, 2022. Zindi. Accessed: 2025-03-22. 3
- [6] Leaf-images-dataset. <https://www.kaggle.com/datasets/hamedetezadi/leaf-image-dataset>, 2022. Kaggle. Accessed: 2025-03-22. 3
- [7] Agriculture crops dataset. <https://www.kaggle.com/datasets/osamajalilhasan/agriculture-crops-dataset>, 2023. Accessed: 2025-03-22. 3
- [8] Maize images dataset. <https://www.kaggle.com/datasets/devanshujoshi01/maize-images-dataset>, 2023. Kaggle. Accessed: 2025-03-22. 3
- [9] Pinotnoirgrapes. <https://www.kaggle.com/datasets/nicolaasregnier/pinotnoirgrapes/data>, 2023. Kaggle. Accessed: 2025-03-22. 3
- [10] Classification maize & cocoa datasets. <https://www.kaggle.com/datasets/clarentjd/classification-maize-and-cocoa-datasets>, 2024. Kaggle. Accessed: 2025-03-22. 3
- [11] Maize phenology images acquired using ndvi camera. <https://www.kaggle.com/datasets/zzzzqi/maize-phenology-images-acquired-using-ndvi-camera>, 2024. Kaggle. Accessed: 2025-03-22. 3
- [12] Wheat segmentation dataset. <https://universe.roboflow.com/wheat-seg-xcxow/wheat-segmentation-sr7mm>, 2024. visited on 2024-07-29. 5
- [13] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 8
- [14] Abdullah Al Mamun, David Ahmedt-Aristizabal, Miaohua Zhang, Md Ismail Hossen, Zeeshan Hayder, and Mohammad Awrangzeb. Plant disease detection using self-supervised learning: A systematic review. *IEEE Access*, 2024. 3
- [15] Emmanuel Asante, Obed Appiah, and Eric Opoku. Maize seed dataset, 2024. <https://www.kaggle.com/dsv/8681789>. 3
- [16] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. 2
- [17] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3478–3488, 2021. 2
- [18] Suchet Bargoti and James Underwood. Deep fruit detection in orchards. *arXiv preprint arXiv:1610.03677*, 2016. 3
- [19] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. 2
- [20] Swami Nisha Bhagirath, Vaibhav Bhatnagar, and Linesh Raja. Winter wheat leaf images dataset. Mendeley Data, V1, 2023. <https://doi.org/10.17632/gj72pjfvb2.1.3>
- [21] Songpol Bunyang, Natdanai Thedwichienchai, Krisna Pintong, Nuj Lael, Wuthipoom Kunaborimas, Phawit Boonrat, and Thitirat Siriborvornratanakul. Self-supervised learning advanced plant disease image classification with simclr. *Advances in Computational Intelligence*, 3(5):18, 2023. 3
- [22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2
- [23] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 1
- [24] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6
- [25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 2, 3
- [26] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 2
- [27] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 2

- [28] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. <https://github.com/open-mmlab/mms Segmentation>, 2020. 6
- [29] Hamish Craze and Dave Berger. Maize in field dataset, 2022. <https://www.kaggle.com/dsv/3603983>. 3
- [30] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. Pecop: Parameter efficient continual pretraining for action quality assessment. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 42–52, 2024. 2
- [31] Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto, Shahameh Shafiee, Izzat SA Tahir, et al. Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods. *Plant Phenomics*, 2021. 3, 5
- [32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1
- [33] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2
- [34] Chengbin Du, Yanxi Li, Zhongwei Qiu, and Chang Xu. Stable diffusion is unstable. In *Advances in Neural Information Processing Systems*, pages 58648–58669. Curran Associates, Inc., 2023. 2
- [35] ErnestMwebaze, Jesse Mostipak, Joyce, Julia Elliott, and Sohier Dane. Cassava leaf disease classification. <https://kaggle.com/competitions/cassava-leaf-disease-classification>, 2020. Kaggle. 3
- [36] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2
- [37] Thomas Mosgaard Giselsson, Rasmus Nyholm Jørgensen, Peter Kryger Jensen, Mads Dyrmann, and Henrik Skov Midtby. A public image database for benchmark of plant seedling classification algorithms. *arXiv preprint arXiv:1711.05458*, 2017. 3
- [38] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4
- [39] Murilo Gustineli, Anthony Miyaguchi, and Ian Stalter. Multi-label plant species classification with self-supervised vision transformers. *arXiv preprint arXiv:2407.06298*, 2024. 3
- [40] Nicolai Häni, Pravakar Roy, and Volkan Isler. Minneapolis: a benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters*, 5(2):852–858, 2020. 5
- [41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [42] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6
- [43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [44] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10086–10096, 2021. 2, 3
- [45] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [46] Aminul Huq, Dimitris Zermas, and George Bebis. Identification of abnormality in maize plants from uav images using deep learning approaches. In *International Symposium on Visual Computing*, pages 583–596. Springer, 2023. 3
- [47] András Kalapos and Bálint Gyires-Tóth. Self-supervised pretraining for 2d medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, pages 472–484, Cham, 2023. Springer Nature Switzerland. 2
- [48] KaraAgro AI Foundation. Drone-based-agricultural-dataset-for-crop-yield-estimation (revision 2530d1d), 2023. <https://huggingface.co/datasets/KaraAgroAI/Drone-based-Agricultural-Dataset-for-Crop-Yield-Estimation>. 3
- [49] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2
- [50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1
- [51] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2
- [52] Yinglun Li, Xiaohai Zhan, Shouyang Liu, Hao Lu, Ruibo Jiang, Wei Guo, Scott Chapman, Yufeng Ge, Benoit Solan, Yanfeng Ding, et al. Self-supervised plant phenotyping by combining domain adaptation with 3d plant model simula-

- tions: Application to wheat leaf counting at seedling stage. *Plant Phenomics*, 5:0041, 2023. 3
- [53] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 1
- [54] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiacong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [55] Federico Magistri, Jan Weyler, Dario Gogoll, Philipp Lottes, Jens Behley, Nik Petrinic, and Cyrill Stachniss. From one field to another—unsupervised domain adaptation for semantic segmentation in agricultural robotics. *Computers and Electronics in Agriculture*, 212:108114, 2023. 1
- [56] Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards Geospatial Foundation Models via Continual Pretraining. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16760–16770, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2
- [57] Patrick Mensah Kwabena, Vivian Akoto-Adjepong, Kwabena Adu, Mighty Abra Ayidzoe, Elvis Asare Bediako, Owusu Nyarko-Boateng, Samuel Boateng, Esther Fobi Donkor, Faiza Umar Bawah, Nicodemus Songose Awarayi, Peter Nimbe, Isaac Kofi Nti, Muntala Abdulai, Remember Roger Adjei, and Michael Opoku. Dataset for crop pest and disease detection. Mendeley Data, V1, 2023. <https://doi.org/10.17632/bwh3zbpkpv.1>. 3
- [58] Ben-Wycliff Mugal, Joyce Nakatumba-Nabende, Andrew Katumba, Claire Babirye, Francis-Jeremy Tsubira, Chodrine Mutebi, Solomon Nsumba, and Gloria Namanya. Makerere university beans image dataset. <https://doi.org/10.7910/DVN/TCKVEW>, 2022. Harvard Dataverse, V2. 3
- [59] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2
- [60] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):2058, 2019. 3
- [61] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2
- [62] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, 2020. 6
- [63] Zongshang Pang, Yuta Nakashima, Mayu Otani, and Hajime Nagahara. Revisiting pixel-level contrastive pre-training on scene images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1784–1793, 2024. 2
- [64] Sarah Parisot, Yongxin Yang, and Steven McDonagh. Learning to name classes for vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23477–23486, 2023. 2
- [65] Petchiammal, Briskline Kiruba, Murugan, and Pandarasamy Arjunan. Paddy doctor: A visual image dataset for automated paddy disease classification and benchmarking. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, pages 203–207, 2023. 3
- [66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2
- [67] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016. 6
- [68] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2
- [69] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, Kurt Keutzer, and Trevor Darrell. Self-Supervised Pretraining Improves Self-Supervised Pretraining. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1050–1060, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2
- [70] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 5
- [71] Roboflow. Detecting diseases dataset. <https://universe.roboflow.com/artificial-intelligence-82oex/detecting-diseases/dataset/6>, 2022. visited on 2025-03-21. 5
- [72] Roboflow. Crops (maize, cassava, sugarcane vs grass) dataset. <https://universe.roboflow.com>.

- [com/school-el5qm/crops-maize-cassava-sugarcane-vs-grass](https://www.kaggle.com/school-el5qm/crops-maize-cassava-sugarcane-vs-grass), 2023. visited on 2025-03-21. 5
- [73] Adrian Salazar-Gomez, Madeleine Darbyshire, Junfeng Gao, Elizabeth I. Sklar, and Simon Parsons. Towards practical object detection for weed spraying in precision agriculture. *arXiv preprint arXiv:2109.11048*, 2021. <https://arxiv.org/abs/2109.11048>. 3
- [74] Thiago T Santos and Luciano Gebler. A methodology for detection and localization of fruits in apples orchards from aerial images. *arXiv preprint arXiv:2110.12331*, 2021. 3
- [75] Sandi Indika Saputra. Corn stalk disease, 2023. <https://www.kaggle.com/dsv/6848682>. 3
- [76] Sandi Indika Saputra. Corn leaf disease, 2023. <https://www.kaggle.com/dsv/6341289>. 3
- [77] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 249–253. 2020. 5
- [78] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 1, 3
- [79] Kaspars Sudars, Janis Jasko, Ivars Namatevs, Liva Ozola, and Niks Badaukis. Dataset of annotated food crops and weed images for robotic computer vision control. *Data in brief*, 31:105833, 2020. 3
- [80] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022. 1
- [81] Vivek Tiwari, Ravi R Saxena, and Muneendra Ojha. Insectbase: Soybean crop insect raw image dataset.v1 with bounding boxes for classification and localization. <https://doi.org/10.6084/m9.figshare.13077221.v4>, 2020. figshare. Dataset. doi:10.6084/m9.figshare.13077221.v4. 3
- [82] Aleksandar Vakanski. A dataset of multispectral potato plants images. https://www.webpages.uidaho.edu/vakanski/Multispectral_Images_Dataset.html, 2024. University of Idaho. Accessed: 2025-03-22. 3
- [83] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022. 1
- [84] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, 2023. 2
- [85] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3024–3033, 2021. 2, 3
- [86] 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. *arXiv preprint arXiv:2409.04038*, 2024. 5
- [87] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, pages 16423–16438. Curran Associates, Inc., 2022. 2, 3, 5
- [88] Jan Weyler, Federico Magistri, Elias Marks, Yue Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley. Phenobench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3
- [89] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6
- [90] Zhenwei Wu, Xinfu Wang, Minghao Liu, and Chengxiu Sun. Tomatoplantfactorydataset. Mendeley Data, V1, 2023. <https://doi.org/10.17632/8h3s6jkjyff.1>. 3
- [91] Shuai Xiang, Pieter M Blok, James Burridge, Haozhou Wang, and Wei Guo. Doda: Diffusion for object-detection domain adaptation in agriculture. *arXiv preprint arXiv:2403.18334*, 2024. 1
- [92] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16684–16693, 2021. 2, 3
- [93] H. Xu. Plantvillage disease classification challenge - color images. <https://doi.org/10.5281/zenodo.1204914>, 2018. Zenodo. doi: 10.5281/zenodo.1204914. 3
- [94] Shivam Yadav, Sanjay Soni, and Sanjay Gupta. Multi-class weeds dataset for image segmentation. <https://doi.org/10.6084/m9.figshare.22643434.v1>, 2023. Dataset, accessed on 2025-03-21. 5
- [95] Jinhui Yi, Lukas Krusenbaum, Paula Unger, Hubert Hüging, Sabine J. Seidel, Gabriel Schaaf, and Juergen Gall. Deep learning for non-invasive diagnosis of nutrient deficiencies in sugar beet using rgb images. *Sensors*, 20(20): 5893, 2020. 3
- [96] Getinet Yilma, Mesfin Dagne, Mohammed Kemal Ahmed, and Ravindra Babu Bellam. Attentive self-supervised contrastive learning (ascl) for plant disease classification. *Results in Engineering*, page 103922, 2025. 3
- [97] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 4

- [98] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021. [2](#)
- [99] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. [1](#)
- [100] Junbo Zhang, Shifeng Xu, Jun Sun, Dinghua Ou, Xiaobo Wu, and Mantao Wang. Unsupervised adversarial domain adaptation for agricultural land extraction of remote sensing images. *Remote Sensing*, 14(24):6298, 2022. [1](#)
- [101] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. [2](#)
- [102] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, pages 19769–19782. Curran Associates, Inc., 2023. [2](#)