AgriCLIP: Adapting CLIP for Agriculture and Livestock via Domain-Specialized Cross-Model Alignment

Umair Nawaz[1], Muhammad Awais[1], Hanan Gani[1], Muzammal Naseer[2], Fahad Shahbaz Khan[1,3], Salman Khan[1], Rao Muhammad Anwer[1]

[1]Mohamed bin Zayed University of AI [2]Khalifa University
[3]Linköping University

Correspondence: umair.nawaz@mbzuai.ac.ae

Abstract

Capitalizing on a vast amount of image-text data, large-scale vision-language pre-training has demonstrated remarkable zero-shot capabilities and has been utilized in several applications. However, models trained on general everyday web-crawled data often exhibit suboptimal performance for specialized domains, likely due to domain shift. Recent works have tackled this problem for some domains (e.g., healthcare) by constructing domain-specialized image-text data. However, constructing a dedicated large-scale image-text dataset for sustainable areas of agriculture and livestock is still open to research. Further, this domain desires fine-grained feature learning due to the subtle nature of the downstream tasks (e.g., nutrient deficiency detection and livestock breed classification). To address this, we present AgriCLIP, a vision-language foundational model dedicated to the domain of agriculture and livestock. First, we propose a large-scale dataset named ALive that leverages a customized prompt generation strategy to overcome the scarcity of expert annotations. Our ALive dataset covers crops, livestock, and fishery, with around 600,000 image-text pairs. Second, we propose a training pipeline that integrates both contrastive and self-supervised learning to learn both global semantic and local fine-grained domain-specialized features. Experiments on a diverse set of 20 downstream tasks demonstrate the effectiveness of the AgriCLIP framework, achieving an absolute gain of 9.07% in terms of average zero-shot classification accuracy over the standard CLIP adaptation via domain-specialized ALive dataset. Our ALive dataset and code can be accessible at Github.

1 Introduction

Recent years have seen the success of large-scale image-text pre-training, e.g., CLIP in general zero-shot capabilities, and their widespread utility (Radford et al., 2021). However, the performance of these models often falters in specialized domains (such as healthcare, geo-sensing, and climate (Wang et al., 2022; Vivanco Cepeda et al., 2024; Mishra-Sharma et al., 2024)) due to the presence of inherent domain gaps and the different nature of downstream tasks in specialized domains. This gap in performance has led to the curation of image-text datasets from existing domain-specific data sources for the training of expert CLIP variants (Wang et al., 2022; Zhang et al., 2023; Xu et al., 2024).

However, adapting vision-text pre-training for agriculture is challenging due to two reasons. First, unlike many other fields, agriculture lacks any comprehensive image-text data sources. Existing agricultural datasets are predominantly designed for narrow tasks (e.g., disease classification) and consist only of images and task-specific information (e.g., class names), restricting their utility in vision-language pre-training. Second, most downstream agricultural tasks require fine-grained feature learning – such as distinguishing subtle differences in rusty patches on visually similar leaves for disease classification – where contrastive learning alone may be insufficient. Some previous works have utilized deep learning for agricultural tasks (Bharman et al., 2022; Farjon et al., 2023), and initial efforts have been made to fine-tune large language models for the field (Arshad et al., 2024). However, to the best of our knowledge, no such effort exists for vision-language pre-training for agriculture.

To address the above-mentioned challenges, we introduce AgriCLIP, which comprises a large image-text dataset called ALive (Agriculture and Livestock) and a vision-language pre-training pipeline that combines the strengths of contrastive and self-supervised learning to learn both global semantic features and fine-grained visual details. This combination of a large-scale dataset and training pipeline enables the vision-language model to achieve strong downstream performance in agricultural tasks. An overview of our method is shown

Proceedings of the 31st International Conference on Computational Linguistics, pages 9630–9639
January 19–24, 2025.

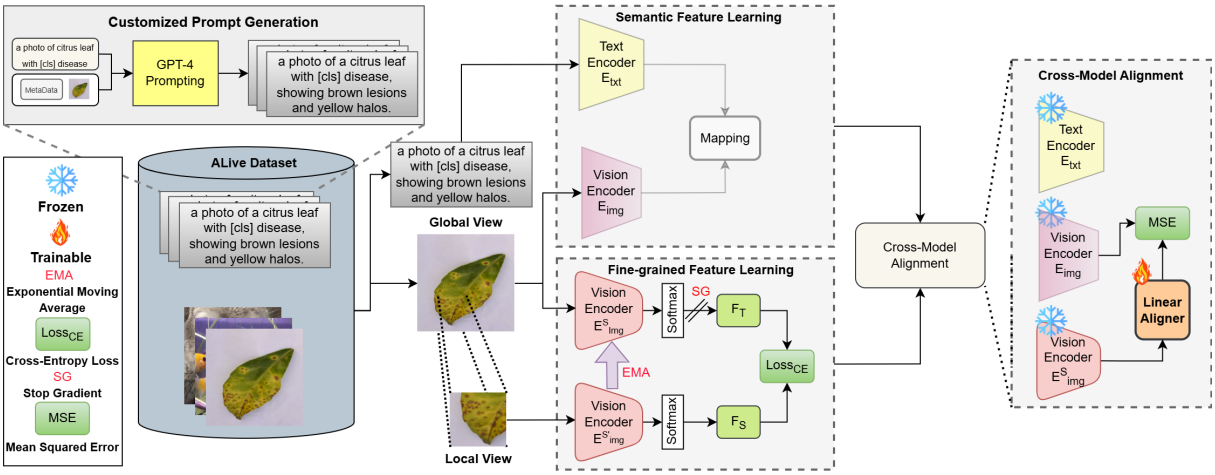

Figure 1: Overview of our proposed framework, consisting of the ALive dataset and the AgriCLIP training pipeline, designed to integrate both global semantic and local fine-grained domain-specialized features. The ALive is an image-text dataset for the agriculture and livestock domain that is constructed by leveraging images and their metadata to prompt GPT-4, generating customized text for each image. The AgriCLIP training pipeline consists of **Semantic Feature Learning**, where contrastive learning is utilized to train image and text encoders; **Fine-Grained Feature Learning**, using a self-supervised approach to train the vision encoder; and **Cross-Model Alignment**, aligning vision encoders from the previous stages to enable zero-shot generalization.

in Figure 1. Our contributions are as follows.

Our primary contribution is the creation of a large, diverse image-text dataset derived solely from vision-based agricultural datasets. To construct the ALive dataset, we carefully selected 25 classification-based datasets covering crops, livestock, and fish, totaling around 600,000 images. The images span various modalities (drone, robotic, RGB), tasks (e.g., nutrient deficiency detection, plant disease identification, livestock breed classification), and environments (indoor, outdoor, and underwater). To generate corresponding text pairs for each image, we use a customized prompt generation strategy that employs metadata and class-specific information to prompt GPT4 (Achiam et al., 2023), producing a diverse set of descriptive text. To evaluate the model's out-of-distribution performance for diverse downstream tasks, we curate a dedicated evaluation set consisting of 300,000 images from datasets entirely disjoint from the ALive training set.

Our second contribution is a training pipeline that combines image-text contrastive and image-only self-supervised learning to boost global semantic features with fine-grained visual details. Our training pipeline consists of three stages. In the first stage, we further pre-trained CLIP (Radford et al., 2021) vision and language encoders on image-text pairs from the ALive dataset using contrastive learning to capture global semantic features. In the second stage, we further pre-trained a separate vision encoder with the DINO-based training (Caron et al., 2021) method, focusing on learning local fine-grained features crucial for downstream tasks. Consequently, we align vision encoders from the first two stages to enable zero-shot classification learning. Our experiments on 20 diverse sets of downstream datasets with around 300,000 images demonstrate the efficacy of our AgriCLIP in terms of zero-shot performance.

2 Related Work

The application of AI in agriculture has been studied, with a focus on tasks such as crop monitoring (Wu et al., 2023), disease detection (Arun and Umamaheswari, 2023; Yousuf and Khan, 2021; Khan and Oberoi, 2019; Khan et al., 2023a), and yield prediction (Meena et al., 2023). Traditional machine-learning approaches have relied heavily on supervised learning, requiring large amounts of labeled data (Kotwal et al., 2023). However, the variability and complexity of agricultural environments often make it challenging to obtain sufficient labeled data, leading to the exploration of zero-shot learning methods.

Zero-shot learning has gained traction in recent years, with models like CLIP (Radford et al., 2021) demonstrating the ability to generalize to unseen categories by leveraging textual descriptions. CLIP's success in various domains has prompted research into its application in specialized fields such as medicine, remote sensing, and astronomy (Zhao

et al., 2023; Li et al., 2023b; Mishra-Sharma et al., 2024). For agriculture, efforts have been made for the specific tasks of plant disease identification in a few-shot manner, but they are mainly restricted to either a single task or limited data variability (Zhou et al., 2024; Zhong et al., 2020; Sun et al., 2024). Self-supervised learning models like DINO (Caron et al., 2021) have also been explored for various applications, including visual recognition tasks (Li et al., 2023a; Liu et al., 2023; Yuen et al., 2024). DINO's ability to learn meaningful representations without labeled data makes it an engaging option for improving zero-shot learning models. Similarly, the weakly supervised approach (Khan et al., 2024a) has also been used for many different applications in medical and agriculture (Bellocchio et al., 2022; Khan et al., 2024b). Despite these advancements, there has been little research focused on applying these techniques to agricultural tasks. This paper aims to bridge this gap by presenting a novel approach that combines CLIP and DINO models for zero-shot classification in the agriculture domain.

3 Method

As discussed earlier, popular vision-language foundational models such as CLIP and its variant have demonstrated impressive zero-shot. However, their applicability diminishes when applied to more specialized domains due to the inherent domain gap (Udandarao et al., 2024). This is likely due to being pre-trained on general-purpose images depicting everyday scenes and objects that lack the fine-grained, domain-specific examples needed for agricultural and livestock tasks, such as identifying subtle distinctions in plant diseases or detecting small variations among different crop species. We introduce AgriCLIP (see Fig. 1), a framework designed to bridge the domain gap in agriculture and livestock tasks. To train AgriCLIP, we construct a large-scale image-text dataset named ALive for agriculture and livestock. AgriCLIP adapts text and vision encoders to learn discriminative, domain-specialized features followed by cross-modal alignment to obtain improved feature representation.

3.1 ALive Dataset for Agriculture and Livestock

Most existing agriculture and livestock datasets are image-only with class-level information. Here, we employ a two-step approach: first, we collect

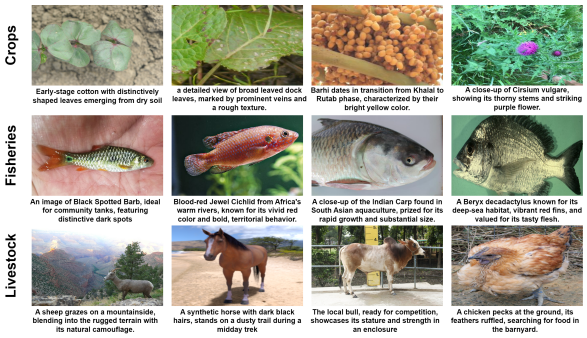

Figure 2: Example images from the ALive dataset, including various crops (such as dates, crop diseases, and plant genera), diverse fish species, and samples from the livestock domain. More examples are in the appendix.

diverse agricultural and livestock data, then synthesize relevant text using dataset and class-level information via customized prompt generation for vision-language contrastive learning.

We gather 25 training datasets across crops, fish, and livestock, creating the Agriculture and Livestock (ALive) dataset with 600k images covering a wide range of conditions. This includes various crop growth stages, classifications, and different farming environments for animals and fish. Next, we design a customized prompt generation strategy where the text based on the dataset and class-level information is leveraged to provide context and fine-grained details for each image. Initially, we crafted simple prompts as are used in the CLIP prompting with the prefix "a photo of [cls]." Then, we passed the image and its metadata to the GPT-4 (Achiam et al., 2023) model and generated more contextually aware prompts. For instance, instead of using a generic CLIP prompt like "a photo of a boron-deficient leaf," we craft prompts like "a photo of a leaf with boron deficiency characterized by yellow patches and curled edges." We then use GPT-4 (Achiam et al., 2023) to generate a diverse variation of these prompts. Table 5 in the Appendix and Figure 2 present examples and details of our ALive dataset. Next, we describe our AgriCLIP pipeline that leverages the ALive dataset.

3.2 Learning Semantic ALive Features

We learn global semantic domain-specialized features via the ALive dataset by utilizing image-text contrastive training. To this end, we adapt CLIP (Radford et al., 2021) by further pre-training it on the ALive dataset using contrastive loss (He et al., 2020). The model consists of a vision encoder E_{img} and a text encoder E_{txt}, and we align their embedding spaces by minimizing the distance between

Dataset	Downstream Tasks	CLIP	CLIP Pre-Training	AgriCLIP
	Fish			
Supermarket Fish (Ulucan et al., 2020)	Local Fish Classification	1.66	22.57	41.38
Aquarium Fish (Moorthy)	Acquarium Fish Classification	22.85	27.13	43.06
FishDataset (Kaur)	Fine-grained Fish Classification	16.04	32.47	49.23
DeepFish (Saleh et al., 2020)	Under-Sea Fish Classification	45.41	55.96	57.78
FishNet (Khan et al., 2023b)	Functional Trait Prediction	0.15	19.52	21.58
Fish Freshness (Rayan)	Fish Freshness classification	29.37	50.05	57.75
Fish Species (Daniel)	Fish Species Classification	13.68	25.89	30.21
	Crops			
Banana Deficiency (Sunitha, 2022)	Nutrients Deficiency Classification	14.08	20.64	23.55
Citrus Fruits (Sharif et al., 2018)	Citrus Fruit Disease Classification	38.09	39.55	40.21
Citrus Leaves (Sharif et al., 2018)	Citrus Leaves Disease Classification	2.18	23.97	34.27
Fruits Diseases (Kour and Arora, 2019)	Native Fruits Classification	68.9	73.26	73.98
PlantDoc (Singh et al., 2020)	Plant Disease Classification	6.02	29.18	35.42
Wheat Rust (Hayit et al., 2021)	Wheat Rust Classification	34.11	53.45	67.38
Bean Lesion (Marquis)	Bean Lesion Classification	18.73	35.47	40.85
	LiveStock			
Chicken Fecus (allandclive)	Chicken Disease Classification	19.28	27.31	34.29
CID (Shagor et al., 2022)	Local Cow Specie Classification	5.62	27.52	49.95
Cow Breed (Hossain)	Cow Breed Classification	31.13	40.23	44.62
Animals-2 (Animals-10)	Livestock Animal Classification	95.48	97.12	98.27
Horses Breed (Belitskaya)	Horses Breed Classification	28.05	48.63	54.50
Average		25.83	39.20	**48.27**

Table 1: Zero-shot classification performance comparison of standard CLIP, further pre-training standard CLIP on ALive dataset and the proposed AgriCLIP in a variety of downstream tasks corresponding to three domains: agriculture, livestock and fishery.

correct image-text pairs (positive pairs) and maximizing the distance for incorrect pairs (negative pairs). The contrastive loss is defined as:

$$L = -\log \frac{\exp(\text{sim}(u, v)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(u, v_k)/\tau)},$$

where $u = E_{\text{img}}(x)$ and $v = E_{\text{txt}}(y)$ are the image and text embeddings, sim represents cosine similarity, τ is the temperature parameter, and N is the number of samples. This adaptation ensures better alignment of vision and text representations for the agriculture and livestock domain.

3.3 Learning Fine-grained ALive Features

In addition to the aforementioned domain-specialized semantic features, different agriculture and livestock problems desire capturing fine-grained visual details, crucial for tasks like identifying subtle variations in disease symptoms (e.g., small color differences in spots), for classification.

To further learn domain-specialized fine-grained features, we complement CLIP's generalization capabilities by employing a DINO-based pre-training strategy to enhance the vision encoder E^S_{img} using the ALive dataset. The self-supervised learning-based technique (Caron et al., 2021) excels at learning detailed visual features. Here, a student-teacher framework is employed, where two randomly augmented views of each image are processed by both models (vision transformers). The student model is trained to match the teacher's representations, enabling it to capture both global and fine-grained details. This combined self-supervised approach enhances the model's ability to handle domain-specialized fine-grained visual features.

3.4 Cross-Model Alignment

The visual encoder E^S_{img}, although a powerful feature extractor, is not inherently aligned with the language encoder and therefore lacks the zero-shot capabilities of vision-language models. To align domain-specialized semantic, fine-grained, and text features, we adopt a vision-language feature alignment approach inspired by (Moayeri et al., 2023), aligning visual features E^S_{img} with CLIP's textual encoder E_{txt} implicitly. Specifically, we apply a learnable affine transformation to map the output of the fine-grained visual encoder E^S_{img} to the space of the semantic visual encoder E_{img}, effectively projecting E^S_{img} to the same space as concept vector for text thereby enabling zero-shot capabilities (Moayeri et al., 2023). This mapping is learned by minimizing the mean squared error (MSE) between the feature space of the two models.

4 Experimental Results

Experimental Setup. For semantic visual encoder and language encoder (stage 1), we use

CLIP's (Radford et al., 2021) open-source implementation called OpenCLIP (Ilharco et al., 2021). For fine-grained feature vision encoder (stage 2), we utilize DINO (Caron et al., 2021). It is trained with global and local crop scales of (0.4, 1) and (0.05, 0.4), respectively. The AdamW optimizer is used with a learning rate of 0.0005 and weight decay of 0.04 for a total of 100 epochs. The rest of the model settings are used as default from the DINO model. For zero-shot evaluation, we follow the original framework utilized by CLIP (Radford et al., 2021). We run all our experiments on a single NVIDIA A100 GPU.

Downstream Tasks and Datasets. To evaluate the performance of AgriCLIP, we assemble a set of 20 datasets to test the model's ability to generalize to unseen concepts. The evaluation set is entirely disjoint from the ALive pre-training set. It contains diverse agricultural and livestock categories, including new types of crops, different disease and nutrient deficiencies, and varied environmental conditions. The downstream datasets include tasks such as crop disease identification, nutrient deficiency classification, animal species or breeds classification, and fish species recognition, providing a comprehensive assessment of the model's zero-shot classification capabilities in practical applications.

Results. We compare the performance of our AgriCLIP with both the original CLIP and its adaptation through further pre-training on the ALive dataset in Table 1 on 20 downstream datasets. To demonstrate the effectiveness of our ALive dataset and training pipeline, we compare three model configurations: the original CLIP model, CLIP further pre-trained on our ALive dataset, and AgriCLIP trained with our proposed training pipeline. The original CLIP model exhibits poor performance across the 20 agriculture-related tasks, with accuracy ranging from 1.66% to 45.41% for fisheries datasets and an overall average zero-shot accuracy of 25.83%. Further pre-training CLIP on the ALive dataset enhances performance, yielding accuracy between 22.57% and 55.96%, with an overall average accuracy of 39.20%. This gain demonstrates the impact of our ALive dataset in bridging the domain gap. Our AgriCLIP, incorporating a fine-grained feature vision encoder, further improves performance on downstream tasks. AgriCLIP achieves an overall classification score of 48.27% over 20 datasets with an absolute gain of 9.07% over its adapted CLIP baseline counterpart. We present more details on the experimental results and ablation studies on the impact of custom prompts, dataset size, and different pre-training for fine-grained encoders in the appendix section.

5 Conclusion

We present a vision-language foundational model, AgriCLIP, for agriculture and livestock domain. To facilitate model pre-training, we introduce a large-scale dataset with 600,000 image-text pairs for agriculture and livestock domain. AgriCLIP learns both semantic and fine-grained domain-specialized features for improved zero-shot classification. Experiments on 20 downstream datasets show the efficacy of AgriCLIP.

Limitations. The current study focuses on a diverse set of classification tasks. Its applicability to other critical downstream dense prediction tasks, such as pest detection, crop yield prediction, and plant disease segmentation, has not been tested. A potential future direction is to expand the model's evaluation to include these tasks, as it would provide a more comprehensive understanding of its practical utility.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

allandclive. Chicken disease image classification. https://www.kaggle.com/datasets/allandclive/chicken-disease-1?select=train_data.csv. (Accessed on 09/12/2024).

Animals-10. Animals-10. https://www.kaggle.com/datasets/alessiocorrado99/animals10. (Accessed on 09/12/2024).

Muhammad Arbab Arshad, Talukder Zaki Jubery, Tirtho Roy, Rim Nassiri, Asheesh K Singh, Arti Singh, Chinmay Hegde, Baskar Ganapathysubramanian, Aditya Balu, Adarsh Krishnamurthy, et al. 2024. Ageval: A benchmark for zero-shot and few-shot plant stress phenotyping with multimodal llms. *arXiv preprint arXiv:2407.19617*.

R Arumuga Arun and S Umamaheswari. 2023. Effective multi-crop disease detection using pruned complete concatenated deep learning model. *Expert Systems with Applications*, 213:118905.

Olga Belitskaya. Horse breeds. https://www.kaggle.com/datasets/olgabelitskaya/horse-breeds. (Accessed on 09/12/2024).

Enrico Bellocchio, Francesco Crocetti, Gabriele Costante, Mario Luca Fravolini, and Paolo Valigi. 2022. A novel vision-based weakly supervised framework for autonomous yield estimation in agricultural applications. *Engineering Applications of Artificial Intelligence*, 109:104615.

Pallab Bharman, S Ahmad Saad, Sajib Khan, Israt Jahan, Milon Ray, and Milon Biswas. 2022. Deep learning in agriculture: a review. *Asian Journal of Research in Computer Science*, 13(2):28–47.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Daniel. fish_classification_dataset. https://www.kaggle.com/datasets/daniel1312412/fish-classification-dataset?select=AlevinosRV. (Accessed on 09/12/2024).

Guy Farjon, Liu Huijun, and Yael Edan. 2023. Deep-learning-based counting methods, datasets, and applications in agriculture: A review. *Precision Agriculture*, 24(5):1683–1711.

Tolga Hayit, Hasan Erbay, Fatih Varçın, Fatma Hayit, and Nilüfer Akci. 2021. Determination of the severity level of yellow rust disease in wheat by using convolutional neural networks. *Journal of plant pathology*, 103(3):923–934.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Shahadat Hossain. Cattle breed data. https://www.kaggle.com/datasets/iamshahadat/cattle. (Accessed on 09/12/2024).

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Jasmeet Kaur. Fishdataset. https://www.kaggle.com/datasets/jasmeetkaur/fishdataset. (Accessed on 09/12/2024).

Asim Khan, Umair Nawaz, Lochan Kshetrimayum, Lakmal Seneviratne, and Irfan Hussain. 2023a. Early and accurate detection of tomato leaf diseases using tomformer. In *2023 21st International Conference on Advanced Robotics (ICAR)*, pages 645–651. IEEE.

Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. 2023b. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20496–20506.

Ufaq Khan, Umair Nawaz, Mustaqeem Khan, Abdulmotaleb El Saddik, and Wail Gueaieb. 2024a. Fetr: A weakly self-supervised approach for fetal ultrasound anatomical detection. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE.

Ufaq Khan, Umair Nawaz, and Abdulmotaleb E Saddik. 2024b. Ultraweak: Enhancing breast ultrasound cancer detection with deformable detr and weak supervision. In *MICCAI Workshop on Cancer Prevention through Early Detection*, pages 144–153. Springer.

Ufaq Khan and Ashish Oberoi. 2019. Plant disease detection techniques: A review. *International Journal of Computer Science and Mobile Computing*, 8(4):59–68.

Jameer Kotwal, Ramgopal Kashyap, and Shafi Pathan. 2023. Agricultural plant diseases identification: From traditional approach to deep learning. *Materials Today: Proceedings*, 80:344–356.

Vippon Preet Kour and Sakshi Arora. 2019. Plantaek: A leaf database of native plants of jammu and kashmir.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. 2023a. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050.

Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. 2023b. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Marquis. Bean leaf lesions classification. https://www.kaggle.com/datasets/marquis03/bean-leaf-lesions-classification. (Accessed on 09/12/2024).

S Divya Meena, Munagala Susank, Tarini Guttula, Srikurmum Hari Chandana, and J Sheela. 2023. Crop yield improvement with weeds, pest and disease detection. *Procedia Computer Science*, 218:2369–2382.

Siddharth Mishra-Sharma, Yiding Song, and Jesse Thaler. 2024. Paperclip: Associating astronomical observations and natural language with multi-modal models. *arXiv preprint arXiv:2403.08851*.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. 2023. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, pages 25037–25060. PMLR.

Moorthy. Aquariumfishes. https://www.kaggle.com/datasets/cmkmoorthy/aquariumfished. (Accessed on 09/12/2024).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Muhammad Abu Rayan. Fish freshness classification. https://www.kaggle.com/datasets/muhammadaburayan/fish-freshness-classification. (Accessed on 09/12/2024).

Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. 2020. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671.

Mobasshir Bhuiya Shagor, Md Zahidul Haque Alvi, and Khandaker Tabin Hasan. 2022. Cid: Cow images dataset for regression and classification. In *Proceedings of the 2nd International Conference on Computing Advancements*, pages 450–455.

Muhammad Sharif, Muhammad Attique Khan, Zahid Iqbal, Muhammad Faisal Azam, M Ikram Ullah Lali, and Muhammad Younus Javed. 2018. Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection. *Computers and electronics in agriculture*, 150:220–234.

Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. 2020. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, page 249–253, New York, NY, USA. Association for Computing Machinery.

Jianqiang Sun, Wei Cao, Xi Fu, Sunao Ochi, and Takehiko Yamanaka. 2024. Few-shot learning for plant disease recognition: A review. *Agronomy Journal*, 116(3):1204–1216.

P. Sunitha. 2022. Images of nutrient deficient banana plant leaves.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip HS Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*.

Oguzhan Ulucan, Diclehan Karakaya, and Mehmet Turkan. 2020. A large-scale dataset for fish segmentation and classification. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2024. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Bingfang Wu, Miao Zhang, Hongwei Zeng, Fuyou Tian, Andries B Potgieter, Xingli Qin, Nana Yan, Sheng Chang, Yan Zhao, Qinghan Dong, et al. 2023. Challenges and opportunities in remote sensing-based crop monitoring: A review. *National Science Review*, 10(4):nwac290.

Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, and Jieping Ye. 2024. Addressclip: Empowering vision-language models for city-wide image address localization. *arXiv preprint arXiv:2407.08156*.

Aamir Yousuf and Ufaq Khan. 2021. Ensemble classifier for plant disease detection. *International Journal of Computer Science and Mobile Computing*, 10(1):14–22.

Kashing Yuen, Jianpeng Zou, and Kaoru Uchida. 2024. Generalized dino: Dino via multimodal models for generalized object detection. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 776–783.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, et al. 2023. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*.

Fangming Zhong, Zhikui Chen, Yuchun Zhang, and Feng Xia. 2020. Zero-and few-shot learning for diseases recognition of citrus aurantium l. using conditional adversarial autoencoders. *Computers and electronics in agriculture*, 179:105828.

Yueyue Zhou, Hongping Yan, Kun Ding, Tingting Cai, and Yan Zhang. 2024. Few-shot image classification of crop diseases based on vision-language models.

A Extended Results

A particularly noteworthy aspect of AgriCLIP is its ability to significantly enhance performance on datasets where CLIP exhibits notably low accuracy. For instance, on the Supermarket Fish dataset (Ulucan et al., 2020), AgriCLIP achieves an accuracy of 41.38%, a substantial improvement over CLIP's 1.66%. Similarly, for the PlantDoc dataset (Singh et al., 2020), a large and fine-grained plant leaf disease dataset, AgriCLIP demonstrates an accuracy of 35.42%, markedly outperforming CLIP's 6.02%. These results highlight AgriCLIP's ability to handle diverse and challenging agricultural datasets, particularly those requiring fine-grained learning and domain-specific knowledge.

A.1 Ablations Studies

Effect of Different Pre-training for Fine-grained Feature Vision Encoders. For the fine-grained feature vision encoder, we experimented with two different self-supervised training frameworks: DINO (Caron et al., 2021) and Masked Auto Encoder (MAE)(He et al., 2022). We pre-trained both models on the ALive dataset and subsequently aligned it with CLIP, following our method. The average accuracy of 22.11% and 42.04% was obtained for 10 downstream tasks as shown in Table 2 of Appendix. Unlike DINO, the MAE model does not perform well on downstream tasks, demonstrating the suitability of DINO for handling the complexities of the ALive dataset compared with MAE.

Effect of Custom Prompts. To show the effectiveness of our prompt design method, which incorporates metadata and class-specific information to synthesize text for the dataset, we perform an ablation study where we compare CLIP's fine-tuning with our prompts and generic, CLIP-style prompts. The results are shown in Table 4.

Impact of Increasing Size of the Dataset. We perform an ablation by increasing the dataset size to understand the effect of adding more data. To this end, the ALive dataset is expanded (ALive++) to nearly 900,000 images, encompassing a broader range of agricultural and livestock scenarios by incorporating segmentation, tracking, and detection datasets. This enhanced dataset is used solely to pre-train the fine-grained feature vision encoder (stage 2) in a self-supervised manner using DINO (Caron et al., 2021). After pre-training, its features are aligned with CLIP, following stage 3 of our method. The results are shown in Table 3, demonstrating the benefits of a larger and more varied dataset.

Dataset	MAE	DINO
Supermarket Fish (Ulucan et al., 2020)	6.20	41.38
Aquarium Fish (Moorthy)	33.33	43.06
FishDataset (Kaur)	16.21	49.23
DeepFish (Saleh et al., 2020)	45.41	57.78
FishNet (Khan et al., 2023b)	0.11	21.58
Banana Deficiency (Sunitha, 2022)	12.04	23.55
Citrus Fruits (Sharif et al., 2018)	26.67	40.21
Citrus Leaves (Sharif et al., 2018)	8.22	34.27
Fruits Diseases (Kour and Arora, 2019)	64.55	73.98
PlantDoc (Singh et al., 2020)	8.45	35.42
Average	22.11	**42.04**

Table 2: Comparison between Masked Autoencoder (MAE) and DINO-based pre-training for fine-grained feature learning. DINO-based training strategy outperforms MAE significantly.

Dataset	ALive	ALive++
Supermarket Fish (Ulucan et al., 2020)	41.38	**41.87**
Aquarium Fish (Moorthy)	43.06	**45.28**
FishDataset (Kaur)	59.23	**62.77**
DeepFish (Saleh et al., 2020)	**57.78**	53.43
FishNet (Khan et al., 2023b)	21.58	**23.57**
Banana Deficiency (Sunitha, 2022)	23.55	**25.34**
Citrus Fruits (Sharif et al., 2018)	40.21	**45.88**
Citrus Leaves (Sharif et al., 2018)	**34.27**	33.74
Fruits Diseases (Kour and Arora, 2019)	73.98	**78.58**
PlantDoc (Singh et al., 2020)	35.42	**39.62**
Average	43.64	45.00

Table 3: Impact of increasing size of ALive dataset.

B Extended Dataset Details

Extended details for the pre-training and downstream datasets are shown in Table 5, including tasks, different sensors used for the collection of data, data variability, number of datasets, and the total number of images. In Table 6, we demonstrate more examples of the customized prompts we constructed for ALive. Also, Figure 3 shows more examples of the ALive dataset.

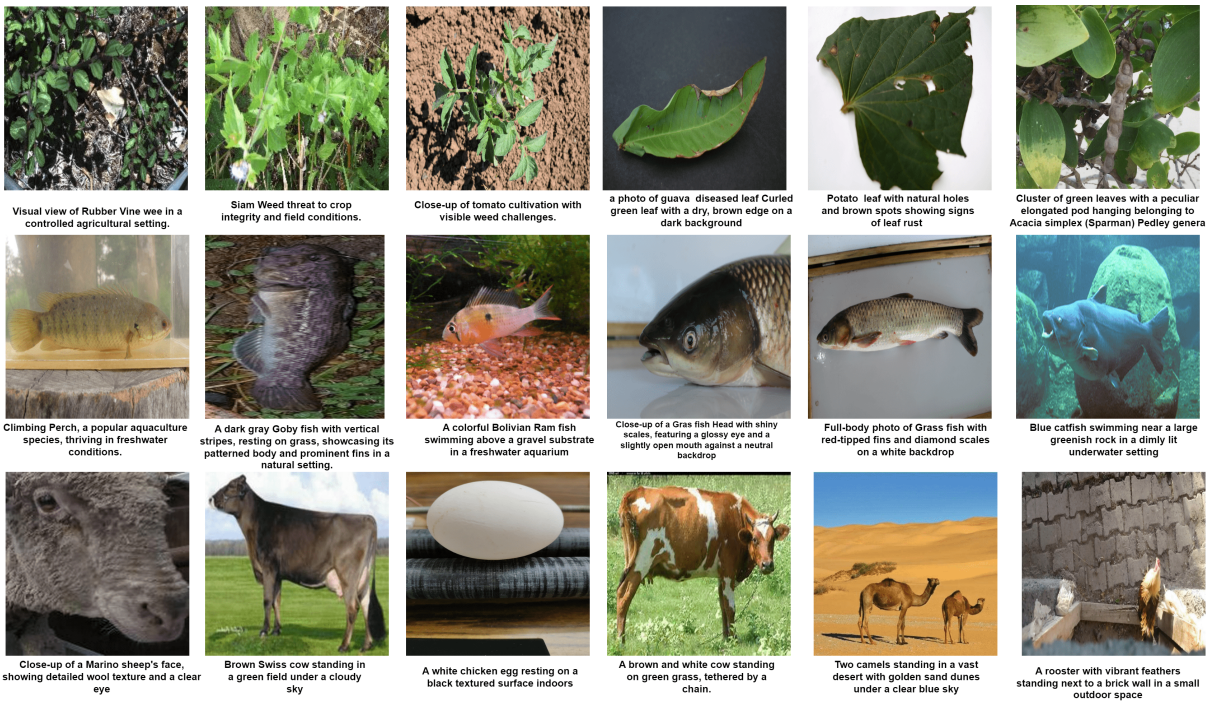

Figure 3: Some more examples of the ALive dataset

Dataset	Normal	Custom
Supermarket Fish (Ulucan et al., 2020)	22.57	23.45
Fish Freshness (Rayan)	50.05	51.25
PlantDoc (Singh et al., 2020)	29.18	31.72
Wheat Rust (Hayit et al., 2021)	53.45	54.82
Bean Lesion (Marquis)	35.47	39.58
Chicken Fecus (allandclive)	27.31	29.54
CID (Shagor et al., 2022)	27.52	30.28
Average	35.07	**37.23**

Table 4: Comparison of generic prompt used by CLIP (Radford et al., 2021) vs. customized prompts utilized in ALive dataset.

Dataset	Tasks	Sensor	Data Variability	# Images	# Datasets
Pre-Training	Nutrients Deficiency, Weed Classification, Plant Disease Classification, Fruits Classification, Plant Specie Classification, Fish Specie Classification, Fish Abundance Classification, Cattle Classification, Cattle Breed Recognition, Horse Classification	Ground Handheld Imagery, Ground vehicle with camera attached, Ground Imagery with Fixed Camera, Ground Imagery (Using Robot), Sea Imagery	Indoor, Outdoor, Underwater, Plain Background	603,626	25
Downstream	Nutrients Deficiency, Weeds Classification, Plant Specie Classification, Plant Disease Classification, Cows Breed Classification, Fish Specie Classification, Functional Trait Prediction, Chicken Classification	Ground Handheld Imagery, Sea Imagery	Indoor, Outdoor, Underwater	301,076	20

Table 5: Overview of different attributes for datasets used in ALive pre-training and downstream evaluation.

Prompt	Customized Prompts using GPT4
a photo with prickly acacia weed specie	Prickly Acacia invasion in agricultural regions, an urgent weed control case Impact assessment of Prickly Acacia on crop health and soil quality Mapping Prickly Acacia spread in critical farming areas Prickly Acacia detection in crop fields, potential for significant yield loss
a photo of rice plant leaf with nitrogen deficiency	A leaf displaying the yellowing characteristics of nitrogen deficiency Early signs of nitrogen deficiency captured in a leaf A close-up of a leaf suffering from lack of nitrogen A photo of rice leaf having yellowish patterns due to nitrogen deficiency
a photo of early stage black nightsade leaves in the field	Black Nightshade presence in crop fields, known for its competitive nature Monitoring aggressive Black Nightshade weed among vegetable crops Impact of Black Nightshade on crop fields, with a focus on containment strategies Field analysis of Black Nightshade weed's effect on adjacent crops
a photo of fish from Freshwater Eel specie	Freshwater Eel, a species known for its elongated, snake-like body Capturing the elusive Freshwater Eel during its nocturnal activity The mysterious life of Freshwater Eels, seen here in a creek Freshwater Eel in a clear stream, showcasing its sleek body
a photo of fish from Big Head Carp specie	Big Head Carp, a large species with a distinctive large head Observing the feeding habits of Big Head Carp in a river setting Big Head Carp, often found in river basins, impacting local ecosystems A snapshot of Big Head Carp, focused on its unique head structure
a photo of a cow from Jersey breed	A Jersey cattle, renowned for its rich, creamy milk The small yet robust Jersey cow, ideal for boutique dairy products Jersey cattle in a dairy setting, noted for high butterfat content in its milk A serene Jersey cow, a favorite among small-scale dairy farmers
a photo of a sheep	A sheep grazing peacefully, a staple of pastoral agriculture Detailed capture of a sheep's wool, essential for textile production A flock of sheep on a sunny day, a vital resource for farmers Sheep in a meadow, representing sustainable agricultural practices

Table 6: Customized prompts for image descriptions used in ALive dataset.