Smart Media Journal / Vol.13, No.7 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2024.13.7.36

Novel Category Discovery in Plant Species and Disease Identification through Knowledge Distillation

Jiuqing Dong[1,2], Alvaro Fuentes[1,2], Mun Haeng Lee[3], Taehyun Kim[4], Sook Yoon[5], Dong Sun Park[1,2]

Abstract

Identifying plant species and diseases is crucial for maintaining biodiversity and achieving optimal crop yields, making it a topic of significant practical importance. Recent studies have extended plant disease recognition from traditional closed-set scenarios to open-set environments, where the goal is to reject samples that do not belong to known categories. However, in open-world tasks, it is essential not only to define unknown samples as "unknown" but also to classify them further. This task assumes that images and labels of known categories are available and that samples of unknown categories can be accessed. The model classifies unknown samples by learning the prior knowledge of known categories. To the best of our knowledge, there is no existing research on this topic in plant-related recognition tasks.

To address this gap, this paper utilizes knowledge distillation to model the category space relationships between known and unknown categories. Specifically, we identify similarities between different species or diseases. By leveraging a fine-tuned model on known categories, we generate pseudo-labels for unknown categories. Additionally, we enhance the baseline method's performance by using a larger pre-trained model, dino-v2. We evaluate the effectiveness of our method on the large plant specimen dataset Herbarium19 and the disease dataset Plant Village. Notably, our method outperforms the baseline by 1% to 20% in terms of accuracy for novel category classification. We believe this study will contribute to the community.

Keywords: novel class discovery | knowledge distillation | plant disease classification | deep-learning

I. INTRODUCTION

Identifying plant species and diseases is crucial for maintaining biodiversity and achieving expected crop yields. Over the past decade, deep learning has shown tremendous success in plant-related areas [1-7]. However, many models heavily rely on the availability of large amounts of labeled data for all relevant categories. This dependence introduces a significant issue: standard classification models may erroneously classify instances that do not belong to any known category as belonging to one of the known categories [8,9]. This phenomenon is particularly common in neural networks when dealing with semantically related inputs.

Recent advancements in open-set recognition and open-world scenarios have addressed this issue [3,9]. Open-set

[1] Department of Electronics Engineering, Jeonbuk National University.
[2] Core Institute of Intelligent Robots, Jeonbuk National University.
[3] Department of Smart Farm, Chungnam State University, South Korea
[4] National Institute of Agricultural Sciences, South Korea
[5] Department of Computer Engineering, Mokpo National University.

* This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) and Korea Smart Farm R&D Foundation(KosFarm) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) and Ministry of Science and ICT(MSIT), Rural Development Administration(RDA)(421005-04)

Manuscript : 2024.06.21
Revised : 2024.07.09
Confirmation of Publication: 2024.07.11
Corresponding Author: Sook Yoon, Dong Sun Park,
e-mail : syoon@mokpo.ac.kr, dspark@jbnu.ac.kr

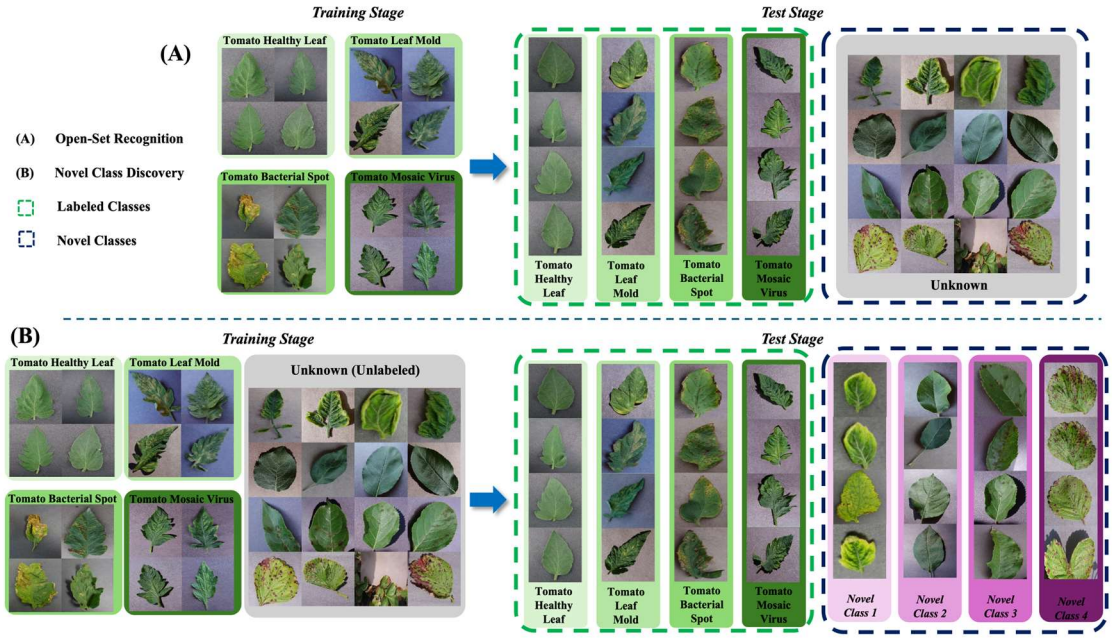

Fig. 1. Comparison of Open-Set Recognition (A) and Novel Class Discovery (B): Training and testing stages involved in each approach. (A) Open-Set Recognition: The training stage involves labeled classes, and the testing stage requires identifying known classes and rejecting unknown classes. (B) Novel Class Discovery: The training stage includes labeled classes and unknown samples, with the testing stage focusing on classifying both known classes and novel classes.

recognition aims to exclude unknown samples through regularization loss or post-hoc out-of-distribution (OOD) detection. For instance, Meng et al.[10] used additive margin softmax loss to make the feature distribution of known categories more compact, facilitating the identification of unknown samples through regularization methods. However, these open-world methods typically do not cluster unknown categories, leaving the unlabeled data underutilized and potentially wasting data resources. In another study, Dong et al. [3] discovered that a model trained on tomato powdery mildew disease could also detect powdery mildew disease on pepper leaves. This experiment inspired our exploration of the relationships between plant diseases or species categories. In other words, effectively utilizing category relationships may assist in classifying unknown samples. This challenge is defined as novel category discovery.

Both open-set recognition and novel category discovery belong to open-world tasks. We illustrate the differences between open-set recognition and novel category discovery in Fig. 1. As shown in Fig. 1, in open-set recognition, the model is typically trained only on labeled categories. During the testing phase, the model is required to classify known categories and identify unknown samples [11]. While in the novel category discovery task, we assume that samples from these unknown categories have already been obtained. The model trains on both labeled samples and these unknown samples, with the understanding that unknown categories do not overlap with known categories. The model then aims to further classify the unknown samples.

We argue that extracting category relationships through knowledge

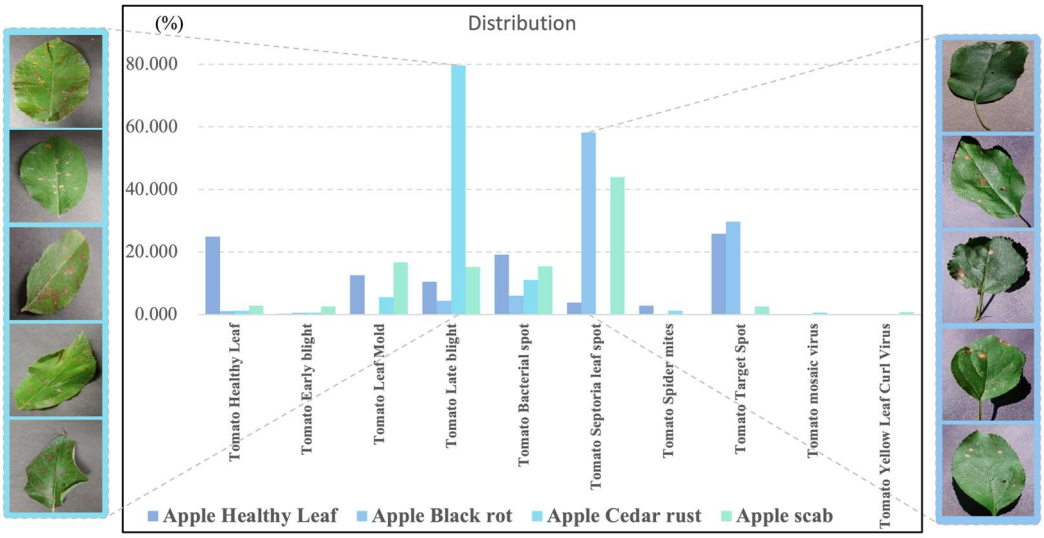

Fig. 2. Class relation distribution of apple leaf disease (test on a model fine-tuned on tomato leaf diseases dataset).

distillation [12] is highly applicable to plant-related research. For example, different plant diseases may exhibit similar symptoms. Therefore, we employ a knowledge distillation framework to achieve novel category discovery related to plants. Additionally, we observed that methods based on large-scale pre-trained models achieve remarkable performance gains in various downstream tasks. Consequently, we utilize state-of-the-art models Dino-v1 [13] and Dino-v2 [14] to initialize the feature extraction network and enhance the system's performance.

II. RELATED WORK

2.1 Novel Class Discovery

Novel Category Discovery (NCD) aims to identify new categories within unlabeled data by leveraging prior knowledge of known categories [8]. The core idea behind NCD is that having a set of known categories allows an appropriate method to enhance its performance by extracting the general concepts that constitute well-defined categories. Traditional NCD problems do not consider the classification of both known and unknown categories, an issue addressed by open-set recognition tasks. In other words, traditional NCD problems assume that only novel categories are classified during testing [8]. UNO [15] proposes a more practical evaluation framework, assuming that the model will encounter both labeled and novel categories during testing. Therefore, the model needs to classify both labeled and novel categories simultaneously. This approach is referred to as an evaluation scheme independent of the NCD task. Moreover, UNO argues that certain three-stage methods—self-training on labeled and unlabeled data, fine-tuning on labeled data, and discovering new classes on unlabeled data—do not yield performance gains in the final novel category discovery. Consequently, they skip the first stage and directly fine-tune the pre-trained model on the labeled dataset to achieve novel category discovery. Our approach follows this two-stage training strategy.

2.2 Knowledge distillation

Knowledge distillation [16] aims to transfer knowledge from a teacher model

to a student model. This process uses the probability distribution from the teacher model to supervise unlabeled samples, a technique commonly employed in semi-supervised and weakly supervised tasks. In the context of plant disease and species identification, we found that features or symptoms of certain diseases may appear similar on different plant leaves. Therefore, we use knowledge distillation to extract relationships between known and unknown categories, thereby improving the representation learning of novel categories.

III. MATERIAL AND METHODS

3.1 Data Split

Table 1 demonstrates the dataset splits for Herbarium19 and Plant Village. Herbarium19 [17] is a plant specimen dataset, while Plant Village [18] is a plant disease dataset. In the Plant Village dataset, subsets A, C, G, P, H, and D correspond to Apple, Corn, Grape, Potato, Healthy, and Other Diseases, respectively. The 10 classes of tomato diseases in Plant Village serve as labeled categories, with the other subsets treated as novel categories.

Table 1. Dataset splits. Note that the Labeled set of Plant Village is a Tomato subset, which includes ten classes.

Dataset	Labeled Set		Novel Set	
	Images	Classes	Images	Classes
Herbarium19	18348	341	18556	342
Plant Village(A)	18,159	10	3,171	4
Plant Village(C)	18,159	10	3,852	4
Plant Village(G)	18,159	10	4,062	4
Plant Village(P)	18,159	10	2,152	3
Plant Village(H)	18,159	10	10,110	7
Plant Village(D)	18,159	10	12,797	6

The labeled set of Plant Village consistently includes 18,159 images across 10 classes, while the novel set contains varying numbers of images and classes, representing different plant diseases or healthy conditions. In contrast, Herbarium19 has 18,348 images in 341 classes for the labeled set and 18,556 images in 342 classes for the novel set. This setup facilitates distinguishing between known and unknown categories, supporting tasks such as novel category discovery and open-set recognition.

3.2 Methods

The premise of our method is that there exists semantic relatedness between unknown and known classes. Fig. 2 supports the validity of this assumption. Initially, we fine-tune a pre-trained model using labeled data and then test the model with unlabeled data. The results indicate that approximately 60% of 'apple black rot' leaves disease are classified as 'tomato septoria leaf spot' leaves, and over 80% of 'apple cedar rust' leaves are classified as 'tomato late blight' leaves. This suggests a degree of correlation in the symptoms of leaf diseases across different species.

Based on this similarity assumption, we employ a knowledge distillation framework for novel category discovery. As illustrated in Fig. 3, our approach involves two main stages: **i.** Fine-tuning the Pre-Trained Encoder: In the first stage, we fine-tune a pre-trained encoder using labeled data. This step encourages the model to learn category-specific knowledge relevant to the domain; **ii.** Novel Category Discovery: In this stage, we use the fine-tuned model from the

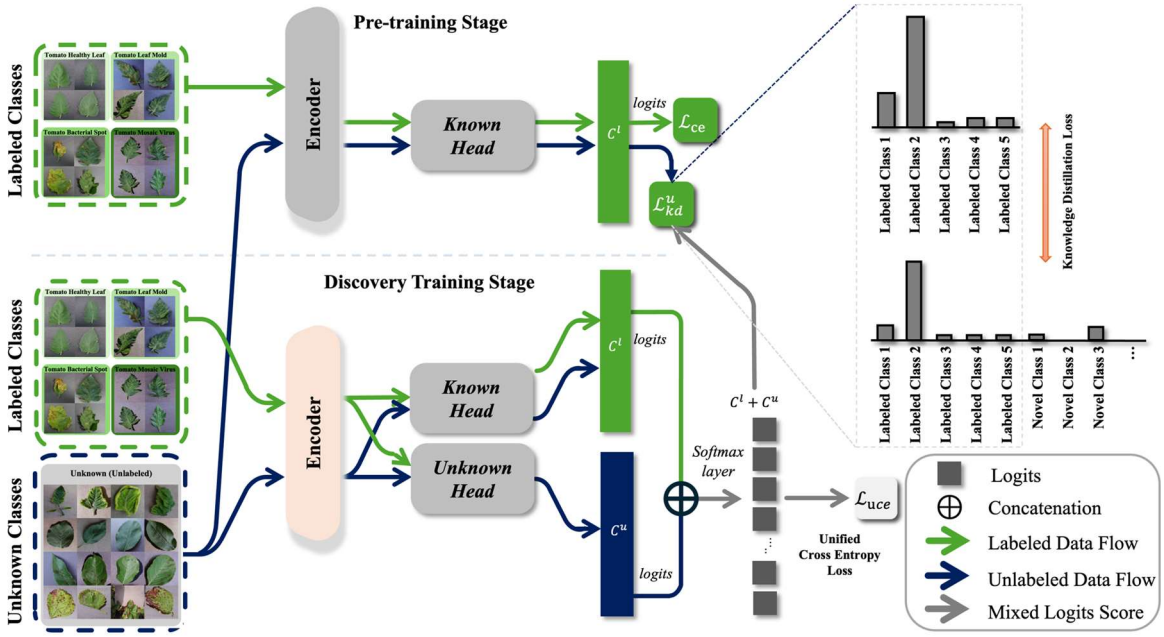

Fig. 3. Framework of our method. In the pre-training stage, labeled classes are encoded and processed through a known head to generate logits and calculate the cross-entropy loss. In the discovery training stage, both labeled and unknown (unlabeled) classes are encoded, producing logits through both known and unknown heads. These logits are combined and processed through a unified cross-entropy loss, with knowledge distillation loss applied to improve representation learning of novel classes.

previous stage as the teacher model to guide the clustering of new categories. For unlabeled data, we use knowledge distillation to establish category relationships between unknown and known samples, denoted as L_{kd}^{u} in Fig. 3. Additionally, we adopt the self-training method proposed in UNO [15] to generate pseudo-labels. This approach achieves weakly supervised learning by constructing a unified loss function. For further details on the self-training method and the implementation of the unified loss function, we refer readers to UNO [15].

This two-stage strategy leverages the relationships between known and unknown categories, enabling effective novel category discovery.

3.3 Protocol and Metrics

We use two evaluation settings to assess our model: task-aware and task-agnostic. As shown in the discovery phase in Fig. 3, we obtain two classification heads after training. In task-aware evaluation, we use the labeled classifier for images belonging to labeled classes and the unlabeled classifier for images belonging to unlabeled classes. This type of evaluation is typically used in traditional NCD tasks. However, in real-world scenarios, this evaluation is less meaningful because it does not assess whether the model can distinguish between labeled and unlabeled classes. Therefore, we also report task-agnostic accuracy. In this evaluation, we concatenate the logits output from both classification heads and predict the most probable output, eliminating the need to differentiate between labeled and novel categories in the test set beforehand. For labeled categories, we use accuracy to evaluate performance. For unlabeled categories, we employ the Hungarian algorithm [19] to find the optimal permutation that matches the true labels of the unknown classes with the predicted

labels. We then calculate the clustering accuracy based on this optimal matching.

IV. EXPERIMENTS

4.1 Implementation details

We provide implementation details in Table 2. Note that PS and DS denote pre-training stage and discovery stage, respectively.

Table 2. Environment and Hyper-Parameters.

Item / Parameter	Value / Details
Framework	Pytorch 3.8
Encoder	ViT-base
Batch size	256
Decay	Cosine annealing
Pre-training Stage	50 epochs
Learning Rate (PS)	0.05-0.001
Discovery Stage	100 epochs
Learning Rate (DS)	0.001-0.0001
UMAP Visualization	UMAP library
Pretraining Model	Dino-v1/v2

4.2 Result

Table 3 presents the main results of our experiments, evaluating the performance of our method on various datasets using Dino-v1 and Dino-v2 pretrained models. The datasets include Herbarium19 and multiple subsets of Plant Village (A, C, G, P, H, D). We report the pretrained accuracy, task-aware novel accuracy, and task-agnostic accuracy for both novel and labeled classes.

The pretrained accuracy for Dino-v1 and Dino-v2 remains consistent across the Plant Village datasets at 99.80%, indicating a high level of feature extraction from the pretrained models. For Herbarium19, the pretrained accuracy is notably lower, with Dino-v1 at 74.31% and Dino-v2 significantly higher at 85.28%.

In the task-aware evaluation, we observe varied performance across datasets. For Herbarium19, the novel accuracy for Dino-v1 and Dino-v2 is 26.77% and 34.71%, respectively. In the Plant Village subsets, the novel accuracy ranges from 57.63% to 82.69% for Dino-v1 and from 56.93% to 86.93% for Dino-v2. This indicates that Dino-v2 consistently outperforms Dino-v1 in identifying novel categories within the task-aware setting.

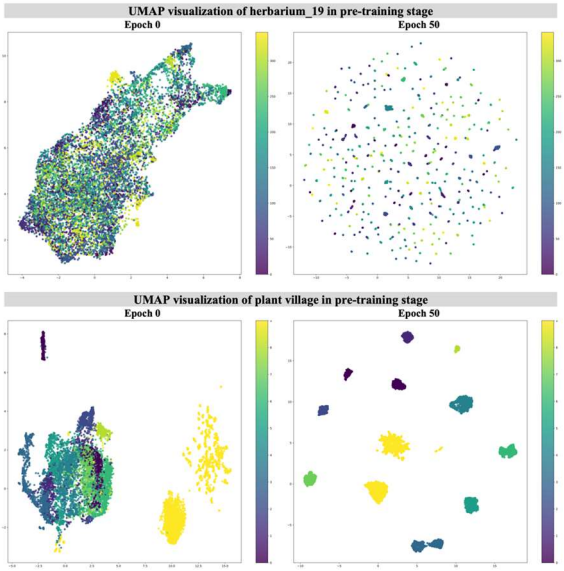

Fig. 4. UMAP visualization of herbarium19 and plant village datasets in the pre-training stage. We display UMAP visualizations of the Herbarium19 and Plant Village datasets at Epoch 0 and Epoch 50 in the pre-training stage. The color scale indicates different classes within the datasets. The pre-training model is dino-v2.

The task-agnostic evaluation shows the overall accuracy (All), novel accuracy, and labeled accuracy. For Herbarium19, the task-agnostic overall accuracy improves to 33.75% for Dino-v1 and 39.03% for Dino-v2, highlighting the advantage of Dino-v2 in a more realistic evaluation scenario. For the Plant Village subsets, Dino-v2 again shows superior performance, with task-agnostic novel accuracy ranging from 54.89% to 88.98% and labeled accuracy consistently high, nearing 100% in most cases. Dino-v1 also

performs well but lags behind Dino-v2 in novel class recognition.

Table 3. Main results. We used Dino-v1 and Dino-v2 to initialize the pre-trained models, respectively.

Dataset	Pretrained model	Pretrained Accuracy	Task-aware	Task-agnostic		
			Novel	All	Novel	Labeled
Herbarium19	Dino-v1	74.31	26.77	48.92	33.74	64.19
	Dino-v2	85.18	34.71	58.47	39.03	78.05
Plant Village(A)	Dino-v1	99.80	63.95	93.91	61.66	99.51
	Dino-v2	99.80	63.98	94.00	61.47	99.65
Plant Village(C)	Dino-v1	99.80	66.91	94.10	67.12	99.67
	Dino-v2	99.80	69.01	94.38	67.54	99.92
Plant Village(G)	Dino-v1	99.80	77.76	95.94	79.16	99.78
	Dino-v2	99.80	89.11	97.84	88.98	99.87
Plant Village(P)	Dino-v1	99.80	82.69	97.81	81.46	99.73
	Dino-v2	99.80	86.93	98.22	85.85	99.67
Plant Village(H)	Dino-v1	99.80	57.63	84.31	56.41	99.74
	Dino-v2	99.80	56.93	83.87	54.89	99.90
Plant Village(D)	Dino-v1	99.80	63.62	85.01	63.84	99.79
	Dino-v2	99.80	67.02	86.40	67.07	99.90

We utilized UMAP to visualize the feature distribution in the pretrained models, as shown in Fig. 4. In the feature distribution plots at epoch 0, both datasets appear very disorganized. However, after fine-tuning for 50 epochs, we observed that most categories have clear boundaries. This is particularly evident in the Plant Village dataset. Such effective feature extraction significantly aids in the clustering of novel categories during the discovery phase.

These results demonstrate the effectiveness of our method, particularly with the Dino-v2 model, in both task-aware and task-agnostic evaluations. The clear feature boundaries achieved after fine-tuning underscore the potential of our approach in practical applications involving novel category discovery and open-set recognition.

V. CONCLUSION

In this study, we found that the relationships between different species or diseases are intuitively important for novel category discovery. We employed a knowledge distillation framework to extract relationships between known and unknown categories, facilitating novel category discovery. Our experimental results demonstrate the effectiveness of our proposed method. Additionally, using Dino-v2 to initialize our model achieved better performance compared to Dino-v1, especially in the accuracy of recognizing novel categories. This indicates that our framework is compatible with larger pre-trained models. UMAP visualizations also confirmed the robustness of our proposed method. Furthermore, the task-agnostic evaluation revealed the practical applicability of our method, confirming its suitability for real-world scenarios where distinguishing between known and novel categories is crucial. Overall, this task is the first to apply novel category discovery

to the field of plant species and disease recognition. We believe this study will contribute to the community.

REFERENCES

1. Dong, J.; Fuentes, A.; Yoon, S.; Kim, T.; Park, D.S, "Towards Improved Performance on Plant Disease Recognition with Symptoms Specific Annotation," *Smart Media Journal*, vol. 11, no.4, pp. 38-45, 2022.
2. Dong, J.; Lee, J.; Fuentes, A.; Xu, M.; Yoon, S.; Lee, M.H.; Park, D.S. "Data-centric annotation analysis for plant disease detection: Strategy, consistency, and performance," *Frontiers in Plant Science,* vol. 13, 2022.
3. Dong, J.; Fuentes, A.; Yoon, S.; Kim, H.; Jeong, Y.; Park, D.S, "A new deep learning-based dynamic paradigm towards open-world plant disease detection," *Frontiers in Plant Science*, vol. 14, 2023.
4. Dong, J.; Fuentes, A.; Yoon, S.; Kim, H.; Park, D.S. "An iterative noisy annotation correction model for robust plant disease detection," *Frontiers in Plant Science*, vol. 14, 2023.
5. Dong, J.; Fuentes, A. "Towards Improved Performance on Plant Disease Recognition with Symptoms Specific Annotation," *Smart Media Journal*, vol. 11, no.4 , pp. 38-45, 2022.
6. Xu, M.; Yoon, S.; Lee, J.; Park, D.S. "Unsupervised transfer learning for plant anomaly recognition," *Smart Media Journal*, vol. 11, no. 4, pp. 30-37, 2022.
7. Xu, M.; Yoon, S.; Park, J.; Baek, J.; Park, D.S. "Predicting Desired Fertigation for Rose Using Internet of Things Sensors and Time-Series Model," *Smart Media Journal*, vol. 13, no. 2, pp. 16-22, 2024.
8. Zhong, Z.; Fini, E.; Roy, S.; Luo, Z.; Ricci, E.; Sebe, N. "Neighborhood contrastive learning for novel class discovery," *In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10867-10875, 2021.
9. Cui, P.; Wang, J. "Out-of-distribution (ood) detection based on deep learning: A review," *Electronics,* vol. 11, no. 21, 2022.
10. Meng, Y.; Xu, M.; Kim, H.; Yoon, S.; Jeong, Y.; Park, D.S, " Known and unknown class recognition on plant species and diseases," *Computers and Electronics in Agriculture,* vol. 215, 2023.
11. Fuentes, A.; Yoon, S.; Kim, T.; Park, D.S. "Open set self and across domain adaptation for tomato disease recognition with deep learning techniques," *Frontiers in Plant Science,* vol. 12, 2021.
12. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789-1819, 2021.
13. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. "Emerging properties in self-supervised vision transformers," *In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650-9660, 2021.
14. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A. "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193 2023.
15. Fini, E.; Sangineto, E.; Lathuilière, S.; Zhong, Z.; Nabi, M.; Ricci, E, "A unified objective for novel class discovery," *In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 9284-9292, 2021.
16. Gu, P.; Zhang, C.; Xu, R.; He, X. "Class-relation knowledge distillation for novel class discovery," *lamp*, 2023.
17. Tan, K.C.; Liu, Y.; Ambrose, B.; Tulig, M.; Belongie, S. "The herbarium challenge 2019 dataset," arXiv preprint arXiv:1906.05372 2019.

18. Hughes, D.; Salathé, M. "An open access repository of images on plant health to enable the development of mobile disease diagnostics," arXiv preprint arXiv:1511.08060 2015.
19. Kuhn, H.W. "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1−2, pp. 83−97, 1955.

Authors

Jiuqing Dong
He received his B.S. and M.S. degree in Shanghai University of Engineering Science. He is currently studying for his Ph.D. in Jeonbuk National University.

Alvaro Fuentes
received his Ph.D. degrees in electronics engineering majoring in artificial intelligence and computer vision from Jeonbuk National University, South Korea, in 2016 and 2019, respectively. He is currently a Postdoctoral Researcher with the Department of Electronics Engineering, Jeonbuk National University.

Mun Haeng Lee
is a professor at Chungnam State University, Republic of Korea. He received his BS from Chungnam State University, Republic of Korea in 2008, and MS and PhD degrees from the Sangmyung University, Republic of Korea in 2013 and 2016. He has published many papers in international conferences and journals. He is a member of The Korean Society for Bio−Environment Control. His research interests include protected horticulture and autonomous greenhouse, fruit and vegetables.

Taehyun Kim
received a B.S.and M.S degree in computer engineering from Sejong University, Seoul, South Korea B.S. in 2006 and M.S. in 2009. He is currently working at the National Institute of Agricultural Sciences after Ph.D candidate at Sejong University. His research interests include wireless communication, artificial intelligence and digital agriculture

Sook Yoon
Sook Yoon received the Ph.D. degree in electronics engineering from Jeonbuk National University, South Korea, in 2003. She is currently a Professor with the Department of Computer Engineering, Mokpo National University, South Korea.

Dong Sun Park
He received his B.S. degree from the Department of Electronic Engineering of Korea University, South Korea in 1979, and his M.S. and Ph.D. degrees from University of Missouri, USA, in 1984 and 1991 respectively. His research interests include deep neural networks, pattern recognition, image processing, digital systems design.