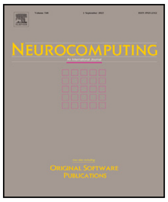
Neurocomputing 610 (2024) 128608

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Beyond supervision: Harnessing self-supervised learning in unseen plant disease recognition

Abel Yu Hao Chai [a],*, Sue Han Lee [a], Fei Siang Tay [a], Pierre Bonnet [b], Alexis Joly [c]

[a] Swinburne University of Technology Sarawak Campus, Q5B, Kuching, 93250, Sarawak, Malaysia
[b] CIRAD, Montpellier, France
[c] INRIA, Montpellier, France

ARTICLE INFO

Communicated by S. He

Keywords:
Unseen plant disease recognition
Self-supervised learning
Supervised learning
Hybrid learning

ABSTRACT

Deep learning models have demonstrated great promise in plant disease identification. However, existing approaches often face challenges when dealing with unseen crop-disease pairs, limiting their practicality in real-world settings. This research addresses the gap between known and unknown (unseen) plant disease identification. Our study pioneers the exploration of the zero-shot setting within this domain, offering a new perspective to conceptualizing plant disease identification. Specifically, we introduce the novel Cross Learning Vision Transformer (CL-ViT) model, incorporating self-supervised learning, in contrast to the previous state-of-the-art, FF-ViT, which emphasizes conceptual feature disentanglement with a synthetic feature generation framework. Through comprehensive analyses, we demonstrate that our novel model outperforms state-of-the-art models in both accuracy performance and visualization analysis. This study establishes a new benchmark and marks a significant advancement in the field of plant disease identification, paving the way for more robust and efficient plant disease identification systems. The code is available at https://github.com/abelchai/Cross-Learning-Vision-Transformer-CL-ViT.

1. Introduction

Plant diseases pose a significant threat to agricultural-producing countries and are primarily caused by pathogenic organisms such as bacteria, fungi, or parasitic plants. The plant disease identification approaches have undergone continuous evolution over the last few decades. Traditionally, this task was exclusively performed by plant pathologists or experts, utilizing labor-intensive laboratory techniques that demanded extensive knowledge and experience. However, the challenge arises from the fact that the same pathogen can infect different host crops, and the visual symptoms of different plant diseases often exhibit similarities, creating difficulties for non-experts in the field. The introduction of automated plant disease identification based on computer vision aims to assist individuals in accurately identifying plant diseases.

Recently, deep learning (DL) models have shown promising results for plant disease identification. Initially, their main focus was on the identification of multiple diseases within a single crop [1]. However, with the advent of DL models in big data analysis, the research community has broadened its scope. Larger datasets have been collected and efforts have expanded to identify diseases across multiple crops [2]. Despite this progress, the samples available in the largest publicly accessible PlantVillage (PV) dataset from [3] remain considerably limited in the global context, and the data collection process is proving prohibitively expensive. As a result, several studies have emerged to transfer knowledge gleaned from training datasets to perform identification tasks on data that is not present in the training dataset [4–6]. This scenario in the plant disease field is formally referred to as "unseen plant disease identification".

In the work by Lee et al. [7], a proposed solution to this issue involves either repurposing the model to concentrate exclusively on disease classes by excluding the crop, or implementing a post-prediction approach using a late fusion method to amalgamate probabilities associated with distinct common disease symptoms. While this work stands as a reference in the field, its relatively poor performance on unseen classes underscores the imperative for further exploration, especially within multi-plant contexts. A recent publication of the same author [8] has taken a step forward by incorporating conditional links to strengthen the contextual relationship between diseases and plants. However, the improvement remains fairly modest and does not completely resolve the major obstacle of disentangling the characteristics

* Corresponding author.
E-mail addresses: aychai@swinburne.edu.my (A.Y.H. Chai), shlee@swinburne.edu.my (S.H. Lee), fstay@swinburne.edu.my (F.S. Tay), pierre.bonnet@cirad.fr (P. Bonnet), alexis.joly@inria.fr (A. Joly).

https://doi.org/10.1016/j.neucom.2024.128608
Received 2 April 2024; Received in revised form 5 July 2024; Accepted 11 September 2024
Available online 14 September 2024

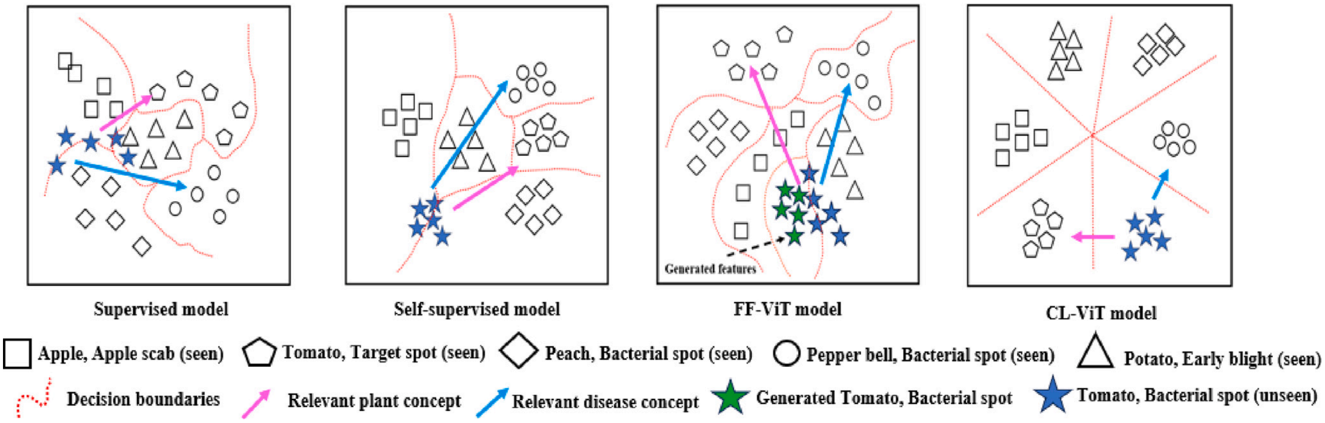

Generated features
Supervised model
Self-supervised model
FF-ViT model
CL-ViT model
Apple, Apple scab (seen)
Tomato, Target spot (seen)
Peach, Bacterial spot (seen)
Pepper bell, Bacterial spot (seen)
Potato, Early blight (seen)
Decision boundaries
Relevant plant concept
Relevant disease concept
Generated Tomato, Bacterial spot
Tomato, Bacterial spot (unseen)

Fig. 1. The diagram illustrates the motivation behind our proposed methods, leveraging the power of self-supervised and supervised learning mechanisms to enhance feature distribution space. This aims to reduce the gap between seen and unseen features while preserving intra- and inter-class separation.

of the plant from those of the disease. To guide the model towards acquiring a more generalized understanding of plant diseases, we propose to incorporate specific inductive biases to constraint the model's parameter space. The idea is to ensure that the disparities in feature distribution between seen and unseen classes are minimized, leading to a more generalized and robust model, while preserving the necessary inter- and intra-class separation.

To achieve this, we propose a novel way of conceptualizing the plant disease identification task. Specifically, we view plant disease identification as a compositional task, where each sample consists of plant and disease concepts. For instance, consider *Tomato* as the plant concept and *Target spot* as the disease concept, which together form the *Tomato_Target spot* composition class. The context of our "unseen" class can be defined as plant samples within the combined composition class that are not present in our training dataset. However, it is important to note that the individual plant or disease concepts comprising this class may already exist within the training dataset. For example, if we introduce *Tomato_Bacterial spot* as an unseen class in our experiment, this specific composition would not be present in our training dataset. Nonetheless, the individual *Tomato* plant and *Bacterial spot* disease concepts might be found within other compositions, such as *Tomato_Target spot* and *Pepper_Bacterial spot* compositions.

When delving into the field of compositional plant disease identification, a practical solution would be to exploit the features from individual concepts and repurpose them to characterize the unseen data through the integration of these concepts. One possible approach is to project the visual images, converting input from the original visual space into low-dimensional embedding spaces that capture the semantic information of the original input [9]. To be more precise, we can begin by disentangling plant and disease concepts into their individual embedding spaces, and these embedded features can subsequently serve a multitude of downstream tasks. For example, by disentangling *Tomato_Early blight*, *Tomato_Late blight*, *Potato_Early blight* and *Potato_Late blight*, we could learn the feature representation of 2 crop species (potato and tomato) and 2 disease types (early blight and late blight). Then, this information can be applied to another composition with a similar individual concept of crop and disease. For example, *Tomato*'s features learned via the aforementioned composition can be used to represent the characteristics of the crop of the new composition of *Tomato_Bacterial Spot*, placing less demand on the training dataset. This can be beneficial for unseen identification tasks where joint compositions are not present in the training dataset, as the model can draw on the knowledge of individual learned concepts from seen compositions. In the work by [9], where we refer to the model as FF-ViT, has been proposed with a pairwise feature generation module. This module is designed to generate joint synthetic features of unseen compositions from seen compositions. Such an approach allows the model to be exposed to both seen and unseen feature distributions.

Given that FF-VIT the model has never seen actual unseen data, minor deviations in the distribution of joint synthetic features from the original plant disease samples can significantly impact the performance of unseen class identification. Experimental results from [9] demonstrate a substantial performance gap between seen and unseen tasks. This essentially creates a bottleneck and drives the search for alternatives capable of learning more robust features that are less sensitive to distribution shifts. Self-supervised learning (SSL) approaches have recently proven to be effective alternatives for learning generalized features that reduce the disparities caused by domain shifts [10–13]. These methods utilize SSL techniques to simultaneously minimizing intra-class variance while maximizing inter-class variance across different domains. As a result, they enable models to learn more robust and discriminative feature representations from different domains, thereby enhancing adaptability to various downstream tasks. Motivated by these studies, our paper adapts SSL principles to develop features that are applicable to both seen and unseen data. Specifically, the model can identify visual appearances or disease symptoms common to different plant species or disease types. For instance, the model can learn the features of spots, which are symptoms common to different plants, as illustrated in Fig. 2. Hence, by using generalized features, the model able to perform on both seen and unseen data.

Inspired by the insights gained from our background study, we introduce our novel model, CL-VIT. The core principle revolves around providing the model with prior knowledge to refine the search space, ultimately minimizing the gap between seen and unseen features. Prior knowledge is injected through SSL mechanisms where the model is able to refine its search space by finding implicit patterns in the actual disease data, independent of the label. Building upon this concept, we have devised a novel approach that eliminates the need for generating synthetic data in FF-ViT model. Instead, the model, equipped with self-supervisory signals, serves as prior knowledge to guide the training process, effectively aligning the feature distribution between seen and unseen classes and demonstrating superior performance compared to our initial model. We illustrate the difference between FF-VIT and CL-VIT in terms of feature distribution space in Fig. 1

To our knowledge, this is the first initiative in tackling the zero-shot setting in plant disease identification, introducing a novel approach to conceptualizing plant disease identification. We introduce a new concept in plant disease identification to recognize unseen classes, with a focus on improving efficiency by incorporating a guided learning mechanism. This mechanism is specifically designed to reduce the feature distribution gap between seen and unseen plant disease data. In summary, the contributions of this paper are outlined below:

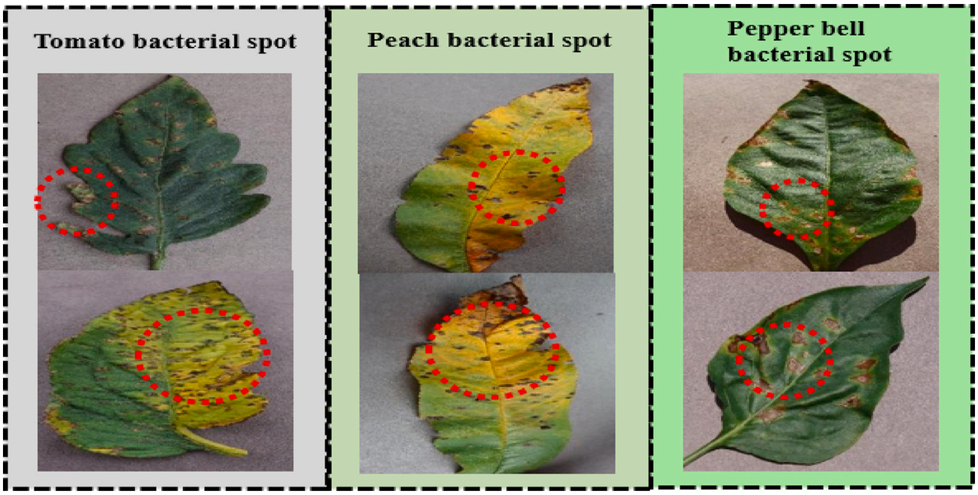


Fig. 2. The diagram illustrates the visual symptoms of bacterial spot disease for different plants. The SSL aims to learn general features that able to represent all plants ignoring other features such as shape or color.

1. We introduce a novel model called CL-ViT, featuring unique conceptual designs, setting a new benchmark in the field of unseen plant disease identification.
2. We demonstrate that the incorporation of a guided learning mechanism surpasses conventional approaches in the multi-plant disease identification benchmark. Furthermore, we show that the CL-ViT model, integrating a SSL approach, outperforms the FF-ViT model employing a purely supervisory learning scheme for unseen plant disease identification tasks.
3. In our qualitative analyses, we illustrate that CL-ViT learns a feature space capable of discriminating between different classes while minimizing the domain gap between seen and unseen data. This underscores the superiority of CL-ViT in implementing a more effective guided learning mechanism.

The rest of this paper is organized as follows. Section 2 discusses the work in the literature related to unseen plant disease identification and domain adaptation through supervised learning (SL) and SSL. In Sections 3 and 4, we present our discussion in detail for our CL-VIT architecture and dataset. In Section 5, we compare the performance of CL-VIT to other existing approaches, and Section 6 conducts ablation studies to evaluate the robustness of our best model. In Section 7, we analyze the models qualitatively in terms of feature distribution visualization to further justify the performance differences between models. Finally, a conclusion aiming at synthesizing the obtained results, highlighting limitations, and suggesting potential future work is presented in Section 8.

2. Related work

In this section, we provide an overview of the pertinent advancements related to our work.

2.1. Deep learning in plant disease identification

Plant disease identification has evolved rapidly over the past few decades. In the past, it was mainly carried out by plant pathologists by time-consuming laboratory techniques that required in-depth knowledge and experience.

Today, it can be easily assisted by DL models trained from visual images [1,14]. Previously, plant disease identification focused on the identification of a single plant species/crop and multiple diseases [1]. Since the emergence of DL technology, the research community has collected new datasets that encompass more classes, including multiple plant species and multiple diseases [2,3]. With these new datasets, models need to learn both species and disease information for multi-species and multi-disease identifications. Convolutional neural networks (CNNs) have been widely used in the field of plant disease identification [3,15]. However, recent studies have revealed that vision transformer (ViT) models tend to perform better than CNNs, with the latter focusing in some cases on regions that are irrelevant in plant disease identification [7,16,17]. Motivated by this growing body of evidence, our study adopts ViT models as the backbone model for our proposed network architecture.

The current DL model heavily relies on the size of the training data; a larger dataset leads to better model performance. However, in the context of plant disease identification, acquiring sufficient data is challenging. This limitation has prompted us to address the core issue where no training samples are available. The recent approach proposed by [9] aims to bridge this gap. Motivated by these studies, we incorporate a new strategy to further enhance the unseen plant disease identification framework.

2.2. Unseen plant disease identification

In plant disease identification, where pathogens attacking different crops are categorized under the same disease class primarily due to their similar visual symptoms [18], researchers utilize the concept of transferring knowledge learned from one crop disease to another. This approach leverages the similarities in disease visual symptoms, allowing models trained on one crop's disease to be effectively applied to others, including both seen and unseen classes. Existing mainstream methods focus on converting unseen plant disease identification tasks into general identification tasks. They aim to directly transfer the knowledge learnt from seen data and perform on unseen data ignoring their distribution shift. For example, [7] proposed to train their models with a disease-oriented classifier that only focuses on disease features and disregards plant/species features. However, this approach, which omits species-specific features, may not be consistent with the perception of expertise in distinguishing plant diseases, particularly when information on host species is often needed to delineate lists of diseases associated with host species [19]. Besides, the disease-oriented classifier also limits the performance of model towards multi-plant species identifications.

Other mainstream methods presented in [8] trained both species-oriented and disease-oriented classifiers which resolved the limitation of previous studies in [7] that were unable to perform on plant identifications. However, their models are only trained with labeled seen data which results in the models tend to bias towards seen data. Although they established a conditional link between the plant features and disease features to encourage knowledge sharing between both features, the performance on unseen identification has not shown significant improvement. Therefore, we notice that it is important to obtain a generalized feature representation so that the knowledge learned from seen data can be applied to unseen data.

The recent proposal, FF-ViT [9] incorporates a new strategy to reduce the disparities between seen and unseen classes via a synthetic feature generation scheme. FF-ViT marked a significant milestone as the benchmark for addressing unseen plant disease identification. Despite its innovation, we observed a potential limitation: the feature distribution of synthetic compositions might not faithfully represent the true feature distribution of unseen compositions. Consequently, a performance gap emerged between seen and unseen identification tasks.

In the field of general unseen identification tasks, [20] enhances seen and synthetic features using its Feature Refinement (FR) module, which learns unified features for seen and synthetic data. Similarly, [21] improves its model with its Multi-Decision Fusion Model (DMFM), which exploits different image views. While previous work has focused on refining their feature learning modules through improvements to the model architecture or the use of different views of the input image, we address this challenge using an alternative approach. We introduce a new CL-ViT model, which integrates SSL and SL to learn a unified feature representation. We show that this novel approach improves the learning process by capturing relevant and generalized features of individual concepts, ensuring a more accurate and complete understanding of both seen and unseen classes, thus closing the performance gap between them.

2.3. Learning feature representation with self-supervised learning

The aim of SSL is to diminish the distance between features of the same class while amplifying the distance between features of different classes from unlabeled data [22,23]. This learning technique is widely applied to a variety of tasks, encompassing not only visual images but also videos [24]. Recently, the Contrastive Language-Image Pre-Training (CLIP) model was introduced by [25], which integrates visual and textual features using SSL. The CLIP model demonstrates outstanding performance on various downstream tasks, including those involving unseen data. However, the model's effectiveness depends on meticulous parameter tuning and the selection of appropriate text prompts. To address these challenges, [26,27] have improved on the original model by introducing techniques such as layer-wise prompt learning and Video Object Segmentation (VOS), respectively, to ease the burden of data annotation. Despite these advances, the field of plant disease identification presents unique challenges, primarily due to the lack of standardized textual descriptions for disease types. Hence, in this study, we focus on visual features, enhancing the visual models.

There are also SSL approaches that leverages auxiliary pretext tasks [28,29], which serve as a means to learn representations through these tasks. These tasks can be thought of as pseudo-labels or labels automatically generated based on the dataset's attributes. Inspired by a recent proposal for unsupervised domain adaptation that employs self-supervision through auxiliary pretext tasks to align the learned representations of two domains in a shared feature space [10], and supported by evidence demonstrating its effectiveness in reducing disparities in plant domain data [11], we further explore this concept. Specifically, we initiate an exploration to assess the potential of leveraging various auxiliary pretext tasks, within plant disease visual data, to learn general features applicable to unseen plant disease classes.

A common approach in SSL through auxiliary pretext tasks involves employing pixel-wise data augmentations [30–33]. The model learns to recognize these augmented versions to capture fine-grained features. It is notable that recent proposals [10,29] have explored alternative SSL techniques that move beyond pixel-wise data augmentations. These approaches focus on non-pixel-wise data augmentations such as rotation and flipping to solve the problems of domain adaptation as they claimed that features learned from these augmentations can induce alignment between the source and target domain.

This becomes particularly crucial in tasks like unseen plant disease identification, where the primary challenge lies in transferring knowledge acquired from seen data to unseen data. Although visual disease symptoms may appear similar, variations in leaf patterns and structures among different plant species can result in differences in the overall visual appearance across various plant species. The use of pixel-wise data augmentations might result in learned features that are overly specific to the seen data, making it challenging to generalize and extract useful information for the unseen data. Therefore, drawing inspiration from recent proposals [10,29], we have designed a novel model with carefully chosen pretext tasks. These tasks are seamlessly integrated into our proposed CL-ViT model, ensuring that the learned features encompass a broader range of plant species, including both the seen and unseen, that share similar disease symptoms.

3. Methods

3.1. Problem formulation

For ease of reference, we provide a list of key abbreviations and notations in Table 1. In this study, the plant disease samples are associated with compositions of plant species/crop concepts, $P = (p_0, p_1, \dots, p_m)$ and disease concepts, $D = (d_0, d_1, \dots, d_n)$ where m and n are denoted as total unique species and disease concepts respectively. The total unique compositions are denoted as $C = m \times n$. Besides, we also prepared two disjoint sets C^s and C^u where both are subset of C. Specifically, C^s and C^u represent the seen compositions and unseen compositions respectively. The goal of our unseen composition plant disease identification is to recognize both compositions from C^s and C^u. Note that, in the FF-ViT model [9], identifications are solely based on the knowledge extracted from C^s. In contrast, the CL-ViT model learns from both C^s (labeled) and C^u (unlabeled).

In the following sections, we first present the framework of our proposed CL-ViT model along with a detailed explanation of its components. Before diving into the experimental results, we also discuss the strategies we discovered for further improving the existing FF-ViT model [9] to better benchmark our novel CL-ViT.

3.2. Cross Learning Vision Transformer (CL-ViT)

Recent publications have highlighted a pervasive bias in models trained with SL, where they tend to favor the distributions of seen data. These models often have the ability to learn very complex decision boundaries tailored to the seen data features, irrespective of the broader feature distribution in the training dataset [34,35]. This observation has prompted our focus on a critical issue: the need for models to learn relevant and generalized individual concept features. This emphasis is essential for transferring knowledge acquired from seen data to unseen data in the context of multi-task plant disease identifications. Inspired by the work in [10,36] that leveraged SSL to reduce disparities between cross-domain data, we envision the deployment of SSL as a means to learn feature distributions that closely align with both seen and unseen plant disease data. This strategic alignment could help to minimize knowledge and domain gaps. Specifically, we introduce a new model known as Cross-Learning Vision Transformer (CL-ViT).

CL-ViT employs multiple self-supervised tasks that can operate concurrently with SL. The essence of CL-ViT lies in its dual purpose: firstly, it unifies the embedding learning process for both plant and disease

Table 1
The key abbreviations and notations.

Abbreviation and Notation	Definition
SSL	Self-supervised learning
SL	Supervised learning
PV	PlantVillage
FF-ViT	Feature Fusion Vision Transformer model
CL-ViT	Cross Learning Vision Transformer model
P	Plant concept
$\hat{P}$	Features of plant concept
D	Disease concept
$\hat{D}$	Features of disease concept
C	Total compositions which consist of both plant and disease concepts
C^s	Seen composition in both training and testing dataset
C^u	Unseen composition (unlabeled) in testing dataset
$\hat{C}$	Features of total compositions
$\hat{R}$	Feature from rotational SSL task
$\hat{F}$	Feature from flipping SSL task
$\hat{I}$	Feature from image patch SSL task
CLS	CLS token from attention output
Patch	Attention output exclude CLS token
L_S	Supervised loss for seen compositions
L_{SS_S}	Self-supervised loss for seen compositions
L_{SS_U}	Self-supervised loss for unseen compositions

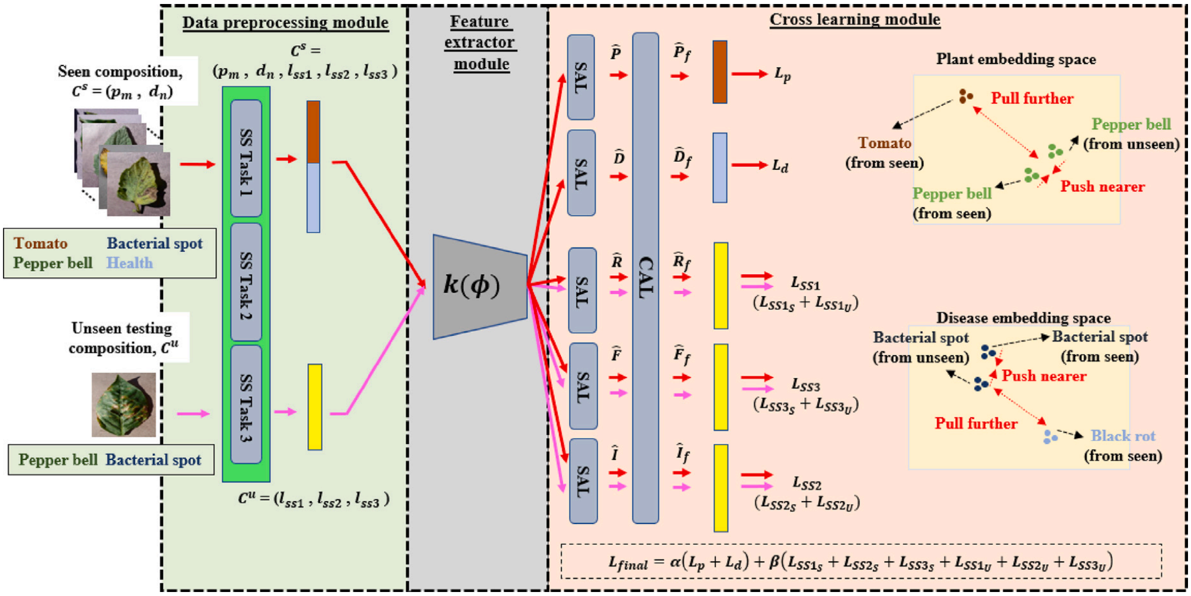

Data preprocessing module
$C^s = (p_m, d_n, l_{ss1}, l_{ss2}, l_{ss3})$
Seen composition, $C^s = (p_m, d_n)$
Tomato
Bacterial spot
Pepper bell
Health
SS Task 1
SS Task 2
SS Task 3
Unseen testing composition, C^u
Pepper bell
Bacterial spot
$C^u = (l_{ss1}, l_{ss2}, l_{ss3})$
Feature extractor module
$k(\phi)$
Cross learning module
SAL
CAL
$\hat{P}$
$\hat{D}$
$\hat{R}$
$\hat{F}$
$\hat{I}$
$\hat{P}_f$
$\hat{D}_f$
$\hat{R}_f$
$\hat{F}_f$
$\hat{I}_f$
L_p
L_d
L_{SS1} $(L_{SS1_S} + L_{SS1_U})$
L_{SS3} $(L_{SS3_S} + L_{SS3_U})$
L_{SS2} $(L_{SS2_S} + L_{SS2_U})$
Plant embedding space
Pull further
Tomato (from seen)
Pepper bell (from unseen)
Pepper bell (from seen)
Push nearer
Disease embedding space
Bacterial spot (from unseen)
Bacterial spot (from seen)
Push nearer
Pull further
Black rot (from seen)
$L_{final} = \alpha(L_p + L_d) + \beta(L_{SS1_S} + L_{SS2_S} + L_{SS3_S} + L_{SS1_U} + L_{SS2_U} + L_{SS3_U})$

Fig. 3. The figure shows the architecture of our proposed CL-ViT model. SAL represents self-attention layer with Gaussian Error Linear Units (GeLU) activation. CAL represent cross-attention layer with GeLu activation.

components, facilitating the establishment of essential contextual relationships vital for characterizing plant disease data. Notably, this approach tackles the intricate challenge of disentangling features when disease and species details intricately interweave in leaf samples. Secondly, CL-ViT guides the learning of the feature distribution space using actual disease data through SSL. By doing so, it effectively reduces the disparities between seen and unseen data, aligning the model's understanding with the actual distribution of plant diseases in the real world.

CL-ViT model can be separated into three main modules which are the data preprocessing module, feature extractor module and cross-learning module. The data preprocessing module will perform data augmentation for all input images based on our carefully designed self-supervised tasks. After the feature extraction process, these detailed features are projected into separate spaces for SL and SSL.

3.2.1. Data preprocessing module

This module plays a crucial role by applying random augmentations to input plant disease images, mapping them with specific pretext labels. The purpose is to facilitate the training of the feature extractor on non-annotated data, enabling the learning of valuable features that can serve as robust priors aligning the feature distribution between seen and unseen classes.

However, it is important to note that not all pretext tasks are suitable for capturing plant disease characteristics. For instance, those that delve into pixel-level details, such as predicting pixel colors or fine-grained brightness adjustments, are not ideal as they do not support high-level visual concepts essential for bridging the similarity gap between seen and unseen classes. The inclusion of such references could lead to a separation between seen and unseen class distributions, ultimately biasing the model towards seen classes due to the influence of supervised signals. Careful consideration in choosing pretext tasks is essential to avoid this undesirable bias. Therefore, in this study, we deliberately choose data augmentations that are able to extract the essential structural information from plant disease images. These augmentations are then transformed into a range of classification-based self-supervised tasks, drawing inspiration from prior research [29].

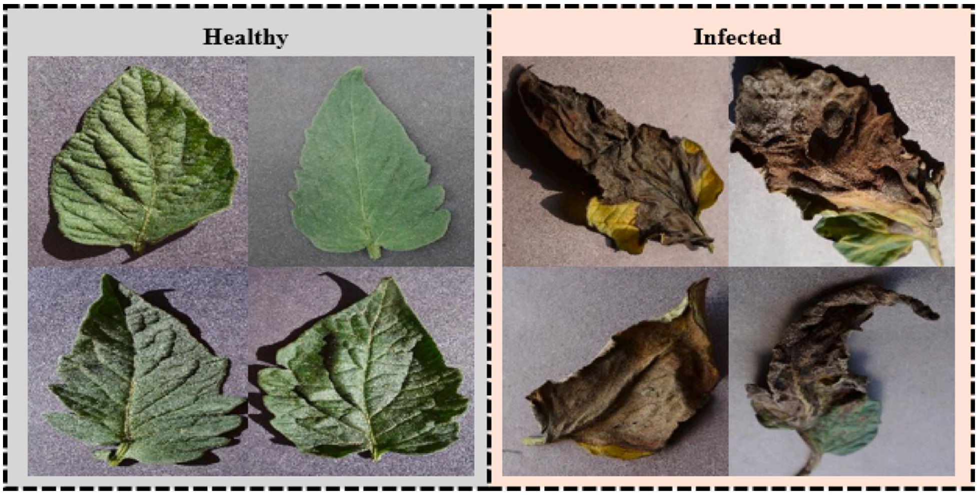


Fig. 4. The figure above shows a comparison between a healthy and an infected *tomato* plant. The pathogens have infected various regions of the leaf samples, obscuring plant details such as venation and leaf shape.

Due to the characteristics of disease symptoms that can occur in distinct leaf regions, the patch location prediction methodology employed in [10] is not suitable for our plant disease identification tasks. The approach consists of cropping the original image into several patches and the cropped images may not necessarily reflect the plant-disease relevant features, thereby hindering the model's effective learning of task-specific features. To address this limitation, we proposed a new self-supervised task called "image patch randomization prediction". Specifically, we avoid constraining the model to look at designated patches but instead, examine holistic patches of an image in a random manner. We present below the detail of each self-supervised task chosen for our study:

- **Rotation prediction**: The input images are rotated into 4 different angles which are 0°, 90°, 180° and 270°. The task is to predict the rotational angle as an identification task.
- **Flip prediction**: The input images are either not flipped, horizontally flipped, or vertically flipped. The task is to predict whether the images are either flipped or not.
- **Image patches randomization prediction**: The input images are either uncropped, cropped into 16 image patches, or cropped into 64 image patches. After that, the order of these image patches is shuffled randomly before being reassembled into their original image size. By using this task, the global or structural information of the images will be distorted in order to ensure the model learns local features while all relevant visual features such as disease symptoms are still retained in the input images. The task is to predict the total number of cropped pieces.

3.2.2. *Feature extractor module*

FF-ViT [9] introduces a methodology that involves disentangling plant and disease features. However, the intricate nature of disease development often results in pathogens infecting diverse regions of leaf samples, obscuring specific plant details and making it impractical to disentangle them into individual concept features, as depicted in Fig. 4. Attempting to force the model to unravel these complex features risks learning irrelevant aspects, deviating from the intended concept. Consequently, the CL-VIT feature extractor adopts a unified approach, consolidating the embedding learning process for both plant and disease components.

This unified strategy is designed to establish essential contextual relationships crucial for characterizing plant diseases effectively. In this module, we utilize similar backbone from FF-ViT model which is ViT model (vit_base_patch16_224) from [37] as plant disease feature extractor, $k(\phi)$ to extract plant disease features, $\hat{C} = k(C^s)$. These extracted features are subsequently passed to the subsequent cross-learning modules to project the features into class categorical distribution space.

3.2.3. *Cross learning module*

Our cross-learning model is designed to acquire a generalized class categorical distribution space that encompasses both seen and unseen classes while preserving the discrimination between classes. It comprises three main components: supervised learning tasks, multiple SSL tasks, and cross-attention mechanisms.

- **Supervised learning**. In the SL tasks, the model is trained to discern the distinctions between plant disease classes, learning the decision boundaries using labeled data. We employ two linear classifiers, each corresponding to species and disease concepts, similar to [8,38]. Attention layers are utilized to enhance the discriminative features of each concept. The features of each concept can be formulated as disentangled plant features, $\hat{P} = Att_1(\hat{C})$ and disentangled disease features, $\hat{D} = Att_2(\hat{C})$ from entangled plant disease features, $\hat{C}$.
- **Self-supervised learning.** One of the key differences from FF-Vit is the incorporation of SSL to reinforce the distribution of learned features, encompassing both seen and unseen data. This SSL acts as prior knowledge, guiding the model to align the distribution between the two. The features learned in SSL are mapped to selected pretext tasks, including rotation, flip, and image patches randomization prediction, detailed in our earlier data preprocessing module. These tasks are specifically designed to capture the intrinsic characteristics of plant disease data. The features of each self-supervised features can be formulated as rotational features, $\hat{R} = Att_3(\hat{C})$, flip features, $\hat{F} = Att_4(\hat{C})$ and image patch features, $\hat{I} = Att_5(\hat{C})$ from entangled plant disease features, $\hat{C}$.
- **Cross attention.** Both SL and SSL play crucial roles in different aspects. SL tasks ensure that the model learned from labeled data. This is essential because the individual concepts learned within that seen space can serve as important clues for unseen classes. On the other hand, SSL ensures that the feature distribution space learned can encompass both seen and unseen data, promoting the generalizability of the model. To ensure that each modality can be optimized without bias towards either SL or SSL tasks, we deploy a cross-attention learning mechanism to reweight the tasks. This mechanism not only reduces bias towards either SL or SSL tasks but also, within the self-supervised tasks, acts as the primary regulator for selecting the most discriminative structural

representation needed to represent plant diseases. This concept is designed to avoid the potential pitfall of forcing the model to rely on representations of irrelevant structural features. Instead, it gives the model the flexibility to discern and select the most relevant structural features on its own as it learns the model. Specifically, The features from SL and SSL will be projected from different embedding spaces into the same embedding space. Each of the features consists of a class token (CLS) and an image patch token (Patch) to represent global features and local features learned from the input. We exchange the CLS token from SL features and SSL features and fuse them with their own patch token following the setting from [39]. The features of each final features can be formulated as below:

$$Final\ plant\ features, \hat{P}_f = CrossAtt(CLS_{R/F/I} + Patch_P) \tag{1}$$

$$Final\ disease\ features, \hat{D}_f = CrossAtt(CLS_{R/F/I} + Patch_D) \tag{2}$$

$$Final\ rotational\ features, \hat{R}_f = CrossAtt(CLS_{P/D/F/I} + Patch_R) \tag{3}$$

$$Final\ flip\ features, \hat{F}_f = CrossAtt(CLS_{P/D/R/I} + Patch_F) \tag{4}$$

$$Final\ image\ patch\ features, \hat{I}_f = CrossAtt(CLS_{P/D/R/F} + Patch_I) \tag{5}$$

3.2.4. Training strategy

In this section. we will discuss in detail all the hyperparameters and training schemes of CL-ViT models. All of the models are trained with an initial learning rate of 0.001 and then decreased by a factor of 10 for at least 1 time when the models reach optimum loss. We use SGD optimizer with a momentum of 0.9 and weight decay of 0.00001. We run the training using an NVIDIA GeForce RTX 3060 graphic card. We trained all models three times and averaged their performance results. This approach balances resource constraints with the need for reliable performance evaluation.

This model consists of three modules which are data preprocessing module, feature extractor module, and cross-learning module. The data preprocessing module will perform image augmentations corresponding to the selected pretext tasks. The feature extractor module will extract plant disease features, $\hat{C}$, from the images formed by the data preprocessing module. The cross-learning module will then combine and regulate the features, $\hat{C}$, by using both SL and SSL. SL learns two linear classifiers of plant and disease by using cross-entropy loss. The supervised loss function can be defined as below:

$$L_S = L_P + L_D \tag{6}$$

$$L_P = \sum_{i=1}^{n} P_i log(\hat{C}_{p\_i}) \tag{7}$$

$$L_D = \sum_{i=1}^{n} D_i log(\hat{C}_{d\_i}) \tag{8}$$

P_i and D_i are the truth label for plant and disease for the i sample in the dataset respectively. On the other hand, the self-supervised loss function can be defined as below:

$$L_{SS} = L_{SS1} + L_{SS2} + L_{SS3} \tag{9}$$

L_{SS1}, L_{SS2} and L_{SS3} are losses from each self-supervised pretext tasks. The loss of each self-supervised pretext tasks can be further dissociate into seen and unseen images losses for example $L_{SS1} = L_{SS1_{SP}} + L_{SS1_{SD}} + L_{SS1_U}$. Therefore, the seen and unseen images self-supervised losses can be defined as below:

$$L_{SS_S} = L_{SS_{SP}} + L_{SS_{SD}} \tag{10}$$

$$L_{SS_{SP}} = L_{SS1_{SP}} + L_{SS2_{SP}} + L_{SS3_{SP}} \tag{11}$$

$$L_{SS_{SD}} = L_{SS1_{SD}} + L_{SS2_{SD}} + L_{SS3_{SD}} \tag{12}$$

$$L_{SS_U} = L_{SS1_U} + L_{SS2_U} + L_{SS3_U} \tag{13}$$

L_{SS_S} and L_{SS_U} represent the total self-supervised losses for seen and unseen images respectively. $L_{SS_{SP}}$ and $L_{SS_{SD}}$ denote the self-supervised losses for seen images corresponding to plant and disease concepts respectively. We assign α and β as weighting coefficients to balance between supervised and self-supervised loss. The final loss function for our CL-ViT can be defined as follow:

$$L_{final} = \alpha(L_S) + \beta(L_{SS_S} + L_{SS_U}) \tag{14}$$

3.3. Feature Fusion Vision Transformer (FF-ViT)

To better benchmark our novel CL-VIT, we revisited the existing FF-ViT [9], as shown in Fig. 5. Basically, the architecture of FF-ViT, can be separated into three main modules: the concept feature extractor module, pairwise feature generation module and composition feature extractor module.

The concepts feature extractor module will extract individual plant, $\hat{P}$ and disease concepts, $\hat{D}$ features into separated embedding space. Specifically, two ViT models (vit_base_patch16_224) from [40] are used as plant feature extractor, $g(\phi)$ and disease feature extractor, $f(\phi)$. Each of the ViT model consists of 12 attention layers (12 attention heads with 768 embedding dimensions) and the final classification head is removed to convert visual images into feature embeddings and project them into plant embedding space and disease embedding space. The disentangled plant features, $\hat{P_m} = g(C^s)$ and disentangled disease features, $\hat{D}_n = f(C^s)$ can be obtained from their respective embedding space.

The pairwise feature generation module will harvest individual plant, $\hat{P}$ and disease concepts, $\hat{D}$ features from previous module and generate synthetic composition features, $\hat{S}$ that include seen, C^s and unseen compositions, C^u. In particular, the module obtains disentangled plant features, $\hat{P_m}$ and disentangled disease features, $\hat{D}_n$ from the concept feature extraction module as inputs, and combines them via feature summation [41] to generate synthetic composition features, $\hat{S} = (\hat{P_m}, \hat{D}_n)$. Unlike conventional feature fusion strategies where only two individual features from the same image are combined to obtain a combined composition, this module combines individual features from different images to generate different unique compositions. The total number of unique synthetic compositions can be formulated as $S \in m \times n$. In short, this module is capable of generating synthetic compositions, S which include seen compositions, C^s and unseen compositions, C^u for the composition feature extraction module.

After that, the final composition feature extractor module will extract features from the generated synthetic composition features, $\hat{S}$ to perform downstream identification tasks. Specifically, multiple self-attention layers [40] are used to exploit the features of synthetic compositions and obtain synthetic disentangled plant, $\hat{S_P} = Attention(\hat{S})$ and synthetic disentangled disease features, $\hat{S_D} = Attention(\hat{S})$.

To preserve entanglement information in the disentangled features and encourage knowledge sharing between the original and synthetic compositions, a residue connection of complementary features from the original disentangled features is introduced for each of the synthetic disentangled features, inspired by the CMTL architecture of [8]. Finally, the final synthetic disentangled plant features, $\hat{S_P} = \hat{S_P} + \hat{D}$ and final synthetic disentangled disease features, $\hat{S_D} = \hat{S_D} + \hat{P}$ will be used to perform seen and unseen plant disease identification tasks.

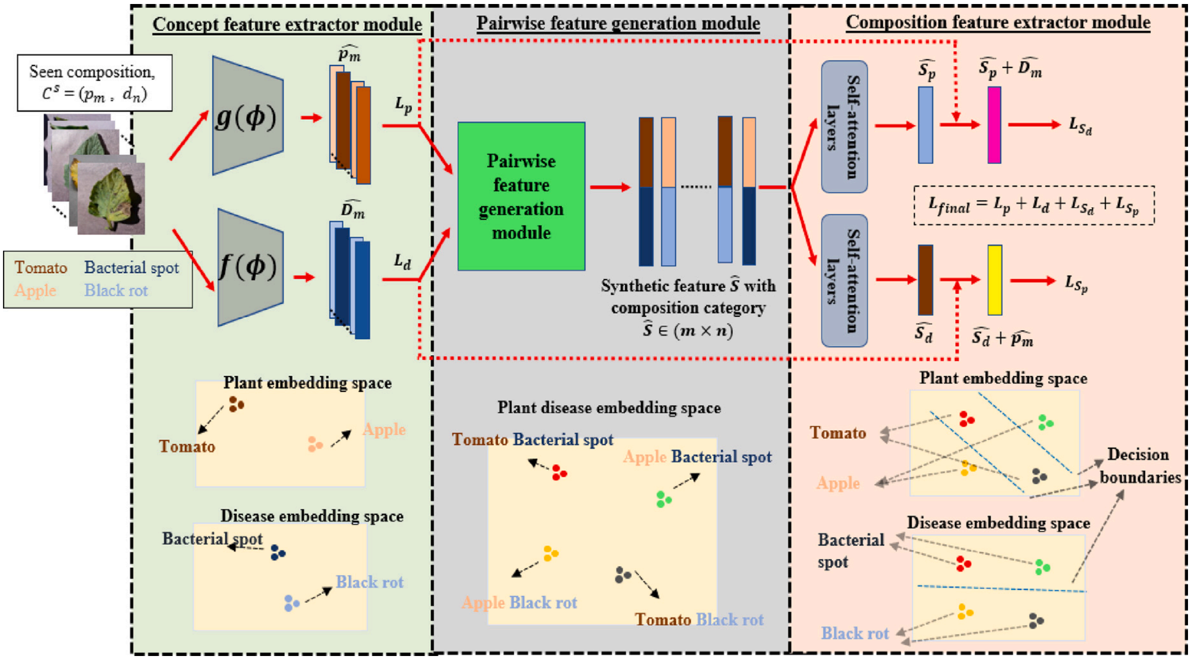


Fig. 5. The diagram shows the architecture of FF-ViT model. The pairwise feature generation module will generate synthetic compositions for both seen and unseen data to enrich the plant and disease embedding space for composition feature extractor module. The FF-ViT model utilizes the distribution of synthetic data to perform both seen and unseen plant disease identification tasks.

3.3.1. Improved FF-VIT

FF-ViT* (PWA + *a*) [9], recognized as the pioneering benchmark in this domain, has been subject to additional refinements driven by our extensive studies and insights into training schemes and design flows.

We enhance the model through a revised training process and a synthetic unique composition generation scheme (*rs*). The original training process proposed in [9] optimizes the concept feature extraction module to a certain weight range before optimizing the composition feature extraction module. However, this may limit the correlation and sharing of features learned for both modules, especially features of unseen synthetic compositions that are only available in the later module. Therefore, we propose to optimize both modules simultaneously by removing the moving weighted sum, *a* in the illustrated loss function, $L_{final} = a(L_P + L_D) + (1 - a)(L_{S_P} + L_{S_D})$, as introduced in the prior method [9]. This is to associate synthetic composition features with the plant and disease embedding space from the concept feature extraction module. This may also improve the generalization capability of the FF-ViT model, as the composition feature extraction module is highly dependent on the concept feature extraction module.

The revised scheme (*rs*) implements feature summation for disentangled plant features, $\hat{P}_m$ and disentangled disease features, $\hat{D}_n$ in the data preprocessing phase. This contrasts with the method demonstrated in [9], which performed this summation in the mini-batch. By relocating this feature summation process, our pairwise feature generation module can generate a scheme based on information from the entire training dataset, rather than relying solely on samples from each minilot. As a result, the scheme ensures a more balanced distribution of total samples for each unique synthetic composition. In addition, it also reduces the bias of FF-ViT model towards any dominant class.

4. Dataset

In this study, we assess the models' performance using two different datasets that represent laboratory-centric and real-world environments. The datasets include (1) PV dataset (largest publicly available laboratory images from [3]) and (2) real-world plant disease dataset (challenging dataset proposed in [8]).

PV dataset consists of 38 plant disease class pairs with 54,305 images. We separated the dataset into 80% training set and 20% testing set according to the previous study in [3]. We further divide the original test set into a seen test set (37 plant disease pairs from PV) and an unseen test set (only *Pepper bell bacterial spot*). Given that the PV dataset is captured in a controlled environment without a noisy background, our analysis is honed in on evaluating the robustness of plant disease-oriented features learned by our proposed models. Specifically, we are keen on understanding how effectively these features can transition from the seen to the unseen data.

The real-world plant disease dataset consists of a mixture of both the laboratory image and the field. It is composed of PV dataset, Digipathos dataset, IPM dataset and Pl@ntNet dataset. The dataset consists of 1146 plant disease class pairs with 10,226 training images and 1951 seen testing images. We also use INRAEdi dataset from [8] which consists of 388 images as additional unseen testing images.

5. Quantitative performance evaluations

In this section, we begin with the performance analysis of the enhanced FF-VIT, which serves as the crucial benchmark for our novel CL-VIT. Subsequently, we assess the performance of CL-VIT against various state-of-the-art (SOTA) models. Following that, we present a comprehensive ablation study on CL-VIT.

5.1. Improvement of FF-ViT

FF-ViT* (PWA + *a*) [9], acknowledged as the pioneering benchmark in this domain, has undergone additional improvements based on our in-depth studies of training schemes and design flows. This section provides a comparative analysis of the performance among various enhanced FF-ViT models.

From Table 2, it can be seen that the FF-ViT* (PWA + *a*) architecture proposed in [9], achieved 99.69% accuracy for seen and 9.26% for unseen plant disease identification. We show that with *rs*, FF-ViT (PWA + *rs* + *a*) outperforms FF-ViT* (PWA + *a*) on the unseen task, achieving an accuracy of 15.74%, with comparable performance in the seen class. The performance disparity shows that FF-ViT with *rs*

Table 2
Performance comparison between SOTA models and our proposed model on seen and unseen plant disease identification for PV dataset.

Model	Seen Top 1	Unseen Top 1	Time per epoch (min)	Total parameters
ViT (PWA)	99.52	4.17	17	86M
ViT (NPWA)	99.43	7.87	17	86M
DINO-ViT (NoFT)	98.42	12.50	7	87M
CMTL-ViT [8]	99.49	6.94	11	89.4M
FF-ViT* (PWA + a) [9]	99.69	9.26	27	200M
FF-ViT (PWA + rs + a)	99.60	15.74	47	200M
FF-ViT (PWA + rs)	**99.75**	19.44	47	200M
FF-ViT (NPWA + rs)	99.67	20.37	46	200M
CL-ViT	99.31	**32.41**	34	125M

PWA and NPWA are pixel-wise and non-pixel-wise data augmentation strategies respectively. NoFT is without fine-tune of DINO backbone. FF-ViT* (PWA + a) is original model architecture and training scheme from [9]. a and rs are the moving weighted sum and revised generation scheme for FF-ViT. The results for SOTA models are reproduced by using our training and testing data.

that learns a balance distribution based on the entire training dataset, is capable of learning features of each unique composition including unseen compositions.

The original training process [9] involved the integration of a weighting coefficient, a to sequentially optimize the concept feature extraction module and the composition feature extraction module. However, through a rigorous performance monitoring of our FF-ViT model, we observed that excluding the weighting coefficient, a, led to a noticeable performance improvement for seen and unseen identification tasks of 0.15% and 3.70% respectively. Henceforth, we deduce that simultaneous optimization of both modules is crucial to strengthen the neurons' co-adaptation, thus improving the overall feature learning capability across the entire network.

5.2. Comparison between different SOTA model

In this section, we compare our novel CL-VIT model against various state-of-the-art (SOTA) models. CL-VIT is designed based on the optimal values derived from our comprehensive studies, highlighting the superiority of NPWA over PWA and the effectiveness of SSL networks in capturing generalized features for unseen plant disease images. Detailed experiments on these sub-modules can be found in Section 5.3.

One notable finding is that our proposed CL-VIT model achieved the highest performance on the unseen task with a 32.41% accuracy. This demonstrates that the hybrid concept of combining SL and SSL can effectively improve feature learning compared to models trained solely with either approach. Such a hybrid model effectively leverages the best features of both training schemes, further enhancing the learned feature distribution to encompass both seen and unseen plant disease classes.

Another observation is that CL-VIT could surpass the best FF-ViT model in the unseen task but achieve comparable performance in the seen task. This shows that the synthetic unique compositions generated from our improved pairwise feature generation module in FF-ViT model are able to further optimize the learned feature distributions to delineate better decision boundaries for seen data. However, the synthetic unseen features in FF-ViT, which may not accurately represent the actual distribution of unseen plant disease features, lead to lower performance in the unseen task.

We also noticed that DINO-ViT (NoFT) outperformed ViT models for unseen identification tasks. This may be due to DINO-ViT's ability to extract more generalized features for downstream identification tasks, as reported in [42]. However, in the field of plant diseases, certain classes of plants or diseases may have notable visual similarities, posing a significant challenge to DINO-ViT's generalized features in distinguishing them effectively. As a result, DINO-ViT achieved the lowest performance among all models for the seen identification task with an accuracy of 98.42%.

Additionally, we perform a comparative analysis of computing resources, evaluating particularly time per epoch and total number of parameters for all models. Our results indicate that the CL-ViT model shows superior efficiency, with a 27.7% reduction in time per epoch and a 37.5% decrease in total number of parameters compared to the FF-ViT model (PWA + rs), which is the best-performing FF-ViT variant. This efficiency is attributed to the design of the FF-ViT model, which learns plant and disease features separately, requiring dual feature extractors. Conversely, the CL-ViT model uses a unified approach, which learns both features simultaneously, improving computational efficiency. Furthermore, the results presented in Table 2 also demonstrate that the CL-ViT model is capable of learning more generalized features than the FF-ViT model.

5.3. Pixel-Wise (PWA) and Non-Pixel-Wise (NPWA) data augmentation

In this section, we conduct a thorough analysis of pretext tasks relevant to plant disease identification. Specifically, we compare two different data augmentation strategies, NPWA and PWA. The NPWA consists of random resize crop, transpose, horizontal flip, vertical flip, shift scale rotation, hue saturation value and random brightness contrast whereas the PWA is only rotation, flipping and image patches randomization. We initially compare these two data augmentation strategies on the baseline model ViT, which serves as the backbone for the FF-ViT and CL-ViT models. Subsequently, we analyze the performance differences in the improved FF-ViT model.

According to Table 2, ViT (NPWA) shows a slight decrease in accuracy on the seen task by 0.09% but demonstrates an improvement of 3.7% in the unseen task when compared to ViT (PWA). A similar trend can be seen in FF-ViT (NPWA + rs), where the seen task suffers a degradation of 0.08%, while the unseen task improves by 0.93% when compared to FF-ViT (PWA + rs). The performance disparity is attributed to the utilization of PWA or NPWA in the model training process. PWA enforces the model to learn class-specific features at a pixel-wise granularity. However, such fine-grained details may not be robust or appropriate for unseen plant disease classes. Despite discriminative features being learned for the seen classes, the characteristics that are learned without incorporating high-level visual concepts fail to generalize across unseen plant disease data, often occurring in different domains with variations in visual appearance. Therefore, we deduce that the importance of retaining the high-level concept, with our proposed NPWA, is much more effective in learning features that are more generalized across both seen and unseen plant disease data. These important findings are incorporated into our new CL-VIT model.

5.4. Supervised vs. self-supervised learning

Motivated by recent developments in the effectiveness of SSL networks, we explore their effectiveness in learning plant disease-related features. Specifically, we compare the recently proposed architecture DINO [42] with the most basic SL model to study the performance disparity between them for the plant disease identification task. Before diving into the comparative evaluation, it is essential to point out that ViT (PWA) and DINO-ViT (NoFT) have the same backbone architecture. In addition, the ViT and FF-ViT models were trained using SL, unlike the DINO-ViT (NoFT) model, which was trained using a SSL approach.

It is evident from Table 2 that for the seen identification task, ViT (PWA) outperformed DINO-ViT (NoFT) with an accuracy of 1.10%. For the unseen identification task, on the other hand, we observe considerable differences in model performance between the two learning approaches. ViT (PWA) displays a decrease in performance with an accuracy of 4.17%, while DINO-ViT (NoFT) performs better, with an accuracy of 12.50%. The observed variance in performance can be

Table 3
Performance comparison between SOTA models and our proposed model on plant disease identification for real-world plant disease dataset. The test set consists of both seen and unseen classes.

Model	Top 1
ViT (PWA)	60.62
DINO-ViT (NoFT)	48.40
CL-ViT	**61.31**

attributed to the different learning methods employed to learn feature distributions.

SL (ViT (PWA)) primarily focuses on learning class-specific features to construct decision boundaries for distinguishing different classes, as evidenced in previous research [43]. However, when unseen features exhibit slight deviations from seen features, pre-established decision boundaries may inadequately capture the feature distributions of both seen and unseen classes. This is in contrast to SSL (DINO-ViT (NoFT)), where the model's feature extraction modules are learned without strict constraints on class labels, resulting in more generalized adaptability to diverse variant tasks. However, it is important to note that the feature extractors, without proper guidance to adapt to the target dataset, may impact their effectiveness. Despite DINO's excellence in unseen tasks, it shows a performance deficit compared with the improved FF-ViT model, which is trained solely with a supervisory approach. The FF-ViT model achieves an accuracy of 19.44%, which is 6.94% higher than the DINO-ViT model.

5.5. Real-world plant disease dataset

In this study, we also assess the models' performance using real-world plant disease dataset (challenging dataset proposed in [8]). The real-world plant disease dataset consists of a mixture of both the laboratory image and the field. The real-world plant disease dataset is used to evaluate the effectiveness of the models in learning generalized features for real-world plant disease identification tasks.

To conduct a thorough comparative analysis, we benchmarked our model against ViT (PWA) representing SL and DINO-ViT for SSL. The results in Table 3 reveal that the CL-ViT model achieves an accuracy of 61.31%, surpassing the ViT (PWA) model by 0.69% and outperforming the DINO-ViT model by a significant margin of 12.91%. Notably, the DINO-ViT model, relying solely on generalized features extracted through SSL, records the lowest accuracy at 48.40%. This underscores the effectiveness of our novel cross-learning module in regulating essential generalized and discriminative conceptual features for real-world plant disease identification tasks. Specifically, SL guides the feature extractor to emphasize plant or disease-specific features, while SSL further reinforces the model to learn generalized features applicable to diverse sets of images captured in varying environments.

6. Ablation study of CL-ViT

In this section, we provide an empirical evaluation of our proposed CL-VIT model. We assess the model performance under various conditions.

6.1. Importance of unseen data in training pipelines

In this experiment, we analyze the impact of incorporating unlabeled data in our training pipelines for the CL-ViT model. The unlabeled data is utilized to leverage the features learned through the self-supervised pipeline task. As depicted in Table 4, we observe that CL-ViT with unseen data in the training pipelines improves seen and unseen identification tasks by 0.09% and 8.8% respectively. This indicates that the inclusion of unseen data features can enhance the feature spaces and domain invariance of our model. The rationale behind this improvement lies in the fact that prior knowledge gained from unseen data contributes to refining the unified feature mapping, thereby reducing the domain gap between seen and unseen data.

Table 4
The ablation studies of CL-ViT model.

Unseen training data	Cross attention	α	β	Feature dimension	Seen Top 1	Unseen Top 1
	✓	1.0	0.5	192	99.22	23.61
✓		1.0	0.5	192	99.40	23.15
✓	✓	1.0	0.5	192	99.31	**32.41**
✓	✓	0.5	1.0	192	98.92	11.57
✓	✓	1.0	1.0	192	98.96	29.17
✓	✓	1.0	0.5	768	**99.41**	21.76
✓	✓	1.0	0.5	384	99.35	16.20
✓	✓	1.0	0.5	96	99.01	31.94

6.2. Importance of cross attention

In this section, we investigate the importance of cross-attention in our cross-learning module in CL-ViT. The cross-learning module aims to learn and regulate SL and SSL features to form unified features for downstream tasks. According to Table 4, CL-ViT with the cross-attention layer achieved comparable performance on the seen identification task and improved on the unseen identification task by 9.26%. This observation clearly demonstrates that the cross-attention layer in the cross-learning module, which uses different SL or SSL task features for each task, is able to effectively optimize unified features for both seen and unseen identification tasks.

6.3. Comparison in feature dimension

In this experiment, we study the impact of the different dimensions of unified features (The features after cross attention (CAL) in Fig. 3). The results show a notable trend: a drop in performance for the seen identification task while observing an improvement in the unseen identification task as the feature dimension decreases. This can be attributed to the fact that a reduced feature embedding size leads to features that are less biased towards the seen class, promoting generalization to the unseen class. A higher feature dimension may contribute to model overfitting on the seen data. However, the performance in the unseen identification task plateaus when the feature dimension is further reduced. This observation suggests that an excessively small feature dimension may inadequately represent the distribution of features learned from the SL and SSL tasks. Consequently, we determine that our CL-ViT model achieves optimal performance when the unified feature dimension is set to 192.

6.4. Comparison in weighting coefficients

In this experiment, we analyze the impact of weighting coefficients, α and β, in our loss function (Eq. (14)) for CL-ViT. The loss function consists of supervised and self-supervised losses, and the weighting coefficients, α and β, serve as additional regulators to optimize the unified features appropriately reweighted between supervised and self-supervised features. Our results indicate that setting $\alpha = 1.0$ and $\beta = 0.5$ achieves the highest performance for both seen and unseen identification tasks.

6.5. Comparison in batch size

In Fig. 6, we compare the performance of CL-ViT for unseen plant disease identification with different batch sizes. Additionally, we include the previously proposed FF-ViT (PWA + *rs*) in this analysis to provide a performance comparison between the two models. The models are trained using the PV dataset, and we intentionally set the upper limit for batch size at 294, corresponding to the total unique compositions derived from the Cartesian product of the plant and disease classes. The results exhibit a notable trend: an increase in batch size corresponds to improved performance for both models. However,

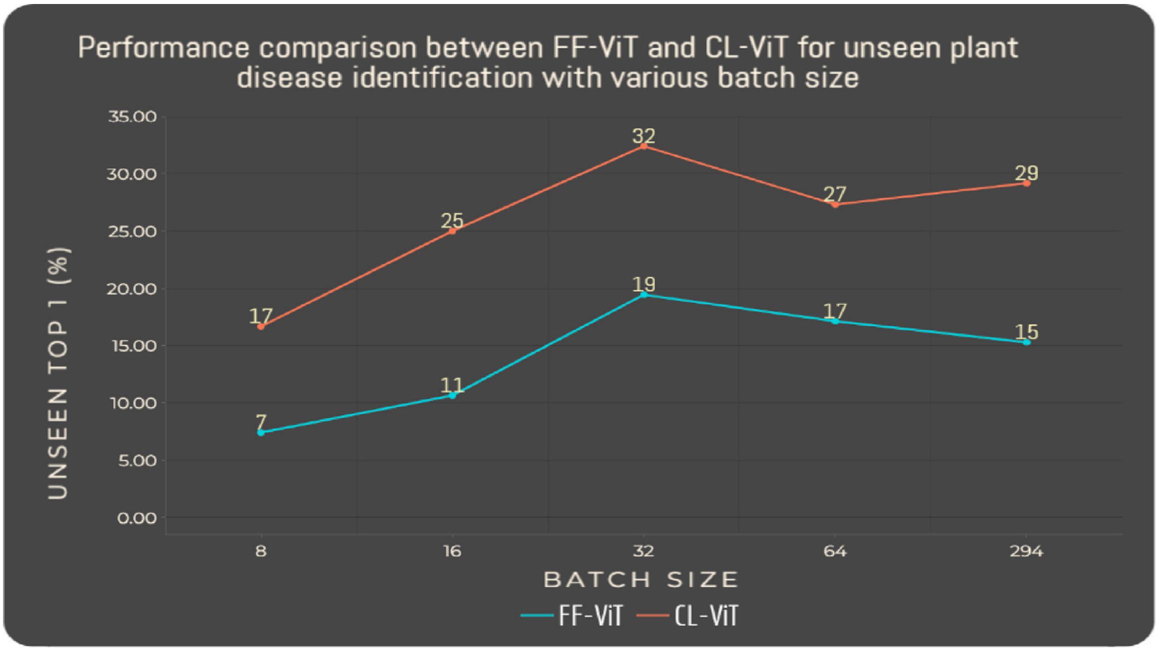


Fig. 6. The performance comparison between FF-ViT and CL-ViT for unseen plant disease identification with various batch size.

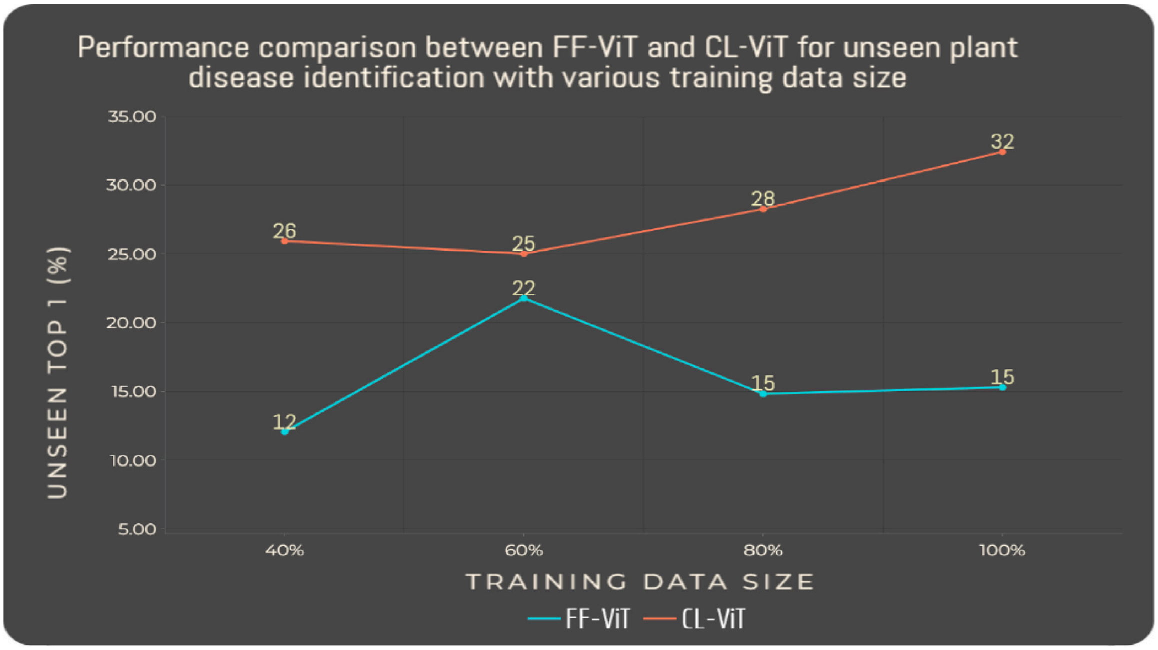


Fig. 7. The performance comparison between FF-ViT and CL-ViT for unseen plant disease identification with various training data size.

the optimal performance for both models is achieved at a batch size of 32. This suggests that the impact of batch size is closely correlated with the characteristics of the dataset. Furthermore, in all batch size configurations, the CL-ViT model consistently outperforms the FF-ViT model, highlighting its superior performance in this context.

6.6. Comparison in size of training data

In Fig. 7, we evaluate the performance of CL-ViT in identifying unseen plant diseases with different training data sizes. We also include the previously proposed FF-ViT (PWA + *rs*) in this analysis to further justify the superiority of the models. We prepare four training datasets corresponding to different ratios of the original training PV dataset (40%, 60%, 80% and 100%). The results show a trend where the performance of both models can be improved by using a larger training data size. However, the increase in performance of the CL-ViT model is greater than that of the FF-ViT model, suggesting that our CL-ViT model is capable of learning a better feature distribution with larger training data. Furthermore, the CL-ViT model regularly outperforms the FF-ViT model in all training data size configurations, underlining its dominance in this context.

7. Visualization analysis

In this section, we present feature distribution visualizations to offer additional justification for the performance differences among various

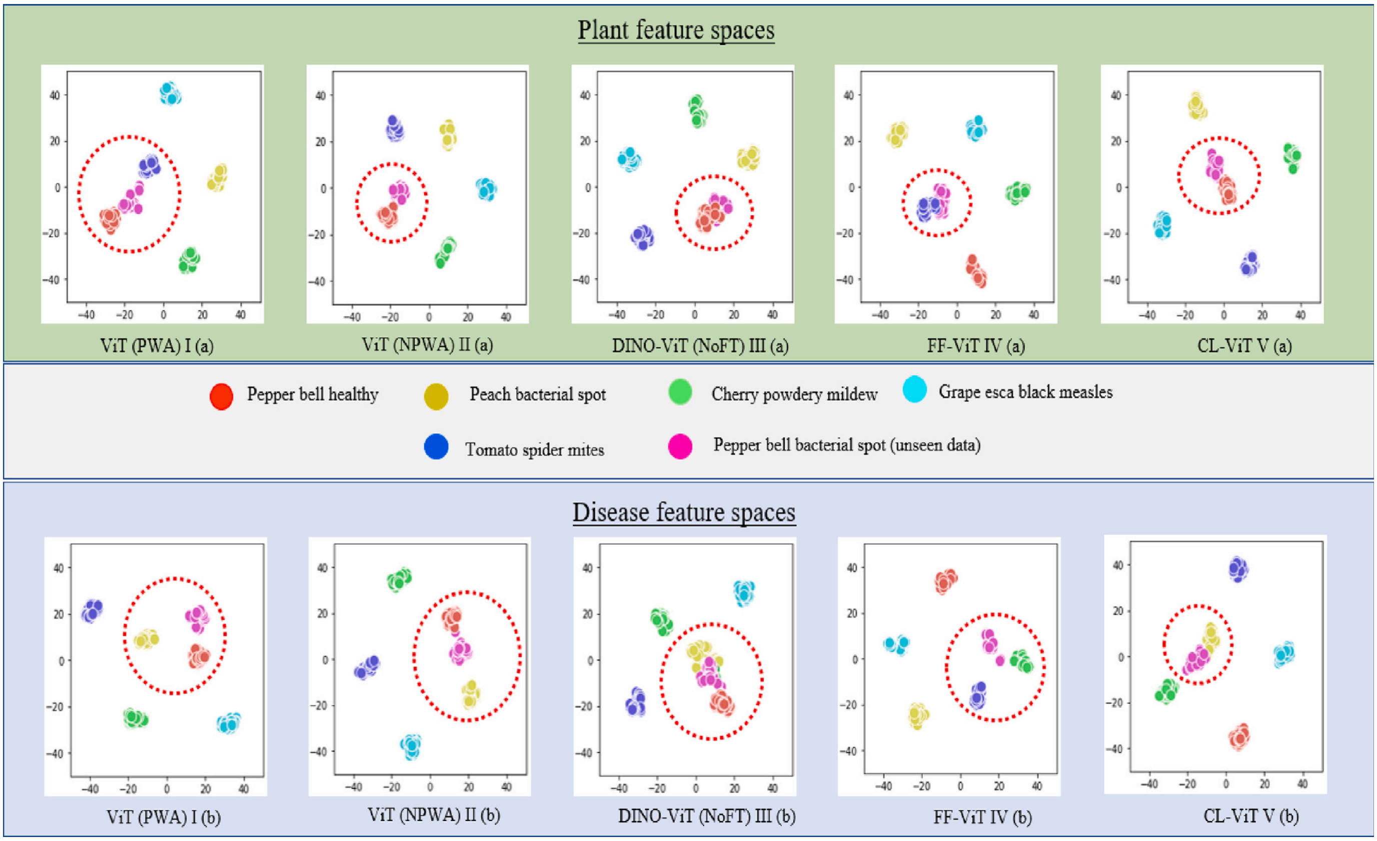


Fig. 8. The figure shows plant (a) and disease (b) feature maps for different models from T-SNE. *Pepper bell bacterial spot* (pink) is our testing unseen data. For plant feature spaces (a), *Pepper bell healthy* is considered as relevant seen data due to similar plant concept (*Pepper bell*). For disease feature spaces (b), *peach bacterial spot* is considered as relevant seen data due to similar disease concept (*bacterial spot*). The vertical and horizontal axes of each t-SNE map denote the two-dimensional embedding coordinates of the sample images, where similar images or features are positioned closer together, indicating their proximity in the high-dimensional feature space.

model variants. We utilized T-SNE [44] to visualize the plant and disease categorical feature embeddings. We randomly select images from five classes of seen data and one class of unseen data. We ensure that the classes selected included ‘relevant’ seen data, i.e. data that shared similar plant or disease concepts to our unseen data.

7.1. Pixel-wise and non-pixel-wise data augmentation

Fig. 8 illustrates the plant and disease embedding space learned by all models using t-SNE [44]. We first explore the impact of employing distinct data augmentation techniques on feature representation learning. To illustrate this, we first direct attention to the features learned by the ViT model shown in Fig. 8’s I and II. The feature distributions I and II correspond to the plant (a) and disease (b) feature spaces for PWA and NPWA respectively.

Firstly, we observe that both models can learn class-specific features where features corresponding to each class exhibit distinct clustering without perceptible overlap, and their performance in terms of seen accuracy is comparable in Table 2. However, when evaluating the feature distances between seen and unseen data, the features corresponding to the *Pepper bell bacterial spot* (unseen data) in feature distributions I(a) exhibit closer proximity to both relevant *Pepper bell healthy* and irrelevant *Tomato spider mites*. Conversely, in feature distributions II (a), the features of *Pepper bell bacterial spot* are primarily aligned with the relevant *Pepper bell healthy*. We also observe a similar pattern in disease feature spaces.

The discrepancy arises from the inherent characteristics of feature learning in the two approaches, PWA and NPWA when faced with data with slight distribution shifts. For example in Fig. 9, the visual appearances of the data exhibit variations even for the same disease. The *Peach bacteria spot* showcases symptoms of brown spots spreading around the leaf blade with substantial deviations in color. In contrast, *Pepper bacteria spot* displays brown spots primarily concentrated on a greenish leaf blade, illustrating the distinct visual characteristics within the dataset. The model employing PWA, with its emphasis on fine-grained features, may concentrate on capturing pixel-wise details, potentially limiting generalization across diseases with similar symptoms. On the other hand, NPWA, which correlates structural information and learns long-range pattern features, offers greater adaptability, making it more suitable for generalizing across plant samples exhibiting similar disease symptoms. Based on this insight, we opt to incorporate NPWA into our CL-ViT model.

7.2. Supervised and self-supervised learning

In this section, we investigate the extent to which SSL methods can reinforce the underlying structure of learned features in plant disease data. First, we draw attention to the features learned using DINO-ViT (NoFT) with SSL, which claims to be able to learn generalized features for the downstream tasks [42] in Fig. 8’s III (a) and (b). From the figure, we can observe that the feature representation learned solely by SSL is able to show distinct clustering between seen data in plant feature space, but is unable to distinguish between seen data in disease feature space. (overlap between seen data in circled region). This may be due to the over-generalization of features learned through SSL, making them insufficiently discriminating for precise identification tasks in the disease domain. Nevertheless, these generalized features are still effective in reducing the gap between relevant seen and unseen data distances between relevant seen and unseen features are closer than those of the ViT model with SL as shown in III (a) and III (b). This prompted us to incorporate SSL into our final CL-ViT model.

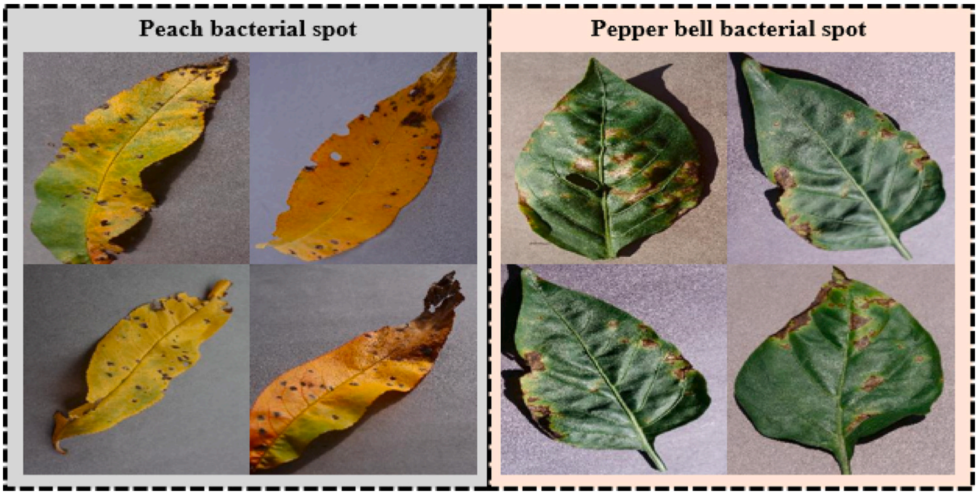


Fig. 9. The figure above shows a comparison of symptoms for *bacterial spot* disease between *peach* and *pepper bell* plant.

When comparing FF-ViT and CL-ViT, distinct clustering is evident among each set of data. Notably, the feature distribution's distance between relevant seen and unseen data is closer for the CL-ViT model than FF-ViT model. This difference is clearly illustrated in the disease feature distributions of FF-ViT from IV(b) and CL-VIT from V(b). It can be observed that the features of *Pepper bell bacterial spot* (unseen) are close to both irrelevant disease classes, *cherry powdery mildew* and *tomato spider mites*, for the FF-ViT model, whereas they are only close to relevant disease class, *peach bacterial spot* for the CL-ViT model. This divergence may stem from FF-ViT generating synthetic unseen data as novel entries into the feature space, learning new decision boundaries without concurrent optimization of data distributions based on the actual plant disease characteristics. In contrast, the CL-ViT model optimizes the distributions between each set of data with our novelty cross-attention module, ultimately reducing the feature distance between seen and unseen data.

8. Conclusion

We introduce a novel approach, the CL-ViT model, which represents a significant advancement through its innovative SSL framework, establishing a new benchmark in the field of unseen plant disease identification. Specifically, we emphasize the crucial role of SSL in guiding the model to learn and leverage not only the seen but also unseen classes, thereby reducing the feature distribution gap. Our CL-ViT outperforms the existing FF-ViT model, which relies solely on a supervisory learning scheme, with a substantial improvement of 23.15%. Our visualization results further validate the effectiveness of CL-ViT in learning a feature space capable of discriminating between different classes while minimizing the domain gap between seen and unseen data compared to other existing approaches. Thorough ablation studies have been conducted to determine the most optimal parameter setting for the best CL-ViT model performance.

Limitations and Future works: It is challenging to determine the most appropriate set of self-supervised tasks capable of learning generalized features for diverse plant disease identification tasks. Consequently, future research may be directed towards the design and formulation of self-supervised tasks capable of adapting to various settings within the domain of plant disease identification. Another limitation is the inadequate analysis of the model's stability and robustness to changing environmental conditions, including weather, soil health, and agricultural practices. Overcoming this requires diverse datasets and extensive data collection, as planned in the Pl@ntAgroEco project.[1]

CRediT authorship contribution statement

Abel Yu Hao Chai: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Sue Han Lee:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Fei Siang Tay:** Writing – review & editing, Validation, Supervision, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Pierre Bonnet:** Writing – review & editing, Validation, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Alexis Joly:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Abel Yu Hao Chai reports financial support was provided by Malaysia Ministry of Higher Education. Sue Han Lee reports financial support was provided by Malaysia Ministry of Higher Education. Sue Han Lee reports financial support was provided by Swinburne University of Technology - Sarawak Campus. Abel Yu Hao Chai reports equipment, drugs, or supplies was provided by NEUON AI SDN BHD. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We appreciate the comments and advice from Hervé Goëau and Yi Lung Then on our study and drafts. This research is supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant (Ref: FRGS/1/2021/ICT02/SWIN/03/2), from the Ministry of Higher Education Malaysia and Swinburne Sarawak Research Supervision Grants (SSRSG) (Ref: SUTS/SoR/RMC/SSRGS/2023). This work also benefited from a government grant managed by the Agence Nationale de la Recherche as part of the France 2030 programme, through the Pl@ntAgroEco project, under the reference "ANR-22-PEAE-0009". We also gratefully acknowledged the support of NEUON AI with the GPU workstation used for this research.

[1] https://plantnet.org/en/plantagroeco-2/.

References

[1] T. Wiesner-Hanks, E.L. Stewart, N. Kaczmar, C. DeChant, H. Wu, R.J. Nelson, H. Lipson, M.A. Gore, Image set for deep learning: field images of maize annotated with disease symptoms, BMC Res. Notes 11 (1) (2018) 1–3.

[2] X. Liu, W. Min, S. Mei, L. Wang, S. Jiang, Plant disease recognition: A large-scale benchmark dataset and a visual region and loss reweighting approach, IEEE Trans. Image Process. 30 (2021) 2003–2015, http://dx.doi.org/10.1109/TIP.2021.3049334.

[3] S.P. Mohanty, D.P. Hughes, M. Salathé, Using deep learning for image-based plant disease detection, Front. Plant Sci. 7 (2016) 1419.

[4] Y. Xian, S. Sharma, B. Schiele, Z. Akata, F-vaegan-d2: A feature generating framework for any-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10275–10284.

[5] D. Huynh, E. Elhamifar, Compositional zero-shot learning via fine-grained dense feature composition, Adv. Neural Inf. Process. Syst. 33 (2020) 19849–19860.

[6] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, L. Shao, Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning, IEEE Trans. Image Process. 29 (2020) 3665–3680, http://dx.doi.org/10.1109/TIP.2020.2964429.

[7] S.H. Lee, H. Goëau, P. Bonnet, A. Joly, New perspectives on plant disease characterization based on deep learning, Comput. Electron. Agric. 170 (2020) 105220.

[8] S.H. Lee, H. Goeau, P. Bonnet, A. Joly, Conditional multi-task learning for plant disease identification, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 3320–3327.

[9] A.Y.H. Chai, S.H. Lee, F.S. Tay, Y.L. Then, H. Goëau, P. Bonnet, A. Joly, Pairwise feature learning for unseen plant disease recognition, in: 2023 IEEE International Conference on Image Processing, ICIP, IEEE, 2023, pp. 306–310.

[10] Y. Sun, E. Tzeng, T. Darrell, A.A. Efros, Unsupervised domain adaptation through self-supervision, 2019, arXiv preprint arXiv:1909.11825.

[11] J. Villacis-Llobet, H. Goëau, P. Bonnet, A. Joly, E. Mata-Montero, Domain adaptation in the context of herbarium collections: A submission to plantclef 2020, in: Conference and Labs of the Evaluation Forum, 2020, URL https://api.semanticscholar.org/CorpusID:225073609.

[12] F.J. Piva, G. Dubbelman, Exploiting image translations via ensemble self-supervised learning for unsupervised domain adaptation, Comput. Vis. Image Underst. 234 (2023) 103745.

[13] R. Zhao, Y. Zhu, Y. Li, CLA: A self-supervised contrastive learning method for leaf disease identification with domain adaptation, Comput. Electron. Agric. 211 (2023) 107967.

[14] A. Picon, A. Alvarez-Gila, M. Seitz, A. Ortiz-Barredo, J. Echazarra, A. Johannes, Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild, Comput. Electron. Agric. 161 (2019) 280–290.

[15] S. Chulif, S.H. Lee, Y.L. Chang, K.C. Chai, A machine learning approach for cross-domain plant identification using herbarium specimens, Neural Comput. Appl. (2022) 1–23.

[16] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, E. Vezzetti, Vision transformer for femur fracture classification, Injury (2022).

[17] S. Cuenat, R. Couturier, Convolutional neural network (CNN) vs vision transformer (ViT) for digital holography, in: 2022 2nd International Conference on Computer, Control and Robotics, ICCCR, IEEE, 2022, pp. 235–240.

[18] D. Vijayreddy, Classification of Plant Diseases, Characteristics and Symptoms on Crop Plants, Golden Leaf Publishers, 2024, pp. 171–183, Ch. 18.

[19] M.B. Riley, M.R. Williamson, O. Maloy, Plant disease diagnosis, Plant Health Instr. 10 (2002).

[20] S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, L. Shao, Free: Feature refinement for generalized zero-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 122–131.

[21] S. Shao, L. Xing, R. Xu, W. Liu, Y.-J. Wang, B.-D. Liu, MDFM: Multi-decision fusing model for few-shot learning, IEEE Trans. Circuits Syst. Video Technol. 32 (8) (2021) 5151–5162.

[22] D. Kim, Y. Yoo, S. Park, J. Kim, J. Lee, Selfreg: Self-supervised contrastive regularization for domain generalization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9619–9628.

[23] S. Chen, J.-H. Xue, J. Chang, J. Zhang, J. Yang, Q. Tian, SSL++: Improving self-supervised learning by mitigating the proxy task-specificity problem, IEEE Trans. Image Process. 31 (2021) 1134–1148.

[24] Z. Qin, X. Lu, X. Nie, Y. Yin, J. Shen, Exposing the self-supervised space-time correspondence learning via graph kernels, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 2110–2118.

[25] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[26] Y. Zhang, B. Sun, J. He, L. Yu, X. Zhao, Multi-level neural prompt for zero-shot weakly supervised group activity recognition, Neurocomputing 571 (2024) 127135.

[27] X. Lu, W. Wang, J. Shen, Y.-W. Tai, D.J. Crandall, S.C. Hoi, Learning video object segmentation from unlabeled videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8960–8970.

[28] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, Springer, 2016, pp. 69–84.

[29] X. Zhai, A. Oliver, A. Kolesnikov, L. Beyer, S4l: Self-supervised semi-supervised learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1476–1485.

[30] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, I. Sutskever, Generative pretraining from pixels, in: International Conference on Machine Learning, PMLR, 2020, pp. 1691–1703.

[31] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.

[32] L. Li, T. Ma, Y. Lu, Q. Li, L. He, Y. Wen, A multi-grained unsupervised domain adaptation approach for semantic segmentation, Pattern Recognit. 144 (2023) 109841.

[33] C. Tang, X. Zeng, L. Zhou, Q. Zhou, P. Wang, X. Wu, H. Ren, J. Zhou, Y. Wang, Semi-supervised medical image segmentation via hard positives oriented contrastive learning, Pattern Recognit. 146 (2024) 110020.

[34] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (3) (2021) 107–115.

[35] S. Arora, R. Ge, B. Neyshabur, Y. Zhang, Stronger generalization bounds for deep nets via a compression approach, in: International Conference on Machine Learning, PMLR, 2018, pp. 254–263.

[36] H. Shen, Z. Tang, Y. Li, X. Duan, Z. Chen, HAIC-NET: Semi-supervised OCTA vessel segmentation with self-supervised pretext task and dual consistency training, Pattern Recognit. (2024) 110429.

[37] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, Int. Conf. Learn. Represent. (2021).

[38] S.H. Lee, C.S. Chan, P. Remagnino, Multi-organ plant classification based on convolutional and recurrent neural networks, IEEE Trans. Image Process. 27 (9) (2018) 4287–4301.

[39] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

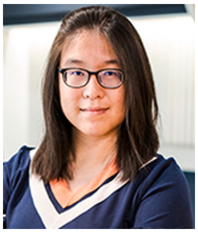
[41] H. Alshammari, K. Gasmi, I. Ben Ltaifa, M. Krichen, L. Ben Ammar, M.A. Mahmood, Olive disease classification based on vision transformer and CNN models, Comput. Intell. Neurosci. 2022 (2022).

[42] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[43] R. Sarussi, A. Brutzkus, A. Globerson, Towards understanding learning in neural networks with linear teachers, in: International Conference on Machine Learning, PMLR, 2021, pp. 9313–9322.

[44] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

Abel Yu Hao Chai earned his B.Eng. (Hons) and Master in research (MARes) from Swinburne University of Technology. He is currently pursuing Ph.D. in Swinburne University of Technology. His current research interest include deep learning, machine learning, computer vision and computational botany.



Sue Han Lee earned her B.Eng. (Hons) from Multimedia University, M.Sc. from Shinshu University, and Ph.D. from the University of Malaya. She specializes in computer vision and machine learning for plant species and disease identification. Currently, she is a lecturer at Swinburne University, actively contributing to computational botany research.

Fei Siang Tay received his BEng (Hons) degree in Electrical and Computer Systems Engineering from Monash University Malaysia, and his Ph.D. in Control Engineering from Swinburne University of Technology, Melbourne, Australia. His current research interests include intelligent control systems, deep learning, smart farming, and dynamic fuzzy modeling. He is a registered Professional Technologist with MBOT and a Senior Member of IEEE.

Pierre Bonnet received Ph.D. in Plant Biology from the University of Montpellier in 2008. He is a researcher in CIRAD. His work focuses on the identification and characterization of plants on a large taxonomic and geographical scale. He has been coordinator of the Pl@ntNet project since 2009.

Alexis Joly received Ph.D. in computer science at University of La Rochelle. He is a researcher at Inria working on the challenges of multimedia information retrieval with related interests in machine learning and computer vision. He created the international research forum LifeCLEF and co-responsible for the Pl@ntNet project.