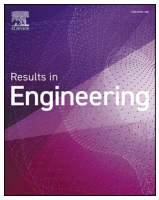
Results in Engineering 25 (2025) 103922

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Attentive Self-supervised Contrastive Learning (ASCL) for plant disease classification

Getinet Yilma, Mesfin Dagne, Mohammed Kemal Ahmed, Ravindra Babu Bellam *

Department of Computer Science and Engineering, Adama Science and Technology University, Ethiopia

ARTICLE INFO

Keywords:
Plant disease classification
Attentive self-supervised representation
Contrastive learning

ABSTRACT

Deep-learning plays a crucial role in large-scale health monitoring of agricultural plants. One of the challenges in plant disease classification is the limited availability of annotated training data, where supervised deep feature learning typically excels. However, traditional deep learning backbones often produce representations that are not sufficiently discriminative or interpretable for detailed plant disease analysis. We propose an Attentive Self-supervised Contrastive Learning (ASCL) framework that leverages transferable representations as supervision signals. The ASCL framework enhances interpretability by incorporating attention mechanisms, such as squeeze-excitation and convolutional block attention module, which highlight key regions in plant images, aiding in transparent decision-making. In the present work, a pre-trained squeeze-excitation ResNet50 Siamese backbone network on the unlabeled PlantVillage dataset was used to validate the generalizability of the learned representations. The pre-trained weights were then fine-tuned on small-scale unseen datasets extracted from the 17-class PlantVillage Taiwan Tomato and Apple datasets. Despite involving fewer than 17 classes, the high variability within each class, such as the disease progression stages, underscores the fine-grained nature of the classification task. Extensive experiments demonstrated that the ASCL framework achieved 89 % accuracy, outperforming a baseline supervised model that scored 88.9 %. Moreover, when the ASCL learned, weights were transferred to an unseen dataset, and the model achieved 93.5 % accuracy, compared to 91.66 % with supervised ResNet50. The framework is scalable to larger datasets with more classes, making it applicable to broader fine-grained classification tasks. Therefore, the proposed ASCL framework demonstrates the generalizability and transferability of the downstream plant disease classification tasks.

1. Introduction

Plant health monitoring is the key to agronomists' success [1], ensuring crop productivity, food security [2], food health, and stable ecological agroecosystems [3] [4]. Biochemical and convolutional neural network (CNNs) solutions have become significant in plant disease-recognition techniques [5]. Biochemical-based diagnosis is the most expensive disease identification method because of chemical reagents, expert requirements, and the difficulty of covering massive agricultural lands. However, relatively cheaper plant disease identification and analysis techniques that extract plant leaf image features through deep CNN techniques have been proposed. This technique helps minimize resources compared to the biochemical methods. Recently, deep learning researchers collected and annotated plant leaf images under controlled environments, such as greenhouses and shades, to train such deep learning techniques, limiting training sample diversity, creating class imbalance, and creating image characteristics that might highly affect disease recognition performance. In addition, annotating images requires domain experts, which are labor-intensive and impossible to scale for diverse plant species, and training deep learning models [6] [7], [8], [9], [10],.

To address the training, as mentioned above, and sample challenges, different authors have attempted to maximize the training sample through augmentation and advanced generation techniques [6,11,8]. However, the sample geometric transformation technique usually duplicates existing training samples, and some transformations, such as cropping, zooming, and extreme distortions, deform the structure of the disease spot in the images. Sometimes, the transformation focuses on the image background instead of the diseased spot [9,8,12]. In contrast, advanced generation techniques, such as generative adversarial networks (GANs) [13], are subjected to mode collapse unless we introduce class labels as a condition to guide the generation process that leads to

* Corresponding author.
E-mail address: ravindrababu4u@yahoo.com (R.B. Bellam).

https://doi.org/10.1016/j.rineng.2025.103922
Received 21 August 2024; Received in revised form 28 December 2024; Accepted 1 January 2025
Available online 19 January 2025

plausible sample generation. However, the generator can still quickly memorize class labels instead of learning the semantic meaning of images [9,12],–[15]. The second mechanism involves learning automatic supervision signals using self-supervised methods from unlabeled images. Self-supervised contrastive learning is critical for the automatic annotation of training samples by learning a generic semantic representation from unlabeled images [14,15,16,17,18,19,20],. This technique helps to learn the inherent structure of unlabeled data and thereby captures representative semantic features of the training dataset, which is crucial for fine-tuning downstream classification tasks. The authors of [18,21] developed self-supervised contrastive learning using a domain adaptation method for plant disease classification. However, the learned representations in such methods are not sufficiently discriminative for localized disease-spot classification tasks. Furthermore, the work in [21] was pretrained and tested on a single dataset [22]. The third is learning discriminative features using attention mechanisms [23]; this technique assists the model in capturing salient intermediate features, which improves the plant disease classification performance in small-scale labeled datasets [23,24]. However, attention mechanisms generalize less to the training samples collected from different environments.

Using the available massive unlabeled dataset to train an attentive self-supervised framework can help extract generic discriminative feature embedding and can be transferred to small-scale plant disease recognition. Hence, this study proposes the ASCL framework, which benefits from generic self-supervised contrastive feature learning that captures similar semantic embedding and attention mechanisms for salient transferable feature learning. The contributions of the present work are as follows. (i) Proposed an ASCL framework to learn semantic feature representations from unlabeled plant disease images. (ii) To further improve the proposed method, we introduced attention blocks into the Siamese ASCL framework to learn attentive feature representations and classifications. (iii) We evaluated the effectiveness of the ASCL framework by pretraining the generic Plant Village (tomato, maize, and potato) dataset [22]. We then transferred the ASCL pre-trained weights to the Plant Village subset of the Apple dataset [22] and a challenging tomato dataset collected from wild agricultural fields [25].

2. Related work

Plant health monitoring and disease recognition are critical challenges for crop productivity, food security, health, and stable ecological agroecosystems [4]. To address these research gaps, supervised deep-learning-based plant disease detection and classification have been proposed [6,7,8]. However, these methods generalize well for large, diverse, balanced human-annotated training samples. To improve training sample challenges, other studies proposed data generation and augmentation mechanisms [4] [6,9,11,8]. On the other hand, recent works have proposed deep-learning architectural design choices that help improve feature learning [9,11]. Furthermore, some studies have introduced multi-stage deep learning architectures to improve feature extraction, bounding box generation, and segmentation [11,23]. Recently, attention mechanisms have been designed to effectively detect fine-grained details in small-scale data via discriminative feature learning [23,26]. However, most of these studies are trained using supervised deep learning methods, which cannot be scaled owing to limited training data. A mechanism to utilize existing large unlabeled images from different sources is necessary for scalable recognition and plant health surveillance [14].

Self-supervised representation learning from unannotated images generates pseudo-labels using a contrastive objective function. The contrastive objective function predicts a pair of similar samples out of a mini-batch of augmented images to learn semantic features from unlabeled images [14,17,27]. Self-supervised Contrastive learning methods have been extensively evaluated for object recognition vision applications and have demonstrated exemplary performance for coarse-level object classification [14]. Clustering plant images based on kernel k-means [16] uses a high-accuracy self-supervision representation learning method to cluster plant images, which still requires systematic analysis for further downstream tasks. Self-supervised contrastive learning using the xResNet34 model as a backbone for agricultural image classification and segmentation target tasks has improved performance [18]. However, these methods have been poorly studied for fine-grained recognition tasks such as plant disease detection and classification and are inefficient in learning discriminative features.

Attention mechanisms enable the model to learn discriminative targets, boosting the representation performance [23,26,28]. For example, in a squeeze-and-excitation (SE) network, the squeeze captures feature maps across $H \times W$ spatial dimensions using global average pooling, and the excitation block passes the squeeze output through a fully connected layer to capture channel-wise dependencies [29]. SE has excellent potential in discriminative feature learning but misses the potential representation power of spatial dimension information. The convolutional block attention module (CBAM) is another attention mechanism that attends channel-wise and spatial-wise at every convolutional block to enhance the CNN in learning discriminative features [30]. CBAM CNN can handle both channel and spatial attention information. However, SE and CBAM have been extensively evaluated for supervised object-recognition tasks. This study assessed self-supervision representation learning enhanced by these attention mechanisms for plant disease recognition.

3. Proposed approach

The proposed ASCL pretraining and downstream fine-tuning techniques are described in detail. The detailed ASCL framework for representation learning and plant disease classification downstream tasks is shown in Fig. 1. Fig. 1(a) illustrates that the ASCL framework extracts generic pseudo labels from large unlabeled training samples. The input sample image x is given to the ASCL framework, and then x is transformed via augmentation methods, producing x_i and x_j . The Siamese backbones generate intermediate features, $\overline{x}_i$ and, $\overline{x}_j$. These features are attentive intermediate features passed through the projection head H, usually a fully connected feature vectors z_i and z_j also called latent embedding, which is given to the contrastive loss. The task of contrastive loss is to optimize the difference between the features generated from the two projection heads. Fig. 1(b) The pre-trained weight from the ASCL framework is used to solve the downstream disease classification task. Fig. 1(c) Single-generic residual attention block. The residual block employed the ResNet50 [31] network, and the attention module was the Squeeze and Excitation module from [29].

3.1. ASCL framework

The ASCL framework structure employed a standard ResNet50 Siamese backbone network with a channel-wise attention module integrated at each ResNet50 basic and bottleneck layer. The ASCL framework assists in capturing generic automatic annotations from unlabeled images via self-supervised contrastive loss. The ASCL framework network structure displays two attentive residual Siamese backbones, taking a batch of transformed inputs to learn the semantic representation, which is then given to an expander projection head. The distance between the same transformed images was minimized, and latent embedding was used to generate a pseudo-cluster in the feature space. All the training samples are clustered in this way, keeping similar features categorized as the same annotation and dissimilar ones farther in the feature space.

Siamese networks are computationally more expensive than single-backbone networks. Unlike attention-augmented residual networks, the proposed ASCL framework employs standard squeeze and excitation modules to capture generic and discriminative features. The intuition

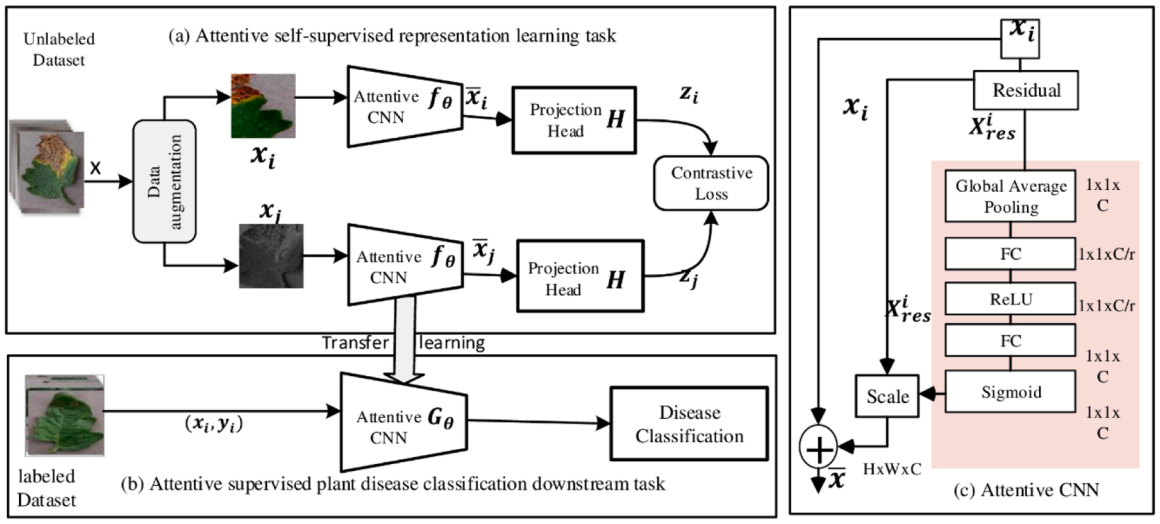


Fig. 1. ASCL Process: (a) ASCL framework uses contrastive loss to learn generic representation from unlabeled images $\boldsymbol{x}$. The $\boldsymbol{x_i}$ and, $\boldsymbol{x_j}$ are augmented samples from $\boldsymbol{x}$ taken as input to the ASCL framework, The $\bar{x}_i$ and $\bar{x}_j$ learned attentive intermediate features projection head $\boldsymbol{H}$, and $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ from the projection head are given to the contrastive loss. (b) The pre-trained weight from the ASCL framework is used to solve the downstream disease classification task. (c) Shows a single generic residual SE block.

behind this framework is to learn transferable supervision signals from the unlabeled images. Each branch in the Siamese backbone takes in a batch *of* 304 images and contrasts one image with 304 remaining images to get its similar pair via self-supervised contrastive loss, resulting in clusters of generic and similar features as pseudo labels to be transferred to downstream tasks. For the pre-training of the proposed method, the attentive ResNet50 Siamese network was used as a backbone to learn generic representations. A composition of robust random augmentation techniques, such as center crop, resize, crop-resize, zoom, color jitter, Gaussian blur, salt-and-pepper noise, flipping, and rotation, are applied to a single input image. This argument created two versions of the same anchor image. By utilizing these augmentations, contrastive learning [32] loss maximizes the feature similarity of the two augmented positive sample pairs, while minimizing the features of antagonistic pairs.

3.1.1. Advantages of the proposed ASCL framework

The proposed ASCL framework offers distinct advantages by integrating the attention mechanisms and self-supervised learning for plant disease classification. Attention mechanisms such as squeeze-and-excitation (SE) and Convolutional Block Attention Module (CBAM) improve feature selection, ensuring that critical areas, such as disease spots, are emphasized, boosting classification accuracy. These mechanisms also highlight disease-specific regions, enhancing both model performance and interpretability and making it easier for human experts to understand the model's predictions. Furthermore, attention modules help reduce the background noise, allowing the model to generalize better across diverse datasets. Self-supervised learning eliminates the need for extensive labeled data and effectively learns from vast numbers of unlabeled samples. This method also enables transferable representations, allowing models trained on one dataset to perform well on other datasets. Additionally, by generating pseudo-annotations, self-supervised learning helps address class imbalance and maintains strong performance, even with uneven class distributions in the data.

3.2. Problem formulation

A mini batch of augmented image samples $\mathrm{x_i}$ and $\mathrm{x_j}$, fed into the attentive residual Siamese backbone network $f_\theta(X_i)$. Which contains a residual function $F(x)$ that performs convolution, Batch Normalization (BN), and ReLU operations. The residual feature map learning process is formulated as follows (Eqs. (1) and (2)):

$$Xresi = F(x_i) + \mathrm{x_i} \tag{1}$$

$$Xresj = F(x_j) + \mathrm{x_j} \tag{2}$$

X^i_{res} and X^j_{res} refers to the feature map produced from the residual Siamese backbone.

Following the residual block, the method used squeeze-and-excitation (also called channel attention) block S_{se} that aggregate the inter-channel relationship to attend meaningful patterns [29]. The squeeze and excitation perform two major tasks; the first is the squeeze operation, which performs global average pooling to squeeze the input tensor C × H × W into C × 1 × 1 and extract the mean value of each channel across H × W spatial dimension. The S_{se} the module takes the output feature map $\{X^i_{res}, X^j_{res}\} \in R^{H \times W \times C}$ as an input and transform them into $\{X^i_c, X^j_c\} \in R^{H \times W \times C}$ feature vectors. The c in the variable X^i_c shows channel, the $S_{se}$is formulated as Eqs. (3)-4:

$$X_{c_i} = S_{se}\left(\sigma\left(W_{2\rho}\left(W_1\left(\frac{1}{W*H}\sum_{k=1}^{H,\,L} WX_{res_i kl}\right)\right)\right)W\right) \tag{3}$$

$$X_{c_i} = S_{se}\left(\sigma\left(W_{2\rho}\left(W_1\left(\frac{1}{W*H}\sum_{k=1}^{H,\,L} WX_{res_j kl}\right)\right)\right)W\right) \tag{4}$$

Where S_{se} is attentive squeeze-and-excitation block, σ is a sigmoid activation function, ρ is a ReLU activation $W_1 \in R^{\frac{c}{r}\times C}$ and $W_2 \in R^{\frac{c}{r}\times C}$ are learned weights from two fully connected layers and added non-linearity. C denotes channel, r is the reduction ratio set to a value <1. if the number of channels is given by C, then the number of channels after reduction would be $\frac{c}{r}\times C$. The attended output features obtained from S_{se} block are re-weighted through re-scale operations such as Eqs. (5)-7:

$$x_i = f_\theta(F_{re-scale}(Xres_i x_{ci}) + x_i) = f_\theta\left(X_{res_i} * x_{c_i} + x_i\right) \tag{5}$$

$$x_j = f_\theta\left(F_{re-scale}\left(Xres_j x_{cj}\right) + x_j\right) = f_\theta\left(X_{res_j} * x_{c_j} + x_j\right) \tag{6}$$

The $F_{re-scale}$ rescale factor is applied elementwise to the original feature maps, multiplying each channel's value by its corresponding weight. This process allows Siamese backbones to selectively amplify or

suppress specific channels based on their importance, thereby providing a mechanism for adaptive feature recalibration.

The encoded attentive features $\bar{x}_i$ and $\bar{x}_j$ are further transformed by the projection head H as $Z_i = H(\bar{x}_i) = MLP(\bar{x}_i)$ and $Z_j = H(\bar{x}_j) = MLP(\bar{x}_j)$ embedding. H is an MLP that contains two fully connected layers, each followed by BN and ReLU layers that assist in mapping the learned representation from the base encoder to 100-dimensional latent space. To compute the similarity between the learned features Z_i and Z_j the l_2 normalized temperature scaled cross-entropy loss [32] is employed, which is given as Eq. (7):

$$l_{ij} = -\log\left(\frac{\exp\left(\frac{sim(Z_i,\ Z_j)}{\tau}\right)}{\sum_{k=1,2,\ldots,N;k\neq i}\exp\left(\text{sim}\ (Z_i, Z_j)\Big/\tau\right)}\right) \tag{7}$$

Where $\{Z_i,\ Z_j\} \in R^n$ represents the high-level attended image representations, $sim(.)$ denotes the cosine similarity function, the $1[k \neq 1] \in \{0,\ 1\}$ evaluates to 1 *iff* $k \neq i$, and τ is a temperature parameter. After completing the pre-training, the projection head H is discarded, and the N-dimensional feature vectors are transferred to fine-tune the downstream task of target plant disease classification. The detailed architecture of the SE-ResNet50 backbone presented shows the input and outputs of the conv blocks, SE layer, and final projection head, as given in Table 1.

3.3. Fine-tuning ASCL for plant disease classification

Fine-tuning on small-scale labeled plant disease datasets was performed to evaluate the generalization capability of the pre-trained ASCL representations for subsequent downstream supervised tasks. Once pre-trained, the ASCL framework was applied to the unlabeled PlantVillage dataset [4] [22]. The projection head H shown in Table 0.2 has been removed, and the 2048 dimensional features vector is taken to fine-tune various small labeled datasets taken from PlantVillage [4] [22], such as Apple data from Plant Village [4] [22] and from different data sources Taiwan Tomato datasets [25].

Given a mini batch of N labeled images $x_i,\ y_i$ encoded via G_θ as $\bar{y}_i = G_\theta(x_i)$. The method freezes all the weights of the ASCL pre-trained backbone during the fine-tuning stage, and only the last layer is initialized randomly. The linear-layer cross-entropy loss function was then used to classify plant diseases. The formal cross-entropy loss is Eq. (8)

Table 1
The proposed ASCL SE-ResNet50 backbone architectural details.

Block	Input-output		# ResBlocks	Block	Input-output		# ResBlocks
conv, 7 × 7, 64, maxpool, 3 × 3			1x	conv 1 × 1	512	256	6x
				conv 3 × 3	256	256	
conv 1 × 1	64	64	3x	conv 1 × 1	256	1024	
conv 3 × 3	64	64		SE	1024	1024	
conv 1 × 1	64	256		conv 1 × 1	1024	512	3x
SE	256	256		conv 3 × 3	512	512	
conv 1 × 1	256	128	4x	conv 1 × 1	512	2048	
conv 3 × 3	128	128		FC	2048	100	Projection Head
conv 1 × 1	128	512		FC	100	100	
SE	512	512					

$$L_{CE} = -\frac{1}{M}\sum_{i=1}^{N} -y_i(\log(\bar{y}_i)) \tag{8}$$

Where y_i is a ground truth, $\bar{y}_i$ is the predicted disease category, and M is the number of class categories.

3.4. Evaluation metrics

We used accuracy Acc, precision Pr, recall Rec, and F1 score to calculate the classification performance of the proposed and baseline methods.

Eq. (9)

$$Acc = \frac{TP + TN}{TP\ +\ \text{TN} +\ FP + FN},\ Pr = \frac{TP}{TP\ +\ FP}, Rec = \frac{TP}{TP\ +\ FP},\ F1 = 2\ \times\ \frac{Pre\ \times Rec}{Pre\ +\ Rec} \tag{9}$$

4. Experiment, result, and discussion

4.1. ASCL self-supervised pre-training settings

Two Tesla P100 16GB GPUs are the hardware and parameter configurations used for the self-supervised pre-training. The experiment employed the Pytorch library with a batch size 310, maximum epoch of 1000, SGD optimizer, lr=0.0001, momentum=0.9, and weight decay=0.0001. To pretrain the ASCL self-supervision framework, the self-supervised Siamese ResNet50 backbone network without attention (without attention) and the self-supervised Siamese CBAM-ResNet50 backbone [30] were used as baselines. The proposed ASCL framework is designed to use a channel-wise attention module in a self-supervised Siamese SE-ResNet50 [29] network. To pre-train the proposed ASCL, the 17-classes PlantVillage (tomato, potato, and maize sub-species only) dataset was used without labels. To evaluate the performance of a downstream task, fine-tuning seen in the ASCL pre-training tomato dataset from PlantVillage was used, unseen in the ASCL pre-training apple dataset from PlantVillage and unnoticed in the ASCL pre-training Taiwan tomato datasets were employed. The resolution of an input image is $128 \times 128 \times 3$ spatial dimension selected for pre-training and fine-tuning. The statistical characteristics of the datasets are shown in Fig. 2.

4.1.1. Datasets description and characteristics

To evaluate the proposed ASCL framework's performance and transferability, we utilized the Apple and Wild Tomato datasets. Below, we provide a detailed description of each dataset, including its sources, characteristics, and relevance to this study.

i. **Apple dataset**

The Apple dataset, a subset of the PlantVillage dataset, is widely recognized for its well-annotated images collected under controlled conditions. This dataset includes four distinct categories representing healthy and diseased apple leaves. The images feature consistent lighting, uniform backgrounds, and minimal noise, making them ideal for testing classification models in a controlled environment. The dataset contains 5120 labeled images, with a minimum of 275 images and a maximum of 1645 images per class. Its balanced nature and high-quality annotations accurately assess the ASCL framework's performance in structured scenarios.

i. **Wild tomato dataset**

The Wild Tomato dataset, collected from real agricultural fields, poses a more challenging testing environment due to its diversity and variability. Unlike the Apple dataset, the images in this dataset include

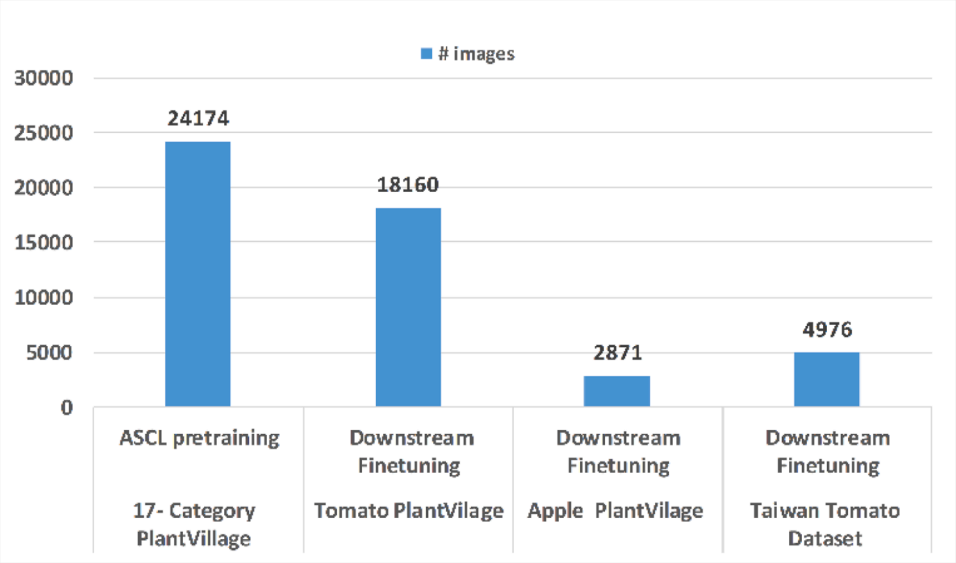


Fig. 2. The 17-classes PlantVillage dataset for ASCL framework pre-training (unlabeled) and downstream classification (labeled) task datasets.

multiple leaves per frame, varied lighting conditions, complex backgrounds, and natural noise. It comprises six categories, with 536 to 1255 labeled images per class, covering healthy and diseased tomato leaves. This dataset is particularly significant for evaluating the framework's generalizability and robustness in unstructured, real-world settings. Its complexity mimics practical agricultural challenges, making it an essential resource for testing the ASCL framework's ability to handle variability and unseen scenarios.

i. **Relevance to the Study**

The combination of the Apple and Wild Tomato datasets enables a comprehensive evaluation of the ASCL framework. The controlled nature of the Apple dataset highlights the framework's accuracy and effectiveness under ideal conditions, while the Wild Tomato dataset tests its adaptability and transferability in real-world scenarios. This dual evaluation approach ensures that the framework is accurate and practical for diverse applications in plant disease classification. This study provides a well-rounded analysis of the ASCL framework's capabilities by including these datasets, addressing fine-grained classification in structured environments, and generalization to more complex, real-world datasets.

4.1.2. Dataset organization and composition

This section describes the data sets used in this study, including how the 17 classes are organized, the distribution of images across these classes, and the preprocessing steps applied. This information is crucial for understanding the structure and diversity of the data utilized to evaluate the ASCL framework.

i. **Definition of the 17 classes**

The datasets used in this study—Apple and Wild Tomato—are organized into 17 classes representing distinct plant health and disease categories. These include various stages of specific diseases and healthy leaves for apple and tomato plants. The class definitions are based on visual and biological characteristics, such as the presence of spots, blight symptoms, and discoloration. For example:

- **Apple dataset**: Includes four classes: healthy, apple scab, black rot, and cedar apple rust.
- **Wild tomato dataset**: This dataset includes six classes: healthy, early blight, late blight, and other disease-specific categories.

The remaining classes are derived from other sub-datasets within the PlantVillage dataset, such as potato and maize species, each contributing multiple classes to form the full set of 17.

i. **Dataset composition**

Table 2 summarizes the composition of the datasets, highlighting the total image count and distribution across classes:

i. **Preprocessing steps**

To ensure uniformity across the datasets, the following preprocessing steps were applied:

- **Image resizing**: To standardize input dimensions, all images were resized to 128 × 128 pixels.
- **Augmentation**: Techniques such as rotation, flipping, and noise addition were used to enhance variability and address class imbalance.
- **Normalization**: Pixel values were normalized to 0 to 1 to optimize input for the ASCL framework.

i. **Dataset characteristics**

The datasets present diverse characteristics:

Table 2
Description of the dataset.

Dataset	Class	Image Count
Apple Dataset	Healthy	635
	Apple Scab	1250
	Black Rot	1450
	Cedar Apple Rust	1785
Wild Tomato	Healthy	920
	Early Blight	1015
	Late Blight	1255
	Other Categories	850

- **Variability in disease stages**: Images capture early and late stages of diseases, providing a comprehensive range for classification.
- **Backgrounds and Lighting**: The Wild Tomato dataset includes natural field conditions with varied lighting and complex backgrounds, while the Apple dataset maintains controlled conditions with uniform backgrounds.
- **Disease Presentation**: The dataset is diverse, with different leaf patterns, discolorations, and spot types represented.

i. **Dataset Access**

The datasets used in this study are publicly available:

- **Apple Dataset**: Accessible from the PlantVillage repository [33].
- **Wild Tomato Dataset**: Available at Mendeley Data repository [34].

4.2. Supervised (downstream fine-tuning) settings

To validate the transferability of the pre-trained ASCL weights, the method was fine-tuned on the 17 classes PlantVillage itself, the Apple dataset, another subset of PlantVillage [4] [22], and the tomato dataset with their labels. The apple dataset extracted from PlantVillage [4] [22] consisted of four categories with a minimum of 275 labeled images and a maximum of 1645 labeled images per category. The method is also evaluated on a tomato dataset collected from agricultural fields [25], which is a challenging dataset collected from wild fields with complex characteristics. This dataset has six categories, each containing a minimum of 536 and a maximum of 1255 labeled images. The fine-tuning training uses an 80 : 20% train test split, 10% of the train as validation, learning *rate* = 0.001, *batch size* = 32, *max epoch* = 400, *momentum* = 0.9, *weight decay* = 0.0001, *optimizer* = *SGD*, cross-entropy loss for classification label prediction.

4.3. Evaluation of downstream plant disease classification tasks

4.3.1. Linear evaluation of the 17-classes PlantVillage dataset

Linear evaluation in the self-supervised technique assesses the quality of the representations learned by an ASCL self-supervised model. After the ASCL model has learned representations through pre-training on unlabeled data, a linear evaluation is performed to evaluate the usefulness of these learned representations for downstream classification tasks by training a linear classifier on top of the pre-trained features for plant disease classification. The learned classifier was tested on the 17-classes PlantVillage dataset. The linear classification performance of the ASCL pre-trained weight was compared with that of the baselines, as presented in Table 3. It is observed that the baseline ResNet50 without (w/o) attention scores of 68.1 %, while the CBAM-ResNet50 achieves 72.78 %, and the proposed ASCL SE-ResNet50 yields a competitive result of 71.12 % with the ASCL CBAM-ResNet50 backbone. The ASCL SE-ResNet50 backbone enabled salient intermediate feature learning, which boosted the contrastive objective function to capture subtle differences. The linear evaluation results show that the proposed method effectively learns discriminative features from which a linear classifier can generalize.

Fine-tuning on Different Datasets: To further validate the generalizability of the learned features by the proposed ASCL method to the unseen dataset, evaluation of the ASCL feature transferability using the apple dataset from PlantVillage [4] [22] and tomato dataset collected from wild fields, the tomato consists of multiple leaves in a single image, single leaf in a single image, and the images have different backgrounds [25]. Neither the apple nor challenging tomato datasets were used during the pre-training (unseen) of the ASCL framework. The fine-tuning results are shown in Table 4. The ResNet50 backbone without attention and attentive SE-ResNet50 backbone pre-trained weights were fine-tuned on these two unseen datasets, and the accuracy, attention heat map, and confusion matrix results were obtained.

As shown in Table 4, the attention-based SE-ResNet50 backbone trained from the ASCL weight has an accuracy of 93.50 % for the Apple dataset, which is high performance. We conjecture that this performance is achieved because the dataset is taken in a controlled environment with a uniform background, in the same fashion as the 17-classes dataset used in the pre-training phase. As in Table 3, the proposed SE-ResNet50 backbone had an accuracy of 83.00 % for a very challenging tomato dataset collected from wild agricultural fields. Compared with Self-supervised ResNet50 without attention, the proposed ASCL SE-ResNet50 backbone achieved +3.98 accuracy improvement. The results show that the proposed ASCL representation learning framework can capture similar generic and transferable embeddings to unseen plant disease and plant species classification downstream tasks.

4.3.2. Qualitative analysis on the 17-classes PlantVillage dataset

The confusion matrix in Fig. 3 also demonstrates the classification performance of the proposed ASCL framework and baselines. As can be seen from Fig. 3, the backbone networks trained from the ASCL framework show competitive classification performance compared to the baseline supervised ResNet50 backbone. The improvement indicates that ASCL encourages these backbones to learn robust image features to classify disease categories with less confusion. Supervised ResNet50 has many confusions in each category, demonstrating that supervised models are not strong enough to capture discriminative information. The self-supervised method without attention has shown better classification performance. However, the attentive self-supervised method showed superior classification compared with the other baseline methods.

In addition to the confusion matrix, visualization was also used as an attention heat map of various plant disease spots, as shown in Fig. 4. As seen in the attention heat map, the proposed SE-ResNet50 backbone trained from ASCL was able to spot the diseased area of the plant images. In contrast, the baseline CBAM-ResNet50 self-supervised ResNet50 (without attention) was less effective at spotting the diseased region of the image. Meanwhile, attention heat maps generated by supervised ResNet50 are visually inconsistent and unable to capture disease spots. The overall qualitative results demonstrated the capability of the proposed ASCL to capture discriminative features that can boost the performance of plant disease visualization.

This method visualizes the embedding space using tSNE on a 4k test dataset to further evaluate the discriminability of the learned image representations. Fig. 5 shows that the weights fine-tuned from the ASCL framework have a more robust semantic clustering capacity on unseen data than those fine-tuned without attention and supervised

Table 3
Linear evaluation and (b) fine-tuning classification accuracy result(%).

Method	Dataset used in ASCL pre-training (seen)	
	(a) Linear evaluation	(b) Fine-tuning
Supervised ResNet50 [31]	–	**89.01**
SSLResNet50 (w/o attention)	68.10	82.72
ASCL CBAM-ResNet50	**72.78**	84.38
ASCL SE-ResNet50(Ours)	71.12	**88.89**

Table 4
Fine-tuning classification results for two unseen datasets (a) and (b) (%).

Method	Fine-tuning accuracy	
	(a) Wild Tomato dataset [25]	(b) Apple dataset [4] [22]
Supervised ResNet50 [31]	86.51	91.66
ResNet50 (w/o attention)	79.02	91.00
ASCL SE-ResNet50	**83.00**	**93.50**

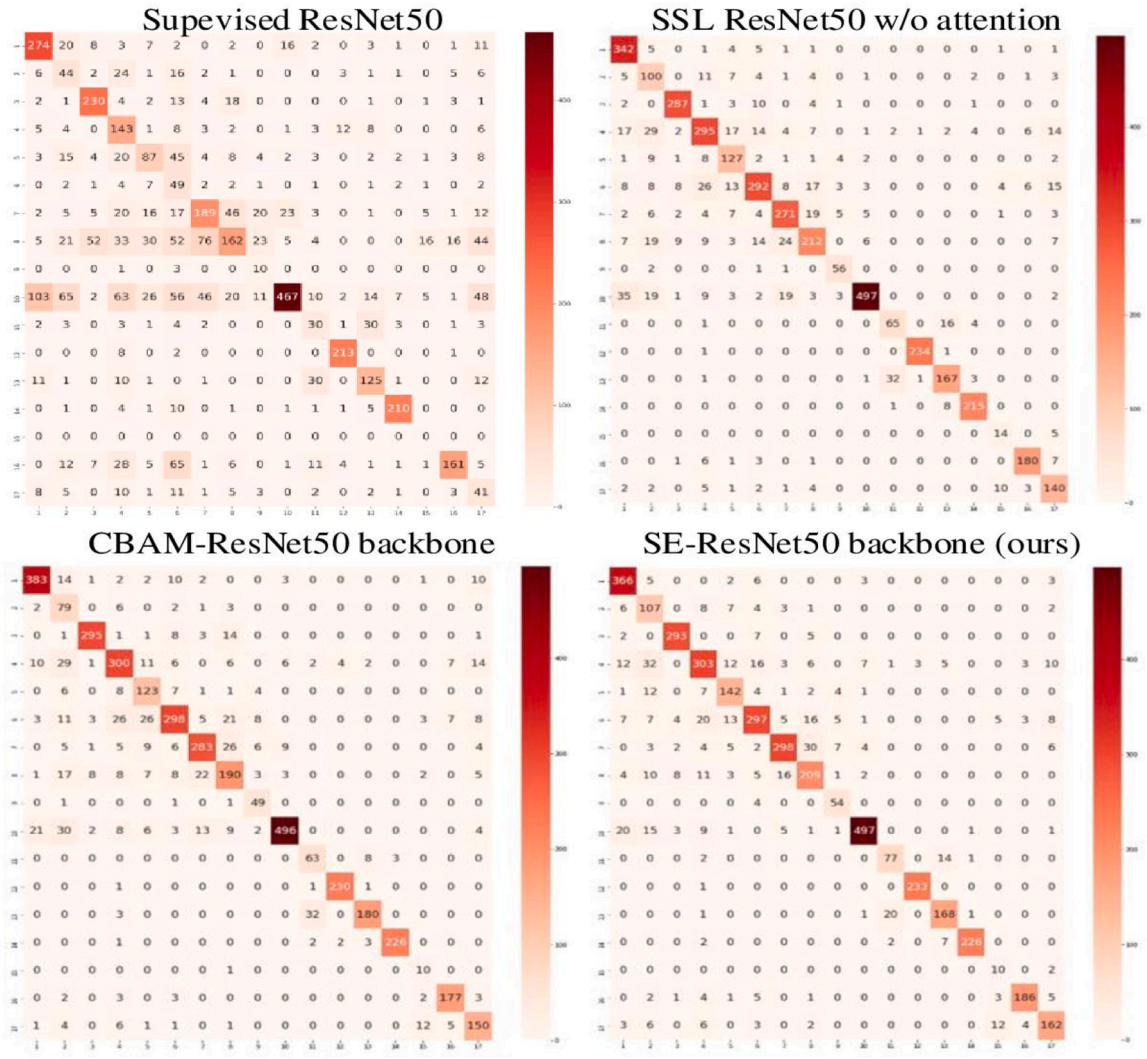


Fig. 3. Confusion matrix for the 4k test samples on the 17-classes PlantVillage dataset.

counterparts. Specifically, the SE-ResNet50 backbone fine-tuned from ASCL weights demonstrated powerful semantic-level clustering performance.

Qualitative Analysis on Unseen Datasets

This subsection presents the confusion matrix and attention heat map results obtained from the proposed ASCL pre-trained framework on apple and tomato datasets, which were not used during pre-training ASCL with the SE-ResNet50 backbone. Fig. 6 shows the confusion matrix for the tomato dataset. As can be seen, the tomato dataset was more confusing than the apple dataset. This was primarily because the nature of the dataset was challenging. On the other hand, as shown in Fig. 7, the confusion matrix for the Apple dataset on 634 holdout test data demonstrates robust disease classification performance with correct 569 image classifications and 65 image confusions. The performance gain in the apple dataset indicates that the attentive representations extracted by ASCL are more generalizable to unseen plant disease datasets than the unseen tomato dataset.

Attention Heat map. Attention heat maps of challenging plant disease images were generated to further evaluate the generalization capability of learned representations on unseen plant diseases. As shown in Fig. 7, it is observed that the SE-ResNet50 backbone trained from ASCL captures the disease spot of the images, which indicates the robustness of the learned representations. Hence, the attention heat map is powerful enough to highlight disease spots in challenging tomato and apple datasets. The Apple dataset was not difficult; it showed better heap map visualization than the challenging tomato dataset. Fig. 7's apple healthy leaf shows the attention heap marking the whole leaf structure. Therefore, the proposed ASCL is a practical approach for leveraging transferable representations from unlabeled images to solve target plant disease classification with a small-scale labeled dataset.

4.4. Discussion on results

Although the supervised ResNet50 model achieved marginally higher accuracy in specific scenarios, such as those shown in Table 3, there are several reasons why the proposed ASCL method may not always surpass this baseline. A primary reason is that supervised ResNet50 directly benefits from having access to well-annotated, labeled data, which allows it to optimize its parameters for specific tasks precisely. This advantage is particularly evident in controlled environments with high-quality labeled datasets, where the model can fully utilize the data for training.

In contrast, the ASCL framework is designed to capitalize on unlabeled data and reduce reliance on large, annotated datasets. This makes it particularly valuable when labeled data are limited or expensive to acquire, such as in agricultural domains. The real strength of ASCL lies in its generalizability and ability to perform well across unseen and less-structured datasets, as evidenced by its successful transferability to new domains.

Although it may not always surpass supervised models in controlled

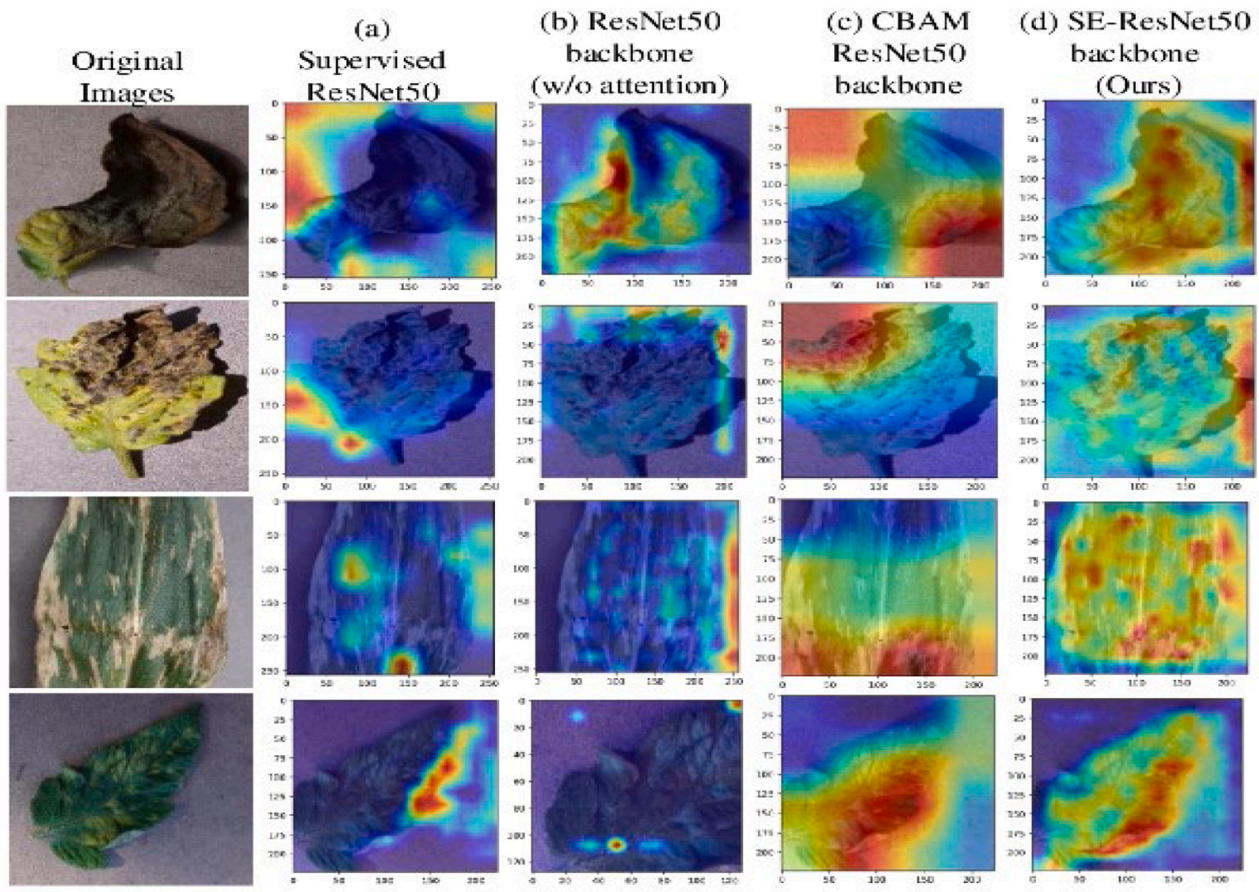


Fig. 4. Attention visualization on the 17-classes PlantVillage dataset.

settings, the ASCL framework offers a competitive alternative in real-world applications, where acquiring vast amounts of labeled data is challenging. Its performance, without extensive annotation requirements, makes it a practical solution for broader and more diverse scenarios.

4.4.1. Model complexity analysis

The ASCL framework is influenced by the ResNet50 backbone and attention mechanisms such as squeeze-and-excitation (SE) and convolutional block attention modules (CBAM). With approximately 25 million parameters from ResNet50, the attention layers add minimal overhead, keeping the model size comparable to the baseline supervised ResNet50, with only a slight increase owing to the attention modules.

Self-supervised pre-training on the unlabeled PlantVillage dataset takes longer to converge, requiring more computational resources and time. However, once pre-trained, fine-tuning on smaller labeled datasets takes significantly less time than fully supervised models, which require labeled data from the start.

4.5. Comparison with related works

We compared the proposed findings with methods that employ self-supervised learning pre-trained on one dataset and fine-tuned on different datasets. As presented in Table 4, The xResNet34 SwAV was pre-trained on the DeepWeeds dataset and evaluated on GrassLand Europe and Areal Farmer land datasets [18]. The ResNet50(4x) SimCLR [14] was pre-trained on the ImageNet dataset and evaluated on the Food dataset. The downstream tasks transfer all the pre-trained weights to the small-scale training dataset. Whereas the proposed ASCL was pre-trained on attentive SE-ResNet50, the baseline self-supervised ResNet50 without attention backbones was pre-trained on unlabeled 17-classes PlantVillage dataset and fine-tuned on the Apple dataset. As shown in Table 5, the ASCL SE-ResNet50 backbone outperforms other related works by a margin of +4.1 %. In summary, the attention and self-supervision contrastive learning methods in the proposed ASCL SE-ResNet50 backbone network improved discriminative feature extraction and effective transferability to downstream unseen disease classification tasks.

5. Conclusion

This work presented the ASCL framework that leverages attention and self-supervision for generic discriminative representation learning using the Channel-wise ResNet50 Siamese backbone network. Self-supervision has reduced massive dataset annotations in supervised network training. With self-supervision, we extracted automatic semantic-level training dataset annotations from massively available unlabeled data. With attentive self-supervision, we can capture discriminative semantic-level training dataset annotation, whereby annotations focus on features that are discriminative enough for fine-grained recognition tasks by contrasting unlabeled plant disease image pairs. The ASCL framework was pretrained on unlabeled plant disease images extracted from the PlantVillage public dataset. The ASCL pre-trained representation weight was transferred to a small-scale labeled plant disease dataset for the downstream plant disease classification task on three different datasets. The ASCL pretrained weight is a supervision signal for small-scale downstream dataset tasks. The feature transferability of the proposed approach to different unseen small-scale datasets justifies the generalization ability of the proposed ASCL framework.

ASCL bases multi-species datasets with common early and late disease characteristics such as tomato Early Blight, tomato Late Blight,

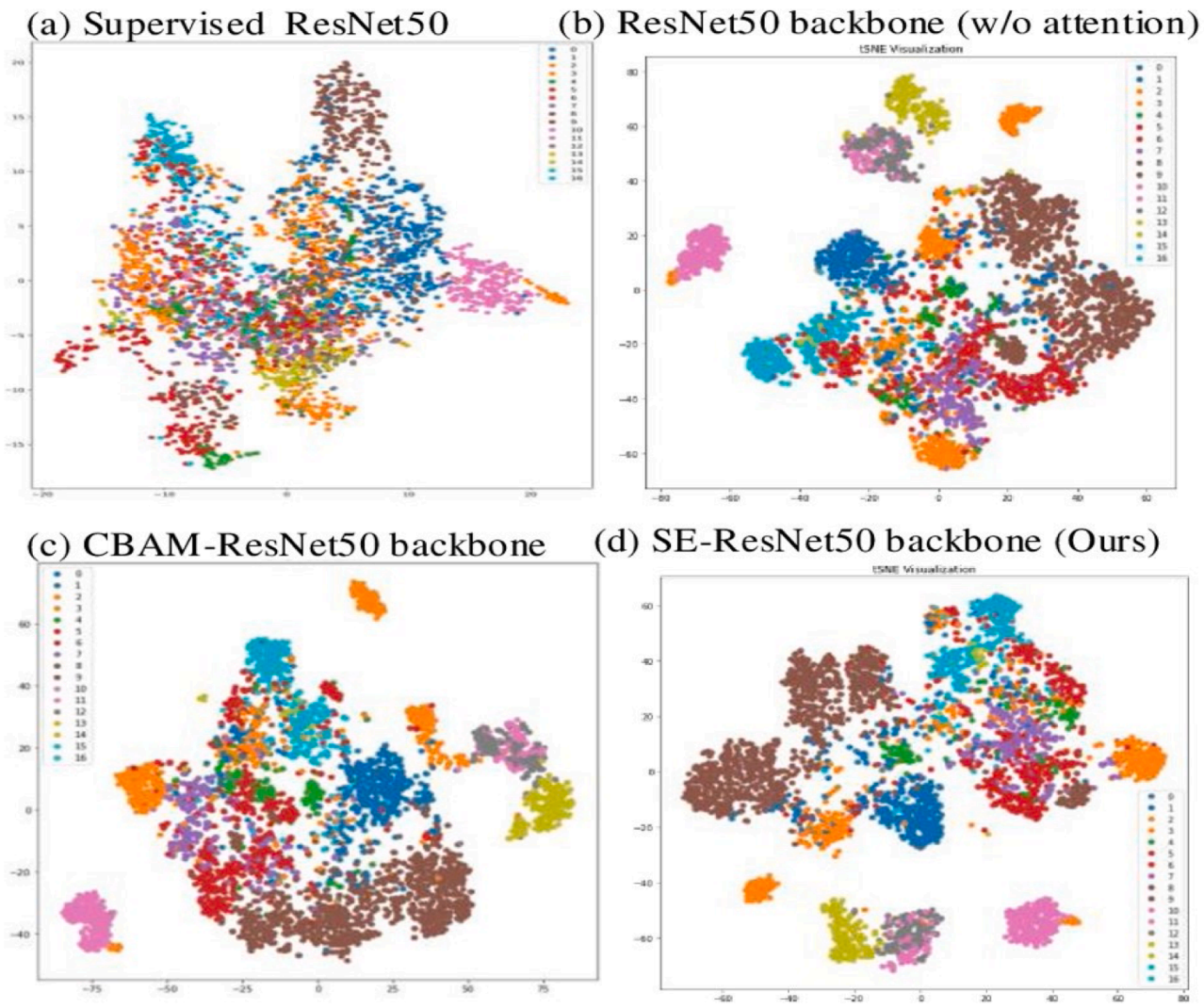


Fig. 5. 2D tSNE visualization of 4k test images on the 17-classes PlantVillage dataset.

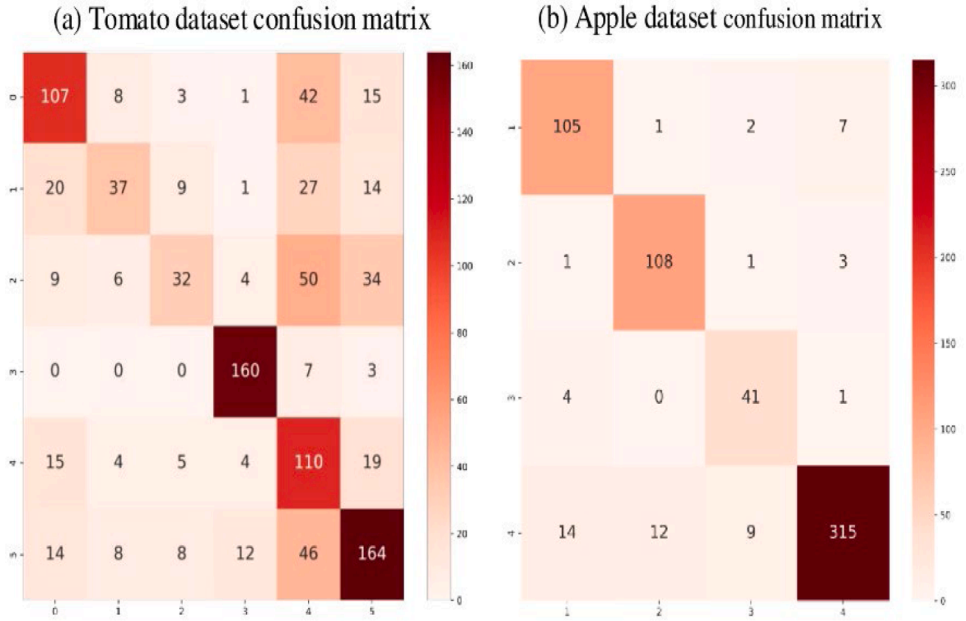


Fig. 6. Confusion matrix for unseen tomato and apple dataset.

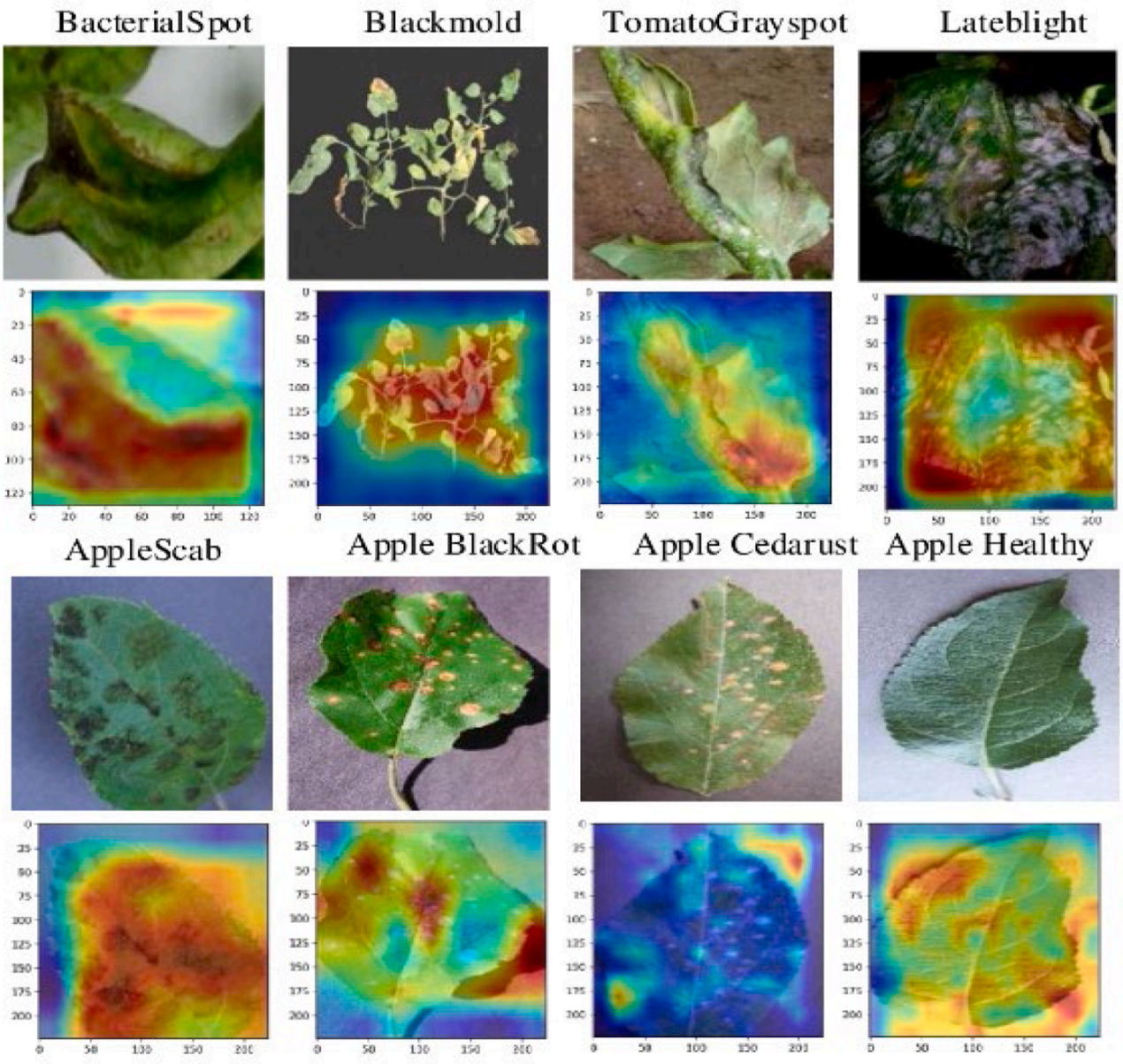


Fig. 7. Attention visualization for the Tomato (upper) and Apple (lower) datasets.

maize Leaf Blight, potato Early Blight, and potato Late Blight. These plant disease categories emulate different disease development stages and the same disease for various plant species. Because only a few plant species with different disease stage characteristics exist, the proposed methods were tested on early and late disease characteristics, which can indicate the generalization capability of the proposed methods. The experimental results demonstrated that the learned representations from ASCL are transferable to improving the performance of plant disease classification downstream tasks by a large margin compared to baseline models. Furthermore, the different visualizations confirmed the ASCL robustness for the generalizability of the learned representations.

6. Limitations and future directions

The present study of the ASCL framework of self-supervised pre-training may not capture all the necessary discriminative features for highly structured datasets, where supervised methods excel owing to access to abundant labeled data. Future research can focus on enhancing contrastive learning and introducing advanced data augmentation techniques to improve pretraining. In addition, hybrid approaches that integrate supervised fine-tuning with self-supervised pre-training may help narrow the performance gap between fully supervised and self-supervised models. Expanding the ASCL framework to more complex datasets with more classes can further its potential for fine-grained classification tasks. These future directions address the current limitations and improve the robustness of the model across various applications.

CRediT authorship contribution statement

Getinet Yilma: Writing – original draft, Project administration, Data curation, Conceptualization. **Mesfin Dagne:** Writing – original draft, Investigation, Data curation, Conceptualization. **Mohammed Kemal Ahmed:** Writing – original draft, Software, Resources, Data curation, Conceptualization. **Ravindra Babu Bellam:** Writing – review & editing, Visualization, Validation, Supervision, Software, Conceptualization.

Table 5
Related work comparisons linear evaluation and fine-tune accuracy (in %).

Related work	Evaluation dataset	Linear evaluation	Fine-tune
CLA [21]	PlantVillage [4] [22]	–	90.52
Plant disease clustering [16]	PlantVillage [4] [22]	–	89.00
InceptionResNet V2 [22]	PlantDoc [22]	70.53	–
GrassLand xResNet34 SwAV [18]	GrassLand Europe	**85.50**	89.00
Farmland xResNet34 SwAV [18]	Areal Farmland	68.40	72.10
ResNet50(4x) SimCLR [14]	Food dataset [14]	–	89.40
InceptionV3 [22]	PlantDoc [22]	71.00	–
Attention Residual CNN [24]	Wild Tomato Dataset [25]	–	82.50
ASCL SE-ResNet50 (ours)	Wild Tomato dataset [25]	–	83.00
SSL ResNet50 w/o attention	Apple Dataset [4] [22]	67.20	91.00
ASCL SE-ResNet50 (ours)	Apple Dataset [4] [22]	81.14	**93.50**

Declaration of competing interest

The Author declaring no conflict of interest.

Data availability

Data will be made available on request.

References

[1] S. Tiwari, A. Gehlot, R. Singh, B. Twala, N. Priyadarshi, Design of an iterative method for disease prediction in finger millet leaves using graph networks, dyna networks, autoencoders, and recurrent neural networks, Results. Eng. 24 (2024) 103301, https://doi.org/10.1016/j.rineng.2024.103301.

[2] F. Rodríguez-Díaz, A.M. Chacón-Maldonado, A.R. Troncoso-García, G. Asencio-Cortés, Explainable olive grove and grapevine pest forecasting through machine learning-based classification and regression, Results. Eng. 24 (2024) 103058, https://doi.org/10.1016/j.rineng.2024.103058.

[3] S. Chakraborty, A.C. Newton, Climate change, plant diseases and food security: an overview, Plant Pathol. 60 (1) (2011) 2–14, https://doi.org/10.1111/j.1365-3059.2010.02411.x.

[4] S.P. Mohanty, D.P. Hughes, M. Salathé, Using deep learning for image-based plant disease detection, Front. Plant Sci. 7 (Sep. 2016) 1–10, https://doi.org/10.3389/fpls.2016.01419. September.

[5] A. Balasundaram, P. Sundaresan, A. Bhavsar, M. Mattu, M.S. Kavitha, A. Shaik, Tea leaf disease detection using segment anything model and deep convolutional neural networks, Results. Eng. 25 (2025) 103784, https://doi.org/10.1016/j.rineng.2024.103784.

[6] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, D. Stefanovic, Solving current limitations of deep learning based approaches for plant disease detection, Symmetry (Basel). 11 (7) (Jul. 2019) 939, https://doi.org/10.3390/sym11070939.

[7] T. Wiesner-Hanks, et al., Millimeter-level plant disease detection from aerial photographs via deep learning and crowdsourced data, Front. Plant Sci. 10 (2019) 1–11, https://doi.org/10.3389/fpls.2019.01550. December.

[8] G. Yilma, S. Belay, Z. Qin, K. Gedamu, and M. Ayalew, "Plant disease classification using two pathway encoder GAN data generation," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP* 2020, 2020. doi: 10.1109/ICCWAMTIP51612.2020.9317494.

[9] G. Yilma, Z. Qin, M. Assefa, G. Alemu, and M. Ayalew, "Attention augmented convolutional neural network for fine-grained plant disease classification and visualization using stochastic sample transformations," in *2021 5th International Conference on Advances in Image Processing (ICAIP)*, New York, NY, USA: ACM, Nov. 2021, pp. 13–19. doi: 10.1145/3502827.3502836.

[10] A.Debasu Mengistu, Image analysis for Ethiopian coffee plant diseases identification, Seffi Gebeyehu Mengistu Dagnachew Melesew Alemayehu Int. J. Biometrics Bioinforma. (10) (2016) 1.

[11] Q.H. Cap, H. Uga, S. Kagiwada, H. Iyatomi, LeafGAN: an effective data augmentation method for practical plant disease diagnosis, IEEE Trans. Autom. Sci. Eng. (Feb. 2020) 1–10, https://doi.org/10.1109/TASE.2020.3041499.

[12] J. Ma, K. Du, F. Zheng, L. Zhang, Z. Gong, Z. Sun, A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network, Comput. Electron. Agric. 154 (Nov. 2018) 18–24, https://doi.org/10.1016/j.compag.2018.08.048. August.

[13] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: a review, Med. Image Anal. 58 (2019), https://doi.org/10.1016/j.media.2019.101552.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *37th International Conference on Machine Learning, ICML 2020*, Feb. 2020, pp. 1575–1585.

[15] Z.M. Lonseko, et al., Semi-supervised gastrointestinal lesion segmentation using adversarial learning, in: 2021 IEEE 3rd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), IEEE, 2021, pp. 63–66, https://doi.org/10.1109/ecbios51820.2021.9510611.

[16] U. Fang, J. Li, X. Lu, L. Gao, M. Ali, Y. Xiang, Self-supervised cross-iterative clustering for unlabeled plant disease images, Neurocomputing. 456 (2021) 36–48, https://doi.org/10.1016/j.neucom.2021.05.066.

[17] M. Assefa, W. Jiang, K. Gedamu, G. Yilma, B. Kumeda, M. Ayalew, Self-supervised scene-debiasing for video representation learning via background patching, IEEE Trans. Multimed. XX (X) (2022) 1–15, https://doi.org/10.1109/TMM.2022.3193559.

[18] R. Güldenring, L. Nalpantidis, Self-supervised contrastive learning on agricultural images, Comput. Electron. Agric. 191 (Dec. 2021) 106510, https://doi.org/10.1016/j.compag.2021.106510. October.

[19] D.K. Getinet Yilma, Machine learning prediction of human activity recognition, Ethiop. J. Sci. Sustain. Dev. p-ISSN 5 (1) (2018) 20–33.

[20] M.K. Ahmed, D.P. Sharma, H.S. Worku, G. Yilma, A. Ibenthal, D. Yadav, Livestock disease data management for e-surveillance and disease mapping using cluster analysis, Adv. Artif. Intell. Mach. Learn. 4 (1) (2024) 1991–2013, https://doi.org/10.54364/AAIML.2024.41114.

[21] 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, https://doi.org/10.1016/j.compag.2023.107967.

[22] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra, "PlantDoc: a dataset for visual plant disease detection," *ACM Int. Conf. Proc. Ser.*, pp. 249–253, 2020, doi: 10.1145/3371158.3371196.

[23] G. Yilma, et al., Attention augmented residual network for tomato disease detection and classification, Turkish J. Electr. Eng. Comput. Sci. 29 (2021), https://doi.org/10.3906/elk-2105-115.

[24] A. Bhujel, N.E. Kim, E. Arulmozhi, J.K. Basak, and H.T. Kim, "A lightweight attention-based convolutional neural networks for tomato leaf disease classification," *Agric.*, vol. 12, no. 2, pp. 1–18, 2022, doi: 10.3390/agriculture12020228.

[25] M.-L. Huang, Y.-H. Chang, Dataset of tomato leaves, in: Mendeley Data, 1, 2020.

[26] R. Karthik, M. Hariharan, S. Anand, P. Mathikshara, A. Johnson, R. Menaka, Attention embedded residual CNN for disease detection in tomato leaves, Appl. Soft Comput. J. 86 (2020), https://doi.org/10.1016/j.asoc.2019.105933.

[27] M. Assefa, W. Jiang, G. Yilma, B. Kumeda, M. Ayalew, M. Seid, Self-supervised multi-label transformation prediction for video representation learning, J. Circuits, Syst. Comput. 31 (9) (2022) 1–20, https://doi.org/10.1142/S0218126622501596.

[28] J. Chen, D. Zhang, M. Suzauddola, Y.A. Nanehkaran, Y. Sun, Identification of plant disease images via a squeeze-and-excitation MobileNet model and twice transfer learning, IET. Image Process. 15 (5) (2021) 1115–1127, https://doi.org/10.1049/ipr2.12090.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[30] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "CBAM: convolutional block attention module," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer International Publishing, Jul. 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2016) 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016-Decem.

[32] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Jun. 2018, pp. 3733–3742, https://doi.org/10.1109/CVPR.2018.00393.

[33] D.P. Hughes and M. Salathe, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," 2015.

[34] Y.-H. Huang, Mei-Ling; Chang, "Dataset of tomato leaves," 2020. doi: 10.17632/ngdgg79rzb.1.