ORIGINAL ARTICLE

An evaluation of a pre-trained transformer-based self-distillation model (DINOv2) for cross-domain plant species identification

Chin Ann Ong[1,2] · Fei Siang Tay[1] · Yi Lung Then[3] · Chris McCarthy[4]

Received: 31 July 2024 / Accepted: 9 July 2025

Abstract
Plant species identification is a fundamental process in botany and agriculture sector. In recent years, deep neural networks have become the primary approach for automating this task, providing valuable insights into biodiversity, ecological systems, and agricultural practices. Along with more discoveries in plant species, training a deep neural network becomes very challenging as the cost required to collect and annotate plant samples is expensive and impractical. Despite the lack of labelled plant samples, recent studies have explored the potential of leveraging publicly available and systematically annotated plant specimens in herbaria coupled with field images for plant species identification through cross-domain adaptation techniques. However, the accuracy of these methods remains unsatisfactory, motivating the exploration of alternative approaches. In this paper, we evaluated the feasibility of employing a pre-trained transformer-based self-distillation model (DINOv2) for cross-domain plant species identification tasks. We trained our model with the PlantCLEF2020 dataset comprised of approximately 320 k herbarium and field images representing 997 plant species. Our approach leverages the advanced feature extraction capabilities of DINOv2, which enhances cross-domain adaptation by effectively bridging the gap between herbarium and field images, achieving a 17.7% improvement over the best model proposed in previous work, that employs ensembles of Siamese network architectures with triplet loss (HFTL-ENS and OSM-ENS).

Keywords Plant species identification · Cross-domain adaptation · Self-supervise learning · DINOv2 · PlantCLEF

1 Introduction

Plant species identification is essential for enhancing productivity, sustainability, and conservation efforts. Accurate classification plays a pivotal role in preserving biodiversity, advancing agricultural practices, and improving crop management by understanding plant responses to environmental changes [1, 2]. Additionally, effective plant species identification helps new discoveries in pharmaceutical [3], smart plantation or farming, [4], and the plant phenological studies [5]. These contributions extend to addressing global challenges such as climate change observation, mitigation, and prediction while improving overall ecosystem health. Furthermore, integrating these technologies into plant disease detection can help farmers address critical issues like declining crop quality and crisis on food shortages [6].

https://doi.org/10.1007/s00521-025-11499-6
Published online: 01 August 2025

A systematic review of computer vision-based methods for plant identification [7] highlights the growing interest in using of deep Convolutional Neural Networks (CNNs) such as Resnet [8] and Inception Network [9]. However, identifying or classifying plant species presents several challenges. First and foremost, there are approximately 220,000 identified plant species globally while many other species remain unidentified. Secondly, training a model to classify some species remains challenging due to the limited labelled data available for the actual life plant or field images. Lastly, variations in shapes, sizes, and colours within the same plant species further complicate the identification process. Consequently, data preparation for model training becomes very challenging, as it is impractical to capture, digitize, and annotate all field images individually. Furthermore, the annotation process can be laborious and expensive, even for skilled subject matter experts (taxonomic impediment) or botanists [1].

Fortunately, research institutes such as iDigBio[1] and e-ReColNat[2] have systematically collected, cataloged, annotated, and stored plant specimens in herbaria form over the past several centuries. These catalogs are publicly accessible through online repositories and cover a wide range of plant and tree species. Figure 1 shows an example of a life plant or field image (left) and the specimen or herbarium sheets (right).

The availability of annotated herbarium data has opened up a new avenue for training models using a combination of herbarium sheet and field image data. This approach was inspired by Carranza-Rojas et al. [11], who suggested a machine can potentially be trained to simulate the drying process from a plant's field image to herbarium specimens. The drying process can be achieved by learning a transfer function between two representation spaces: one dedicated to the herbarium image and another to the field image. This approach presents an exciting opportunity for research in plant species identification and the goal is to train a model to transfer common features between herbarium and field images through a technique known as domain adaptation.

In general, domain adaptation, or specifically cross-domain adaptation in deep neural networks is a kind of transfer learning technique that makes use of knowledge learned from one task or domain to improve performance on another task or domain [12]. The idea is to leverage knowledge from one context (source domain) and apply it to another (target domain), maximizing the overlap between source and target distributions in the shared space (domain confusion). This technique minimizes the need for extensive training data and computational resources in the target domain. In the plant species identification context, domain adaptation technique enables a model to learn and apply knowledge from herbarium images (the source domain) to field images (the target domain), despite differences in their visual characteristics.

On the other aspect, the advent of the Vision Transformer (ViT) [13] has revolutionized image classification tasks. The ViT's growing popularity has led to its widespread adoption in image detection and classification tasks due to the robustness of the model along with superior performance compared to state-of-the-art convolutional neural networks while significantly reducing computational demands [14]. The self-distillation model DINOv2, a ViT-based framework [15], is particularly innovative for this task due to its ability to capture long-range dependencies and its robustness in handling complex visual patterns. Key features such as self-supervised learning, an efficient and rich pre-training process using a large dataset (142 million images), make DINOv2 highly effective for cross-domain adaptation in plant species identification. These features significantly enhance generalization from herbarium to field images, addressing challenges posed by the limited availability of field images.

Building on advancements in domain adaptation and vision transformer models, we propose a transformer-based approach for cross-domain plant species identification. This model leverages a large volume of herbarium images to efficiently overcome the scarcity of field data. Our main contributions can be summarized as follows:

- We conducted an in-depth background study and summarized six recent related yet understudied works for cross-domain plant species identification.

[1] https://www.idigbio.org/.

[2] https://www.recolnat.org/en/.

Fig. 1 Sample images of the same plant capture from field (left) vs. archived herbarium sheets (right) (Goëau, Bonnet, and Joly [10])

- We explored, reviewed, and highlighted a list of recent, emerging, and prominent self-supervised learning (SSL) techniques for linear classification tasks.
- Finally, we proposed a novel application of the transformer-based self-distillation model (DINOv2) and yielded a 17.7% accuracy improvement against the state-of-the-art model proposed by others in this field.

2 Related works

In the context of plant species identification, adversarial learning and deep metrics learning techniques have been widely adopted in cross-domain adaptation works. These works involve training a model with herbarium images or a combination of herbarium with field images (source domain) and using it to classify species with the field images (target domain). This section first provides an overview of commonly used datasets and standard measurement metrics, followed by an in-depth discussion of related works.

2.1 Dataset for plant species identification and measurement metric

A public large dataset (PlantCLEF2020) containing both herbarium and field images consolidated by Goëau, Bonnet, and Joly [10] was among the most commonly used benchmark dataset by many researchers for the cross-domain plant species identification task. The training set contains 321,270 herbarium images relating to a selection of 997 species of Amazonian plants mainly centred on French Guianaand and 6316 photos of plants on the field for 502 plant species. These smaller amounts of field images (i.e. 6316 or approximately 2%) are provided as part of the training set to enable the learning of a mapping between the herbarium and field domain. Among all the 997 plant species, 502 of them have both herbarium and field images in the training set and only 354 out of the 502 species are from the same individual plant to allow a more precise mapping between the two domains. This leaves 495 (997-502) plant species containing herbarium samples only. On the other hand, the testing set contains 3186 field images from 408 species. Given the same plant species, the samples in the testset are captured from a completely different individual plant as compared with the sample in the training set. The summary of dataset distribution is shown in Table 1.

It is worth noting that the evaluation has been classified into two categories, namely the wholeset (test-all) and subset (test-difficult). The test images groupped under wholeset consist of plant species where both herbarium and field images are made available during the training process whereas for the subset test images includes plant species where only herbarium images were made available, but with limited (less than 11) or even no field images provided during the training process. This set-up simulate real-world scenario and to encourage the development of methods that are as generic as possible. It also helps evaluate a model's effectiveness to transfer knowledge

Table 1 PlantCLEF 2020 dataset distributions

Group Set	Domain	Number of samples	Number of classes
Trainset	Herbarium	321,270	997
Trainset	Field	6316	502
Testset	Field	3186	408

Source: Goëau, Bonnet, and Joly [10]

between the two domains (herbarium to field), even when no field examples are provided during the training for the subset species.

Figure 2 illustrates the estimated number of herbarium and field images for each species used for the training. Each data point in the graph represents a single species. The number of herbarium and field training images for each species are represented in the x and y axes respectively. The data points with a single blue circle represent plant species made available for training but not included in the testset (neither wholeset nor subset). The data points with two circles (blue-orange) represent species trained with both herbarium and a substantial amount of field images evaluated under the wholeset category. Lastly, the data points with three circles (blue-orange-red) represent the species trained with herbarium images but with limited or even no field images evaluated under the subset (difficult) category.

The Mean Reciprocal Rank (MRR) is used as the standard measurement metric for model performance against the test set. The MRR is defined as the mean of the multiplicative inverse of rank for the correct answer as shown in Eq. (1) below:

$$\mathrm{MRR} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\mathrm{rank}_q} \tag{1}$$

where Q is the number of plant observations and rank_q is the predicted rank for the true label (class) for the qth plant observation.

2.2 Adversarial learning based domain adaptation

The use of adversarial learning was mainly inspired by the Generative Adversarial Network (GAN) [16], which effectively learns feature representations to minimize the discrepancy between the source and target domains.

Fig. 2 Estimated number of herbarium and field training images for each species. Data points with two circles (blue-orange) represent species evaluated as wholeset category, while data points with three circles (blue-orange-red) represent species evaluated as the subset (difficult) category [10]

GAN uses two neural networks, known as the generator network and the discriminator network, which engage in a game-like competition setting (minimax). The goal of the generator network is to produce realistic outputs that can deceive the discriminator network, while the goal of the discriminator network is to correctly distinguish between real and generated outputs.

2.2.1 Adversarial consistent learning model (ACL)

A partial domain adaptation with Adversarial Consistent Learning model (ACL) was proposed by Zhang and Davison [17]. The model employs adversarial loss following GAN approach to learn a mapping from the source domain to the target domain. The network trains a discriminator to distinguish between the two domains while the generator learns to confuse the discriminator. The goal of learning (domain confusion) is achieved when the discriminator can no longer determine whether a feature generated by the generator originates from the correct domain.

The ACL model employs pre-trained deep feature extractor, NASNetLarge (Φ) [18], which is jointly trained with source ($\mathcal{X}_S$) and target ($\mathcal{X}_T$) features as shown in Fig. 3. The emphasis of this architecture is the application of three loss functions namely the source classification loss ($\mathcal{L}_S$, a typical cross-entropy loss function for feature extractor f), the adversarial domain loss ($\mathcal{L}_A$) supporting the discriminator (D), and the feature consistency loss ($\mathcal{L}_{Con}$) where both $\mathcal{L}_A$ and $\mathcal{L}_{Con}$ are responsible to distinguish the target domain from the source domain while preserving the fine-grained feature transition between the two domains. During the training, the weights of feature extractor (Φ) are updated by the backpropagation from all the three loss functions i.e. $\mathcal{L}_S$, $\mathcal{L}_A$ and $\mathcal{L}_{Con}$.

Although the model effectively maintains domain-invariant features and reduces negative transfers between domains, the model only achieves an accuracy of 0.032 MRR and 0.016 MRR for the wholeset and the subset categories, respectively. This was due to the significant difference between the herbarium domain and photo domain in the dataset, as well as the highly imbalanced data distribution between the two domains.

2.2.2 Weighted pseudo-labelling refinement network (WPLR)

To overcome the imbalance dataset problem and improve domain-invariant features, Zhang and Davison redesigned another unsupervised domain adaptation model known as the weighted pseudo-labelling refinement network, WPLR [19].

The main component of WPLR architecture remains similar to ACL as shown in Fig. 4, where the feature extractor (G) uses NASNetLarge coupled with three linear layers (F) and three loss functions. To address the issue of imbalanced numbers of samples of each class in the source domain, the conventional cross-entropy loss function for the source domain in ACL was replaced with a weighted cross-entropy loss ($\mathcal{L}_{WS}$). Additionally, the adversarial domain loss ($\mathcal{L}_A$) and feature consistency loss ($\mathcal{L}_{Con}$) were substituted with CORrelation ALignment ($\mathcal{L}_{CORAL}$) loss [20], a distance-based loss function for minimizing the difference between source and target

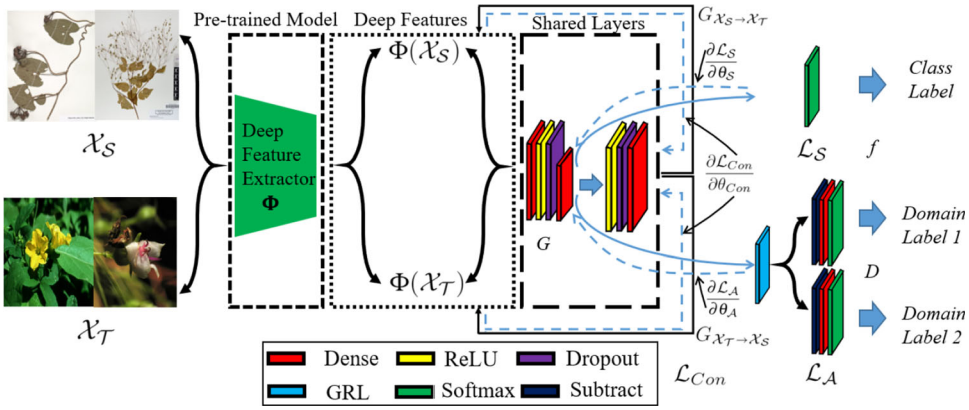

Fig. 3 Adversarial Consistent Learning (ACL) network architecture proposed by Zhang and Davison [17]

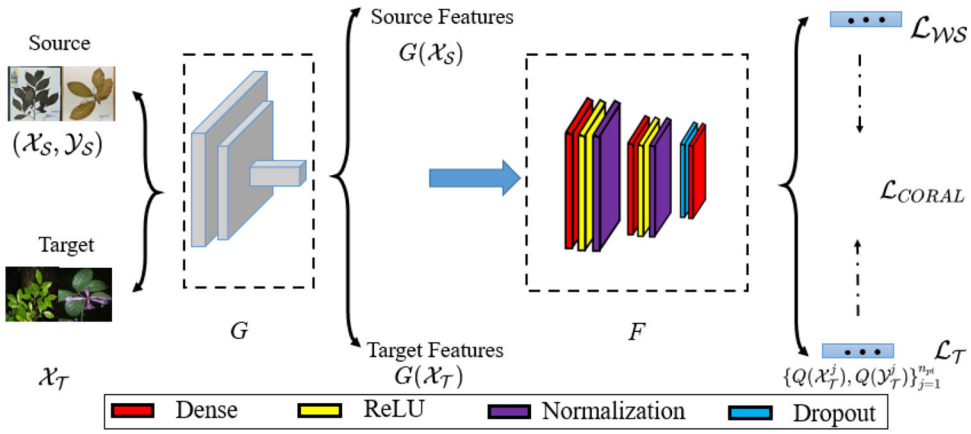

Fig. 4 Weighted pseudo-labelling refinement (WPLR) network architecture proposed by Zhang and Davison [19]

domain, as well as pseudo-labelling refinement on the target domain ($\mathcal{L}_T$) which further reduce source-target domain discrepancies.

The WPLR outperformed the formal model (ACL), with MRR improving from 0.032 to 0.065 for the wholeset and from 0.016 to 0.037 for the subset categories. Despite this improvement, WPLR's performance remains relatively low compared to similar works, as discussed in the next sections.

2.2.3 Few-shot adversarial domain adaptation (FSADA)

Villacis et al. [21] proposed another approach for domain adaptation for plant species identification using few shot adversarial domain adaptation strategy (FSADA) [22]. FSADA leverages adversarial training to learn an embedded subspace that maximizes domain confusion while semantically aligning their embeddings between the herbarium (source) and field image (target) domains (Fig. 5).

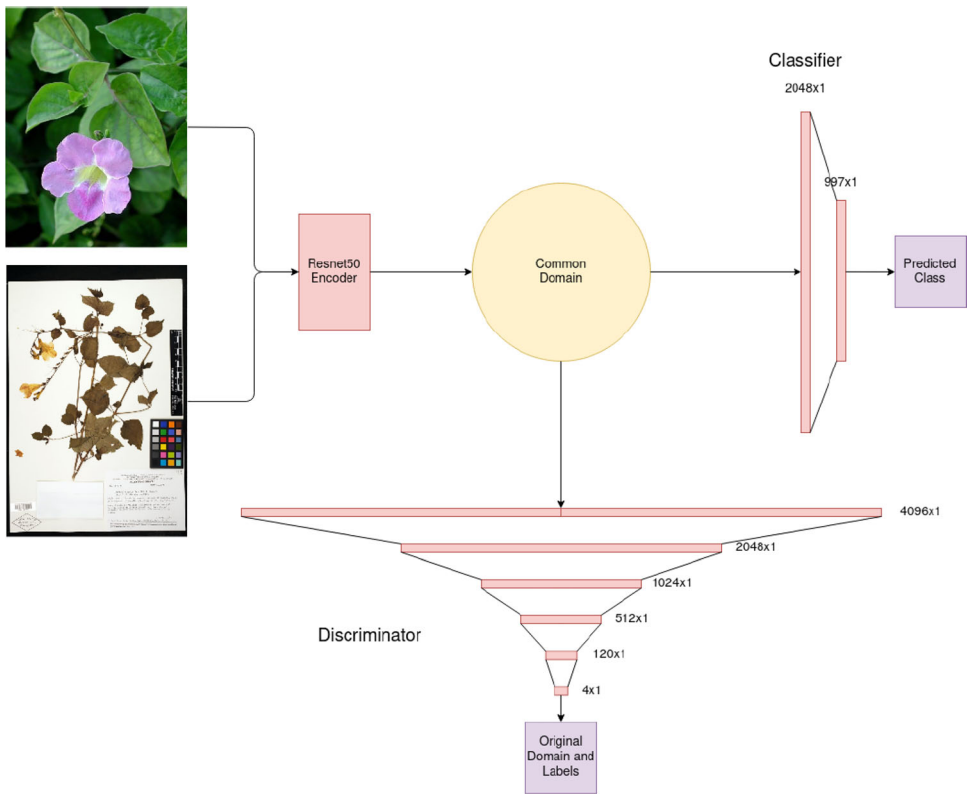

Fig. 5 Few shot adversarial domain adaptation (FSADA) architecture with Resnet50, a classifier, and a discriminator proposed by Villacis et al. [21]

The model uses Resnet50 [8] pre-trained on ImageNet as the encoder, coupled with another encoder containing 2,048 neurons. Unlike typical adversarial learning with binary discriminator, FSADA uses a domain-class discriminator with four outputs to find a shared feature space. Given a pair of images from either the two domains i.e. herbarium and field domain, the discriminator predicts if the pair of images is either from:

- the same domain and same class → (0)
- different domain but same class → (1)
- the same domain but different class → (2)
- different domain and different class → (3)

The FSADA model is trained using a large amount of herbarium images combined with a limited number of field images to create a shared space where the original domains cannot be distinguished. This approach also encourages the semantic alignment between classes, even when the image pairs are from different domains.

An additional self-supervised task is incorporated during the training process to address the challenges of limited or imbalanced data, particularly with field images using the jigsaw puzzle solving [23] operation. This operation divides the original image is into tiles, rearranged randomly into one of 64 possible orderings, and then the network is tasked with determining the correct arrangement of the puzzle pieces. The objective of this additional operation is to extract robust visual information from the image, which can be useful either as initial weights or to assist the main model during training. The final model weights are updated based on the loss functions from both the original classifier and the newly introduced predictor for the jigsaw puzzle solving as shown in Fig. 6.

It is important to note that this work incorporates additional taxonomical information, such as the plant genus and family name, along with the species name when training the model. The model's overall performance achieves 0.180 MRR for the wholeset and 0.052 MRR for the subset category. However, the model's performance decreases when taxonomical information is excluded, with the whole set scoring only 0.148 MRR and 0.039 MRR for the subset category. To facilitate comparison with other works, this paper presents the performance and results of models trained without taxonomical information, as the focus of this study is to evaluate the model's robustness based solely on image data.

2.2.4 Extended few shot adversarial domain adaptation (FSADA)

Building on the approach proposed by Villacis et al. [21], Goëau, Bonnet, and Joly [24] extended the FSADA architecture by integrating additional classifiers related to the plant taxonomical information/traits i.e. plant lifeform, woodiness, and binary plant growthform (climber, herb, shrub, and tree). The extended model with the additional classifiers achieved MRR score of 0.198 and 0.093 for the wholeset and subset categories, respectively. Additionally, the extended FSADA model achieved MRR of 0.153 and 0.056 for the wholeset and subset without the taxonomical information/traits.

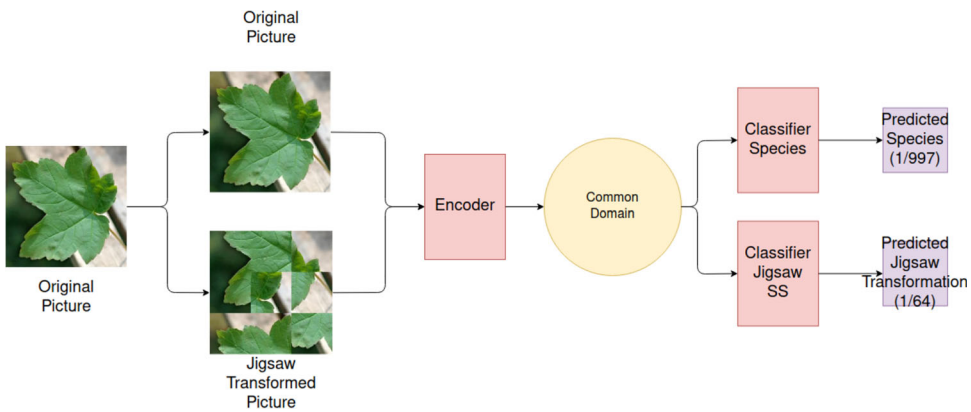

Fig. 6 FSADA training scheme with the incorporation of self-supervised jigsaw puzzle solving technique

When comparing the evaluation results of both experiments conducted without taxonomical information, the model proposed by Goëau, Bonnet, and Joly performs slightly better compared to the original FSADA model proposed by Villacis et al. (0.148 and 0.039). However, Goëau, Bonnet, and Joly did not provide any specific changes to the architecture for the extended work. The MRR scores for both works are included in Table 10 for further discussion.

2.3 Deep metrics learning based domain adaptation

The techniques discussed so far are primarily based on adversarial learning concepts with enhancements to loss functions and partial self-supervised learning strategies on data augmentation for cross-domain adaptation tasks, particularly when dealing with limited or imbalanced data. Another commonly used approach for cross-domain adaptation tasks is deep metrics learning. The deep metric learning strategy relies on a distance metric that measure the similarity or dissimilarity between two or more data [25]. This approach trains a model to learn feature embeddings or the distance between data (e.g. images) and classify new samples based on its distance to a labelled samples or classes.

Several deep metrics learning strategies have been proposed with the contrastive loss coupled Siamese network being one of the most widely employed strategies in the visual understanding research community [26]. The Siamese network is a convolutional neural network architecture that consists of two identical subnetworks with shared weights. It is trained to learn feature space where similar images are represented with identical or similar embeddings. In common practices, a Siamese network with contrastive loss function is trained using pairs of images, either from the same class (positive pair) or from different class (negative pairs). The network is trained with a label annotated as (1) if both images are from the same class or (0) otherwise.

Besides contrastive loss, triplet loss can also be used along with Siamese network to form a triplet network for the same purpose. The triplet network consists of three identical subnetworks with shared weights, each taking different images annotated as the anchor sample, positive sample, and negative sample. The anchor and positive samples are images from the same class, while the negative sample is from any class other than the class the anchor sample belongs to. Similar to the Siamese network, the embeddings generated by the three subnetworks will be used to calculate the loss.

2.3.1 Herbarium-field triplet loss network (HFTL)

Chulif and Chang [27] proposed a two-steamed herbarium-field triplet loss network (HFTL) based on Siamese network architecture with triplet loss as shown in Fig. 7 for cross-domain adaptation task by measuring the similarity between herbarium-field pairs. The network employs two Inception-v4 networks [9] known as a

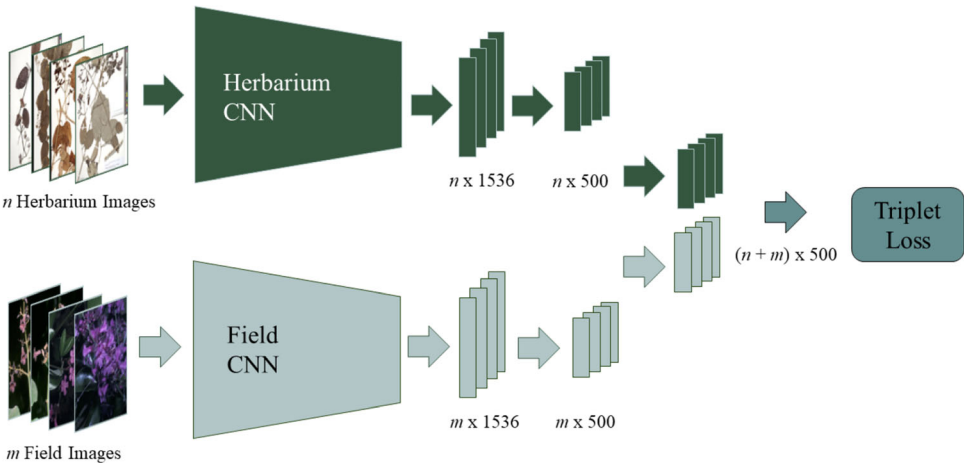

Fig. 7 Herbarium–Field triplet loss network n and m denoted the batch size from the two networks [27]

"stream", each handling herbarium and field generalization. The output from the two streams are combined into an embedding i.e. $(n + m) \times 500$ and passed to a triplet loss function, allowing the network to learn the feature similarity between the herbarium and field embeddings.

Once the two-stream network is trained, random herbarium samples from each species are augmented and processed through the herbarium stream (Fig. 8). The resulting embeddings are averaged to create a single embedding representation for each species, which is stored as a dictionary for all the 997 species. For evaluation, the test field images undergo similar augmentation but through the field stream and obtain an averaged embedding representation. The test field embedding is compared to the herbarium dictionary using cosine similarity to measure the distance metric between the herbarium (source) and field (target) embedding.

For this work, five experiments with different set-ups are conducted, and the performance of each experiment is summarized in Table 2.

Among the five experiment configurations, the NFL-AUG-ENS model achieved the highest MRR score at 0.121 MRR for wholeset and 0.107 MRR for subset which is comparable to the FSADA model and outperforms other works (ACL and WPLR), especially on the subset category. The improvement can be attributed to the use of ensemble networks instead of a single network, as seen in the performance contrast between NFL-AUG-ENS and NFL-ENS. Additionally, the use of a generated herbarium dictionary and test image embedding for comparison also distinguishes this approach.

2.3.2 Herbarium-field triplet loss-OSM network (HFTL-OSM)

Following the findings from HFTL experiments, Chulif and Chang proposed an improved version of HFTL ensemble network [28]. The key difference between this work and the former is the introduction of the one-streamed mixed subnetwork (OSM), supplemented with the combination of different configurations for network training and evaluation.

The OSM subnetwork (stream) resembles a conventional convolutional neural network that uses Inception-v4 or Inception-ResNet-v2 [9], with an additional feature vector reduced from 1536 to 500 (Fig. 9) where q is the batch size. Instead of being used for classification tasks, the stream is trained for weights initialization and subsequently generates a dictionary as well as the embedding of test data, allowing similarity comparison, serving a similar objective as the HFTL network discussed earlier. Its goal is to map the test data to its learned features since the training data (herbarium images) and test data (field images) share different feature distributions. The learned features on OSM stream enable the network to predict classes with missing field images, similar to the triplet network.

Table 3 outlines the performance of HFTL-OSM network across various configurations. The ensemble network (ENS1), comprising two sub-ensemble networks (HFTL-ENS and OSM-ENS), achieved the highest MRR scores of 0.181 and 0.158 for the wholeset and subset categories, respectively. This signifies that ensemble networks

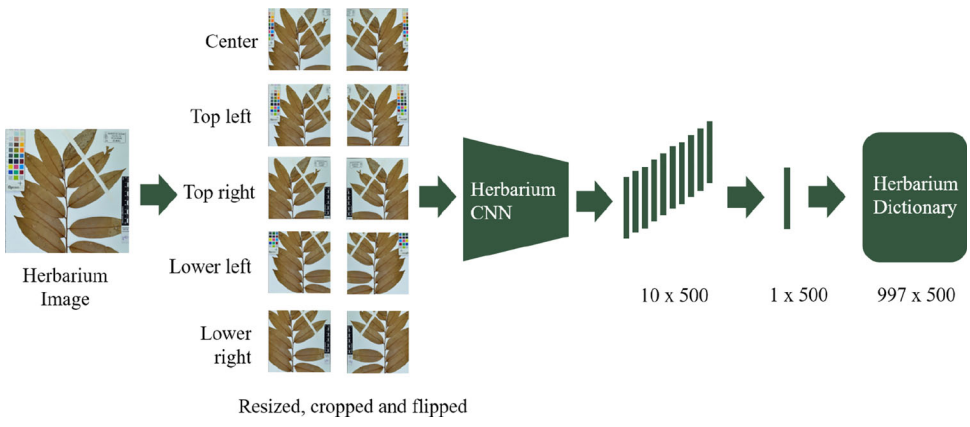

Fig. 8 Dictionary generation process using herbarium dataset over herbarium stream

Table 2 Overall evaluation of HFTL network

Network	Description	Wholeset MRR	Subset MRR
NFL-AUG-ENS	Ensemble of NFL-ENS + NFL-AUG	**0.121**	**0.107**
NFL-ENS	Ensemble of 3 NFL models	0.111	0.108
NFL-AUG	NFL model with data augmentation	0.093	0.066
NFL	Train both feature extractor and fully connected layers	0.103	0.094
FL	Freezes the feature extractor while training the fully connected layers only	0.081	0.061

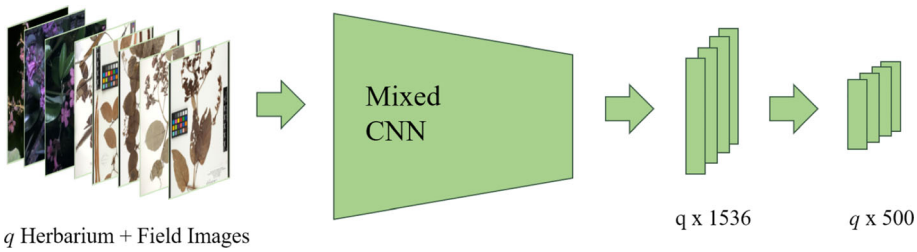

Fig. 9 One-steamed mixed network architecture [28].

Table 3 HFTL-OSM evaluation result

Network	Wholeset MRR	Subset MRR
ENS1 [HFTL-ENS + OSM-ENS]	**0.181**	**0.158**
ENS2 [HFTL-I (Field) + HFTL-I-IR (Field) + HFTL-IR-AUG (Field) + OSM-ENS]	0.176	0.153
HFTL-ENS [HFTL-I (Field) + HFTL-I-AUG (Field) + HFTL-IR (Field)]	0.169	0.150
OSM-ENS [OSM-I + OSM-IR]	0.152	0.117
ENS3 [HFTL-I + HFTL-I-AUG + HFTL-IR + OSM-ENS]	0.147	0.129
HFTL-I-AUG + OSM-ENS	0.143	0.126
HFTL-I + OSM-ENS	0.137	0.116
HFTL-I-21 + OSM-ENS	0.088	0.073
HFTL-I	0.071	0.066
HFTL-I-21	0.060	0.056

enhance performance in cross-domain plant species identification, particularly with imbalanced datasets. The performance comparison between ENS1, HFTL-ENS, and OSM-ENS, as well as between HFTL-ENS and HFTL-I, highlights the impact of different model configurations, such as training with mixed herbarium-field images (OSM) or sequential training (HFTL).

2.4 Motivation-emerging of self-supervised learning

The review from related works in Sects. 2.2 and 2.3 highlighted several key insights for the plant species identification task. First, Zhang and Davison [17, 19] demonstrated the impact of loss function choice and combination on model learnability, as shown by their ACL and WPLR models. Meanwhile, deep metric learning strategies, such as Siamese networks combined with contrastive or triplet loss and mixed data training schemes [27, 28], offer promising direction for further exploration of integrating self-supervised and deep metric learning approaches. Villacis et al. [21] further emphasized the effectiveness of integrating self-supervised techniques within the FSADA architecture to address issues on imbalanced dataset. Additionally, these studies employ

various feature extractors (or backbone architectures) as fundamental building blocks. For examples, the ACL and WPRL models uses NASNetLarge [18], the FSADA model incorporates ResNet50 [8], while the HFTL and OSM models implement both Inception-v4 and Inception-ResNet-v2 [9]. Lastly, the MRR score of the best model (HFTL and OSM) remains low, signifying a lack of robustness in effectively transferring or extracting information from the herbarium domain to the field domain.

Driven by this insight, we explored self-supervised learning (SSL) architectures related to deep metric or contrastive learning strategies for downstream classification as well as domain adaptation tasks [29, 30]. Among the numerous SSL architectures proposed in recent years, we discuss a few prominent and emerging models, including Momentum Contrast (MoCo) [31], a simple framework for contrastive learning (SimCLR) [32], Bootstrap Your Own Latent (BYOL) [33], Swapping Assignments between Views (SwAV) [34], Simple Siamese (SimSiam) [35], Barlow Twins [36], Variance-Invariance-Covariance Regularization (VICReg) [37], and Self-Distillation with No Labels (DINO) [38].

The year 2020 marked a significant advancement in SSL development as shown in Fig. 10. He et al. from Facebook AI Research (FAIR), introduced Momentum Contrast (MoCo) [31], an unsupervised learning technique crafted on top of contrastive learning and the unsupervised representation learning strategy used in natural language processing such as BERT [39] and GPT [40]. This work trains a visual representation encoder by matching an encoded query (Fig. 10a), q i.e. an image to a dictionary of encoded keys (memory bank). The dictionary keys $k0$, $k1$, $k2$, etc. are defined on-the-fly by a set of data samples and then encoded by a slowly progressing encoder driven by a momentum update with the query encoder. The query and dictionary keys are then compared using contrastive loss. Similarly, Chen et al. [32] from Google Brain proposed SimCLR, a simplified constrastive learning framework (Fig. 10b). SimCLR learns representations by maximizing agreement between differently augmented views (z) of the same image (i.e. random crop and rescaling, colour distortion, and Gaussian blur) using contrastive loss in the latent space without requiring a memory bank. While measuring the similarity of two augmented samples from the same images, SimCLR also repels or compares an augmented sample from a positive image with the other $2(N-1)$ augmented examples within a minibatch as negative examples where N is the batch size.

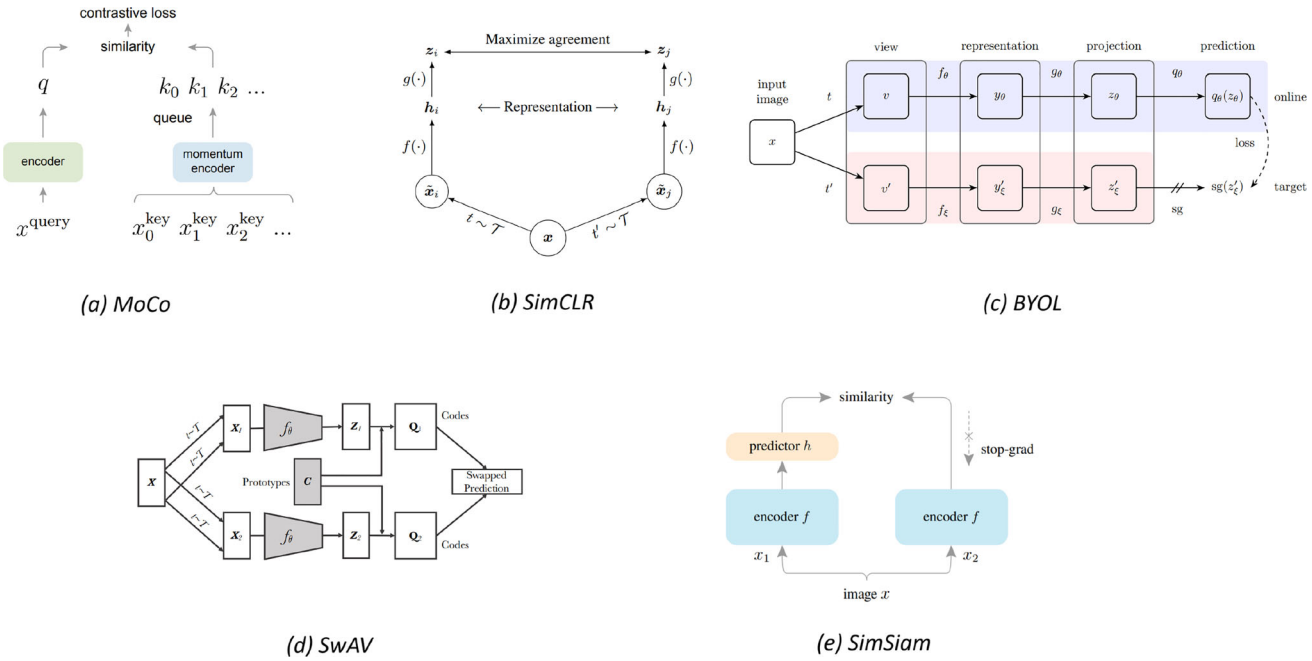

Fig. 10 Collections of Self-Supervised Learning (SSL) proposed since year 2020: **a** Momentum Contrast, **b** Simple Contrastive Learning, **c** Bootstrap Your Own Latent, **d** Swapping Assignments between Views, and **e** Simple Siamese

Following the models MoCo and SimCLR, Grill et al. proposed a hybrid model called Bootstrap Your Own Latent (BYOL) [33], which integrates the elements from momentum contrastive learning and data augmentation while eliminating negative samples. Generally, BYOL relies on two neural networks, the online and target networks as shown in (Fig. 10c), that interact and learn from each other. The online network is trained with an augmented image (v) to predict the target network's representation of a different augmented view of the same image (v'). Meanwhile, the target network is trained with a slow-moving average of the online network, which provides smoother changes in the target network representation. This model has surpassed the two models i.e. MoCo and SimCLR tested on the ImageNet (linear evaluation) using ResNet50. Additionally, Caron et al. proposed Swapping Assignments between Views (SwAV) [34], an online algorithm which takes advantage of contrastive methods without requiring pairwise comparisons on a large dataset, large memory bank, or a special momentum network. It uses a scalable online clustering loss to learn features by Swapping Assignments between multiple Views of the same image. On top of that, SwAV also coupled with a multi-crop strategy to increase the number of views of an image while maintains the memory or computational requirements during training. When compared with BYOL, SwAV performs slightly better (i.e. 1%) and also surpasses the two models (MoCo and SimCLR) discussed previously when tested in the same settings.

The SSL models (except MoCo) discussed so far require large batch sizes and negative samples, while MoCo and BYOL rely on momentum encoder to prevent collapse. SwAV uses an online clustering algorithm for the same purpose. These components can be inefficient and cumbersome. To address these issues, Chen and He from Facebook AI Research (FAIR) revamped MoCo and introduced SimSiam [35], a simpler Siamese network with smaller batch sizes and eliminate the need for negative samples, the momentum encoder, and the online clustering algorithm (Fig. 10e). A stop-gradient operation is integrated into one of the networks, enabling the model to learn asymmetrically and prevent collapse after the removal of the momentum encoder.

In 2021, Zbontar et al. introduced Barlow Twins [36], yet another simpler SSL model that eliminates the need for large batch sizes, negative samples, momentum encoders, online clustering algorithms, or stop gradients required in SimSiam. Instead, it employs an objective function, $\mathcal{L}_{\mathcal{BT}}$ (Fig. 11a) that avoids collapse by measuring the "cross-correlation matrix" between outputs of two identical networks fed with distorted versions of a sample. This approach aims to equate the diagonal elements of the cross-correlation matrix to 1, making the embedding

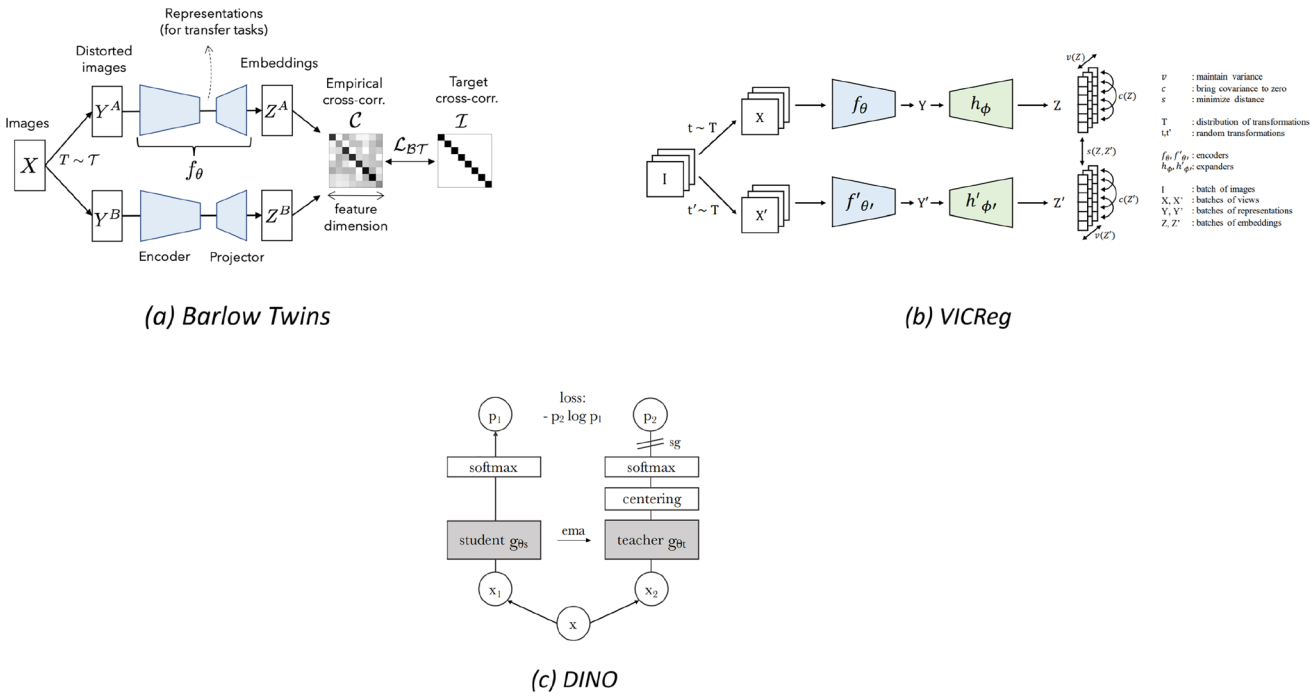

Fig. 11 Collections of Self-Supervised Learning (SSL): **a** Barlow Twins, **b** VICReg, and **c** Self-Distillation with No Labels

invariant to the distortions applied and concurrently to equate the off-diagonal elements of the cross-correlation matrix to 0, to decorrelates the different vector components of the embedding. This decorrelation reduces redundancy and enhances representation quality, especially with high-dimensional output vectors. In general, Barlow Twins performs comparably to other SSL models but slightly lags behind BYOL and SwAV. However, it introduced a novel direction for SSL design, inspiring extensions like Variance-Invariance-Covariance Regularization (VICReg) [37]. VICReg retains Barlow Twins' decorrelation mechanism, penalizing off-diagonal covariance terms, and trains a symmetric architecture based on the principle to preserve the information content of the embeddings using three terms namely: (i) Variance, a hinge loss ensuring embedding variables maintain a minimum standard deviation, (ii) Invariance, minimizing the mean squared distance between embeddings, and (iii) Covariance, enforcing decorrelation to preserve information content.

Like Barlow Twins, VICReg eliminated additional components such as weight sharing between the branches, momentum encoders, negative samples, online clustering, stop gradients, and memory banks but suffers slight performance degradation compared to BYOL and SwAV. To address this, Caron et. al. [38] introduced the *Self-Distillation with No Labels* (DINO), a foundational models producing universal features suitable for image-level visual tasks using a structure similar to [31–35] with knowledge distillation. This model processes two random transformations of an image through student and teacher networks (similar to the online-target networks architecture of BYOL), with the teacher's output centred with a mean computed over the batch and compared to the student's output using cross-entropy loss. A stop-gradient (sg) operator is integrated into the teacher network, restricting gradient flow only through the student network, while the teacher network is updated via exponential moving average (EMA) from the student network. The performance of DINO matches SwAV and surpasses other models. For a fair comparison, Caron et. al. replaced ResNet50 on BYOL and SwAV models with ViT-small architecture. The results show that DINO outperforms both models by a significant margin under the same experimental conditions, which further demonstrating its superior performance. Later, Oquab et al. [15] introduced an enhanced version of DINO, called DINOv2, which drastically improved self-supervised learning (SSL) performance, achieving results comparable to weakly supervised features. DINOv2 was tested and pre-trained on a massive, curated dataset of 142 million images, 100 times larger than ImageNet ILSVRC, using variations of Vision Transformers (ViT) without supervision. The pre-trained DINOv2 model's visual features are compatible with simple classifiers like linear layers, indicating that the learned representations are readily accessible without fine-tuning for downstream tasks. Moreover, DINOv2 demonstrated the superior performance of transformer-based networks over CNNs like ResNet due to its effectiveness to capturing long-range dependencies and contextual information [41]. This advantage is particularly evident in tasks requiring global understanding, such as plant species identification through leaf texture, veins, and shape analysis.

To the best of our knowledge at the point of writing, we have not found any related cross-domain adaptation tasks using DINOv2, specifically for plant species identification. Therefore, we are inspired to adopt DINOv2 as a feature extractor and evaluate its effectiveness for this purpose.

3 Method

The following subsections outline and describe our experimental procedure, network architecture, dataset organization, and evaluation method. Results and analysis are presented in Sects. 4 and 5 respectively.

3.1 Network architecture

We constructs our model as shown in Fig. 12 using of DINOv2 with pre-trained weights (ViT Base with 14 patches on 142 million images, *vit_base_patch14_dinov2.lvd142m*) as a feature extractor (f) and replaced its projection head with a single linear classifier (g). We trains the classifier (g) while freezing the pre-trained feature

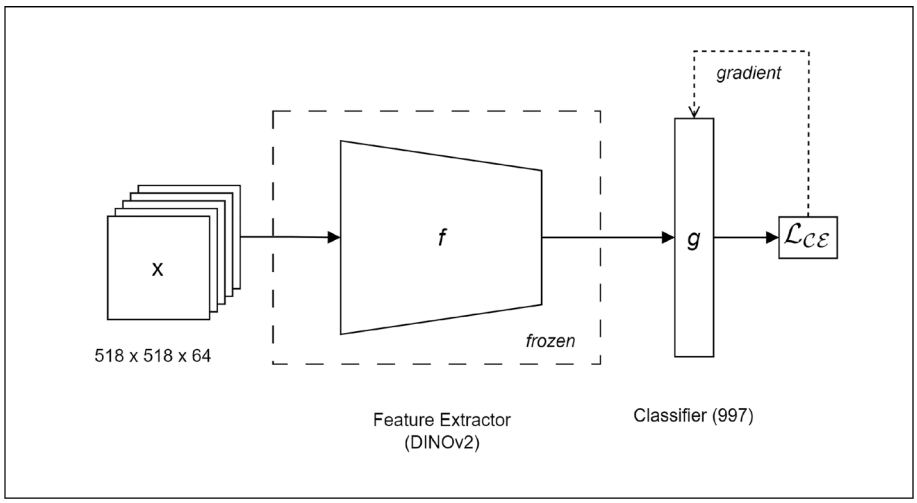

Fig. 12 Proposed network architecture comprising DINOv2 (f) and a single linear classifier (g)

extractor (f) as it has demonstrated strong performance across various domains [15]. The cross-entropy loss ($\mathcal{L}_{CE}$) was employed to calculate the loss and update the weights for the classifier (g) defined as:

$$\mathcal{L}_{CE} = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(\hat{y}_i) \tag{2}$$

3.2 Data preparation

To provide a fair comparison between ours and the state-of-the-art methods, we use the same dataset obtained from the PlantCLEF2020/21 AI Crowd[3] repository. The data were partitioned into four groupsets as listed in Table 4. Three out of the four groupsets are trainsets namely "herbarium" *(D-H)*, "photo" *(D-P)*, and "herbarium_photo_associations" *(D-HPA)*. The last groupset is the testset *(T-P)* consists of 3,186 field images of 408 plant species.

The "herbarium" trainset contains the vast majority of the data: it is a collection of approximately 320 k herbarium scans or images of the 997 plant species while the "photo" trainset contains 4482 images of 376 species. A separate set of images from both the "herbarium" and "photo" groups are combined into the "herbarium_photo_associations" trainset, which contains 1816 images from approximately 224 species, each with field images and one or more herbarium images. The objective of this grouping is to provide researchers with an additional approach for designing the training procedure using the information from either the "herbarium" or "field" images, as well as to explore associations between the two domains.

To better measure and evaluate our model's performance, particularly on the subset (difficult) category, we incorporated an additional test set prepared by Chulif et al. [42] as the Evaluation set *(E-P)*. This groupset focused on the plant species that trained exclusively with herbarium, without field images, and consisting 166 field images from 84 species.

3.3 Evaluation metric

We use Mean Reciprocal Rank (MRR), as defined in Eq. (1) to evaluate our model's performance, ensuring consistency with related works discussed in Sects. 2.2 and 2.3. For each experiment, the model processes all test

[3] https://www.aicrowd.com/challenges/lifeclef-2021-plant.

Table 4 Experimental data preparation

Group set	Image type	Abbreviation	Number of samples	Number of classes
Trainset	Herbarium only	*D-H*	320,752	997
Trainset	Photo only	*D-P*	4482	376
Trainset	Herbarium photo association	*D-HPA*	1816	224
Testset	Field (PlantCLEF2020)	*T-P*	3186	408
Evaluation set	Field (Chulif et. al.[42])	*E-P*	166	84

images and generates 997 records per image that include the *observation ID, predicted class (0–996), probability*, and *ranking*. Given a test set of 3186 images, this process yields a total of 3,176,442 raw records. These records are aggregated based on 638 unique *observation IDs* and 997 *classes*, followed by re-ranking using the mean probability score, producing a final "runfile" with 636,086 entries. Finally, the "runfile" is submitted to the dataset owner for the wholeset and subset MRR score computation.

4 Ablation study and result

In this section, we provide an empirical evaluation of our experiments under various domain-specific settings. Specifically, we ablated four series of experiments using the model specified in Sect. 3.1, in order to evaluate the impact of different configurations on overall performance. The initial experiments was designed to evaluate the model's sensitivity towards different training procedures using single, mixed, and sequential data (trainsets) as inspired by [28] and trained over 10 epochs. Subsequent experiments extend the training to 50 epochs, incorporating variations in training sequences, dataset types, and optimizers. To ensure consistency, all input images were scaled to a size of 518×518, and the batch size was fixed at 64. No data augmentation was applied in any experiment.

4.1 Experiments over different trainsets [A]

A total of 6 initial experiments (*A1-A6*) as shown in Fig. 13 were conducted on 10 epochs with Adam optimizer at the learning rate of 0.001 and weight decay of 0.0001. We first conducted two experiments with a single trainset i.e. *D-H* only and *D-P* only followed by an experiment where we sequentially trained the model with *D-H* first and then fine-tuned with *D-P*. Lastly, we conducted three additional experiments where we mixed and shuffled the trainsets in the form of MixHP (*D-H* + *D-P*), MixHHPA (*D-H* + *D-HPA*), and MixAll(*D-H* + *D-P* + *D-HPA*). We exclude the possible training procedures with (a) A single trainset on *D-HPA* only and (b) Sequential training on *D-H* then on *D-HPA* as the amount of sample in *D-HPA* is relatively small i.e. 1816 which might not produce any interesting findings.

The evaluation results for this series of experiments are shown in Table 5. Most notably, the model trained with *D-H* only achieved the best MRR score with 0.2006 and 0.1255 for the wholeset and subset categories, respectively. In comparison, the model trained with photos only (*D-P*) obtained the lowest for all three categories. The three experiments trained with the Mixed trainset scored the best for the MRR (evaluation category) but lower as compared with the model trained *D-H* only trainset on both the wholeset and subset categories. Lastly, the model trained with *D-H* and *D-P* sequentially does not exhibits significant outcome for the three MRR scores. Hence, we eliminated possible training procedures using *D-P* alone as well as the training with *D-H* and *D-P* sequentially for subsequent experiments.

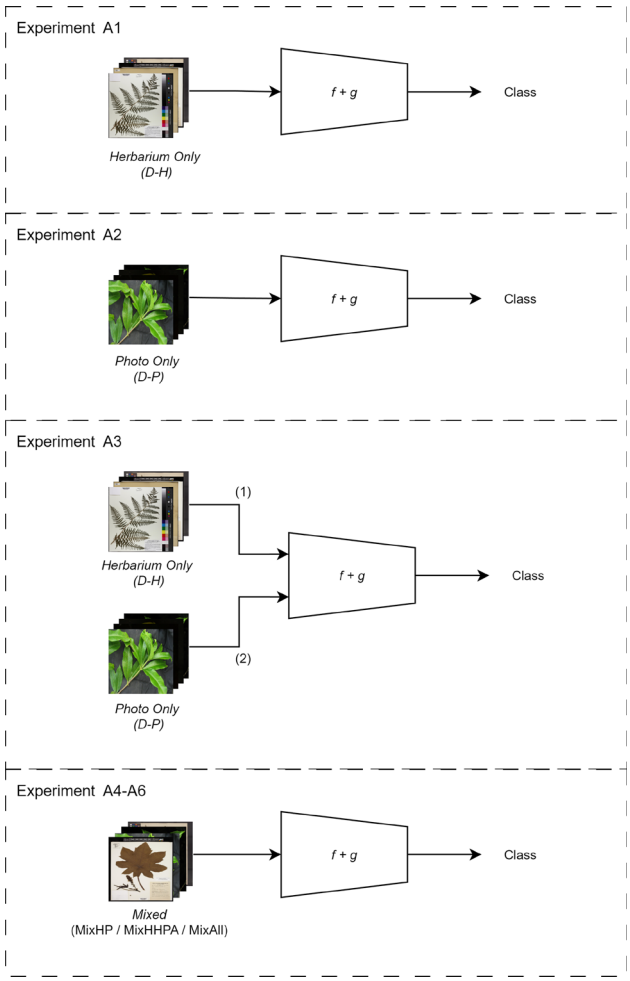

Fig. 13 Training scheme and procedures

Table 5 Initial experimental results with different trainsets and training sequence for 10 epochs on Adam optimizer

Exp	Training setting	MRR[42] (Evaluation)	MRR Wholeset (test-all)	MRR Subset (test-difficult)
A1	Train D-H only	0.2641	**0.2006**	**0.1255**
A2	Train D-P only	0.0021	0.0021	0.0020
A3	Train D-H Then D-P (sequential)	0.1727	0.0979	0.0467
A4	Train MixHP (D-H + D-P)	0.2772	0.1845	0.1211
A5	Train MixHHPA (D-H + D-HPA)	**0.2791**	0.1848	0.1000
A6	Train MixAll (D-H + D-P + D-HPA)	0.2742	0.1730	0.0923

4.2 Experiments over different optimizers [B]

In this series of experiment, we adjust the training procedure along with additional optimizers i.e. Stochastic Gradient Descent (SGD) and Adam with decoupled Weight Decay (AdamW). We also increase the training epochs from 10 to 50 to evaluate if this adjustment helps improve the model's performance. The model is trained with MixAll trainset which is the lowest performance among the three experiments with mixed trainsets conducted previously. The learning rate and weight decay for each optimizer used remain unchanged i.e. 0.001 (lr) and 0.0001 (wd) respectively.

The evaluation results for this series of experiments are shown in the table 6 and the results exhibit two key observations:

(1) *Performance decline with increasing epochs*: Comparing the model trained with MixAll trainset using Adam Optimizer for 50 epochs (B1) to its counterpart which trained for 10 epochs (A6), we observed a decline in performance across all three MRR scores. This decline suggests training with Adam for higher epochs led to overfitting, reducing model's performance in all MRR categories.
(2) *Optimizer sensitivity towards different testset*: For models trained on 50 epochs, the optimizer choice significantly influence performance on different test set categories. For example, the model trained with the SGD optimizer (B2) achieved the highest MRR score for the wholeset (0.1787), while the model trained with the AdamW optimizer (B3) attained the best MRR score for the subset (0.0927). Both results slightly outperformed or matched the initial Adam-trained model (A6).

These findings demonstrate the computational efficiency of DINOv2-based models, as training for 10 epochs achieved performance comparable to training for 50 epochs. This contradict with classical convolutional neural networks such as ResNet50, which typically require 100 or more epochs for optimal performance. Further analysis of this phenomenon will be presented in Sect. 5.

4.3 Experiments over different trainsets and optimizers [C]

The findings from experiments [A] and [B] above indicates a model's performance can be influnced by different training data i.e. Herbarium only (*D-H*), MixHP (*D-H* + *D-P*), MixHHPA (*D-H* + *D-HPA*), and MixAll (*D-H* + *D-P* + *D-HPA*) along with the choices of optimizer (Adam, SGD, and AdamW). To identify the optimal combination for maximizing MRR scores in wholeset and subset categories, we conducted a grid search over 12 settings, combining four trainsets with three optimizers.

To avoid overfitting as observed in experiment A6 vs. B1 (50 epochs), we incorporated the *ReduceLROnPlateau* learning scheduler for subsequent experiments to detect and reduce the learning rate when a model stops learning. This configuration will adjusts the learning rate by a factor of 0.1 when no improvement is detected after 5 epochs (*patience* = *5*), preventing gradient explosion and performance degradation. This configuration was excluded for earlier experiments as we presume the learning scheduler may not be activated and subsequently updating the learning rate due to a lower number of epochs i.e. 10 or 50. We utilized the photo-only trainset (*D-P*) as the validation set to track accuracy and loss along with the training accuracy and loss during the training process.

Table 7 outlines the training configuration, while the result (C1−C12) are detailed and visualized in Table 8 and Fig. 14, respectively.

In this series of experiments, the model trained with *D-H* trainset using SGD optimizer (C1) achieved the highest MRR scores and outperformed previous experiments for both the wholeset (0.2126) and evaluation categories (0.3264). The model trained with *MixHHPA* trainset on AdamW optimizer (C9) achieved the best result for subset MRR (0.1083) but still underperformed compared to the best model trained with *D-H* trainset on Adam optimizer (A1) trained for 10 epochs (0.1255). On the other hand, SGD optimizer generally outperformed

Table 6 Experimental result trained with MixAll trainset for 50 epochs on Adam, SGD, and AdamW optimizers

Exp	Training setting (Optimizer)	Epochs	MRR[42] (Evaluation)	MRR wholeset (test-all)	MRR subset (test-difficult)
A6	Adam (Initial)	10	0.2742	0.1730	0.0923
B1	Adam	50	0.2627	0.1554	0.0753
B2	SGD	50	**0.2679**	**0.1787**	0.0726
B3	AdamW	50	0.2321	0.1325	**0.0927**

Table 7 Experimental set-up and configurations

Properties	Settings	Remarks
Model architecture	DINOv2 (Pre-trained) + Classifier	Freeze DINOv2 and train Classifier only
Epochs	50	
Trainsets used	*D-H* only, MixHP (*D-H* + *D-P*), MixHHPA (*D-H* + *D-HHPA*), MixAll (*D-H* + *D-P* + *D-HHPA*)	
Optimizer	Adam, SGD, AdamW	Learning rate: 0.001, weight decay: 0.0001
Scheduler	ReduceLROnPlateau	Factor: 0.1, patience: 5

Table 8 Experimental result trained with different trainsets for 50 epochs on Adam, SGD, and AdamW optimizers with scheduler (sorted by higher MRR wholeset to lower MRR wholeset)

Exp	Trainset	Optimizer	MRR[42] (Evaluation)	MRR wholeset (test-all)	MRR subset (test-difficult)
C1	D-H	SGD	**0.3264**	**0.2126**	0.0930
C2	MixHHPA	SGD	0.3035	0.2082	0.0894
C3	MixHP	SGD	0.2952	0.1824	0.0638
C4	MixAll	SGD	0.2866	0.1799	0.0718
C5	D-H	Adam	0.2839	0.1695	0.1015
C6	MixHHPA	Adam	0.2615	0.1631	0.0920
C7	MixHP	Adam	0.2581	0.1578	0.0836
C8	D-H	AdamW	0.2134	0.1480	0.1008
C9	MixHHPA	AdamW	0.2161	0.1471	**0.1083**
C10	MixAll	Adam	0.2593	0.1465	0.0778
C11	MixAll	AdamW	0.2119	0.1312	0.0721
C12	MixHP	AdamW	0.2155	0.1284	0.0787

Fig. 14 Comparison of MRR Score for models trained with different trainsets (x-axis) against different optimizers with scheduler group by testsets (subplots **a**, **b**, and **c**)

its counterparts for evaluation and wholeset categories (Fig. 14a, b), but worst for the subset category (Fig. 14c). There is no clear indicator emerged for the optimal trainset-optimizer combination in the subset category.

Further analysis of training and validation accuracy indicates the learning scheduler only updated once for models trained with Adam optimizer across trainsets (Fig. 15, second row), while models trained with SGD and AdamW were unaffected. This led to further experiments using a validation loss-based learning schedule to refine performance.

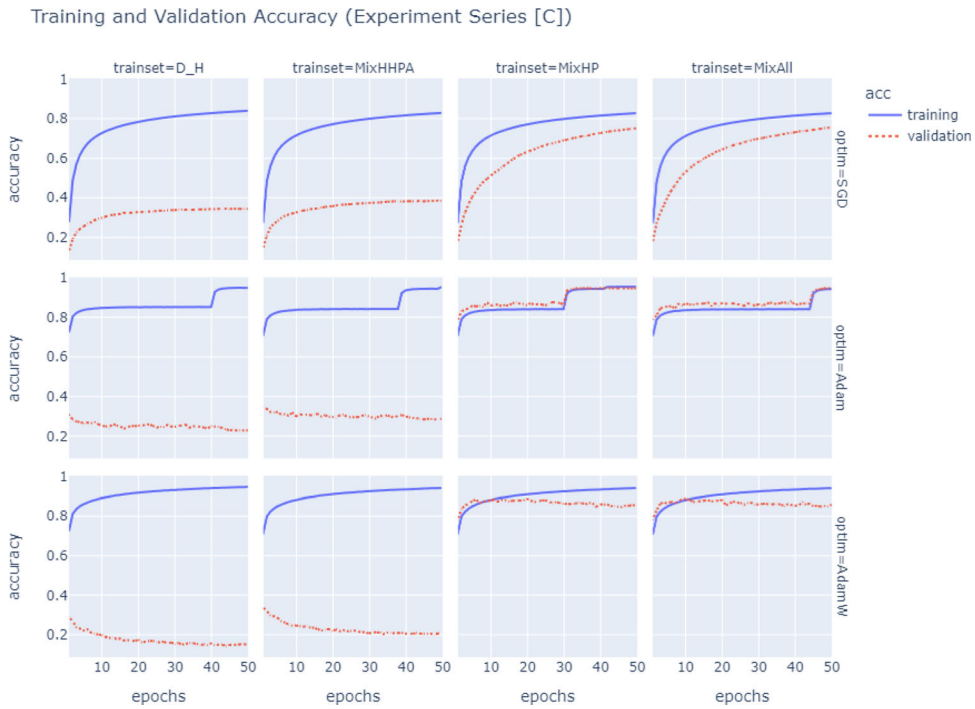

Fig. 15 Training and validation accuracy

4.4 Extra experiments [D]

To simulate real-world condition enabling machine to learn from herbarium or mixed data and classifying plant species using field image, we repeated the previous experimental set-up [C] but configured the learning scheduler to be updated based on validation loss instead of training loss. The result for these experiments (D1-D12) are presented in Table 9 and Fig. 16.

Consistent with prior findings, models trained with the SGD optimizer generally outperformed others for both the wholeset and evaluation categories except the model trained with *MixAll* on evaluation set (D5) in this series of experiment. Specifically, the model trained with *MixHHPA* scoring 0.1996 (D1) and 0.3228 (D2) for the

Table 9 Experimental result trained with different trainsets for 50 epochs on Adam, SGD, and AdamW optimizers with scheduler updated by validation loss (sorted by higher MRR wholeset to lower MRR wholeset)

Exp	Trainset	Optimizer	MRR[42] (Evaluation)	MRR wholeset (test-all)	MRR subset (test-difficult)
D1	MixHHPA	SGD	0.2976	**0.1996**	0.0807
D2	D-H	SGD	**0.3228**	0.1953	0.0782
D3	D-H	Adam	0.2884	0.1878	0.1094
D4	MixHP	SGD	0.2803	0.1819	0.0797
D5	MixAll	SGD	0.2520	0.1772	0.0684
D6	MixHHPA	Adam	0.2883	0.1738	0.0915
D7	D-H	AdamW	0.2415	0.1734	**0.1150**
D8	MixHP	Adam	0.2596	0.1683	0.0863
D9	MixAll	Adam	0.2697	0.1633	0.0868
D10	MixHHPA	AdamW	0.2458	0.1591	0.0902
D11	MixHP	AdamW	0.2199	0.1287	0.0730
D12	MixAll	AdamW	0.2149	0.1185	0.0668

Fig. 16 Comparison of MRR Score for models trained with different trainsets (x-axis) against different optimizers with scheduler (update based on validation loss) group by testsets (subplots **a**, **b**, and **c**).

wholeset and evaluation categories, respectively, while the model trained with *D-H* using AdamW optimizer (D7) scores 0.115 and achieved the highest score for the subset category. When comparing the two series of experiments, only models trained with Adam optimizer (D3, D6, D8, and D9) showed slight improvement over the previous series (C5, C6, C7, and C10), but none surpassed the earliest models from initial experiments (A1 and A4).

5 Discussion

This section discusses four insightful findings revealed from the four series of experiments conducted:

1. *Train on lower epochs over high epochs*: A key observation throughout experiments was the impact of epochs configuration on model performance. Figure 17 illustrates the transition of MRR scores for Adam optimizer across all trainsets in subplots (a)–(d).

The models trained on 10 epochs during the initial experiments [A] generally outperformed those trained on 50 epochs for the wholeset (red dotted line) and subset (green dashed line) categories, except for the evaluation test set (blue solid line). This suggested the DINOv2's feature extractor enables faster convergence for linear classifiers compared to other state-of-the-art models, such as ResNet or InceptionNet, which require more epochs for effective feature extraction.

2. *Optimizer sensitivity to test set categories*: The findings from experiment [C] and [D] suggest that models trained using SGD optimizer generally perform better across all trainsets on the evaluation and wholeset category while the opposite for the subset category. This findings are obvious as illustrated in Fig. 18 (concatenation of Figs. 14 and 16). Althought no consistent pattern observed for the optimal combination between trainset-optimizer pairs for the subset category, some noticeable spikes are observed for models trained with *MixHHPA* (subplot in row 1, col 3) and *D-H* (subplot in row 2, col 3) trainsets using AdamW optimizer, which excels on the subset category in both series of experiments.

This suggests the possibility of combining these top-performing models into an ensemble, as proposed by Chulif et al. [42] to improve performance through majority voting during the inference process.

Fig. 17 Transition of MRR scores for the model trained on Adam optimizer during different experiments series (x-axis) against different trainsets (subplots **a**, **b**, **c**, and **d**)

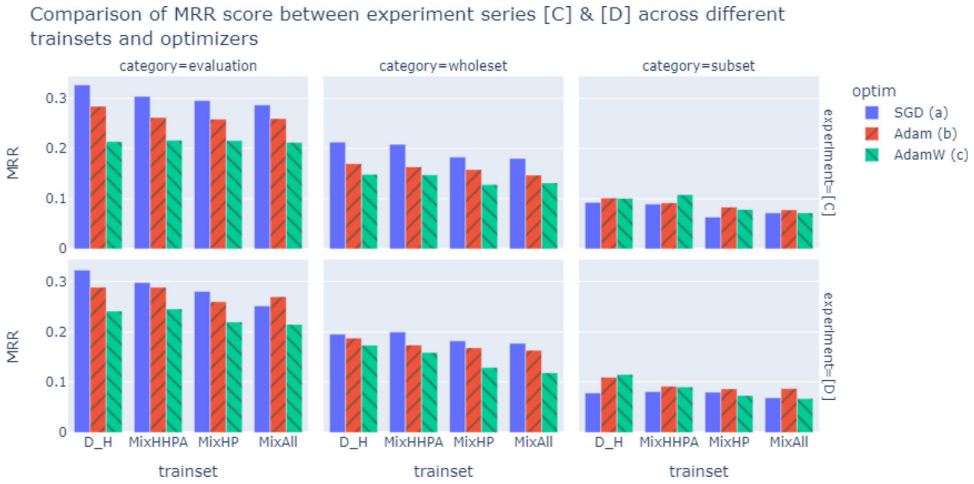

Fig. 18 Comparison of MRR between experiment series [C] and [D] across different trainsets on different optimizers

3. *Strength of DINOv2 as feature extractor*: Table 10 and Fig. 19 presents a performance comparison between our models and prior works discussed in Sects. 2.2 and 2.3 for both the wholeset and subset categories. Our best models achieved MRR scores of 0.213 (C1) and 0.201 (A1), outperforming the previous state-of-the-art model, HFTL-ENS and OSM-ENS (0.181) [27] by 17.7% (+0.032 MRR), showcasing effective cross-domain adaptation from herbarium data to field images.

Our observation found that HFTL-ENS and OSM-ENS model used hybrid feature transfer learning with an ensemble of mixed networks but relies on the InceptionNet v4 pre-trained on the ILSVRC 2012 dataset (1.28M training images, 1000 classes) [43], which limiting robustness compared to DINOv2. DINOv2 on the other hand, was pre-trained on LVD-142 M, a larger, more diverse dataset, enhancing its feature extraction for image classification.

Although a comparison based on dataset size may seem biased, it is justified. First, DINOv2 achieves superior results with large datasets and outperforms traditional convolutional neural networks (CNNs) like ResNet or InceptionNet while requiring less computational power [13, 38]. Secondly, CNNs perform well on smaller datasets but struggle with larger ones, explaining why earlier CNN-based models discussed in Sects. 2.2 and 2.3 underperform even with substantial pre-training data. This distinction subsequently justifies how our models that leveraging DINOv2 outperforms previous works (CNN-based methods), particularly given the higher train-test

Table 10 Comparison of MRR score from previous and ours works on PlantCLEF dataset

Model	MRR wholeset (test-all)	MRR subset (test-difficult)	MRR drop whole minus subset
HFTL-ENS and OSM-ENS [28]	0.181	**0.158**	13%
FSADA [24]	0.153	0.056	63%
FSADA [21]	0.148	0.039	74%
HFTL [27]	0.121	0.107	**11%**
WPLR [19]	0.065	0.037	43%
ACL [17]	0.032	0.016	50%
Resnet50 (Baseline) [21]	0.002	0.002	0%
DINOv2 (ours, C1)	**0.213**	0.093	56%
DINOv2 (ours, A1)	0.201	0.126	37%
DINOv2-baseline (ours-No Herbarium, A2)	0.002	0.002	0%

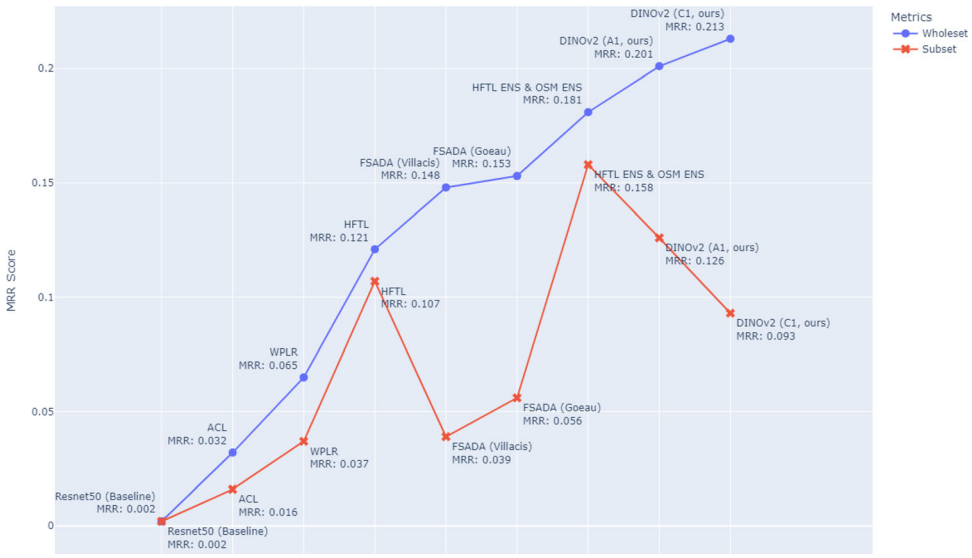

Fig. 19 Comparison of previous and our models using MRR Scores on Wholeset (blue) and Subset (red), highlighting performance improvements across different approaches

image ratio (103:1) in PlantCLEF2020 compared to ImageNet-1K (13:1), crucial for cross-domain plant species identification.

Lastly, the vision transformer-based model architecture utilizes a multi-head self-attention mechanism, which enables the model to understand the image globally. This is in contrast to the CNN-based model, which relies on local receptive fields and long-range dependencies to classify images [41, 44, 45]. To further investigate the effect of receptive fields, we perform a visual analysis using gradient-weighted class activation mapping (GradCAM) [46], a gradient-based visualization technique that identifies the critical regions influencing the classification decision by quantifying the sensitivity of the target class scores to the feature maps of a neural network.

Figure 20 shows the feature maps generated by GradCAM for the "Monstera adansonii" plant (class label 27760) on various models trained in our experiments [C] and additional models for ResNet50. The feature maps are visualized using a heatmap colour scheme, where red, yellow, and blue regions represent areas of high,

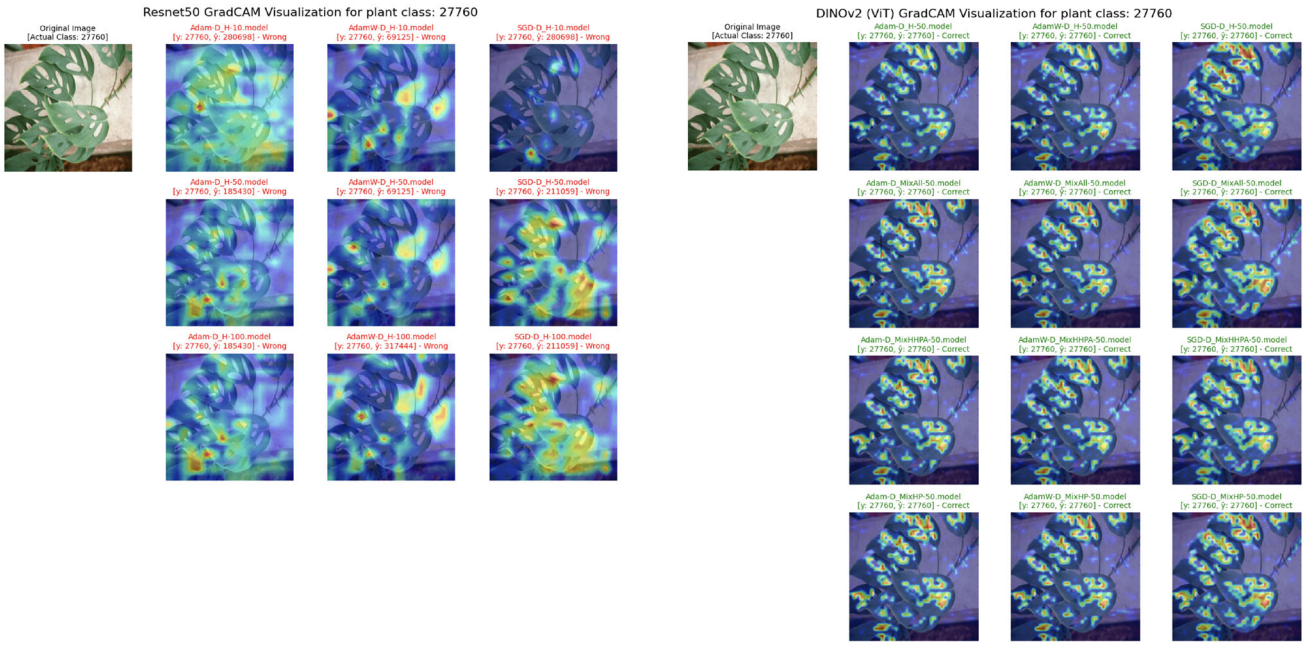

Fig. 20 Visual analysis of feature maps between ResNet50 (CNN-based, left) and DINOv2 (ViT-based, right) models on various training configurations for "Monstera adansonii" (plant class label 27760)

intermediate, and low activation, respectively. These activation patterns indicate the regions that the models focus on most intensively during the classification process. The left figure illustrates the feature maps for the ResNet50 (CNN-based) model, which appear scattered and fragmented. This may have contributed to the model's misclassification of the plant. For example, the feature maps generated from the ResNet50 model trained with Adam optimizer generally lost their focus, while the model trained with AdamW optimizer seems to focus on the background instead of the plant itself. In contrast, the DINOv2 (ViT-based) feature maps on the right exhibit a more consistent focus region and multiple attention patterns, which likely aid the classification task. Notably, the shape of the attention regions in the DINOv2 feature maps is more pronounced (e.g. the leaf region near the hole), suggesting that the ViT-based architecture is more sensitive to the overall shape, vein, and other organs/characteristics of the plant compared to the local features captured by the CNN-based model.

4. *Limitation of DINOv2 as feature extractor*: We appraised DINOv2's strength and robustness in extracting discriminative features from herbarium data and its capability to transfer knowledge to field images, as demonstrated by our best models in the wholeset category. Conversely, DINOv2 still struggles with domain generalization. Using our best models, it only outperformed some previous works in the subset category, achieving MRR scores of 0.126 (A1) and 0.093 (C1), which were 20.3% (−0.032) and 41.1% (−0.065) lower than the HFTL-ENS and OSM-ENS model (0.158), ranking 2nd and 4th, respectively (Fig. 19).

Our models' performance declined as field images were incrementally introduced into the training dataset, transitioning from the *D-H* trainset (0 field images) to *MixedHHPA* (1342 field images), *MixedHP* (4482 field images), and *MixedAll* (5824 field images) in experiment [C] and [D] (Fig. 18). Additionally, DINOv2 remains less effective in few-shot learning, as reflected in the significant MRR disparities between the wholeset and subset categories (C1 56%, A1: 37%). Species in the subset category often lack sufficient herbarium and field images, limiting the model's ability to capture subtle variations and train effective linear classifiers. Analysis of the herbarium-field training image distribution (Fig. 2) shows that over half the species (on the right side) have at least 50 up to 3700 herbarium images, regardless of with or without field images. In contrast, species in the subset category (represented by datapoints with three circles located on the left side) not only lack field images but also have fewer herbarium images, further hindering the model's performance. This further justify why DINOv2 being a transformer-based model, excels with large datasets but struggles with limited data and impacting its ability to generalize [13].

These limitations underscore the need for improvements in DINOv2 to enhance its robustness, particularly for imbalanced datasets and for crossdomain applications.

6 Conclusion

In this paper, we conducted a comprehensive study on cross-domain plant species identification on PlantCLEF2020 dataset, which consists of approximately 320k herbarium and field images representing 997 plant species. We highlighted the challenges of this task, primarily due to the limited and imbalanced data for the field domain data, as collecting and annotating field images is costly and impractical. In addition, the identification accuracy of previous work remains low, with Mean Reciprocal Rank (MRR) scores of 0.181 and 0.158 for two evaluation categories (wholeset and subset, respectively).

Our key contribution exhibits a novel application of DINOv2 that leverages the strengths of transformer architecture and uncovers its superior feature extraction capabilities for this task. Specifically, our best model shows substantial improvement for the wholeset category, achieving a 17.7% increase in performance (+0.032 to 0.213 MRR score) compared with the best model from previous works (HFTL-ENS and OSM-ENS). This signifies the effectiveness of DINOv2 in transferring knowledge by extracting important and discriminative features from the source domain (i.e. herbarium images) and training a classifier for prediction in the target domain (i.e. field images). Furthermore, our model converged faster than models which employ conventional

CNNs like ResNet50 or InceptionNet, even with fewer training epochs (e.g. 10 or 50). However, DINOv2 exhibits limitations in few-shot learning scenarios and when training data is scarce, as demonstrated by our evaluation. Our model struggled to accurately differentiate species with limited or no field images and fell marginally short of 20.3% (−0.032) for the subset category as compared to the best model from previous work, achieving 0.126 MRR score only. These findings suggest that pre-trained DINOv2 as a feature extractor may not yet reach optimal efficiency for domain generalization tasks.

The broader impact of our findings is substantial and significant. Our results demonstrate the viability of transformer-based models and self-supervised learning for cross-domain plant species identification. This approach not only improves identification accuracy but also offers a scalable solution for handling diverse and complex plant species data. The advancements made in this study can be applied to other domains facing similar challenges with limited and imbalanced data. We hope this work offers valuable insight for the research community on the potential and area for improvement of DINOv2 along with the self-distillation approach in the self-supervised learning (SSL) arena.

In future, we strive to explore further in this direction by addressing the limitations posed by DINOv2 and SSL for cross-domain adaptation and few-shot learning tasks. We aim to enhance model performance by incorporating additional data augmentation or segmentation, and advanced techniques such as Generative AI or metrics learning.

Acknowledgements The authors want to express utmost sincere gratitude to Dr. Hervé Goëau from CIRAD, UMR AMAP, Montpellier, Occitanie, France for his diligent efforts in validating the results and ensuring the accuracy and reliability of the experiments. The authors also wish to thank Dr. Abel Yu Hao Chai, from Swinburne University of Technology for his support and advice when preparing the revised manuscript.

Author Contributions Chin Ann Ong (A), Fei Siang Tay (B), Yi Lung Then (C), Chris McCarthy (D) (A), (B), and (C) conceived of the presented idea. (A) and (C) collected the data. (D) contributed and supported on HPC facilities set-up for experimentation. (A) and (B) developed the theoretical formalism and performed the computations. (A) conducted the experiment and preliminary analysis, and wrote the manuscript with support from (B) and (C). (B), (C), and (D) proofread and contributed to the manuscript. All authors discussed the results and contributed to the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research/first author is funded by the Swinburne University of Technology HDR Scholarship. No other form of funding was received while producing this work.

Data availability The PlantCLEF dataset can be found at the PlantCLEF2020/21 AI Crowd Repository https://www.aicrowd.com/challenges/lifeclef-2021-plant.Experimentation results and findings are presented in the tabular formatted within the manuscript.

Declarations

Conflict of interest We declare that there is no Conflict of interest of any kind to the best of our knowledge.

References

1. Cope JS, Corney D, Clark JY, Remagnino P, Wilkin P (2012) Plant species identification using digital morphometrics: a review. Expert Syst Appl 39(8):7562–7573. https://doi.org/10.1016/j.eswa.2012.01.073

2. Dwivedi P, Sharma DK (2023) Plant species classification using information measure and deep learning for an actual environmental problem. In: Recent developments in machine and human intelligence, pp 208–227. IGI Global, Hershey, PA, USA
3. Mulugeta AK, Sharma DP, Mesfin AH (2024) Deep learning for medicinal plant species classification and recognition: a systematic review. Front Plant Sci 14:1286088
4. Saiz-Rubio V, Rovira-Más F (2020) From smart farming towards agriculture 5.0: a review on crop data management. Agronomy 10(2):20207. https://doi.org/10.3390/agronomy10020207
5. Katal N, Rzanny M, Mäder P, Wäldchen J (2022) Deep learning in plant phenological research: a systematic literature review. Front Plant Sci 13:805738
6. Lee SH, Goëau H, Bonnet P, Joly A (2021) Conditional multi-task learning for plant disease identification. In: 2020 25th international conference on pattern recognition (ICPR), pp 3320–3327. https://doi.org/10.1109/ICPR48806.2021.9412643
7. Wäldchen J, Mäder P (2018) Plant species identification using computer vision techniques: a systematic literature review. Arch Comput Methods Eng 25(2):507–543
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
9. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 31
10. Goëau H, Bonnet P, Joly A (2020) Overview of lifeclef plant identification task 2020. CEUR-WS
11. Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of herbarium specimens. BMC Evol Biol 17(1):1–14
12. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76
13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: transformers for image recognition at scale
14. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A (2021) Do vision transformers see like convolutional neural networks? Adv Neural Inf Process Syst 34:12116–12128
15. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A et al (2023) Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144
17. Zhang Y, Davison BD (2020) Adversarial consistent learning on partial domain adaptation of plantclef 2020 challenge. arXiv preprint arXiv:2009.09289
18. Zoph B, Vasudevan V, Shlens J, Le QV (2018) Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8697–8710
19. Zhang Y, Davison BD (2021) Weighted pseudo labeling refinement for plant identification
20. Sun B, Saenko K (2016) Deep coral: Correlation alignment for deep domain adaptation. In: computer vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14, pp. 443–450. Springer
21. Villacis J, Goëau H, Bonnet P, Joly A, Mata-Montero E (2020) Domain adaptation in the context of herbarium collections. a submission to plantclef 2020. CEUR-WS
22. Motiian S, Jones Q, Iranmanesh S, Doretto G (2017) Few-shot adversarial domain adaptation. Advances in neural information processing systems **30**
23. Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision, Springer, pp 69–84
24. Goëau H, Bonnet P, Joly A (2021) Overview of plantclef 2021: cross-domain plant identification. In: CLEF (Working Notes), pp 1422–1436
25. Kaya M, Bilge HŞ (2019) Deep metric learning A survey. Symmetry 11(9):1066
26. Lu J, Hu J, Zhou J (2017) Deep metric learning for visual understanding: an overview of recent advances. IEEE Signal Process Mag 34(6):76–84
27. Chulif S, Chang YL (2021) Herbarium-field triplet network for cross-domain plant identification. In: Experimental IR meets multilinguality, multimodality, and interaction: 12th international conference of the CLEF association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, pp 173–188
28. Chulif S, Chang YL (2021) Improved herbarium-field triplet network for cross-domain plant identification: Neuon submission to lifeclef 2021 plant
29. Balestriero R, Ibrahim M, Sobal V, Morcos A, Shekhar S, Goldstein T, Bordes F, Bardes A, Mialon G, Tian Y et al (2023) A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210
30. Gui J, Chen T, Cao Q, Sun Z, Luo H, Tao D (2023) A survey of self-supervised learning from multiple perspectives: algorithms, theory, applications and future trends. arXiv preprint arXiv:2301.05712

31. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
32. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp 1597–1607, PMLR
33. Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M et al (2020) Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 33:21271–21284
34. Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A (2020) Unsupervised learning of visual features by contrasting cluster assignments. Adv Neural Inf Process Syst 33:9912–9924
35. Chen X, He K (2021) Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15750–15758
36. Zbontar J, Jing L, Misra I, LeCun Y, Deny S (2021) Barlow twins: self-supervised learning via redundancy reduction. In: International conference on machine learning, PMLR, pp 12310–12320
37. Bardes A, Ponce J, LeCun Y (2021) Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906
38. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers
39. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
40. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
41. Liu Z, Qian S, Xia C, Wang C (2024) Are transformer-based models more robust than CNN-based models? Neural Netw 172:106091
42. Chulif S, Lee SH, Chang YL, Chai KC (2023) A machine learning approach for cross-domain plant identification using herbarium specimens. Neural Comput Appl 35(8):5963–5985
43. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115:211–252
44. Tuli S, Dasgupta I, Grant E, Griffiths TL (2021) Are convolutional neural networks or transformers more like human vision? arXiv preprint arXiv:2105.07197
45. Naseer MM, Ranasinghe K, Khan SH, Hayat M, Shahbaz Khan F, Yang M-H (2021) Intriguing properties of vision transformers. Adv Neural Inf Process Syst 34:23296–23308
46. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626

Authors and Affiliations

Chin Ann Ong[1,2] · Fei Siang Tay[1] · Yi Lung Then[3] · Chris McCarthy[4]

✉ Chin Ann Ong
cong@swinburne.edu.my; chinann.ong@ntu.edu.sg

Fei Siang Tay
fstay@swinburne.edu.my

Yi Lung Then
ylthen@unimas.my

https://doi.org/10.1007/s00521-025-11499-6

Chris McCarthy
cdmccarthy@swin.edu.au

[1] Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Kuching, Malaysia

[2] College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

[3] Faculty of Engineering, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia

[4] School of Science, Computing and Emerging Technologies, Swinburne University of Technology, Melbourne, Australia