Neural Computing and Applications (2023) 35:5963–5985
https://doi.org/10.1007/s00521-022-07951-6

ORIGINAL ARTICLE

A machine learning approach for cross-domain plant identification using herbarium specimens

Sophia Chulif[1,2] · Sue Han Lee[1] · Yang Loong Chang[2,3] · Kok Chin Chai[2]

Received: 23 March 2022 / Accepted: 12 October 2022 / Published online: 12 November 2022

Abstract

The preservation of plant specimens in herbaria has been carried out for centuries in efforts to study and confirm plant taxa. With the increasing collection of herbaria made available digitally, it is practical to use herbarium specimens for the automation of plant identification. They are also substantially more accessible and less expensive to obtain compared to field images. In fact, in remote and inaccessible habitats, field images of rare plant species are still immensely lacking. As a result, rare plant species identification is challenging due to the deficiency of training data. To address this problem, we investigate a cross-domain adaptation approach that allows knowledge transfer from a model learned from herbarium specimens to field images. We propose a model called Herbarium–Field Triplet Loss Network (HFTL network) to learn the mapping between herbarium and field domains. Specifically, the model is trained to maximize the embedding distance of different plant species and minimize the embedding distance of the same plant species given herbarium–field pairs. This paper presents the implementation and performance of the HFTL network to assess the herbarium–field similarity of plants. It corresponds to the cross-domain plant identification challenge in PlantCLEF 2020 and PlantCLEF 2021. Despite the lack of field images, our results show that the network can generalize and identify rare species. Our proposed HFTL network achieved a mean reciprocal rank score of 0.108 and 0.158 on the test set related to the species with few training field photographs in PlantCLEF 2020 and PlantCLEF 2021, respectively.

Keywords Plant identification · Herbarium · Triplet loss · Convolutional neural networks · Computer vision

1 Introduction

Plants make up the majority of the earth's life-form, taking up 80% of the global biomass [1]. They provide us with food, raw materials, medicine, regulate the air we breathe, protect us from extreme climates, prevent soil erosion, support wildlife habitat and diversity, and even help our physical and mental health. The World Checklist of Vascular Plants records around 350,000 known vascular plant species. It is the most extensive and frequently updated species list of its kind [2]. The publishing of correctly documented plant records is essential for reference, research, breeding programs, conservation, and plant reintroduction [3]. An accurate and reproducible way of plant identification enables the reliable application of plants such as the integration of digital datasets and traditional medicine [4]. Internationally, efforts to name plants are being made and around 2000 new plant species are discovered annually. The morphological and molecular characteristics of plants and their similarity to plants

✉ Sophia Chulif
schulif@swinburne.edu.my

Sue Han Lee
shlee@swinburne.edu.my

Yang Loong Chang
yangloong@neuon.ai

Kok Chin Chai
kc@neuon.ai

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

[2] Department of Artificial Intelligence, NEUON AI SDN. BHD., Kota Samarahan, Malaysia

[3] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

already discovered are used to classify new species [5]. However, this task is challenging and time-consuming due to the laborious identification process. First, a scientist has to acquire the suspected unknown plant specimen to keep in a herbarium. Then, they need to ensure it has not already been reported by comparing it to reference specimens of similar species. Finally, a name has to be chosen and published with its characteristics to the scientific literature [2].

Over the years, different methods have been investigated to alleviate and automate the plant identification process. Many researchers have adopted the use of computer vision in plant recognition due to the advancement in computational infrastructure. In early studies, handcrafted feature approaches dominated computer vision techniques. They are used to extract important features from an image to characterize the properties of the plant. These properties are also known as descriptors. Various feature descriptors such as scale-invariant feature transform [6], histogram of gradients [7], and speeded-up robust features [8] are employed together with conventional machine learning classification algorithms like K-nearest neighbors [9] and support vector machines [10] for plant identification. The extracted features from the image form definitions, also known as "bag-of-words", which are then used to classify the image. Examples of these early studies include the identification of tree species [11], medicinal plants such as herbs and shrubs [12], flowers [13], house plants [14], and weed [15]. However, this approach can be challenging as the important features are, as its name implies, handcrafted. The computer vision engineer may have to face a long trial and error process of fine-tuning the parameters to suit the classification of many different classes [16].

Since the winning deep learning architecture in the ImageNet Challenge, AlexNet [17] was first introduced in 2012, the development of deep learning architectures for computer vision tasks has significantly improved by researchers [18]. Deep learning methods consist of a training phase in which the classifier learns to distinguish classes from a given dataset. Since they are trained instead of programmed, they require less fine-tuning and expert analysis. In addition, the employment of deep learning convolutional neural networks (CNNs) in plant identification has shown tremendous achievement compared to conventional handcrafted feature approaches as demonstrated in [19–22]. This achievement has led to the rise of automated plant identification applications such as LeafSnap [23], Pl@ntNet [24], iNaturalist [25], and Flora Incognita [26]. Furthermore, to reduce crop loss, machine learning models such as plant disease detection [27–30], invasive plant detection [31], and pest detection [32] have been implemented. In regard to plant identification, it has been shown that deep learning approaches even outperform human experts [33–35]. Nevertheless, these applications tend to only work well on species with sufficient training data [18, 36]. The major challenges of automated systems include the variability in view and quality of images taken by users in the field, the plant characters distinguishing between species not present in the images available, and the poorly represented (or absent) species in the reference or training images used for identification [37].

It requires a tremendous effort and cost to collect fresh specimens or new field images, either known or unknown to science, to extend existing data. With that in mind, it is worth investigating the existing data for plant identification, i.e., plant specimens in herbaria. According to Index Herbariorum [38], over 3000 active herbaria in the world are accommodating approximately 390 million specimens. Plant specimens are dried and systematically stored in herbarium sheets, which contain crucial information such as their taxon, habitat, and color. These preserved collections of plants have long-yielded valuable information, but no investigation has been thoroughly conducted to utilize this resource [2]. Herbaria are often crowded, leaving specimens unprocessed and inaccessible for study. Moreover, the lack of experts in taxa results in a delay in recognition. Due to this reason, new species are overlooked, misplaced, or assigned to unidentified material at the end of each family [39].

Nonetheless, recent joint efforts are facilitating the aggregation and digitization of herbarium specimens. These include the Botanical Information and Ecology Network, the Australian Virtual Herbarium, and JSTOR Global Plants. Much effort is underway to increase the accessibility of the herbaria collection. Furthermore, the Global Biodiversity Information Facility has made available many herbarium specimens digitally accessible to all [2]. The study by [39] demonstrated that significant numbers of undescribed species have already been collected and are housed in herbaria. To discover new species in herbaria, prudent and continuous examination of all specimens across the range of a taxon is required. In essence, herbaria serve as a key enabler in species discovery.

In correspondence to the cross-domain plant identification challenge of PlantCLEF 2020 and PlantCLEF 2021, we present our approach in implementing a two-streamed Herbarium–Field Triplet Loss Network (HFTL network) to classify field species from the training data made primarily of herbarium images. This achievement could help identify not only species already present in herbaria but also adapt to the identification when the species is rare or threatened.

2 Related works

The authors in [40] presented one of the earliest works of computer vision on herbarium specimens. They used a combination of leaf shape and vein features in their study to identify 26 herbarium species. They first segment the leaves and then apply normalization techniques to counteract leaf shape distortion. Finally, the features were extracted using Fourier descriptors [41] from the normalized image and fed into a support vector machine for classification. Their results show that it is possible to classify herbarium species with overlapping leaves and few training images with satisfactory accuracy.

On the other hand, the authors in [42] have demonstrated the potential of deep learning on herbarium species identification on a larger dataset with thousands of images from herbaria. By using CNNs, they have shown that it is possible to do transfer learning on herbarium specimens even from different regions. However, they did show that it is not advantageous to do transfer learning from herbarium data to leaf scan images, and it is even unfavorable to do transfer learning from herbarium data to field images with CNNs. It is due to the morphological difference between the dried herbarium specimens and field (real-world) plant images. The herbarium specimens which are dried and flattened significantly differ from their field image counterparts. Not only do they differ in color, but also their available plant organs. The pressed organs (e.g., flowers and fruits) in a herbarium specimen are completely transformed and often overlapped by leaves [42]. Furthermore, the unfavorable results of transfer learning from herbarium data to field images with classical CNNs were also shown in the PlantCLEF 2020 Challenge [43–45].

In addition, the use of higher plant taxonomy (i.e., Family and Genus) was shown to improve herbarium species identification in [46]. The authors employed several CNN architectures, i.e., multi-task network and hierarchical network with higher taxonomic data, and showed improvements over a single species classifier. Moreover, in recent years, several competitions in automated herbarium identification, such as the Herbarium Challenge 2019, 2020, 2021 [47] have been created to encourage the development of better models for herbarium identification. With the increase of more comprehensive datasets and training methods, it provides an opportunity for plant experts and computer scientists to address difficult species identification problems.

2.1 Motivation

Herbarium specimens are commonly used in automated herbarium identification. However, utilizing herbarium specimens for automated identification of plants in the field is considerably new. Considering the significant variation between the morphological features of herbarium sheets and field images, it is not appropriate to use only herbarium images for generalization. Therefore, this paper studies the cross-domain plant identification between herbarium and field images. We adopt the triplet loss function introduced by [48] in our network to optimize the embeddings of herbarium and field images instead of directly classifying them as conventional CNNs. The triplet loss, L can be mathematically represented as follows:

$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right] \quad (1)$$

$f(x) \in \mathbb{R}^d$ denotes the embedding that embeds an image x into a d-dimensional Euclidean space. This embedding is constrained to live on the d-dimensional hypersphere that is $\|f(x)\|_2 = 1$. i denotes the ith input. Meanwhile, x_i^a represents the anchor image, x_i^p represents the positive image, and x_i^n represents the negative image. α is the margin imposed between the positive and negative pairs [48]. A positive pair here denotes the herbarium and field of the same species. On the other hand, a negative pair here denotes the herbarium and field of different species. The goal is to maximize the embedding distance of different pairs and minimize the embedding distance between the same pairs.

Typically, a CNN uses a softmax function for classification, which is used to calculate the probability of a vector, i.e., embedding belonging to a class. Nevertheless, studies [49, 50] have shown that this method usually works well when the training sample per class is large or balanced. In our case, where the field images are limited and even absent in some classes, it is not favorable to use the softmax function.

On the other hand, the triplet loss function is considered a deep metric learning approach in which it serves as a similarity function. Instead of focusing on class probabilities, the learned distance metric forms a new data representation that separates the class embeddings by their similarities and dissimilarities. Our main objective is to model common features between the herbarium and field pairs of the same species. In that way, the herbarium and field pairs of the same species share similar feature embeddings, consequently facilitating species identification.

Essentially, deep metric learning is commonly applied in k-shot learning, where limited training data are available. Therefore, we have resolved to this deep metric learning approach using triplet loss.

3 Materials and methods

3.1 Materials

3.1.1 Training data

We used the training dataset from PlantCLEF 2021. It is currently the largest benchmark dataset for cross-domain plant identification. The dataset consists of 997 species focused on the Guiana Shield and the Northern Amazon rainforest. Our downloaded collection is composed of 321,226 herbarium images and 5824 field photographs. A valuable asset of this training dataset is that a set of plant observations are provided with both herbarium and field images of the same individual plant [51]. Therefore, this allows the learning of mapping between the herbarium and field domains. In addition, since the field training images are significantly less than the herbarium images, we utilized the dataset from PlantCLEF 2017, which contains 10,000 plant species, to generalize field images better for the initialization (pre-training) of the HFTL network. The PlantCLEF 2017 dataset is used solely in the training of the Field stream in the HFTL network.

3.1.2 Test data

3.1.2.1 Experiment test sets (closed and open) A subset of species was segregated from the training dataset to form two test sets. The first test set (closed test set) is focused on the species with field images in the training data (species present in the trained 435 classes). Meanwhile, the second test set (open test set) is focused on the species that have no field training data (species not present in the trained 435 classes). Since triplet learning requires herbarium and field pairs to operate, the classes with missing field images are unable to be trained. Consequently, these classes are unseen by the HFTL network models, making them an open set. To create the open test set, we obtained its field images from various reliable online resources other than the data provided by PlantCLEF. This test set was not used in the training process. The experiment test sets[1] are detailed in Table 1.

3.1.2.2 PlantCLEF 2020, 2021 The test sets from PlantCLEF 2020 and PlantCLEF 2021 are used. They are identical and composed of 3186 field photographs related to 638 plant observations. A plant observation refers to a collection of field photographs of the same plant. The test set was carefully constructed to ensure that the plant observations of each species were not found in the training data. Furthermore, priority was given to the species with few or no field training data at all. This makes the task extremely difficult but was to encourage the capability of transferring knowledge from the herbarium domain to the field domain. Secondly, it was to avoid classical CNNs to perform well due to the abundance of field training images rather than the use of herbarium images [51].

[1] The experiment test sets and training scripts are available at https://github.com/NeuonAI/hftl_osm_visuals.

Table 1 Closed test set (with field training data) and open test set (without field training data)

Dataset	Number of images	Number of classes
Closed test set	1,139	338
Open test set	166	84

3.2 Methods

3.2.1 Network architecture

The networks implemented are based on the Inception-v4 or Inception-ResNet-v2 architecture introduced in [52]. Inception-v4 and Inception-ResNet-v2 share a similar network architecture. However, Inception-ResNet-v2 is constructed with residual connections and can be trained much faster while achieving slightly higher final accuracy.

3.2.1.1 Inception-v4 The network architecture of Inception-v4 can be divided into three parts: the stem, inception layers, and classification layers. The stem, which is the initial layers, consists of convolutional layers and max-pooling layers that convert the input image of size $299 \times 299 \times 3$ to $35 \times 35 \times 384$. Meanwhile, the inception layers consist of different inception modules: four Inception-A, seven Inception-B, and three Inception-C modules. Connecting Inception-A and Inception-B is a Reduction-A module that converts the $35 \times 35 \times 384$ image to $17 \times 17 \times 1024$. On the other hand, connecting Inception-B and Inception-C is a Reduction-B module that converts the $17 \times 17 \times 1024$ image to $8 \times 8 \times 1536$. In the inception modules, 1×1, 3×3, 1×7, and 7×1 filters are applied. Besides, averaged pooling is used instead of max-pooling to convolute the features of the image. The entire inception layers serve as the feature extractor of the network. Lastly, the classification layers consist of an averaged pooling layer, dropout layer, and softmax layer, which outputs the probabilities over the predicted classes. In the average pooling layers, instead of converting the image from $8 \times 8 \times 1536$ to 1×1536 as in the original architecture, we modified the layer to output the final image to 1×500. Then, the image is passed to the dropout layer, where 20% of the neurons are removed, and finally,

classified with the softmax layer over N classes, where N represents the number of classes being trained.

3.2.1.2 Inception-ResNet-v2 Likewise, in Inception-v4, the network architecture of Inception-ResNet-v2 can be divided into three parts: the stem, inception layers, and classification layers. However, the output of the stem in Inception-ResNet-v2 is $35 \times 35 \times 320$. In addition, the average pooling layers in their inception layers are replaced with residual layers. Moreover, the inception layers consist of ten Inception-ResNet-A, twenty Inception-ResNet-B, and ten Inception-ResNet-C modules. Each inception module is followed by a 1×1 convolution without activation through the filter expansion layer. Likewise, 1×1, 3×3, 1×7, and 7×1 filters are applied. Furthermore, batch normalization is not used on top of the summations but only the traditional layers. Subsequently, connecting Inception-ResNet-A and Inception-ResNet-B is a Reduction-A module that converts the image to $33 \times 33 \times 1088$. On the other hand, connecting Inception-ResNet-B and Inception-ResNet-C is a Reduction-B module that converts the image to $8 \times 8 \times 2080$. Similar to the Inception-v4 network, the classification layers consist of an averaged pooling layer, dropout layer, and softmax layer, which outputs the probabilities over the predicted classes. In the average pooling layers, instead of converting the image from $8 \times 8 \times 1536$ to 1×1536 as in the original architecture, we modified the layer to output the final image to 1×500. Then, the image is passed to the dropout layer, where 20% of the neurons are removed, and finally, classified with the softmax layer over N classes, where N represents the number of classes being trained.

3.2.2 Herbarium–Field triplet loss network

The Herbarium–Field Triplet Loss Network (HFTL network) is composed of two CNNs: herbarium network and field network, which share the same network architecture, i.e., Inception-v4 or Inception-ResNet-v2. Each network is constructed to cater for the generalization of herbarium and field features, respectively. A batch normalization layer is added at the final embedding layer of each network, and the output is fed into a fully connected layer. The feature vector of the fully connected layer is then reduced to a size of 500 and L2 normalized. Subsequently, they are concatenated to give an output size of $(n + m) \times 500$ whereby n and m is the batch size of the herbarium and field networks, respectively. For the ease of implementation, we set the values of n and m to be the same. The concatenated feature embedding is then passed to the network's triplet loss layer[2] in which the network optimizes the herbarium and field embeddings to their species. This network is illustrated in Fig. 1A. Note that this network does not have any classification layers.

3.2.3 One-streamed mixed network

The one-streamed mixed network (OSM network) is a single-stream CNN based on the Inception-v4 or Inception-ResNet-v2 architecture. Unlike the HFTL network, a single feature embedding is trained with both herbarium and field data. Specifically, instead of having two different feature embedding trained with plant data from different domains: field and herbarium, the OSM network projects all features into a single domain. The OSM network can be considered the baseline for testing the generalization of herbarium and field features. As with the HFTL network, its feature vector is reduced to a size of 500. Unlike the HFTL network, the network is trained with the softmax cross-entropy loss function instead of the triplet loss. This network is illustrated in Fig. 1B. Note that although we used softmax cross-entropy as our loss function during the training of this network, the prediction probabilities obtained in measuring the performances are not calculated using softmax cross-entropy during inference. We simply utilized its feature vector for feature comparison so that it is evaluated in the same way as the HFTL network.

3.2.4 Training setup

3.2.4.1 HFTL network As mentioned in Sect. 3.2.2, the HFTL network is composed of two CNNs: herbarium and field network. To construct the HFTL network, the herbarium and field networks are first trained separately. Both networks are initialized on pre-trained weights from ImageNet [54] and fine-tuned with PlantCLEF 2021 herbarium images (in the herbarium stream) and PlantCLEF 2017 field images (in the field stream), respectively. Fine-tuning the model with datasets containing exclusively plant images is to obtain a pre-trained model in the plant domain before knowledge transfer to our targeted dataset, which is PlantCLEF 2021. Once this pre-trained network is ready, the HFTL network is then fine-tuned with the PlantCLEF 2021 dataset with 435 plant species, containing herbarium and field images. Note that an HFTL network pre-trained solely from PlantCLEF 2021 images was also trained for comparison. This particular network is "HFTL-I-21" in Table 3.

3.2.4.2 OSM network The OSM network is initialized on weights pre-trained from ImageNet [54] and fine-tuned on the herbarium and field images solely from PlantCLEF

[2] The triplet loss is computed using triplet_semihard_loss function provided in TensorFlow 1.12 [53]

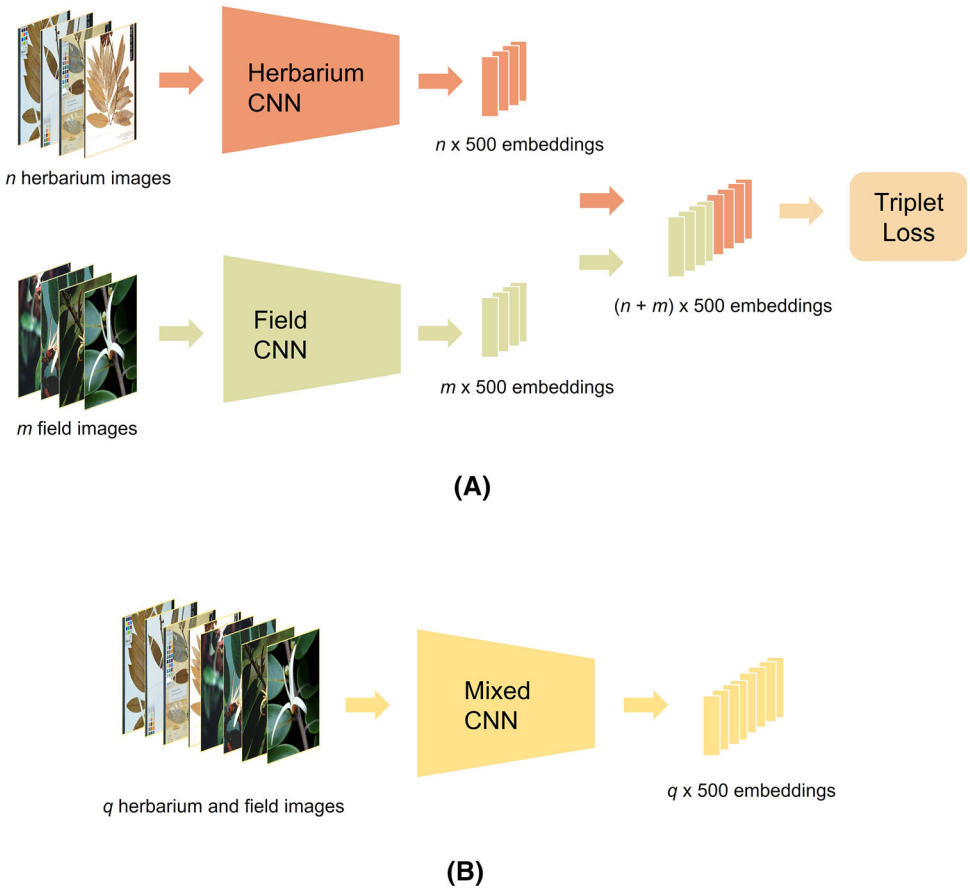


Fig. 1 The network architecture of the implemented networks. **A** the Herbarium–Field Triplet Loss network architecture. **B** the one-streamed mixed network architecture

2021. Two different kinds of OSM networks were experimented with, one trained on the whole 997 species and another on the 435 species with both herbarium and field images.

3.2.4.3 Data augmentation Data augmentation was applied during training of all the networks. Random cropping, horizontal flipping and color distortion (brightness, contrast, saturation and hue) were executed on the images to increase training data. This allows the features that are invariant to their original locations to be learned, improving network generalization and reducing the chances of overfitting.

3.2.4.4 Network parameters The networks trained are set up using TensorFlow 1.12 [53] and TensorFlow-Slim library [55] with hyperparameters as described in Table 2. Note that the batch size for the HFTL network is 16, which is significantly smaller than the rest of the networks with 256. Due to the lack of field images to fit in a larger batch size, we set the batch size of the HFTL network in this way. In addition, we trained our networks using the different machines that we have. Apart from the HFTL networks that were trained on a 16GB graphics processing unit (GPU), the rest of the networks were trained with a 32GB GPU. We set the rest of the networks' batch sizes to 256 to utilize the available GPU for efficient training. Therefore, the training speed of the networks can be improved.

3.3 Inference procedure

3.3.1 Herbarium dictionary construction

For inference, the embeddings from 997 species are first extracted using the trained networks to form a herbarium dictionary that serves as a reference. Random herbarium samples from each species are picked and fed to the network for extraction. The extraction is done with *Center and Corner Cropping* whereby the center, top-left, bottom-left, top-right, and bottom-right regions of the image are cropped, resized, flipped, and passed to the network. The extracted embeddings are then averaged to get a single embedding representation for each species and subsequently saved in a dictionary. Note that instead of only herbarium images, field images were also used to form the herbarium dictionary. This is to add field visual representation of the species in hopes to facilitate species identification. The experiment networks that involve the use of

Table 2 Network training parameters

Parameter	Herbarium network, Field network, OSM network	HFTL network
Batch size	256	16
Input image size	$299 \times 299 \times 3$	$299 \times 299 \times 3$
Optimizer	Adam optimizer [56]	Adam optimizer [56]
Initial learning rate	0.0001	0.0001
Weight decay	0.00004	0.00004
Loss function	Softmax cross-entropy	Triplet loss

Table 3 Variations of experimented HFTL and OSM networks

Network	Architecture	Class trained	Data source (PlantCLEF)	Additional augmentation
HFTL-I-21	Inception-v4	435	2021	No
HFTL-I	Inception-v4	435	2017, 2021	No
HFTL-I-AUG	Inception-v4	435	2017, 2021	Yes
HFTL-IR	Inception-ResNet-v2	435	2017, 2021	No
HFTL-IR-AUG	Inception-ResNet-v2	435	2017, 2021	Yes
OSM-I-435	Inception-v4	435	2021	No
OSM-IR-435	Inception-ResNet-v2	435	2021	No
OSM-I	Inception-v4	997	2021	No
OSM-IR	Inception-ResNet-v2	997	2021	No

Table 4 Performance of the experimented networks on the closed test set (seen classes). Values in bold indicate the best performances

Network	Class trained	MRR	Top-1	Top-5
HFTL-I-21	435	**0.872**	**0.803**	**0.959**
HFTL-I-21 (Field)	435	0.119	0.091	0.135
HFTL-I	435	0.568	0.466	0.680
HFTL-I (Field)	435	0.08	0.048	0.093
HFTL-I-AUG	435	0.61	0.484	0.767
HFTL-I-AUG (Field)	435	0.092	0.057	0.120
HFTL-IR	435	0.777	0.663	0.918
HFTL-IR (Field)	435	0.108	0.076	0.129
HFTL-IR-AUG	435	0.528	0.400	0.674
HFTL-IR-AUG (Field)	435	0.073	0.035	0.090
OSM-I-435	435	0.831	0.757	0.923
OSM-I-435 (Field)	435	0.866	0.808	0.943
OSM-IR-435	435	0.874	0.814	0.947
OSM-IR-435 (Field)	435	**0.908**	**0.866**	**0.959**
OSM-I	997	0.539	0.440	0.655
OSM-I (Field)	997	0.659	0.572	0.752
OSM-IR	997	0.566	0.461	0.687
OSM-IR (Field)	997	**0.692**	**0.609**	**0.788**

Table 5 Performance of the experimented networks on the open test set (unseen classes). Values in bold indicate the best performances

Network	Class trained	MRR	Top-1	Top-5
HFTL-I-21	435	0.062	0.012	0.084
HFTL-I-21 (Field)	435	0.08	0.024	0.120
HFTL-I	435	0.125	0.054	0.205
HFTL-I (Field)	435	0.144	0.072	0.211
HFTL-I-AUG	435	0.132	0.048	0.211
HFTL-I-AUG (Field)	435	0.164	0.072	0.253
HFTL-IR	435	0.133	0.066	0.187
HFTL-IR (Field)	435	0.158	0.078	0.241
HFTL-IR-AUG	435	0.149	0.084	0.199
HFTL-IR-AUG (Field)	435	**0.185**	**0.108**	**0.247**
OSM-I-435	435	0.131	0.054	0.199
OSM-I-435 (Field)	435	0.077	0.024	0.096
OSM-IR-435	435	**0.152**	**0.084**	**0.205**
OSM-IR-435 (Field)	435	0.084	0.024	0.127
OSM-I	997	**0.125**	**0.036**	**0.193**
OSM-I (Field)	997	0.05	0.018	0.054
OSM-IR	997	0.122	0.048	0.175
OSM-IR (Field)	997	0.049	0.018	0.042

field images are indicated with "(field)" in Tables 4 and 5.

The overall process of the herbarium dictionary construction is illustrated in Fig. 2.

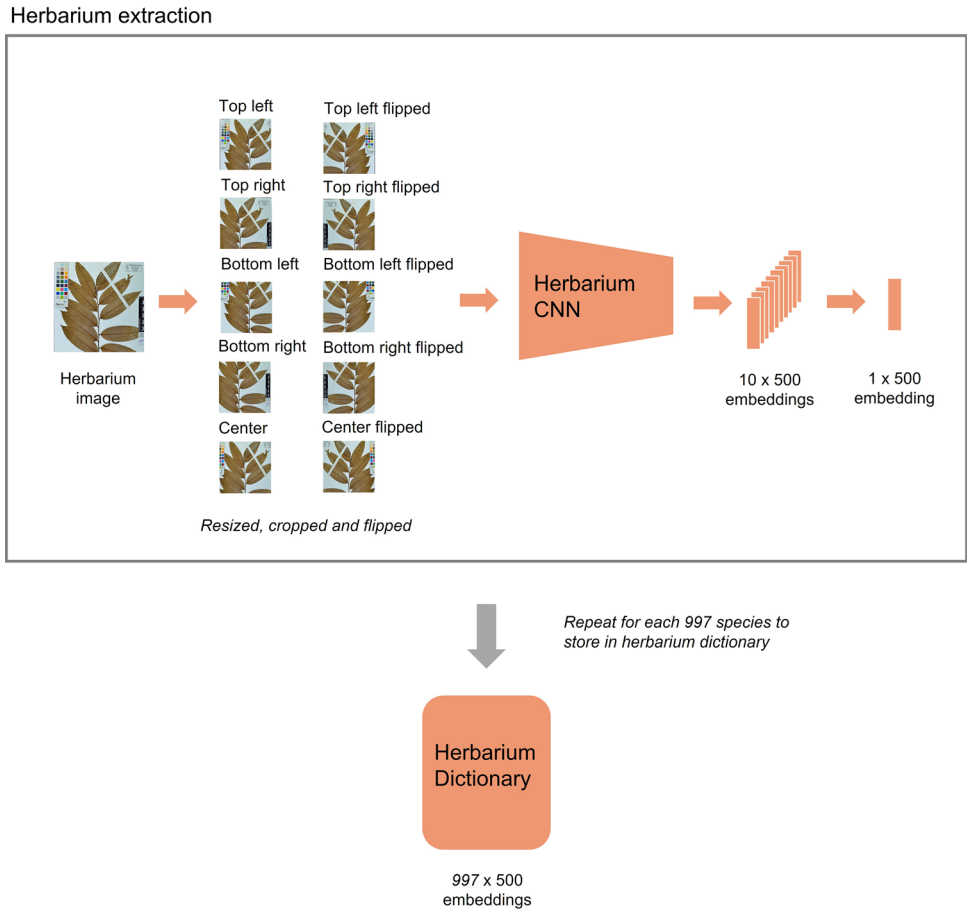


Fig. 2 The process of herbarium dictionary construction

3.3.2 Embedding similarity comparison

After obtaining the herbarium dictionary containing the single embedding representation of each species, the field test image embeddings are extracted. This enables the feature similarity comparison between the herbarium dictionary and test images. Likewise, in the herbarium extraction, the field embedding is also extracted with *Center and Corner Cropping* whereby the center, top-left, bottom-left, top-right, and bottom-right regions of the image are cropped, resized, flipped, and passed to the network. This process is similar to the herbarium extraction done in Fig. 2. The extracted embeddings are then averaged to get a single embedding representation for each image. Cosine similarity is used as the distance metric in measuring the herbarium and field embedding similarity:

$$cos(x, y) = \frac{x \cdot y}{x \cdot y} \tag{2}$$

where x is the embedding vector of the herbarium dictionary and y is the embedding vector of the field test images. It is then transformed with inverse distance weighting into probabilities for ranking the classes. The weights for the transformed embedding vector are calculated as:

$$P_i = \frac{\left(\frac{1}{d_i}\right)}{\sum_k \left(\frac{1}{d_k^n}\right)} \tag{3}$$

where k i is the index of the class, d is the embedding distance, and n is the power in positive real number. The process is illustrated in Fig. 3.

3.4 Evaluation metric

3.4.1 Mean reciprocal rank

Mean reciprocal rank (MRR) is used as the main evaluation metric of the test sets. We used MRR because it is the standard metric used in the PlantCLEF 2020 and 2021 challenge, which constitute the largest benchmark for cross-domain plant identification. It allows us to compare our work with others and the PlantCLEF benchmark. The MRR is a statistic measure for the rank of the correct answer in a list of predictions. It is the average of the reciprocal ranks of the whole test set, with $|Q|$ being the total number of plant occurrences and $rank_i$ being the

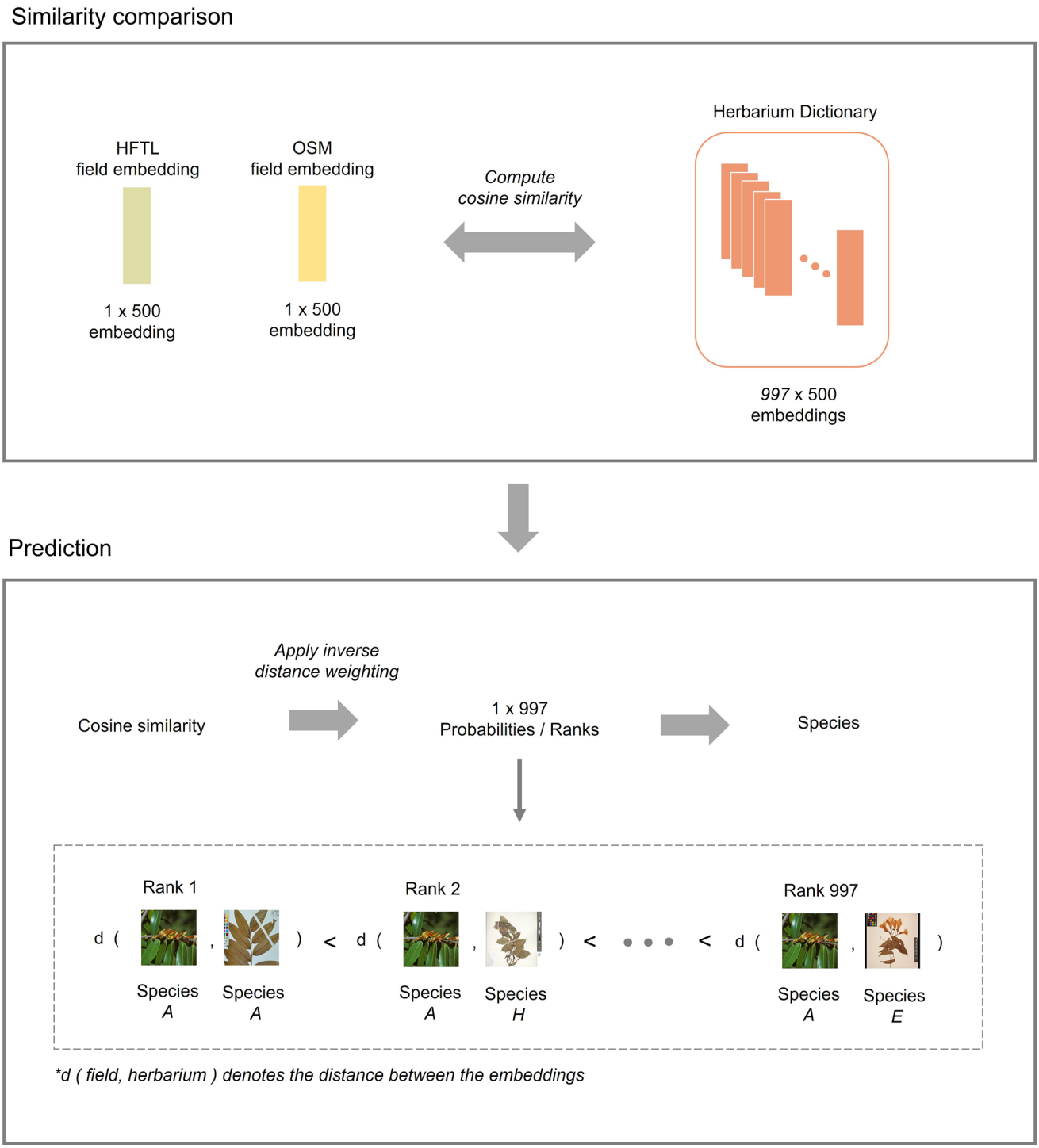


Fig. 3 The process of comparing the similarity between the herbarium–field embeddings

predicted rank of the ground truth label for the *i*th plant occurrence. It is mathematically represented as:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{4}$$

3.4.2 Top-N accuracy

The second evaluation metric used is the top-N accuracy. The top-N accuracy is the fraction of the ground truth class being equal to any of the N highest probability classes predicted by the model. It is defined as the formula below where *TP* represents the true positives of the N highest

probability classes, *TN* represents the true negatives, *FP* represents the false positives, and *FN* represents the false negatives:

$$\text{Top-N Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (5)$$

3.5 Networks

We implemented various training strategies to assess the performance of our models using different network choices, training data, and additional data augmentation. The variations in these networks are described in Table 3. Note that for the experimented HFTL networks, the PlantCLEF 2017 dataset was only used during the training of the field network to initialize the field weights in the HFTL network. Moreover, the additional augmentation in Table 3 refers to increased data augmentation during training (the training images were pre-processed with more transformations and augmentation).

Since the HFTL networks can only be trained with valid herbarium–field pairs, they were trained with 435 classes as these classes have both herbarium and field training data. On the other hand, the OSM networks do not require valid herbarium–field pairs to be trained; hence, they were trained with the full 997 training classes. Additionally, the OSM networks were also trained with the same 435 classes as HFTL networks to compare the network performances.

4 Quantitative results

4.1 Experiment test sets

4.1.1 Closed set

We compare the networks' performance on the closed experimental test set in which the networks are evaluated on the classes they were trained on. Based on the results obtained in Table 4, the OSM networks trained on 435 classes performed the best in the closed test set and even outperformed the HFTL networks trained on 435 classes. On the other hand, the OSM networks trained on 997 classes did not perform as well. This is likely because the OSM models trained on 435 classes are more biased towards this Closed test set which is focused on the 435 classes as compared to the OSM models trained on 997 classes. Notably, adding field images in the herbarium dictionary used during inference improved the prediction performance of the OSM Networks but worsen the prediction for the HFTL networks. Since HFTL networks were trained without field images in their herbarium stream, introducing field images has likely altered the learned herbarium features. This, in turn, reduced the matching probability of the herbarium–field pairs. Unlike the HFTL networks, OSM networks were trained with both herbarium and field images in their single stream. Therefore, their learned features were not disrupted, and they were able to make use of the additional field image data in the herbarium dictionary, thus improving accuracy.

4.1.2 Open set

On the other hand, in Table 5 whereby the networks are evaluated on the classes they were not trained on (open test set), the HFTL networks performed the best. Comparing the HFTL and OSM networks trained on the same 435 classes, HFTL networks performed better. Meanwhile, the OSM networks trained on the whole 997 classes did not perform as well. In contrast to the performance of the closed test set, the use of field images in the herbarium dictionary did not improve the accuracy for the OSM networks. However, it did improve the accuracy of the HFTL Networks. Since the open test set contains unseen classes, the use of field images does not disrupt the learned features of the HFTL networks as they did in the closed test set. Instead, due to this exclusion, the unseen classes' overall predicted rank has moved up. Although this happened not as intended in our design, it is worth noting that it could be further improved if we could rework the embedding generation (for future work) to encapsulate both types of embeddings in the seen and unseen classes. On the other hand, using field images in the herbarium dictionary did not help the OSM networks as they did in the closed test set. This is likely because the learned herbarium–field pairs do not match the unseen herbarium–field pairs.

4.2 PlantCLEF 2020, 2021 challenge

We submitted several runs to PlantCLEF 2020 and PlantCLEF 2021, and our results are tabulated in Table 6. Our HFTL ensemble networks performed the best in the difficult species test set and gave fairly high, and more importantly, equivalent values for both the whole test set and difficult species MRR measures. It indicates that our method is very robust to the lack of training field photos and can generalize rare difficult species in the test set, which was the underlying purpose of the challenge. Our HFTL ensemble networks have shown to outperform traditional CNN [44] and the domain adaptation methods which include: Few-Shot Adversarial Domain Adaptation (FSADA) [45, 51], Adversarial Consistent Learning on Partial Domain Adaptation (ACL) [57] and Weighted Pseudo Labeling Refinement (WPLR) [58], especially in the difficult species test set. Furthermore, comparing the

Table 6 MRR scores of PlantCLEF 2020 (top six rows) and PlantCLEF 2021 (bottom eight rows) submissions. Values in bold indicate the best performances

Method	MRR (Whole test set)	MRR (Difficult species)
FSADA	**0.180**	0.052
FSADA	0.134	0.062
ACL	0.032	0.016
Traditional CNN	0.008	0.003
HFTL-ENS (ours)	0.121	0.107
HFTL-ENS (ours)	0.111	**0.108**
FSADA (Organizer's submission)	**0.198**	0.093
WPLR	0.065	0.037
WPLR	0.057	0.042
FSADA (Organizer's submission)	0.052	0.042
HFTL-ENS + OSM-ENS (ours)	0.181	**0.158**
HFTL-ENS (ours)	0.169	0.150
OSM-ENS (ours)	0.152	0.117
HFTL-21 (ours)	0.060	0.056

methods which did not use external data in Table 7, our HFTL-21 network performed the second-best in the whole test set, while the best in the difficult species test set.

5 Qualitative results

To interpret our model predictions, we analyze the features learned by our models through their activation maps and used t-distributed stochastic neighbor embedding method to analyze the dataset embeddings.

5.1 Activation maps

The class activation mapping technique in which the class discriminative region of an image is highlighted in a heatmap using the global average pooling layer from a CNN was introduced in [59]. These highlighted heatmaps are known as class activation maps and indicate the most focused feature of an image used by the model to identify a class. Similarly, we employ this concept to visualize the most discriminative part of the plant as interpreted by our models. Since our triplet learning model aims to optimize the embedding distance between the herbarium and field domains instead of directly classifying plants, we do not investigate the class activation maps but the activation maps (AMs) using the last layers before the average pooling layers. In other words, we investigate what features our models focus on that foster the prediction of species instead of what species it is predicted as when given a plant image. These features are portrayed in the AMs and represented in feature vectors.

For ease of reference in this paper, we will be visualizing the Top-5 species predictions together with their probabilities. Subsequently, the training samples of the predicted Top-5 species are fed to the network to extract their AMs. The AMs are drawn at the last layers of the CNN, just before their average pooling layers. For comparison, we examine the AMs drawn by the experimented HFTL-IR and OSM-IR networks. Both networks are based on the Inception-ResNet-v2 models. Therefore, the AMs are extracted at their layer: "Conv2d_7b_1x1". The drawn AMs are then normalized, and the activated map with the highest feature vector value is obtained. We obtained its highest activated map index because it demonstrates the most activated part of the image.

In the following activation map (AM) figures, i.e., Figs. 4, 5, 6, 7, 8, the activated map index is shown above the AMs. Since the HFTL network is composed of two streams, the map index shown with a "H" label indicates

Table 7 MRR scores of plantCLEF 2020, 2021 submissions trained without external data. Values in bold indicate the best performances

Method	MRR (whole test set)	MRR (difficult species)
WPLR	**0.065**	0.037
HFTL-21 (ours)	0.060	**0.056**
WPLR	0.057	0.042
FSADA (Organizer's submission)	0.052	0.042
ACL	0.032	0.016
Traditional CNN	0.008	0.003

Fig. 4 Samples of high confidence correct Top-1 prediction AMs from the HFTL-IR and OSM-IR networks. **A** the HFTL-IR network maps from Allamanda cathartica L. and its Top-5 predictions, **B** the OSM-IR network maps from Hibiscus rosa-sinensis L. and its Top-5 predictions

that the AMs were generated from the Herbarium stream of an HFTL network. If "H" is not labeled, the AMs were drawn from either the field stream of an HFTL network or the mixed stream of an OSM network. For example, in Fig. 5A, the Top-1 prediction shows a row of Croton matourensis Aubl herbarium training samples. Below them are their respective AMs with their activation map index. The first image shows an activation index of "H-416". This means the AM was generated from the herbarium stream of the HFTL network at index 416. From the plant samples shown in the Top-5 predictions, priority is given to field images if there are present in the training data. Otherwise, only herbarium images are shown as samples.

5.1.1 High confidence AMs

In this section, we look into the AMs from the models that have high confidence in their predictions. Figure 4 shows samples of test images with their correct high confidence Top-1 predictions from the HFTL-IR and OSM-IR networks. The predicted species AMs from both models are similar to the test image AMs. We can observe that the activated regions are mainly the plant's region of interest,

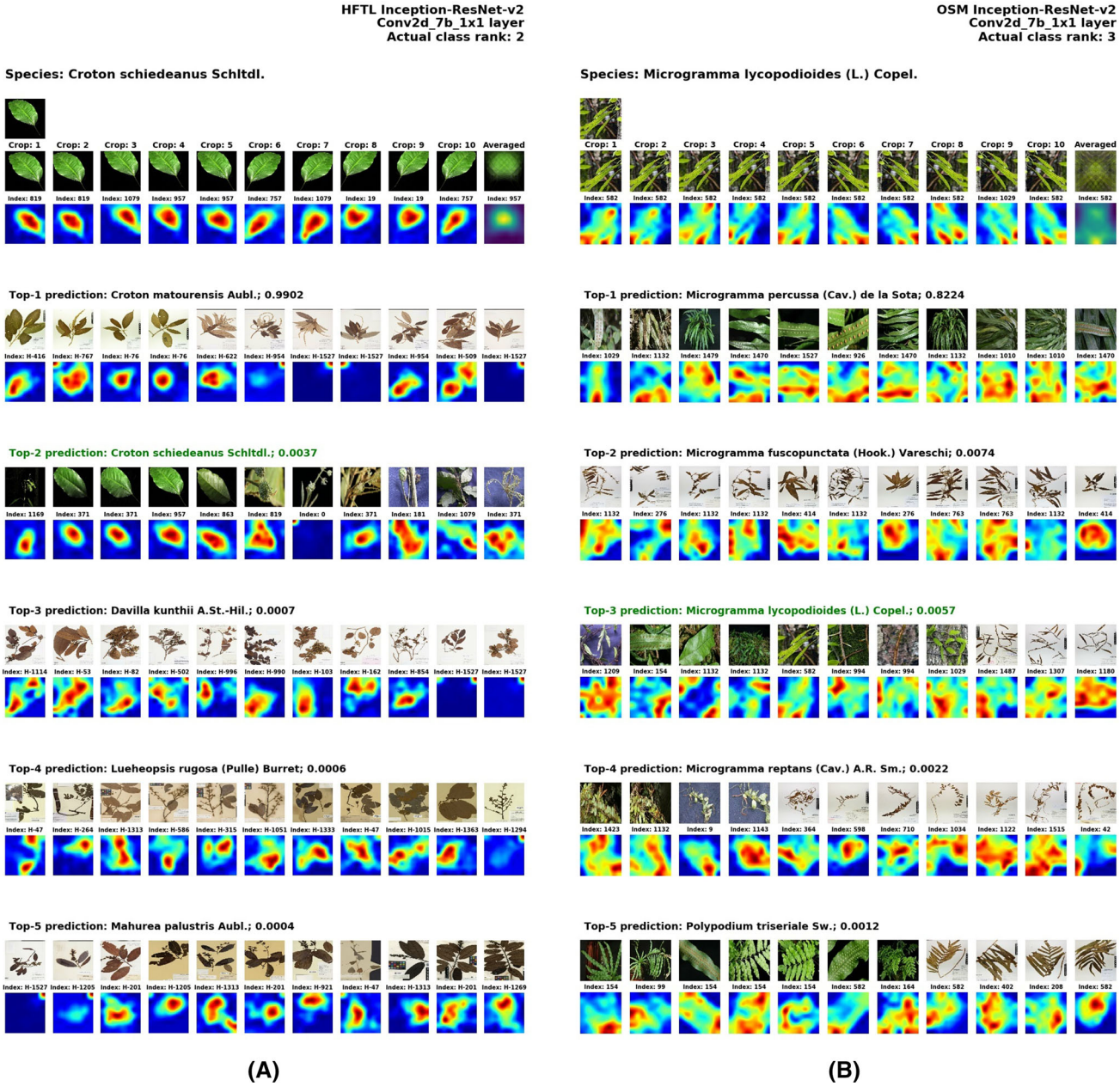


Fig. 5 Samples of high confidence incorrect Top-1 prediction AMs from the HFTL-IR and OSM-IR networks. **A** the HFTL-IR network maps from Croton schiedeanus Schltdl. and its Top-5 predictions, **B** the OSM-IR network maps from Microgramma lycopodioides (L.) Copel. and its Top-5 predictions

i.e., the flower. In Fig. 4A, the Allamanda cathartica L. test image and its Top-1 prediction share a common activated map which is map 1190. The Hibiscus rosa-sinensis L. test image in Fig. 4B, on the other hand, shares a common activated map 181 with its Top-1 prediction. A similar example can be found in the Supplementary Figures. The discriminative regions in the test image (represented and highlighted by the AMs) lead the models to predict the species which share the common class-specific discriminative regions. In other words, features that are activated on the same map index are visually similar and likely belong to the same species. Nevertheless, high confidence predictions do not necessarily equate to correct predictions. In some cases, the high confidence predictions fail to predict the correct species. For example, in Fig. 5, the models produced high confidence predictions but failed to produce the correct Top-1 species. In Fig. 5A, Croton schiedeanus Schltdl. was mispredicted as Croton matourensis Aubl. with high confidence (0.9902). Its actual class was predicted in its Top-2 prediction instead with low

confidence (0.0037). This may be due to the similarity of leaf features between both Croton schiedeanus Schltdl. and Croton matourensis Aubl. since they belong to the same Genus: Croton L. and Family: Euphorbiaceae group. Similarly, Microgramma lycopodioides (L.) Copel. in Fig. 5B was mispredicted as Microgramma percussa (Cav.) de la Sota with high confidence (0.8224). Note that all Top-4 predictions belong to the same Genus: Microgramma C.Presl and Family: Polypodiaceae group. The AMs show

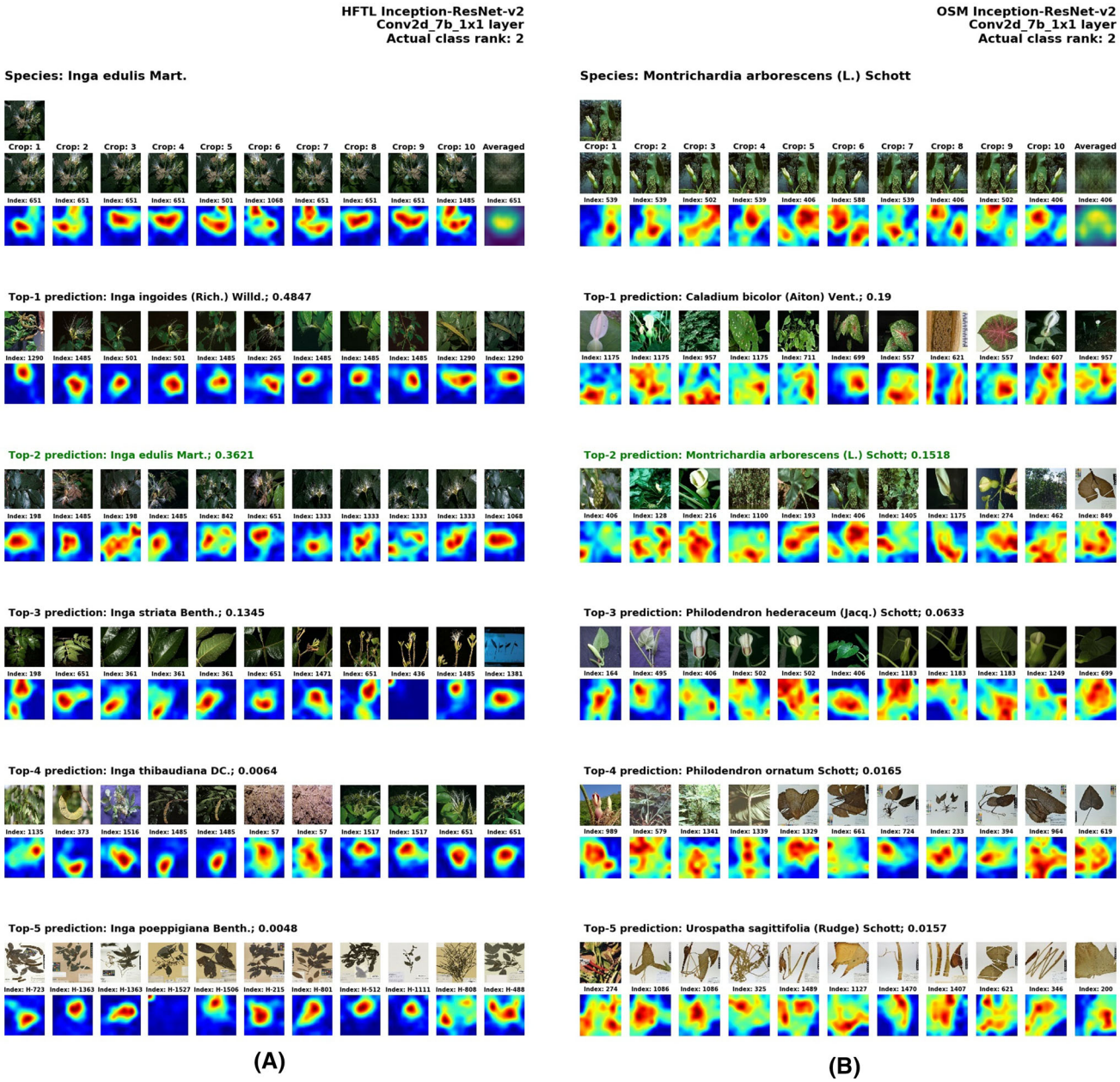
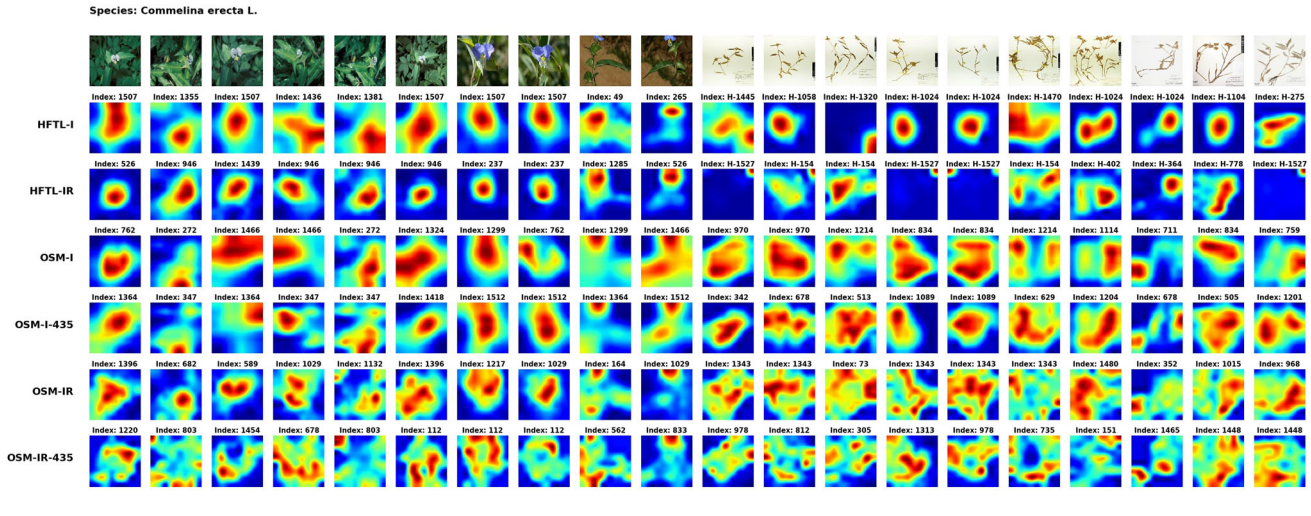
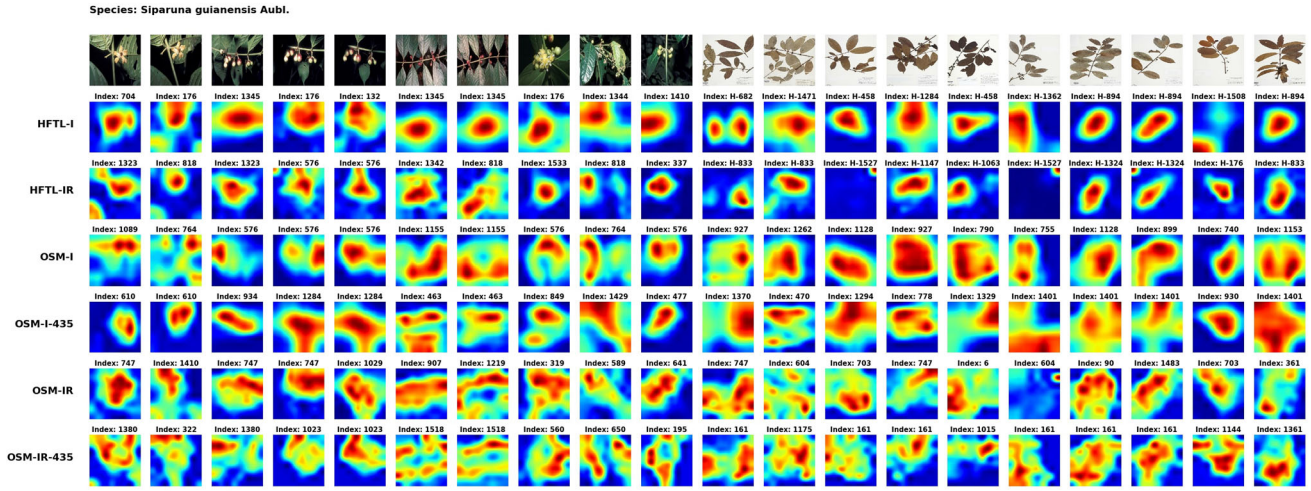


Fig. 6 Samples of low confidence incorrect Top-1 prediction AMs from the HFTL-IR and OSM-IR networks. **A** the HFTL-IR network maps from Inga edulis Mart. and its Top-5 predictions, **B** the OSM-IR network maps from Montrichardia arborescens (L.) Schott and its Top-5 predictions

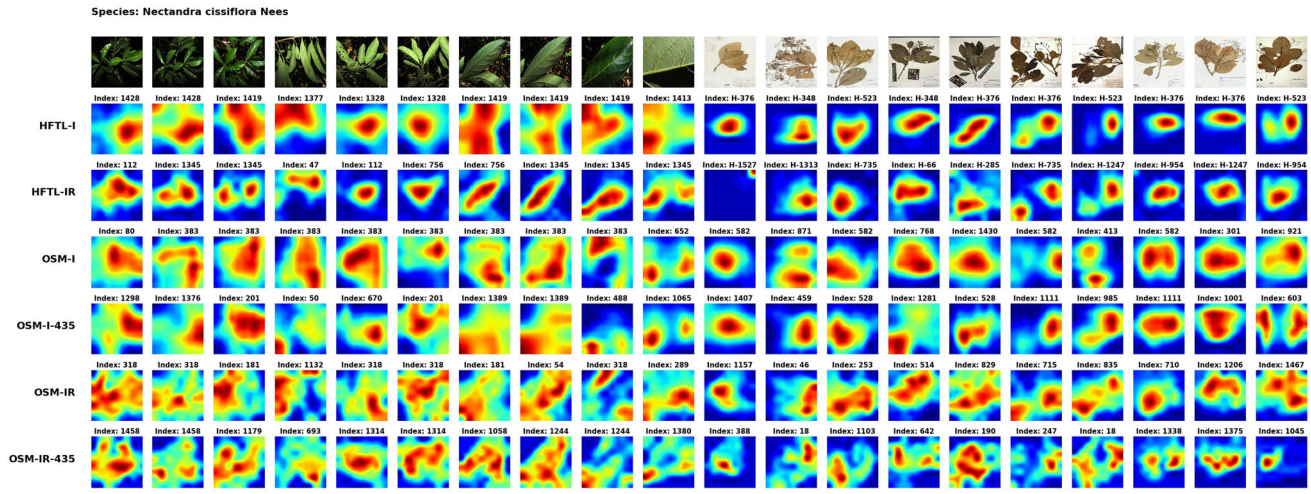
◀**Fig. 7** The species-dominant organs of the plants visualized with the models' activation maps. **A** The activation maps of Commelina erecta L. field and herbarium images drawn from the experimented networks. Its dominant organ is its flower. **B** The activation maps of Siparuna guianensis Aubl. field and herbarium images drawn from the experimented networks. Its dominant organ is its fruit. **C** The activation maps of Nectandra cissiflora Nees field and herbarium images drawn from the experimented networks. Its dominant organ is its leaves


Species: Commelina erecta L.
HFTL-I
HFTL-IR
OSM-I
OSM-I-435
OSM-IR
OSM-IR-435

(A)


Species: Siparuna guianensis Aubl.
HFTL-I
HFTL-IR
OSM-I
OSM-I-435
OSM-IR
OSM-IR-435

(B)


Species: Nectandra cissiflora Nees
HFTL-I
HFTL-IR
OSM-I
OSM-I-435
OSM-IR
OSM-IR-435

(C)

that the model has learned the leaf features of these species with common Family and Genus; however, these common leaf features are difficult to distinguish and hence activate similar AMs, which in turn leads to the wrong prediction. A similar example can be found in Supplementary Figures.

5.1.2 Low confidence AMs

Next, we look into the AMs from the models that have low confidence in their predictions. Figure 6 shows samples of test images with low confidence incorrect Top-1 predictions from the HFTL-IR and OSM-IR networks. In Fig. 6A, the Inga edulis Mart. test image has a Top-1 probability of 0.4847 for Inga ingoides (Rich.) Willd, while

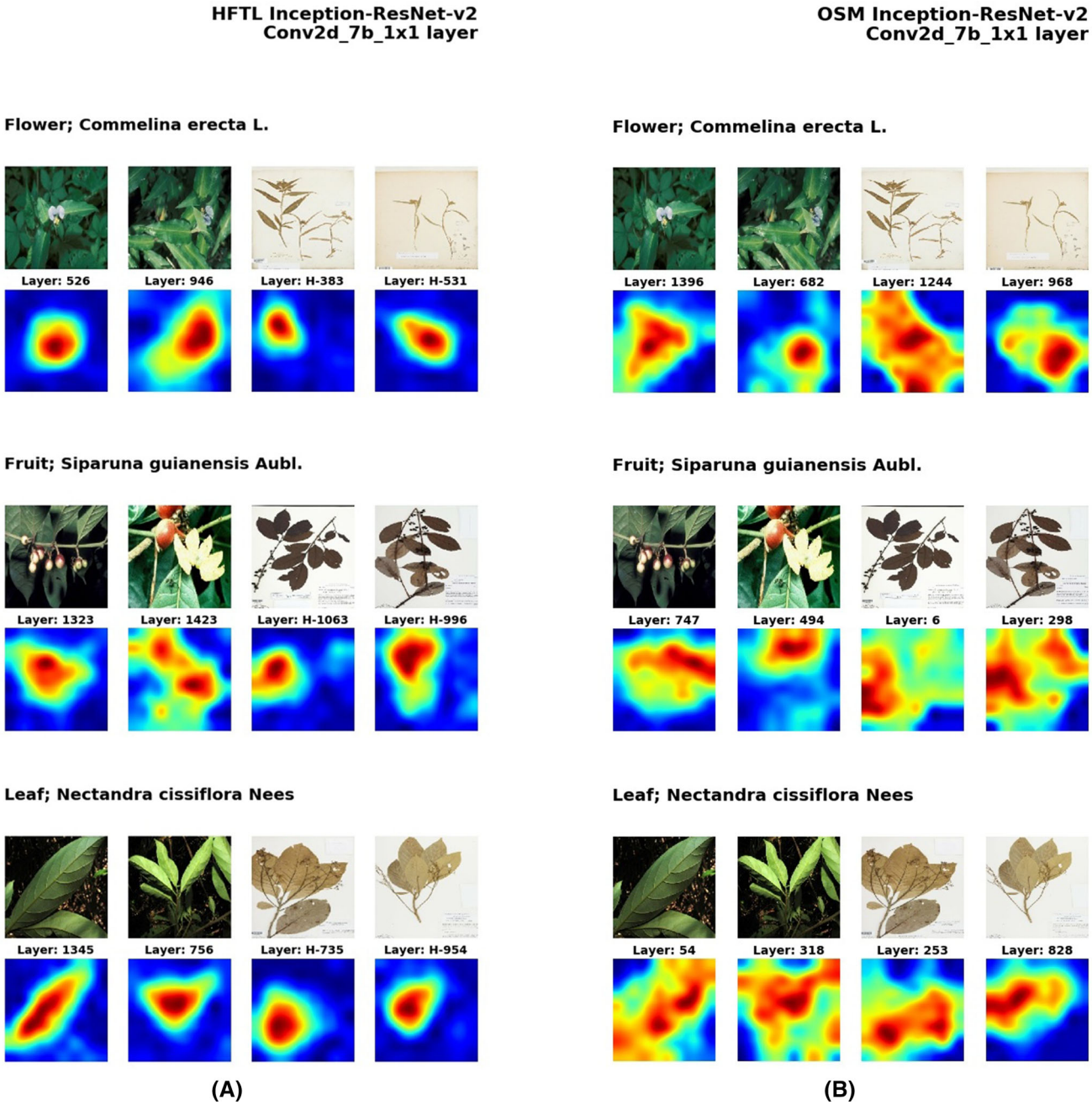


Fig. 8 The activation maps of the dominant flower, fruit, and leaf organs from the field and herbarium samples. **A** is drawn from the HFTL-IR network, **B** is drawn from the OSM-IR network

its actual class was predicted at Top-2 with a probability of 0.3621. Both the Top-1 and Top-2 predictions are close in probability and share the activation map 1485. Moreover, they belong to the same Genus: Inga Mill. and Family: Fabaceae. Likewise in Fig. 6B, Montrichardia arborescens (L.) Schott was mispredicted as Caladium bicolor (Aiton) Vent. with a Top-1 probability of 0.19. Its actual class was predicted at Top-2 with a probability of 0.1518. Furthermore, all of its Top-5 predictions are from the same Family: Araceae. It can be seen that due to their close visual similarities, being from the same Genus and Family, the models have low confidence in distinguishing between these species.

5.1.3 Species-dominant organs

We observed that varying species have varying dominant organs activated. The 3 major types of dominant organs include flower, fruit, and leaf. Figure 7A shows the activation maps of the dominant flower features from the field and herbarium samples. It can be seen that the species' flowers are distinctively localized compared to other parts of the plant. Similarly, Figure 7B shows the activation maps of the dominant fruit features from the field and herbarium samples. Meanwhile, Figure 7C shows the activation maps of the dominant leaf features from the field and herbarium samples. A comparison of dominant organ activation maps from the HFTL-IR network and OSM-IR network is shown in Fig. 8. Comparing the activation maps of both HFTL-IR and OSM-IR networks, the localization of plant organs is more precise in HFTL-IR network, while it is more widespread in the OSM-IR network, covering a large part of the plant. It can be suggested that HFTL network is better at generalizing plant features equally in both field and herbarium images compared to the OSM network. More examples of the species' dominant organs can be found in Supplementary Figures.

In addition, we look into the performance of the models on the dominant plant organs. We segregated our closed and open test sets based on their organs and obtained the results from the HFTL-IR and OSM-IR models to evaluate whether the HFTL model performs better, as seen in the activation maps. Based on our results, as tabulated in Table 8, we found that the HFTL model is indeed generally better at predicting the test sets as compared to the OSM model. The HFTL model performs better at predicting the plants' flowers and fruits. Meanwhile, it performs slightly poorer on the leaf organ (MRR: 0.136) compared to the OSM model (MRR: 0.157).

5.2 t-SNE visualizations

t-distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique that aims to visualize high-dimensional datasets such as images [60]. Using t-SNE, we visualize the embeddings of the PlantCLEF 2021 field dataset to evaluate how well our models model the class embeddings. The entire field dataset was used to generate the visualization using the same HFTL-IR and OSM-IR models as in the activation maps. To generate these visualizations, we used the t-SNE class from the scikit-learn library (version 0.23.2.) [61]. We then set the perplexity parameter to 50 and the number of optimization iterations to 5,000. The rest of the parameters are set to their default values.

Figure 9 shows the t-SNE visualization of the HFTL-IR and OSM-IR models. The top row illustrates the embeddings by their images. Meanwhile, the bottom row illustrates the embeddings by their class IDs. Figure 9A shows the t-SNE embeddings of the HFTL-IR model. On the other hand, Figure 9B shows the t-SNE embeddings of the OSM-IR model. It can be seen that the HFTL model produces more distinct embeddings as compared to the OSM model. Their embeddings are more separable and distinctly localized in clusters. It can be suggested that the embeddings from HFTL networks are more easily classifiable compared to the embeddings from the OSM networks.

For ease of comparison, we randomly segregated 1,000 field images from the field dataset and look into their close-up t-SNE visualizations. In most of the images visualized, the HFTL-IR and OSM-IR models can correctly group the samples accordingly. Figure 10 shows a close-up t-SNE of the Species: Combretum indicum (L.) DeFilipps. In addition, Figure 11 shows a close-up t-SNE of the Family: Costaceae.

Comparing both HFTL-IR and OSM-IR models in these visualizations, we observe that the HFTL model is generally better at grouping species under the same family. An

Table 8 Performance of the models on the closed (top two rows) and open (bottom two rows) test sets in terms of the plant organs

Model	Flower			Fruit			Leaf		
	MRR	Top-1	Top-5	MRR	Top-1	Top-5	MRR	Top-1	Top-5
HFTL-IR	0.844	0.751	0.958	0.823	0.720	0.966	0.690	0.537	0.881
OSM-IR	0.699	0.598	0.815	0.543	0.44	0.669	0.511	0.388	0.657
HFTL-IR	0.163	0.074	0.241	0.074	0	0.167	0.136	0.083	0.167
OSM-IR	0.105	0	0.204	0.045	0	0.083	0.157	0.095	0.179

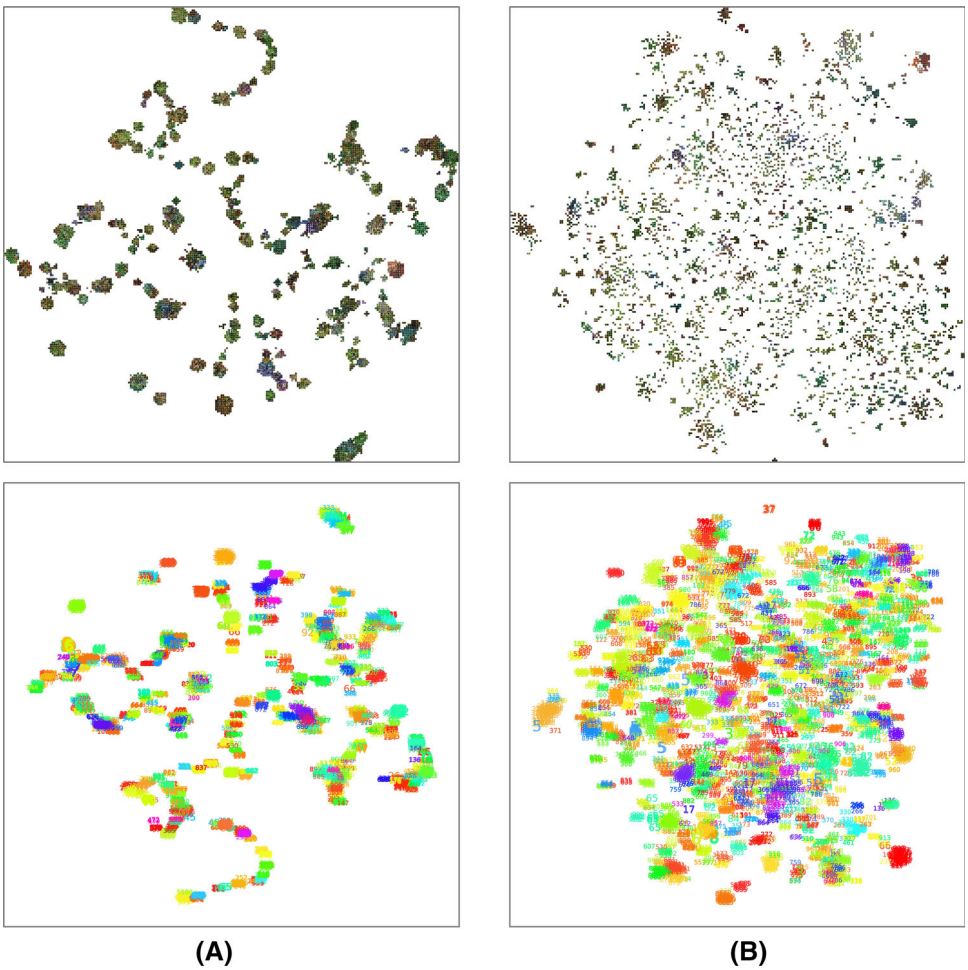


Fig. 9 The t-SNE visualizations on the PlantCLEF 2021 field dataset. It represents the embeddings of 5824 field images and 435 species. **A** is drawn from the HFTL-IR network, **B** is drawn from the OSM-IR network. The t-SNE visualizations from the HFTL-IR network is more distinct and separable compared to the OSM-IR network. It can be suggested that the embeddings produced by the HFTL networks are more easily classifiable compared to the embeddings produced by the OSM networks

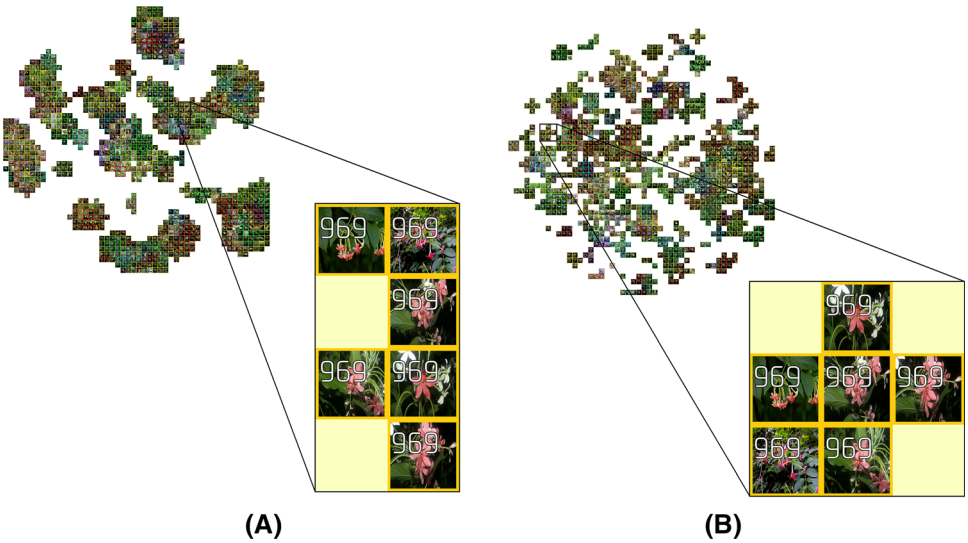


Fig. 10 The close-up t-SNE visualizations of the correctly grouped **Species**: Combretum indicum (L.) DeFilipps. **A** is drawn from the HFTL-IR network, **B** is drawn from the OSM-IR network

example is illustrated in Fig. 12 whereby the HFTL model correctly grouped the Annonaceae family into an obvious large cluster. Meanwhile, the Annonaceae family of the OSM model is distributed loosely across the embedding space. Nevertheless, the OSM model can group species under the same family in some cases. As seen in Fig. 13, the OSM model can group the Urticaceae family better than the HFTL model. All the close-up visualizations in

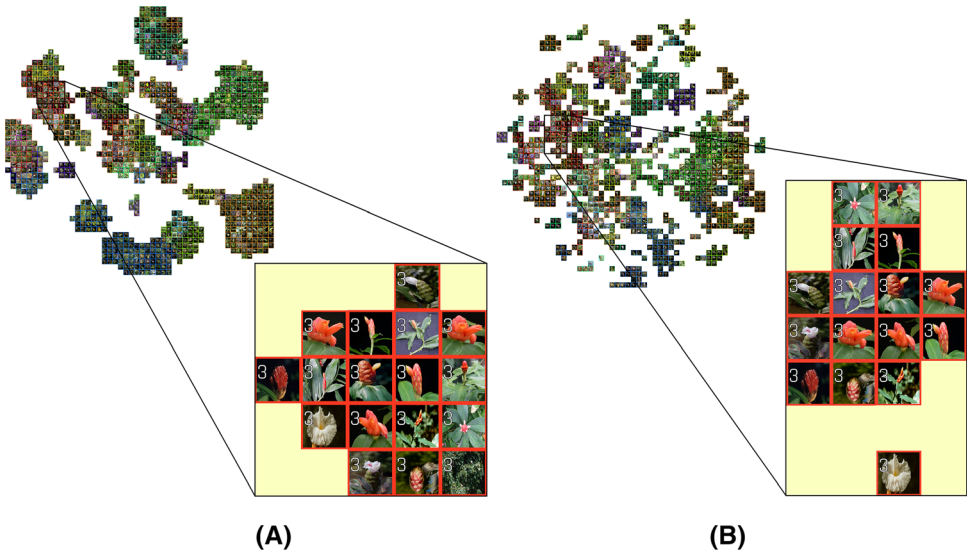


Fig. 11 The close-up t-SNE visualizations of the correctly grouped **Family**: Costaceae. It is composed of the Species, Costus scaber Ruiz Pav., and Costus spiralis (Jacq.) Roscoe. **A** is drawn from the HFTL-IR network, **B** is drawn from the OSM-IR network

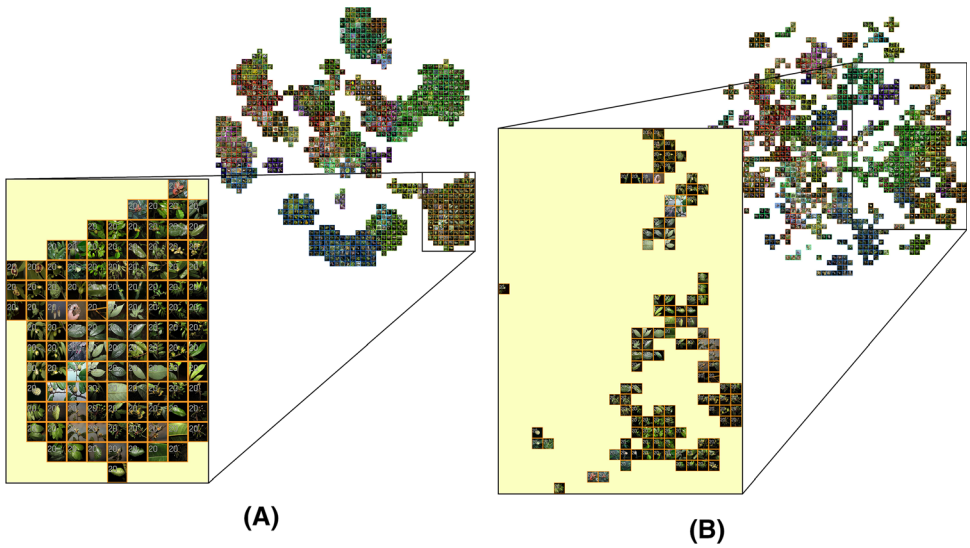


Fig. 12 The close-up t-SNE visualizations of the Family: Annonaceae. **A** is drawn from the HFTL-IR network, **B** is drawn from the OSM-IR network. It was found that the HFTL-IR network is generally better at grouping the species under the same family compared to the OSM-IR network

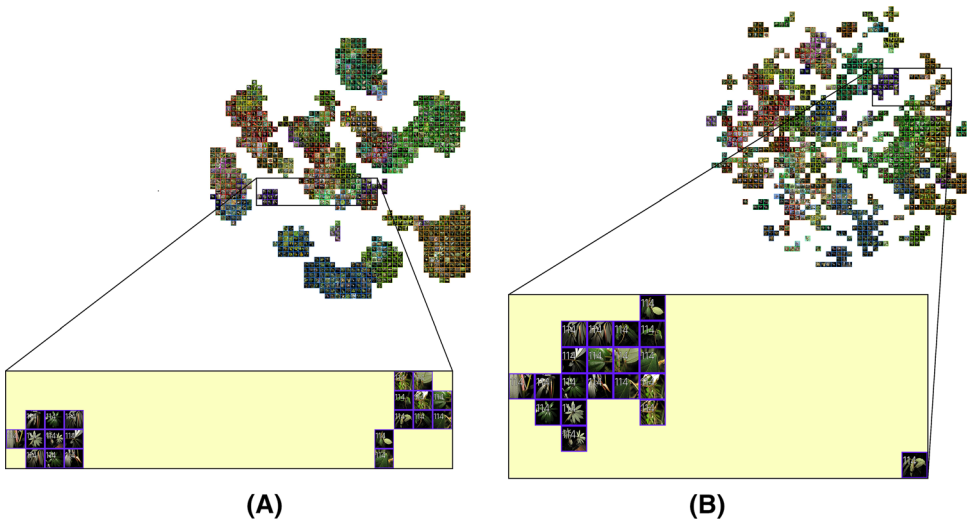


Fig. 13 The close-up t-SNE visualizations of the Family: Urticaceae. **A** is drawn from the HFTL-IR network, **B** is drawn from the OSM-IR network. Although HFTL-IR is generally better at grouping the species under the same family, this example shows the OSM-IR network being better at grouping the species under the same family compared to the HFTL-IR network

higher resolution and full view can be found in Supplementary Figures.

6 Discussion

From our experiment results as tabulated in Tables 4 and 5, the OSM models performed better in seen classes (closed test set), while the HFTL models performed better in unseen classes (open test set). In contrast, based on the official results from PlantCLEF 2021 [51] (as summarized in Table 6), the HFTL models performed better in both the seen (whole test set) and unseen (difficult species test set) classes. Since the PlantCLEF test set is more well-balanced and inclusive compared to our experiments, where data is limited, it can be concluded that the HFTL models (HFTL-ensembled models, HFTL-ENS) performed better than the OSM models (OSM-ensembled models, OSM-ENS). Even though our submitted OSM-ENS in PlantCLEF 2021 was trained with the whole 997 species, it did not perform as well as its HFTL-ENS counterpart which was trained on 435 species, or more specifically, the 435 species that have both herbarium and field training data. This proves that triplet learning is better in handling unknown classes, which further implies that triplet learning produces more generalized features compared to traditional CNNs.

In addition, the ability of HFTL models to generalize features better was reflected in the AMs analyzed in Sect. 5.1. As shown in Fig. 8, the dominant organs of the species were emphasized more clearly and precisely in the AMs of HFTL models as compared to the OSM models. It was further shown in the results tabulated in Table 8 that the HFTL model is generally better at predicting the correct species from the dominant plant organs. Moreover, it was observed that HFTL models produced generally higher prediction probabilities compared to OSM models. In other words, HFTL models are generally more confident when predicting species. This can be supported by the visualizations in Sect. 5.2 (Fig. 9) whereby the t-SNE embedding visualizations show that the dataset embeddings are more separable in the HFTL model.

Although our models did not achieve the best overall performance in PlantCLEF 2020 and 2021. They achieved better genericity on the test set less represented in the training data or rare species (difficult test set) [43, 51]. The best-performing model on the difficult test set was our HFTL network and OSM network ensemble model. Our network variations of different CNN architectures (Inception-v4 and Inception-ResNet-v2), the mixture of field embedding in the herbarium dictionary, and additional augmentation helped improve our approach from an MRR score of 0.108 (in PlantCLEF 2020) to 0.158 (in PlantCLEF 2021).

Furthermore, as shown in Table 7, our HFTL-21 single model which was trained with no additional external data (solely PlantCLEF 2021 data) outperformed the rest of the benchmarks that also used no external data which include the traditional CNN [44], and domain adaptation methods: FSADA [45, 51], WPLR [58], ACL [57] in the difficult species MRR measure. This indicates that our method is suitable to identify plants when their field samples are limited but herbarium specimens are available. This triplet learning mechanism is better at predicting observations of species with missing field images in the training set than conventional classification networks. It has been shown to be able to bridge the gap between the herbarium and field domains.

Despite that, we found that our models perform poorer when the predicted species are similar to one another. This is especially the case when the predicted species belong to the same Genus and Family group as seen in Sects. 5.1.1 and 5.1.2. Since species of the same genus and family may share common features, this makes it difficult to distinguish between them. To improve the identification between these species, it would be better if we utilized the taxonomy data and traits (which cover: plant growth form, habitat, plant lifeform, trophic guild, and woodiness) provided in PlantCLEF 2021 for further experiments. The use of these metadata has shown to be useful in improving prediction accuracy as demonstrated in the organizer's submission [51].

Moreover, although our HFTL model has shown to achieve genericity, it is limited in the number of species being trained. Unlike its OSM model counterpart or the rest of the PlantCLEF 2020 and 2021 benchmarks, it is trained with 435 species instead of the whole 997 classes. This is a setback to the model and could be further improved if the remaining species without herbarium–field pairs can be fully utilized by the model.

7 Conclusion

In this paper, we presented the implementation and performances of our Herbarium–Field Triplet Loss Network (HFTL network) and one-streamed mixed network (OSM network) in the context of the cross-domain plant identification challenges, PlantCLEF 2020 and 2021. It aims to tackle the cross-domain plant identification between plant species in their real-world (field) images from a collection of herbarium specimens. Our results show that despite the lack of field training data, the HFTL network can generalize rare species as equally as species with many training data.

We have shown that our HFTL network performed relatively equal in the whole test set and difficult species

test set regardless if few or no field training images were available. The HFTL network method has demonstrated to be more effective in classifying species with limited or no herbarium-field data as compared to conventional CNNs (such as the OSM network and traditional CNNs) and even outperformed the state of the art in the PlantCLEF 2020 and PlantCLEF 2021 challenges. Nevertheless, the limitation of our proposed network is that it requires both herbarium and field pairs to work. Moreover, it only works better than other supervised methods when the identification has fewer samples for training. This method is designed to cater more to an open set problem, whereby unknown classes are present.

In addition, concerning the lower prediction confidence between species of the same Genus and Family groups, the use of taxonomy can be adopted in future experiments. Since species belonging to the same genus and family may share similar feature characteristics, it would be advantageous to distinguish between the genus and family group to reduce the likelihood of species misclassification. Furthermore, the use of plant traits data can be investigated to improve the overall HFTL predictions. Plants that share the same traits are likely visually similar. Therefore, the plant identification with these extra discriminators would add valuable information to the training of the networks. Finally, for the inference procedure, we would like to rework our embedding generation in the future to better encapsulate the embeddings in the seen and unseen classes.

Although its absolute performance remains low for practical usage, the HFTL approach offers a step in alleviating the tedious task of automated plant identification with few field image samples, specifically rare species, which require high-level expertise. In data-deficient countries, species are under-represented in training data and it remains a challenge to acquire new field images in remote areas. This approach could contribute to the realistic usage of herbarium specimens on automated plant identification to identify not only species already present in herbaria but also adapt to the identification when the species is rare or threatened. On top of that, like various plant identification mobile applications (e.g., Pl@ntNet, Leafsnap), the predictions from this approach can serve as a reference that gives a ranking of the top predictions. Instead of providing a single classification, a list of species can be used as a reference. For example, instead of providing the Top-1 or Top-5 prediction of the model, which have lower accuracies, the Top-20 or Top-30 can also be used to narrow down the actual species. It can serve as the initial reference list for taxonomists, botanists, naturalists, and those who are interested. More importantly, our method performed the best in the test set that focuses on species with less or no training samples (difficult test set) in both PlantCLEF 2020 and 2021 (which is currently the largest benchmark dataset for cross-domain plant identification). We achieved MRR scores of 0.108 and 0.158, respectively. Meanwhile, the best-performing method from the other approaches achieved MRR scores of 0.062 and 0.093, respectively, in both years. Our method can serve as a benchmark for those who are interested in further improving this cross-domain classification, specifically on data-deficient species.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-022-07951-6.

Acknowledgements This research was supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant No. (Ref: FRGS/1/2021/ICT02/SWIN/03/2), from the Ministry of Higher Education Malaysia and NEUON AI SDN. BHD. We thank Prof. Dr. Chan Chee Seng for providing professional guidance and comments that greatly improved this manuscript. We would also like to thank our anonymous reviewers for their time and effort in giving us valuable comments and suggestions for improving the quality of this manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Author Sophia Chulif has received research support from NEUON AI SDN. BHD. Author Kok Chin Chai is the director of NEUON AI SDN. BHD. Author Sue Han Lee has received grant support from the Ministry of Higher Education Malaysia and is a lecturer in Swinburne University of Technology Sarawak Campus.

Code Availability https://github.com/NeuonAI/hftl_osm_visuals

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bar-On YM, Phillips R, Milo R (2018) The biomass distribution on earth. Proceed Nat Acad Sci 115(25):6506–6511. https://doi.org/10.1073/pnas.1711842115
2. Antonelli A, Fry C, Smith R, Simmonds M, Kersey P, Pritchard H, Abbo M, Acedo C, Adams J, Ainsworth A, et al (2020) State of the World's Plants and Fungi 2020. Royal Botanic Gardens, Kew. https://www.kew.org/sites/default/files/2020-10/State%20of%20the%20Worlds%20Plants%20and%20Fungi%202020.pdf
3. Jackson DW (2003) Plant record keeping in 2003. Bot Gard Conserv News 3(10):42–43

4. Kew RBG (2016) State of the World's Plants 2016. https://stateoftheworldsplants.org/2016/report/sotwp_2016.pdf
5. Willis KJ, et al (2017) State of the World's Plants 2017. Royal Botanic Gardens. https://stateoftheworldsplants.org/2017/report/SOTWP_2017.pdf
6. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE international conference on computer vision, vol. 2, pp. 1150–1157. https://doi.org/10.1109/ICCV.1999.790410
7. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177
8. Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In: European conference on computer vision, pp. 404–417. Springer, Berlin & Heidelberg. https://doi.org/10.1007/11744023_32
9. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Info Theory 13(1):21–27. https://doi.org/10.1109/TIT.1967.1053964
10. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018
11. Fiel S, Sablatnig R (2010) Automated identification of tree species from images of the bark, leaves and needles. Technical report, Vienna University of Technology, Faculty of Informatics
12. Anami BS, Nandyal SS, Govardhan A (2010) A combined color, texture and edge features based approach for identification and classification of Indian medicinal plants. Int J Comp Appl 6(12):45–51
13. Guru D, Sharath Y, Manjunath S (2010) Texture features and KNN in classification of flower images. IJCA (Special Issue on RTIPPR) 1:21–29
14. Kebapci H, Yanikoglu B, Unal G (2011) Plant image retrieval using color, shape and texture features. Comp J 54(9):1475–1490. https://doi.org/10.1093/comjnl/bxq037
15. Tellaeche A, Pajares G, Burgos-Artizzu XP, Ribeiro A (2011) A computer vision approach for weeds identification through support vector machines. Appl Soft Comp 11(1):908–915. https://doi.org/10.1016/j.asoc.2010.01.011
16. O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, Riordan D, Walsh J (2019) Deep learning vs. traditional computer vision. In: Science and Information Conference, pp. 128–144. Springer, Cham. https://doi.org/10.1007/978-3-030-17795-9_10
17. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neur Infor Process Sys 25:1097–1105
18. Wäldchen J, Rzanny M, Seeland M, Mäder P (2018) Automated plant species identification-trends and future directions. PLoS Comput Biol 14(4):1005993. https://doi.org/10.1371/journal.pcbi.1005993
19. Lee SH, Chan CS, Wilkin P, Remagnino P (2015) Deep-plant: Plant identification with convolutional neural networks. In: 2015 IEEE International conference on image processing, pp. 452–456. https://doi.org/10.1109/ICIP.2015.7350839
20. Grinblat GL, Uzal LC, Larese MG, Granitto PM (2016) Deep learning for plant identification using vein morphological patterns. Comp Electr Agricult 127:418–424. https://doi.org/10.1016/j.compag.2016.07.003
21. Lee SH, Chan CS, Mayo SJ, Remagnino P (2017) How deep learning extracts and learns leaf features for plant classification. Patt Recognit 71:1–13. https://doi.org/10.1016/j.patcog.2017.05.015
22. Sun Y, Liu Y, Wang G, Zhang H (2017) Deep learning for plant identification in natural environment. Computat Intell Neurosci. https://doi.org/10.1155/2017/7361042
23. Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez IC, Soares JV (2012) Leafsnap: A computer vision system for automatic plant species identification. In: European conference on computer vision, pp. 502–516. Springer, Berlin & Heidelberg. https://doi.org/10.1007/978-3-642-33709-3_36
24. Goëau H, Bonnet P, Joly A, Bakić V, Barbe J, Yahiaoui I, Selmi S, Carré J, Barthélémy D, Boujemaa N, et al (2013) Pl@ntnet mobile app. In: Proceedings of the 21st ACM International conference on multimedia, pp. 423–424. https://doi.org/10.1145/2502081.2502251
25. Heberling JM, Isaac BL (2018) Inaturalist as a tool to expand the research value of museum specimens. Appl Plant Sci 6(11):01193. https://doi.org/10.1002/aps3.1193
26. Mäder P, Boho D, Rzanny M, Seeland M, Wittich HC, Deggelmann A, Wäldchen J (2021) The flora incognita app-interactive plant species identification. Meth Ecol Evol. https://doi.org/10.1111/2041-210X.13611
27. Karthik R, Hariharan M, Anand S, Mathikshara P, Johnson A, Menaka R (2020) Attention embedded residual cnn for disease detection in tomato leaves. Appl Soft Comp 86:105933. https://doi.org/10.1016/j.asoc.2019.105933
28. Hernández S, López JL (2020) Uncertainty quantification for plant disease detection using bayesian deep learning. Appl Soft Comp 96:106597. https://doi.org/10.1016/j.asoc.2020.106597
29. Saeed F, Khan MA, Sharif M, Mittal M, Goyal LM, Roy S (2021) Deep neural network features fusion and selection based on pls regression with an application for crops diseases classification. Appl Soft Comp 103:107164. https://doi.org/10.1016/j.asoc.2021.107164
30. Uğuz S, Uysal N (2021) Classification of olive leaf diseases using deep convolutional neural networks. Neur Comp Appl 33(9):4133–4149. https://doi.org/10.1007/s00521-020-05235-5
31. Guo Y, Du C, Zhao Y, Ting T-F, Rothfus TA (2021) Two-level k-nearest neighbors approach for invasive plants detection and classification. Appl Soft Comp 108:107523. https://doi.org/10.1016/j.asoc.2021.107523
32. Kasinathan T, Uyyala SR (2021) Machine learning ensemble with image processing for pest identification and classification in field crops. Neur Comp Appl 33(13):7491–7504. https://doi.org/10.1007/s00521-020-05497-z
33. Joly A, Goëau H, Glotin H, Spampinato C, Bonnet P, Vellinga W-P, Planqué R, Rauber A, Palazzo S, Fisher B, et al (2015) Lifeclef 2015: multimedia life species identification challenges. In: International conference of the cross-language evaluation forum for European languages, pp. 462–483. https://doi.org/10.1007/978-3-319-24027-5_46. Springer
34. Anubha Pearline S, Sathiesh Kumar V, Harini S (2019) A study on plant recognition using conventional image processing and deep learning approaches. J Intell Fuzzy Sys 36(3):1997–2004. https://doi.org/10.3233/JIFS-169911
35. Goëau H, Bonnet P, Joly A (2019) Overview of lifeclef plant identification task 2019: diving into data deficient tropical countries. In: CLEF 2019-Conference and labs of the evaluation forum, vol. 2380, pp. 1–13. CEUR
36. Chen Q, Abedini M, Garnavi R, Liang X (2014) Ibm research australia at lifeclef2014: Plant identification task. In: CLEF (Working Notes), pp. 693–704
37. Jones HG (2020) What plant is that? Tests of automated image recognition apps for plant identification on plants from the British flora. AoB Plants 12(6):052. https://doi.org/10.1093/aobpla/plaa052
38. Thiers BM Index Herbariorum. (updated continuously). http://sweetgum.nybg.org/science/ih/
39. Bebber DP, Carine MA, Wood JR, Wortley AH, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, Robson NK et al (2010) Herbaria are a major frontier for species discovery.

Proceed Nat Acad Sci 107(51):22169–22171. https://doi.org/10.1073/pnas.1011841108
40. Unger J, Merhof D, Renner S (2016) Computer vision applied to herbarium specimens of german trees: testing the future utility of the millions of herbarium specimen images for automated identification. BMC Evolut Biol 16(1):1–7. https://doi.org/10.1186/s12862-016-0827-5
41. Persoon E, Fu K-S (1977) Shape discrimination using fourier descriptors. IEEE Trans Syst, Man, Cybern 7(3):170–179. https://doi.org/10.1109/TSMC.1977.4309681
42. Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of herbarium specimens. BMC Evolut Biol 17(1):1–14. https://doi.org/10.1186/s12862-017-1014-z
43. Goëau H, Bonnet P, Joly A (2020) Overview of the lifeclef 2020 plant identification task. In: CLEF Working notes 2020, CLEF: conference and labs of the evaluation forum, Sep. 2020, Thessaloniki, Greece
44. Krishna NH, Rakesh M, Ram Kaushik R (2020) Plant species identification using transfer learning-plantclef 2020. CLEF working notes
45. Villacis J, Goëau H, Bonnet P, Mata-Montero E, Joly A (2020) Domain adaptation in the context of herbarium collections: a submission to plantclef 2020. CLEF working notes
46. Carranza-Rojas J, Joly A, Goëau H, Mata-Montero E, Bonnet P (2018) Automated identification of herbarium specimens at different taxonomic levels. In: Multimedia tools and applications for environmental & biodiversity informatics, pp. 151–167. Springer, Cham. https://doi.org/10.1007/978-3-319-76445-0_9
47. Little DP, Tulig M, Tan KC, Liu Y, Belongie S, Kaeser-Chen C, Michelangeli FA, Panesar K, Guha R, Ambrose BA (2020) An algorithm competition for automatic species identification from herbarium specimens. Appl Plant Sci 8(6):11365
48. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823
49. Siripibal N, Supratid S, Sudprasert C (2019) A comparative study of object recognition techniques: softmax, linear and quadratic discriminant analysis based on convolutional neural network feature extraction. In: Proceedings of the 2019 international conference on management science and industrial engineering, pp. 209–214. https://doi.org/10.1145/3335550.3335584
50. Horiguchi S, Ikami D, Aizawa K (2019) Significance of softmax-based features in comparison to distance metric learning-based features. IEEE Trans Patt Anal Mach Intell 42(5):1279–1285. https://doi.org/10.1109/TPAMI.2019.2911075
51. Goëau H, Bonnet P, Joly A (2021) Overview of plantclef 2021: cross-domain plant identification. In: Working Notes of CLEF 2021–Conference and Labs of the Evaluation Forum
52. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI conference on artificial intelligence. https://doi.org/10.1609/aaai.v31i1.11231
53. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/
54. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comp Vis 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
55. Sergio Guadarrama, Nathan Silberman (2016) TensorFlow-Slim: a lightweight library for defining, training and evaluating complex models in TensorFlow. https://github.com/google-research/tf-slim
56. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
57. Zhang Y, Davison BD (2020) Adversarial consistent learning on partial domain adaptation of plantclef 2020 challenge. arXiv preprint arXiv:2009.09289
58. Zhang Y, Davison BD (2021) Weighted pseudo labeling refinement for plant identification. Working Notes of CLEF
59. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929
60. Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11):2579–2605
61. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G (2013) API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for data mining and machine learning, pp. 108–122