MASKED AUTOENCODER-BASED SELF-SUPERVISED LEARNING FOR FOREST PLANT CLASSIFICATION

Luu Van Huy
The University of Danang,
University of Science and Technology,
Viet Nam
lvhuy@dut.udn.vn

Nguyen Huy Tuong
The University of Danang,
University of Science and Technology
Viet Nam
huytuong010101@gmail.com

Le Hoang Ngoc Han
The University of Danang,
University of Science and Technology
Viet Nam
hanlehngoc080601@gmail.com

Nguyen Van Hieu*
The University of Danang,
University of Science and Technology
Viet Nam
nvhieuqt@dut.udn.vn

Article history:
Received 12.03.2024, Accepted 09.06.2024

Abstract

Forest plant identification and classification play a pivotal role in various domains, encompassing biodiversity conservation, agricultural advancement, and beyond. Conventional plant identification methods often rely on expert botanists or manual identification approaches, which can be time-consuming and subjective. Deep learning models have emerged as a promising approach to automatically classify plants, offering high accuracy and efficiency. However, these models often rely on convolutional neural networks (CNNs) and their variants to extract features, which may fail to capture the complex relationships among plant characteristics. This paper proposes a novel feature extraction method using semi-supervised learning techniques combined with Masked Autoencoder architecture to enhance the feature extraction of plant data, applicable to problems with limited datasets. The proposed model, named MAE_SGD, achieves an accuracy of nearly 94% on the QuangNam-ForestPlant - a dataset collected by our research team in Quang Nam province, Central Vietnam, comprising 24,314 images of 710 different forest plant species. Future research directions will focus on expanding the forest plant dataset and improving the recognition model to increase the model's accuracy and overall performance in identifying forest vegetation.

Key words

Forest Plant Classification, Vision Transformer, Masked Autoencoder, Self-supervised learning, Deep learning

1 Introduction

Identification and classification of plants play an immensely crucial role in various fields, including biodiversity conservation, agricultural development, and ecological research. However, while the importance of plants in general has been widely recognized, the significance and clear challenges surrounding forest plants are often overlooked. Unlike easily encountered plant species, forest plants often inhabit remote, inaccessible areas, harboring many mysteries and complexities. This inherent difficulty in accessibility creates barriers to comprehensive data collection and imaging, leading to their relative neglect in scientific research. Accurately determining plant species is vital for understanding the diversity, distribution, and ecological interactions of plants, thereby promoting effective management and conservation strategies. Traditional plant recognition methods, which rely on botanical experts or utilize manually crafted identification features, are often time-consuming, subjective, and labor-intensive, hindering efforts in monitoring and conserving vegetation over large areas. The emergence of deep learning has revolutionized the field of plant classification, offering a promising approach to automate plant identification. Convolutional Neural Network (CNN) models, especially their deeper variants such as VGG-19, ResNet 152, have demonstrated outstanding accuracy in various image classification tasks, including plant recognition. However, CNNs often encounter challenges in capturing the complex and diverse features of plant images, especially when dealing with multiple species exhibiting complex morphological variations due to environmental, temporal factors, or limitations in the quantity of region-specific characteristic plant data. To address these challenges, we propose a novel vegetation feature extraction method using self-supervised learning techniques on the Masked Autoencoder architecture with a core model being the Vision Transformer (ViT). Self-supervised learning enables models to learn features from unlabeled data effectively, thereby reducing the need for large amounts of manually labeled training data. Furthermore, ViT is a recent advancement in deep learning architecture that achieves superior performance in capturing contextual features of plant images, overcoming the limitations of CNNs. This paper presents related research in section 2, followed by the proposed models in section 3. Experiments with a dataset collected by the authors in the Central Highlands region of Vietnam evaluate the accuracy of the proposed models and compare them with other models detailed in section 4. In section 5, an application called QuangNam Plant Id has been developed to help forest plant lovers and related individuals look up information and identify forest plants accurately. Finally, conclusions and future development directions of the research are presented in section 6.

2 Related work

2.1 Convolutional Neural Networks (CNNs) for Plant Classification

Previously, automatic plant identification was often addressed by requiring photographs of specific plant organs, such as leaves [[Kumar et al., 2012], [Fiel et al., 2011], [Sulc et al., 2014]], flowers [[Mattos et al., 2014], [Nilsback, 2009], [Angelova et al., 2013]] or tree bark [[Hieu et al., 2020], [Hieu et al., 2020], [Chen et al., 2021]]. Moreover, some classification systems imposed additional constraints on the input images, such as a white background behind the leaf images. The advent of Convolutional Neural Networks (CNNs) has been successful in several computer vision tasks, particularly those related to recognizing and detecting complex objects. The ability to extract hierarchical features from images makes them highly suitable for capturing patterns and complex structures of plant leaves as well as other morphological characteristics. CNN models tested on Plant CLEF 2015 [Goëau et al., 2015] have significantly outperformed the combination of previous models.

Not stopping there, scientists have further developed deeper CNN models with the concept of ”residual blocks” to address common issues encountered with traditional CNN models such as vanishing gradients [Hochreiter, 1998] when backpropagating a loss function through a specific model. A prime example is the ResNet model, introduced to the public in 2015 in the publication by Kaiming He and colleagues [He et al., 2016], which secured the first place position in the ILSVRC 2015 competition [Russakovsky et al., 2015] with a top-5 error rate of only 3.57%. Since its initial introduction in 2015, a plethora of variations of this architecture have emerged, such as ResNet 50, ResNet 101, ResNet 152 - with a depth of up to 152 layers, over 8 times deeper than VGG yet still maintaining lower complexity. ResNet quickly became the most popular architecture in computer vision, especially in object classification tasks such as identifying plants. For instance, Bodhwania and colleagues [Bodhwania et al., 2019] employed a deep learning model based on ResNet-50 to classify 185 plant species collected at Columbia University, University of Maryland, and the Smithsonian Institution with an accuracy of 93.09%. Similarly, Zhou and colleagues [Zhou et al., 2017] employed a deep residual network model with 152 layers pre-trained on ImageNet. They replaced the original fully connected layers with two randomly initialized fully connected layers and achieved third place in the PlantClef2016 [Goëau et al., 2016] vegetation classification competition. Despite its success, CNN and its variants still encounter certain limitations in plant classification. Firstly, their ability to grasp contextual information may be restricted, which could hinder their performance in identifying plant species with unusual characteristics or complex morphological changes. Moreover, CNNs often require a large amount of labeled training data, which can be difficult to obtain for certain species or specific regions of vegetation. Finally, CNNs acknowledge positional biases, such as translation invariance [Kornblith et al., 2019], where the models do not produce different results if the input is shifted, for example, by slightly moving the camera to the left or right. Due to positional biases, the capability of CNN-based models in predicting species of plants within the same family or species (with very similar features) is limited. In this context, the Vision Transformer (ViT) model developed by Google research groups has shown remarkable results in the field of computer vision in general and plant classification in particular.

2.2 Vision Transformer

The Vision Transformer (ViT) model has been experimented and evaluated by Google research teams[Dosovitskiy et al., 2020], demonstrating up to a four-fold improvement in computational efficiency and image classification accuracy compared to CNN architectures. What distinguishes this model is its capability to utilize self-attention layers to aggregate information from the entire image (Refer to Figure 1). Additionally, the model can learn the relative positions of image patches based on training data, thereby efficiently reconstructing the image structure. The operation of the Vision Transformer model involves partitioning the image into fixed-size patches and then flattening them to generate lower-dimensional embedding features from the flattened image arrays. In addition to patch partitioning, the model also sequences the arrays to

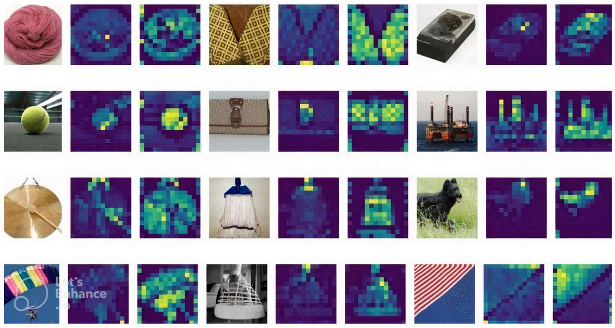


Figure 1. Illustration of the attention mechanism on images of the ViT model[Chen et al., 2021]

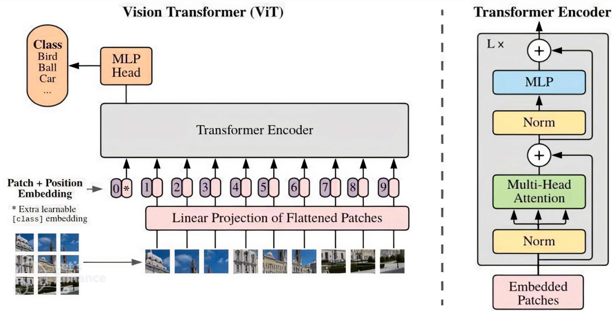


Figure 2. Illustration of the fundamental structure of the Vision Transformer model[Dosovitskiy et al., 2020]

ensure the model retains the positional information of the patches in the original image (Refer to Figure 2). These embedding features are used as inputs for the transformer encoder network. This encoding scheme consists of three main components. The first layer is Multi-Head Self-Attention(MSP) [Huy et al., 2023], which linearly connects all the outputs of the attention mechanism to generate outputs of the same size. Next is the Multi-Layer Perceptrons (MLP) layer, comprising two layers with Gaussian Error Linear Unit functions [Goëau et al., 2022]. The final layer is the Feedforward transition layer added before each block, as it does not have any local dependencies between images in the training process. This feature helps improve training time and overall performance. The best-performing Vision Transformers model achieved an accuracy of 88.55% on the ImageNet dataset.

2.3 Self-supervised learning for plant classification problem

Identifying different species is one of the prerequisites for maintaining biological diversity; however, the process of collecting and labeling images of plants requires substantial resources, involving field research, classification expertise, and meticulous data annotation. These requirements demand significant time, expertise, and financial investment [[Goëau et al., 2022], [Joly et al., 2022]]. Simultaneously, a vast amount of unlabeled and unstructured botanical data containing valuable information remains untapped. Self-supervised learning has emerged as a promising approach to overcome the limitations of supervised learning in plant classification, especially the scarcity of labeled training data. By leveraging unlabeled plant images, self-supervised learning enables models to understand meaningful features without the need for extensive manual labeling. This approach offers several advantages for plant classification, such as significantly reducing the time and effort required for data preparation by eliminating the need for manual labeling of large datasets. Furthermore, self-supervised learning allows models to learn prominent and more generalized feature representations of the botanical world, which can enhance the model's ability to identify plants across different environments and conditions. This is particularly useful when dealing with plant data from specific regions, as models can adapt to local plant species and morphological variations.

The masked language model and its autoregressive counterparts, for instance, BERT [Devlin et al., 2018] and Large Language Models [Minaee et al., 2024], are highly successful for pre-training in Natural Language Processing. These methods retain a portion of the input sequence and train models to predict missing content. The field of computer vision inherits that idea; one of the self-supervised learning tasks receiving community attention is masked image modeling (depicted in Figure 3). The basic idea of masked image modeling is to conceal a portion of image data and task the model with learning to predict the masked portion using information from the unmasked portion. This method is commonly used to learn a general representation of image data. Kaming He et al.'s research introduced a promising model architecture called the Masked Autoencoder, with the vanilla ViT-Huge model achieving the highest accuracy (87.8%) among methods using only ImageNet-1K data. Masked image modeling finds applications in various domains, including image inpainting, generating high-quality images from low-quality ones, and learning automatic representations of image data for tasks such as image classification, object detection, or entity localization within images.

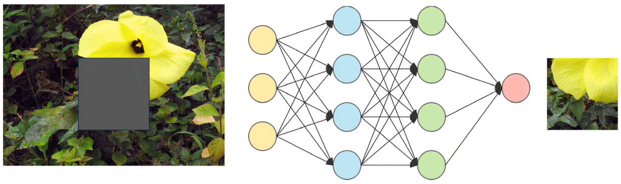


Figure 3. Illustration of the masked image modeling problem

Kaiming He - the progenitor of the residual block-based neural network model ResNet and his colleagues at the Facebook AI Research group, on the premise of the self-supervised task concept of masked image modeling, have proposed an MAE (Masked Autoencoder) masking encoder [He et al., 2022]. The MAE approach is straightforward, involving the masking of random patches of the input image and then reconstructing the missing pixels. The authors developed an asymmetric encoder-decoder architecture: an encoder that transforms observed signals into a latent representation and a decoder tasked with reconstructing the original signal from this latent representation. As a result, the model achieved the highest accuracy (87.8%) compared to other traditional supervised learning methods on ImageNet-1K data. The transfer performance in target tasks surpassed that of supervised training methods, demonstrating promising expansion potential.

In this study, we identify a key objective of developing a pre-task auxiliary task by incorporating self-supervised learning methods into the traditional VIT model. We believe this model can learn the underlying features of the general botanical world by training on a large dataset, thereby significantly improving the performance of final plant classification tasks.

3 Materials and Methods

3.1 Feature Extraction using Masked Autoencoder-Based Self-Supervised Learning

Our model is a masked image modeling architecture, designed to predict occluded patches in the encoded representation space. We train a Masked Autoencoder model on a large dataset of plant images (without using their labels) to learn a vast array of plant features.

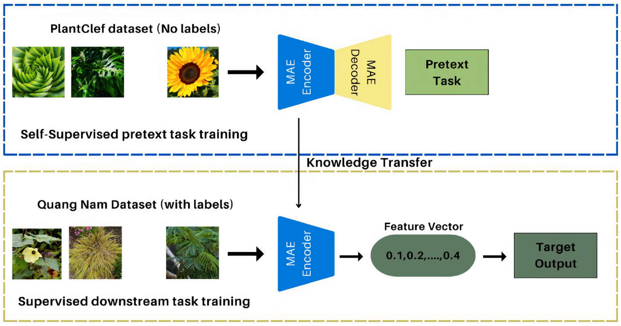

PlantClef dataset (No labels)
MAE Encoder
MAE Decoder
Pretext Task
Self-Supervised pretext task training
Knowledge Transfer
Quang Nam Dataset (with labels)
MAE Encoder
Feature Vector
0.1,0.2,....,0.4
Target Output
Supervised downstream task training

Figure 4. The architecture of masked image modeling in plant classification

More specifically, with the VIT-based core model, we partition images into non-overlapping patches. Then, we randomly sample and mask subsets of these patches (Figure 4). Similar to standard autoencoder architectures, our architecture comprises two main components:

(1) Encoder: This component maps the observed signals into latent representations. In our case, we employ the Vision Transformer (ViT) architecture [Dosovitskiy et al., 2020]. However, unlike traditional autoencoders, our encoder operates only on unobscured signal patches, significantly reducing computational costs.

(2) Decoder: This component utilizes latent representations and mask tokens (vectors learned to represent missing patch positions) to reconstruct the original signal. We employ a lightweight decoder, distinct from the computationally complex encoder, to efficiently handle the entire token stream.

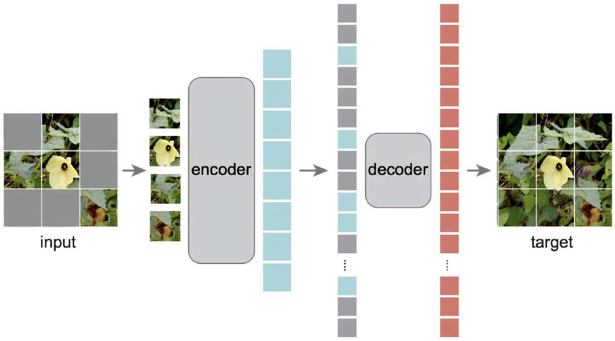


Figure 5. The architecture of MAE concealing vegetation imagery through patch arrays

Our Mean Absolute Error (MAE) reconstructs the input by predicting pixel values for each occluded patch. We employ Mean Squared Error (MSE) between the reconstructed image and the original one as the loss function, computed solely over the occluded arrays.

$$L = \frac{1}{M} \sum_{i=1}^{M} \|x_i - \hat{x}_i\|^2$$

In which: M is the number of obscured patches in the image, x_i is the original pixel value of the i_{th} obscured patch, $\hat{x}_i$ is the predicted pixel value by the decoding algorithm for the i_{th} obscured patch.

3.2 Plant Classification

After the pretext task, we utilize only the encoder part of the MAE model with weights updated for our final task, which is to extract features from the plant classification image dataset. We embed the entire image set of our QuangNamPlant dataset, yielding a 1024-dimensional feature vector for each image, summarizing the learned information from the dataset. These compact feature vectors represent various plant species and serve as inputs for subsequent classification tasks. After extracting features of the plant dataset, we use three algorithms for the final classification stage: the Multi-SVM

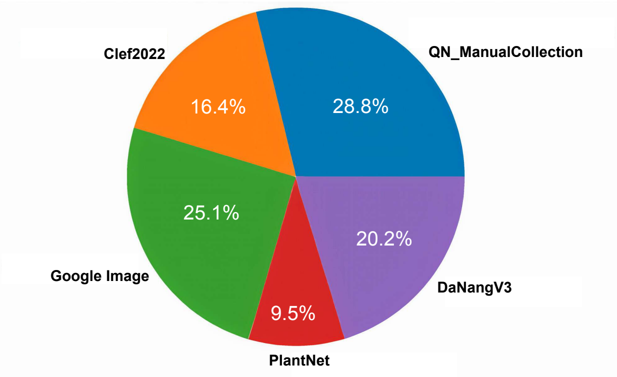


Figure 7. Percentage of images collected from 5 different sources

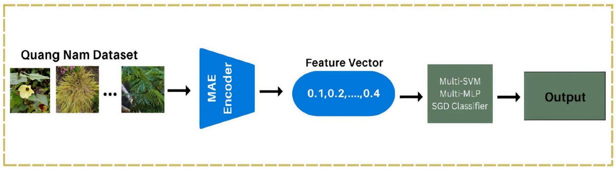


Figure 6. The encoding scheme of the model retained the post-pretext task for the primary task of plant classification.

algorithm [Chamasemani et al., 2011], the multi-layer perceptron (MLP) neural network, and the SGDClassifier - a linear classifier optimized for learning through small batches, is highly effective for large-scale problems due to its ability to learn incrementally.

4 Experiments and Results

As outlined in section 3, our experiment consists of two main phases: self-supervised training for MAE models to extract feature vectors, followed by employing models for plant classification tasks.

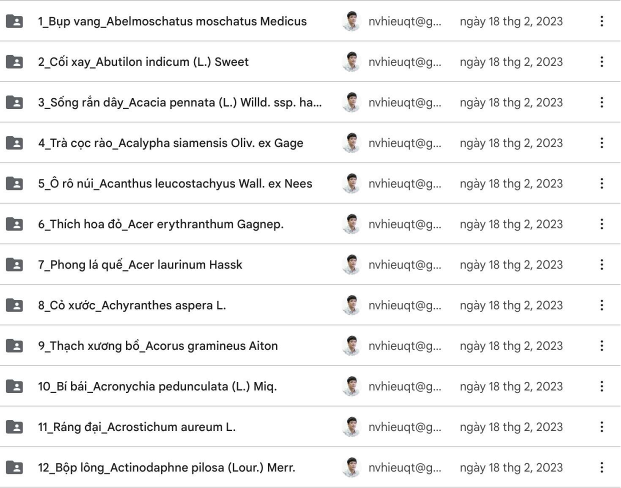
4.1 Data collection

We utilized the training and testing datasets from the PlantClef 2022 competition for the self-supervised learning task for the MAE model. This dataset encompasses 80,000 plant species, a total of over 2.9 million images collected and labeled by reliable experts. Due to hardware constraints, only about one-third of the data above (approximately 1 million plant images) was utilized. Each class comprises an average of 33 images, and to alleviate the imbalance issue, no class contains more than 40 images (Table 1).

Table 1. Statistics of the PlantClef 2022 dataset for MAE model training

Number of Classes	Total amount	Average	Max
30.000	1,000,048	33,3	40

For the main task of plant classification, we utilized a self-collected dataset from Quang Nam region, Vietnam, comprising 24,314 images representing 710 distinct plant species manually gathered from the area. The manually collected data was insufficient for us to conduct model training. Hence, we supplemented it by manually collecting data for each plant species from abundant online resources. The plant image dataset was augmented through collection from various sources such as Google Images, PlantClef 2022, PlantNet, and the Danang Plant Project [Hien et al., 2020]. Figure 7 illustrates the proportion of images gathered from different datasets within our dataset. After collecting data from manual and online sources, images of plants and information about individuals will be aggregated and categorized into a database. The collected data consists of raw data stored in Excel files, folders containing detailed images, and information for each individual (see Table 1, Figures 8, 9).



1_Bụp vang_Abelmoschatus moschatus Medicus	nvhieuqt@g...	ngày 18 thg 2, 2023
2_Cối xay_Abutilon indicum (L.) Sweet	nvhieuqt@g...	ngày 18 thg 2, 2023
3_Sống rắn dây_Acacia pennata (L.) Willd. ssp. ha...	nvhieuqt@g...	ngày 18 thg 2, 2023
4_Trà cọc rào_Acalypha siamensis Oliv. ex Gage	nvhieuqt@g...	ngày 18 thg 2, 2023
5_Ô rô núi_Acanthus leucostachyus Wall. ex Nees	nvhieuqt@g...	ngày 18 thg 2, 2023
6_Thích hoa đỏ_Acer erythranthum Gagnep.	nvhieuqt@g...	ngày 18 thg 2, 2023
7_Phong lá quế_Acer laurinum Hassk	nvhieuqt@g...	ngày 18 thg 2, 2023
8_Cỏ xước_Achyranthes aspera L.	nvhieuqt@g...	ngày 18 thg 2, 2023
9_Thạch xương bồ_Acorus gramineus Aiton	nvhieuqt@g...	ngày 18 thg 2, 2023
10_Bí bái_Acronychia pedunculata (L.) Miq.	nvhieuqt@g...	ngày 18 thg 2, 2023
11_Ráng đại_Acrostichum aureum L.	nvhieuqt@g...	ngày 18 thg 2, 2023
12_Bộp lông_Actinodaphne pilosa (Lour.) Merr.	nvhieuqt@g...	ngày 18 thg 2, 2023

Figure 8. Collecting detailed data information about individual specimens of plant species

4.2 Data preprocessing

After obtaining raw data, we only applied preprocessing to our self-collected dataset. The preprocessing steps include:

- Cropping the botanical images, focusing on the central portion of the image to extract characteristics more effectively;
- Adjusting the image to a fixed size $(224, 224)$. The selection of a fixed image size is based on previous studies on plant classification [Goëau et al., 2015];

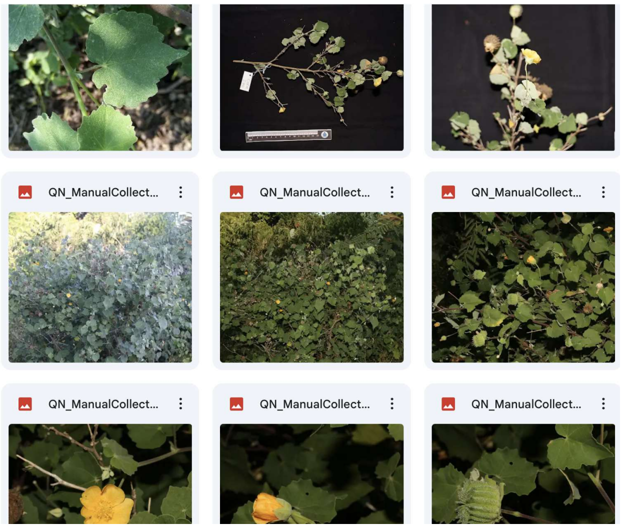
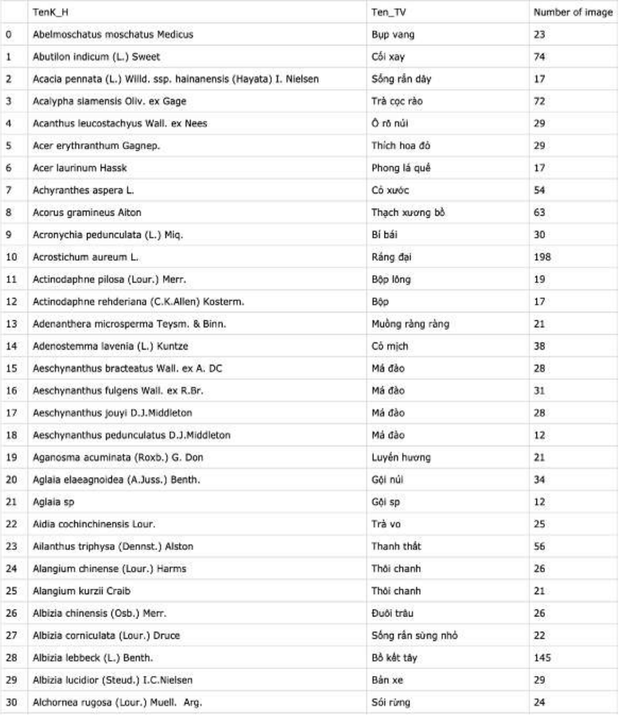


Figure 9. Example of gathering detailed data on the plant species "Abutilon indicum (L.) Sweet"



	TenK_H	Ten_TV	Number of image
0	Abelmoschatus moschatus Medicus	Bụp vang	23
1	Abutilon indicum (L.) Sweet	Cối xay	74
2	Acacia pennata (L.) Willd. ssp. hainanensis (Hayata) I. Nielsen	Sống rắn dây	17
3	Acalypha siamensis Oliv. ex Gage	Trà cọc rào	72
4	Acanthus leucostachyus Wall. ex Nees	Ô rô núi	29
5	Acer erythranthum Gagnep.	Thích hoa đỏ	29
6	Acer laurinum Hassk	Phong lá quế	17
7	Achyranthes aspera L.	Cỏ xước	54
8	Acorus gramineus Aiton	Thạch xương bồ	63
9	Acronychia pedunculata (L.) Miq.	Bí bái	30
10	Acrostichum aureum L.	Ráng đại	198
11	Actinodaphne pilosa (Lour.) Merr.	Bộp lông	19
12	Actinodaphne rehderiana (C.K.Allen) Kosterm.	Bộp	17
13	Adenanthera microsperma Teysm. & Binn.	Muồng ràng ràng	21
14	Adenostemma lavenia (L.) Kuntze	Cỏ mịch	38
15	Aeschynanthus bracteatus Wall. ex A. DC	Má đào	28
16	Aeschynanthus fulgens Wall. ex R.Br.	Má đào	31
17	Aeschynanthus jouyi D.J.Middleton	Má đào	28
18	Aeschynanthus pedunculatus D.J.Middleton	Má đào	12
19	Aganosma acuminata (Roxb.) G. Don	Luyến hương	21
20	Aglaia elaeagnoidea (A.Juss.) Benth.	Gội núi	34
21	Aglaia sp	Gội sp	12
22	Aidia cochinchinensis Lour.	Trà vo	25
23	Ailanthus triphysa (Dennst.) Alston	Thanh thất	56
24	Alangium chinense (Lour.) Harms	Thôi chanh	26
25	Alangium kurzii Craib	Thôi chanh	21
26	Albizia chinensis (Osb.) Merr.	Đuôi trâu	26
27	Albizia corniculata (Lour.) Druce	Sống rắn sừng nhỏ	22
28	Albizia lebbeck (L.) Benth.	Bồ kết tây	145
29	Albizia lucidior (Steud.) I.C.Nielsen	Bản xe	29
30	Alchornea rugosa (Lour.) Muell. Arg.	Sói rừng	24

Figure 10. CSV file storing the names and quantities of plant species

- Enriching the dataset: Generate additional plant images by rotating and flipping images horizontally and vertically;
- Data normalization is performed to ensure that they fall within a specific range of values $(0, 1)$;
- Label encoding: converting labels of plant species into numerical representations. For instance, the label "Abelmoschatus moschatus Medicus" could be encoded as "1", "Abutilon indicum (L.) Sweet" as "2".

We also removed some plant classes with too little data, keeping those with eight or more images. The preprocessed dataset had an average of around 33 images per class and approximately 710 species with eight or more images (refer to Figure 10 and Table 2).

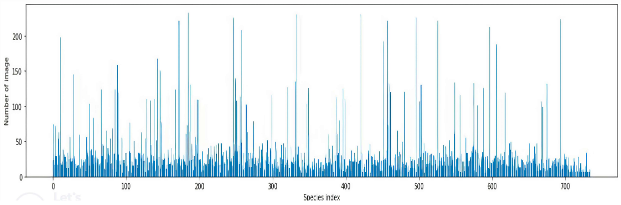


Figure 11. Distribution of the total number of images per species after preprocessing

Table 2. Statistical summary of training data

Total of spices	Total amount	Average	Max	Min
710	24314	33	233	8

4.3 Training and testing data

The vegetation data of Quang Nam is allocated for plant classification tasks with a $60 - 20 - 20\%$ split, where 60% is for the training set, 20% is for the validation set and 20% is for the test set. All three sets for training, validation and testing sets have the same distribution in terms of species quantity (Figure 12, 13). Image data is stored in directories named after the species. The division of training and validation data is recorded in a CSV file. Species names are encoded into integers for ease of training and will be reverse-inferred back to species names after prediction.

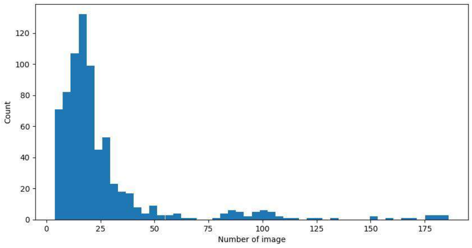


Figure 12. Data distribution of the training set

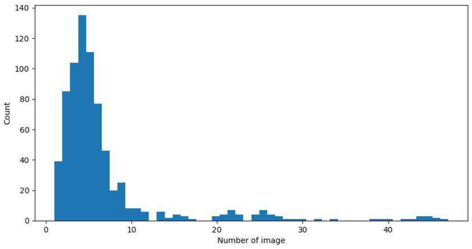


Figure 13. Data distribution of the testing set

4.4 Fine-tuning the model

Pretext task - Self-supervised MAE: We employ a pre-trained ViT-large MAE model on the ImageNet 1k dataset. Random cropping and random horizontal flipping techniques are utilized for data augmentation. The masking rate is set to 75%, following the study by Kaiming He et [He et al., 2022]. High masking rates prevent simple extrapolation from neighboring patches, enhancing the learning process. The model is trained with a batch size 512 for 100 epochs, with a learning rate of 0.005.

Plant classification task: We fine-tuned the last layer of the MAE encoding with a classification layer of 710 plant species. Our models were all trained on a machine with a CPU—Intel Xeon Processor and GPU—Tesla K80 configuration. The initial learning rate was set to 0.001 in the experiments and adjusted and monitored during training to reach optimal values. When performed with 100 epochs, the early-stopping mechanism was also utilized to halt the training process when the model's accuracy measured on the validation set did not improve for three consecutive epochs. The loss function employed for supervised fine-tuning was a standard categorical cross-entropy.

4.5 Evaluating metrics

For the plant classification task, we compared our model with previous plant recognition models using F1 score, Precision, top @1 accuracy (accurate prediction of plant species), and top @5 accuracy (providing five predictions with at least one accurately predicting the plant species).

4.6 Evaluation of results

We compared our three models with popular models such as ResNet and ConvNeXt, as well as a model of our previous research called PlantKViT [Hieu et al., 2023]. Table 3 demonstrates that our proposed model architecture significantly outperforms Resnet and ConvNeXt models in plant classification when using the same dataset. Specifically, the lowest F1 score of our proposed model is significantly higher at 0.81 compared to 0.77 and 0.56 of ConvNeXt and Resnet models, respectively. Similar results are also achieved from our proposed model with accuracy (Top @1 and Top @5 accuracy) being better by 3 to 5%. Moreover, the SGD classifier model shows the highest efficiency with a top @1 accuracy of 87% and top @5 accuracy of approximately 94%, surpassing the application of classifiers of the same level. The results also indicate that our models outperform the previously studied KNN-based model (by approximately 3% in metrics).

The MAE_SGD model loses its advantage only in the prediction index when its index is slightly lower by about 3% compared to the MAE_MLP model. The explanation is that the dataset has underlying non-linear structures, and MAE_MLP with multiple hidden layers allows the model to learn complex, non-linear relationships between the features and the plant classes. Meanwhile, MAE_SGD is a linear classifier that might struggle to capture the intricate relationships between features extracted from plant images and their corresponding classes.

Table 3. Comparison of the accuracy and speed of plant classification models

Model	F1 score	Precision	Top @1	Top @5	Inference time(s)
Resnet 152	0.56	0.56	59.02	76.11	0.567
ConvNeXt	0.77	0.79	77.12	89.01	0.0209
PlantKViT	0.81	0.83	81.73	92.2	0.0881
MAE_SVM	0.82	0.84	86.11	92.45	0.134
MAE_MLP	0.82	**0.85**	86.76	93.74	0.129
MAE_SGD	**0.86**	0.82	**0.89**	**93.89**	0.131

4.7 Data Imbalance in the Plant Classification Problem

Data imbalance is a prevalent issue in plant classification problems, where the number of samples for different plant species varies significantly. In our project, there is a small number of species (majority classes) are well-represented with numerous samples, each with more than 200 images, whereas a large number of species (minority classes) have significantly fewer images (Figure 11). This imbalance poses a significant challenge for our models, which tend to be biased towards the majority classes, leading to poor performance in the minority classes. To overcome this address, in evaluating our plant classification models, we used the F1 score and Precision metrics alongside the traditional Accuracy metric to ensure a fair and comprehensive assessment. While Accuracy alone can be misleading, especially in the context of imbalanced datasets, F1 score and Precision provide a more nuanced evaluation. This

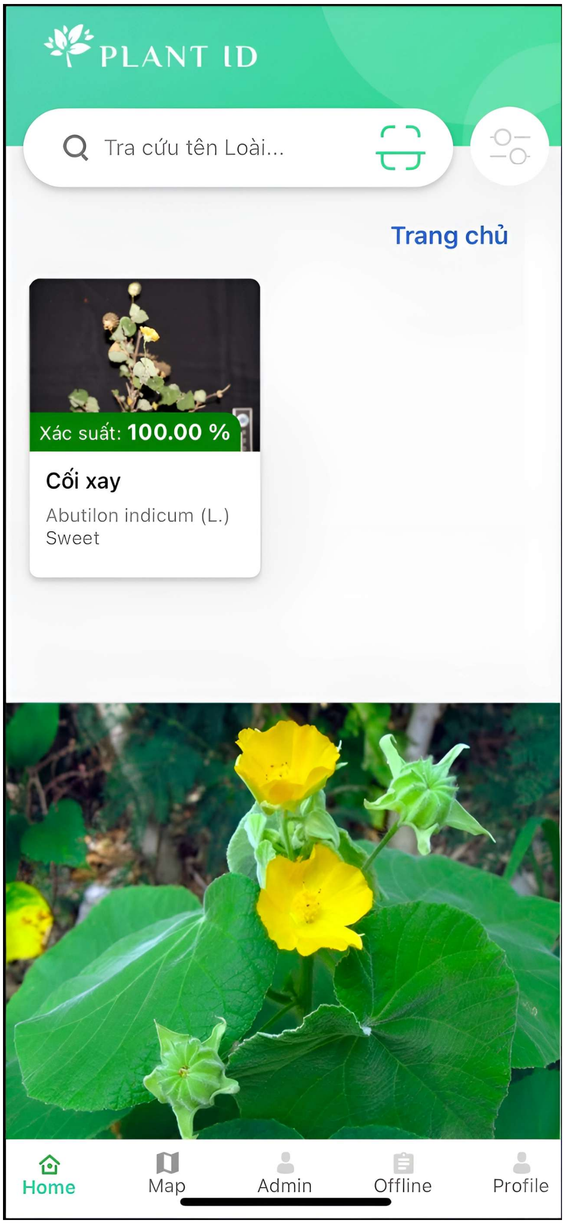


Figure 14. The image of the ”Abutilon indicum (L.) Sweet” species taken from the internet and the prediction generated by the QuangNam Plant Id application

is particularly evident in the performance of the ConvNeXt model, which achieved a high Accuracy of 89%. However, this model’s effectiveness is questionable as its F1 score and Precision were relatively low, at 0.77 and 0.79, respectively. These metrics indicate that the ConvNeXt model struggles with correctly identifying minority classes despite its overall high Accuracy. In contrast, our final model, MAE_SGD, demonstrated robust performance across all three metrics, achieving an Accuracy of 94%, an F1 score of 0.86, and a Precision of 0.82. This balanced performance underscores the importance of using the F1 score and Precision to evaluate models on imbalanced datasets, ensuring that the model not only performs well in majority classes but also maintains high performance in minority classes, leading to a more reliable and fair assessment.

We also tried some loss function tuning experiments to address the data imbalance problem. Specifically, we did explore incorporating a weighted loss function [Fernando et al., 2022]. However, our experimentation revealed that this adjustment provides a negligible improvement of about 0.5% to 1% in the accuracy of the models.

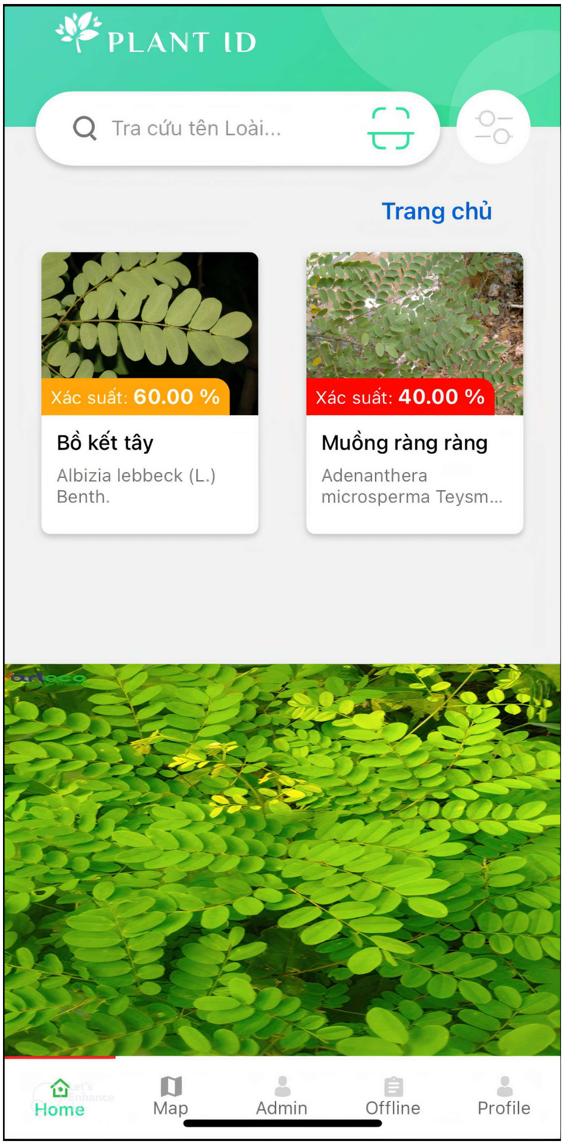
5 Application

With the impressive results presented above, we employed the MAE_SGD model in the Quang Nam Plant ID application, an application designed for classifying plant species in Quang Nam province, Vietnam. It encompasses various functions such as plant image lookup, habitat distribution lookup, and marking plant locations on maps. Currently, the application is available on the Google Play and App Store [Ha, 2023]. Our plant classification model is utilized in the ”plant image recognition via images” feature.

The application helps users identify plant images by requiring them to provide input images of plants. There are two ways to do this: Users can either directly capture images of the plant species they want to identify or upload their images. Once the image is uploaded, the MAE_SGD model is used to classify the plant and provide the confidence level of the prediction.

In Figure 14, with plant species that have abundant and well-trained data, the model provides predictions with 100% confidence for a single output. However, when dealing with plant species with highly similar leaf details or other features, the model may still exhibit errors (Figure 15). Nevertheless, the correct results remain within the model’s output prediction list. These errors present challenges but also opportunities for me to develop models with even higher accuracy in the future. To effectively utilize the Quang Nam Plant ID application, users should take note of the following guidelines:

- Focused imaging: Ensure that the image of the plant species is centrally positioned within the frame of the picture you intend to classify. The plant species should be of sufficient size, and their main portion should be within the frame, which aids in accurately identifying the plant species by the recognition model.
- Limit capturing multiple types of plants in the same photo: In a photograph, users should avoid capturing multiple different plant species, which helps avoid confusing classification models between species and ensures accurate results.
- Removing noise factors: When taking photos, photographers attempt to eliminate noise factors such as background clutter, strong lighting, or unrelated objects from the frame to improve the accuracy of plant recognition.



Xác suất: 60.00 %
Bồ kết tây
Albizia lebbeck (L.) Benth.

Xác suất: 40.00 %
Muồng ràng ràng
Adenanthera microsperma Teysm...

Figure 15. The image of the "Adenanthera microsperma Teysm. & Binn" species taken from the internet and the prediction generated by the Quang Nam Plant Id application

6 Conclusion

Our MAE architecture, trained via self-supervision on a large dataset and then utilized in the primary classification task, demonstrates superior effectiveness over popular deep learning models such as ResNet, ConvNext, or our previous study. This research represents a potential development direction for achieving highly accurate plant identification on a dataset of small or medium size. Moreover, we have manually collected a unique dataset on forest vegetation in Quang Nam province, Vietnam. Unlike the typical plant species gathered in the PlantClef dataset, species in our dataset are often found in hard-to-reach locations and showcase the unique biodiversity of the region.

Plant classification using deep learning models is a promising and innovative approach to automating the identification and classification of different plant species. However, this method still faces significant challenges, such as the influence of other factors in the images, natural variations, and the diversity of plants in terms of shape and color. These complexities pose a significant challenge in developing increasingly robust deep learning models for accurately classifying these species. In the future, we plan to develop plant classification models in several key directions. The first is to enhance and improve the quality of the dataset. The plant retrieval and identification model needs a dataset with a larger number of classes, more balanced sample sizes to accurately assess the model's capabilities. Additionally, integrating data sources in various forms, such as leaf structures and spectral information, may allow the model to characterize plant features more comprehensively. Furthermore, the model's scalability across different plant regions is also a noteworthy consideration.

Acknowledgements

This study is funded and implemented for the project 'The forest plant classification model using deep learning techniques with number T 2023-02-29'. This work is supported by the University of Science and Technology, University of Danang.

References

Angelova, A., Zhu, S., and Lin, Y. (2013). Image segmentation for large-scale subcategory flower recognition. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pp. 39–45.

Bodhwania, V., Acharjyaa, D. P., and Bodhwania, Umesh. (2019). Deep Residual Networks for Plant Identification. In *Proceedings of the International Conference on Pervasive Computing Advances and Applications - PerCAA 2019*, pp. 186–194.

Chamasemani, F. F., and Singh, Y. P. (2011). Multi-class Support Vector Machine (SVM) Classifiers – An Application in Hypothyroid Detection and Classification. In *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 351–356.

Chen, X., Hsieh, C. J., and Gong, B. (2021). When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fiel, S., and Sablatnig, R. (2011). Automated identification of tree species from images of the bark, leaves, and needles. In *Proc. of 16th Computer Vision Winter Workshop*, pp. 1–6.

Fernando, K. R. M., and Tsokos, C. P. (2022). Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. In *IEEE Transactions on Neural Networks and Learning Systems*, 33(7), 2940-2951.

Goëau, H., Bonnet, P., and Joly, A. (2015). Lifeclef plant identification task 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, pp. 8–11.

Goëau, H., Bonnet, P., and Joly, A. (2016). Plant Identification in an Open-world (LifeCLEF 2016). In *CLEF: Conference and Labs of the Evaluation Forum*, pp. 428–439.

Goëau, H., Bonnet, P., and Joly, A. (2022). Overview of PlantCLEF 2022: Image-based plant identification at global scale. In *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, pp. 1916–1928.

Ha, N. (2023). Quang Nam Plant ID (Version 1.0) [Mobile application]. Retrieved from `https://apps.apple.com/vn/app/quangnam-plant-id/id6450864796?l=vi`.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009.

Huy, L. V., Hien, N. L. H., Phuong, N. T. H., and Hieu N. V., (2023). Deep learning model with hierarchical attention mechanism for sentiment classification of Vietnamese comments. *Cybernetics and Physics*, 12(2), 111-120.

Hien, N. L. H., Tien, T. Q., and Hieu, N. V.(2020a). RWeb crawler: Design and implementation for extracting article-like contents. *Cybernetics and Physics*, 9(3), 144-151.

Hieu, N. V., and Hien, N. L. H. (2020a). Recognition of Plant Species using Deep Convolutional Feature Extraction. *International Journal on Emerging Technologies*, 11(3), pp. 904–910.

Hieu, N. V., and Hien, N. L. H. (2020b). Automatic plant image identification of Vietnamese species using deep learning models. *arXiv preprint arXiv:2005.02832*.

Hieu, N. V., Hien, N. L. H., Huy, L. V., Tuong, N. H., and Thoa, P. T. K. (2023). PlantKViT: A Combination Model of Vision Transformer and KNN for Forest Plants Classification. *JUCS - Journal of Universal Computer Science*, 29(9), pp. 1069–1089.

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02), pp. 107–116.

Joly, A., Goëau, H., Kahl, S., Picek, L., Lorieul, T., Cole, E., ... Hruz, M. (2022). Overview of LifeCLEF 2022: An evaluation of machine-learning-based species identification and species distribution prediction. In *CLEF 2022 - 13th International Conference of the CLEF Association*, pp. 257–285.

Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., and Soares, J. V. (2012). Leafsnap: A computer vision system for automatic plant species identification. In *Computer Vision–ECCV 2012*, pp. 502–516.

Mattos, A. B., Herrmann, R. G., Shigeno, K. K., and Feris, R. S. (2014). Flower classification for a citizen science mobile app. In *Proceedings of International Conference on Multimedia Retrieval*, p. 532.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large Language Models: A Survey. *arXiv preprint arXiv:2402.06196*.

Nilsback, M. E. (2009). An Automatic Visual Flora-Segmentation and Classification of Flower Images. Oxford University, Oxford.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, pp. 211-252.

Sulc, M., and Matas, J. (2014). Texture-based leaf identification. In *Computer Vision-ECCV 2014 Workshops*, pp. 185–200.

Zhou, L., Li, Q., Huo, G., and Zhou, Y. (2017). Image Classification Using Biomimetic Pattern Recognition with Convolutional Neural Networks Features. *Computational Intelligence and Neuroscience*.