A guided multi-scale categorization of plant species in natural images

Jonas Krause, Kyungim Baek, and Lipyeow Lim
Dept. of Information and Computer Sciences, University of Hawai'i at Manoa
1680 East-West Rd, Honolulu, HI 96822
krausej,kyungim,lipyeow@hawaii.edu

Abstract

Automatic categorization of plant species in natural images is an important computer vision problem with numerous applications in agriculture and botany. The problem is particularly challenging due to the large number of plant species, the inter-species similarity, the large scale variations in natural images, and the lack of annotated data. In this paper, we present a guided multi-scale approach that segments the regions of interest (containing a plant) from a complex background of the natural image and systematically extracts scale-representative patches based on those regions. These multi-scale patches are used to train state-of-the-art Convolutional Neural Network (CNN) models that analyze a given plant image and determine its species. Focusing specifically on the identification of plant species in natural images, we show that the proposed approach is a very effective way of making deep learning models more robust to scale variations. We perform a comprehensive experimental evaluation of our proposed method over several CNN models. Our best result on the Inception-ResNet-v2 model achieves a top-1 classification accuracy of 89.21% for 100 plant species which represents a 5.4% increase over using random cropping to generate training data.

1. Introduction

Traditionally, botanists analyze different characteristics of a plant to categorize its species. But identifying a plant species accurately based on visual characteristics requires considerable expertise [27], and is almost impossible for the general public. Therefore, an automated system to classify plants has important implications for the society at large not only in the preservation of ecosystem biodiversity including public education, but also in agricultural activities such as automatic crop analysis, species variability analysis, analysis of phylogenetic relationships, identification of pests and diseases, and identification of invasive species.

Computer vision approaches for plant identification using controlled images have shown promising results [12, 14, 17]. However, a real-world plant categorization system needs to deal with natural images, which is a major challenge for any automated method. While there are dozens of plant identification apps available, they typically do not perform well on unconstrained natural images. The analysis of natural images can be extremely difficult due to complex backgrounds, objects appearing in any scale, occlusions, and the presence of numerous different objects in the same image. While the human visual system deals with those factors with ease, an equivalent computational model for plant categorization using natural images is still an open problem.

In this paper, we propose a deep learning-based approach that, given a natural image of a plant, categorizes its species and deal with problems such as: *i)* If there is a plant in the image, where is it located? *ii)* What is the most representative area of the image for the plant categorization problem? *iii)* How to handle the same plant in different scales? And *iv)* how to improve the use of state-of-the-art computer vision methods to categorize the plant species? In particular, we focus on a guided multi-scale approach that is used to train existing CNN models for the fine-grained categorization of plants. Our method exploits current deep learning models to identify regions of interest, i.e., regions containing at least one plant in the natural image. Multiple patches of different sizes are then extracted from those regions. These patches are further post-processed and resized for the particular CNN model used for categorization. To the best of our knowledge, no one has proposed a method designed to extract multi-scale representative patches of plants from natural images. The contributions of this paper are:

- We propose a new approach to make CNNs more robust to scale variations when analyzing plants in natural images.
- We implemented our method with different CNN models for the fine-grained categorization of plants.
- We performed a comprehensive experimental validation and evaluation of the proposed method on different data sets containing natural images. Our results

show a considerable improvement in accuracy when the proposed approach is used.

In the next section, we present the related work by briefly describing methods proposed to address the multi-scale issue of plants in natural images and other fine-grained categorization problems that also require a similar scale analysis. The proposed guided multi-scale approach, as well as its step-by-step process, is explained in Section 3. A detailed analysis and discussions of the experimental results are presented in Sections 4 and 5. We conclude the discussion in Section 6.

2. Related Work

Focusing on multi-scale approaches that try to handle the analysis of objects in natural images, we survey some relevant previous work and organize them as per their implemented method to address the scale issue.

2.1. Human-in-the-Loop

To improve accuracy and address the multi-scale issue in the fine-grained categorization of plant species, a visual analysis (human-in-the-loop) is implemented in [5] where the correct annotation is done by human labelers, allowing the incorrectly classified images to be reintegrated into the dataset after this laboring classification. Another approach using human-in-the-loop is proposed by Wah *et al.* [26], which is designed for the fine-grained categorization of bird species. Their visual recognition system is composed of a machine and a human user, who is asked to provide additional information by clicking on the object parts and answering binary questions. Using a dataset called CUB-200 [28] of 200 bird species and their annotated parts, Wah *et al*. propose to solve the bird classification problem by analyzing specific areas of the image with the assistance of a human user, who can easily indicate the bird parts (head, beak, body, wing, and tale) independent of the image scale.

2.2. Different Feature Representations

Other computer vision techniques have been proposed to solve the multi-scale issue though they are not specifically designed for plants. Nevertheless, some of these ideas can be adapted to the fine-grained categorization of plant species and may help the training process of the CNNs. For example, Yang and Ramanan [29] propose an approach that combines different feature representations to address the multi-scale recognition problem. Their method seeks to extract features from multiple layers of a single deep network. Essentially, a directed acyclic graph (DAG) structured CNN is used to learn a set of multi-scale features at each level, which are shared with the final output predictor simultaneously. Experiments are performed using a wide variety of environmental scenes with different backgrounds and objects in them. Reported results suggest that encoding scale-specific features may be beneficial for training CNNs both for general image classification and fine-grained categorization tasks.

2.3. Multi-Scale Fusion

Back to plant species categorization, Hu *et al.* [7] propose a CNN with multi-scale fusion designed for leaf recognition. Using the MK Leaf [13] and the LeafSnap Plant Leaf [12] datasets, a customized CNN is trained by slowly infusing images of multiple resolutions with the list of bilinear interpolation operations used to sample them. In this way, down-sampled images are progressively fed to the CNN, concatenating extracted features at each level of the network to perform a multi-scale analysis. Nevertheless, their method is designed to work with leaf images taken in controlled backgrounds only, limiting its application.

Another fusion method is presented by Karpathy *et al.* [8] where a multi-resolution CNN architecture is proposed for video classification. In their approach, each frame of the video is fed into two separate streams that converge to fully connected layers at the end. The first stream models low-resolution images, while the second stream processes high-resolution patches cropped at the center of the input frames. As a result, a fast and dual-scale classification on each frame of the videos is enabled through the two streams of the CNN. A similar approach is presented by Mo *et al.* [16] where patches cropped at the center of the images are analyzed by a dedicated CNN while the entire input image passes through another identical network. The outputs of the two networks are concatenated and a third CNN is used at the top of the extracted features for a deeper representation and classification. Both methods [8, 16] implement their preprocessing stage of cropping representative areas focusing on the center of the images. However, when dealing with plants in natural images, it should be considered that plants may not be centered, making a guided approach necessary to indicate where the plant is located.

2.4. Pose Normalized Feature Spaces

An approach for categorization of bird species guided by the selection of useful parts is introduced by Branson *et al.* [2]. Their approach employs pose normalized CNNs and a graph-based clustering algorithm is used to learn a compact pose normalization space. In this case, cropped patches of the bird's head and body, as well as the entire image are used to train each CNN. The multi-scale issue is addressed by randomly extracting cropped image patches with arbitrary sizes to be used by the CNNs to learn scale-invariant features. However, all cropped areas have to be resized to a square to fit the first layer of the CNNs, changing its aspect ratio. Paying attention to this detail, Liu *et al.* [15] present a similar multi-scale approach with additional CNNs called

attention networks. These auxiliary networks are independent CNNs incorporated to identify representative square sample areas at two different scales. The two-scale features extracted at the identified locations are combined with the entire input image for classification. As a result, this approach focuses on three main scales to extract and classify the bird's head, its body, and the entire scene outperforming previously described methods in the fine-grained categorization of birds. However, this method relies on annotated parts to construct the match between parts and classes, which makes it difficult to apply it in the categorization of plant species. To the best of our knowledge, there is no available dataset with annotated plant species and their respective parts (leaf, flowers, bark, steam, fruit, etc.). Therefore, alternative methods to extract representative patches for fine-grained categorization of plants have to be designed.

2.5. Segmentation-based Approaches

An interesting idea that inspired our guided approach is proposed by Krause *et al.* [9]. Although being developed for the fine-grained categorization of bird species and using annotated bounding boxes for training, their approach does not use part annotations in its classification process. Instead, segmentation and alignment are used to generate part images that are combined to represent the entire bird. Nevertheless, annotated bounding boxes are used to train this approach, limiting its application to datasets such as the CUB-200 [28]. Even so, the idea of segmenting the object to extract representative patches can be adapted for fine-grained categorization of plant species.

A systematic object detection and segmentation approach for fine-grained categorization of flowers is presented by Angelova and Zhu [1]. Their method first detects low-level regions that could potentially belong to the object of interest and then performs a full-object segmentation within those regions. They also zoom-in on the object, center it, and normalize its size to a single scale discounting the effects of the background. To understand the benefits of the segmentation step for fine-grained categorization tasks, Angelova and Zhu compare their approach with a baseline model. The baseline model does not use segmentation and is outperformed by their model in all tested datasets, suggesting that the segmentation step helps to improve the recognition performance. For plants in natural images, it is difficult to correctly segment all the details of a plant from a complex background and other similar plants that may be in the same image. However, as suggested by previous studies [1, 9, 10, 11, 19, 20], a segmentation step can help guide the extraction of representative samples from a natural image. Section 3 details how the segmentation is incorporated for plant species categorization using CNNs.

3. Multi-Scale Plant Categorization

The proposed guided multi-scale approach exists as part of a larger plant categorization system and framework, called *WTPlant (What's That Plant?)* [10, 11].

3.1. The WTPlant Framework

The *WTPlant* framework is a system of CNN models for categorizing plants in natural images that is designed to address the challenges of segmenting the plant from a complex background and dealing with large variation in scale of natural images. *WTPlant* addresses these issues by:

- Using stacked convolutional blocks for scene parsing.
- Implementing a preprocessing stage for multi-scale analysis.

Figure 1 shows a diagram of the basic building block or pipeline of the *WTPlant* framework. In general, the depicted pipeline first extracts a various number of multi-scale patches with the guidance of the segmentation process, predicts the plant species at various scales by classifying the patches individually, and combines those predictions to make a final decision on plant species. In the *WTPlant* system, multiple of such pipelines can be constructed to classify specific parts of plants (e.g., flower, fruit, and bark) and the classification results of each pipeline can be combined using ensemble-like methods or a final soft-max layer. The framework is also designed to be modular and extensible, so that newer and better deep learning models can be incorporated with minimal re-customization. Each component of the framework (e.g., segmentation methods, patch extraction processes, and CNN models) can be independently upgraded or swapped with different implementations. A byproduct of this extensibility is that we are able to use *WTPlant* to train and evaluate different CNN models.

3.2. Multi-Scale Approach Guided by Segmentation

In the following, we describe the proposed multi-scale approach where images of plants are first segmented to extract multi-scale representative patches used for training CNNs. This is an important process when working with natural images due to the possible presence of other objects in the scene that may adversely affect the plant categorization.

Segmenting Plant Region: Given a natural image, a broad analysis of the entire scene is performed to detect the presence of plants. This is one of the major problems in computer vision and it is called scene parsing, or segmentation and recognition of objects in an image. Using a CNN architecture with stacked convolutional blocks, Zhou *et al.* [30] developed a cascade segmentation module for the scene parsing problem (henceforth referred to as MIT Scene Parsing). They trained a three-level stacked CNN

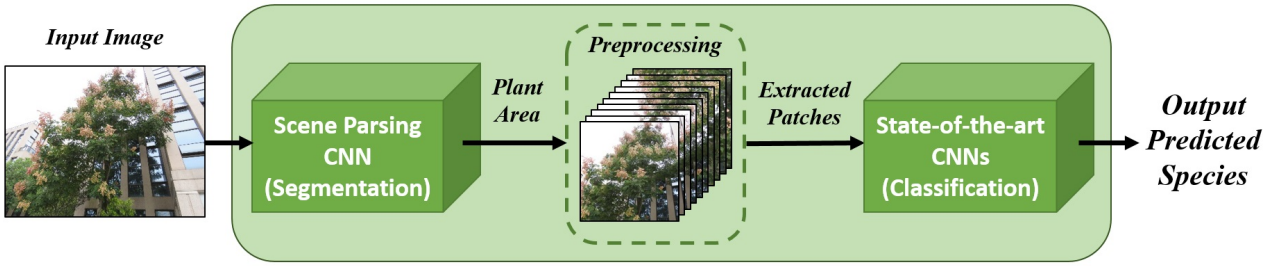


Figure 1. *WTPlant* framework.

using a dataset called ADE20K to segment common background objects (sky, road, building, etc.), foreground objects (car, people, plant, etc.) and object parts (car wheels, peoples head and torso, etc.). The MIT Scene Parsing is trained to segment 150 different objects from a scene, including plants. Due to the highly accurate results reported on the segmentation of plants and the usage of stacked convolutional blocks in their process, we have incorporated this method in the *WTPlant* framework for segmentation process to localize plants in natural images.

The segmentation process produces Regions of Interest (RoI) delimitating the plants' areas in the input image. If more than one RoI is detected, only the largest area is chosen to represent the plant in the image. (We leave the identification of multiple plants in an image to the future work.) If any RoI is collected, meaning the potential presence of plants in the image, the RoI is assumed to contain the most representative information of the plant and is further processed to predict the plant species. If no RoI is identified during the segmentation process, the image is considered as "No Plant Image".

Extracting Multi-Scale Patches: Once the RoI (i.e. plant region) is identified (green boundary in Figure 2(c)), we first define a square bounding box of the RoI based on the minimum and maximum x and y coordinate values of the RoI (red square enclosing green boundary in Figure 2(c)). This bounding box forms the largest patch to be extracted. Secondly, the centroid (center of mass) of the RoI is calculated (red dot in Figure 2(c)) and we define a close-up area centered at the centroid using the input size of the classification CNN (224x224 or 299x299 pixels). It is the most "zoomed in" patch at a minimum resolution, called the close-up patch (smaller red square in Figure 2(c)). Finally, we extract patches with various scales by placing multiple squares evenly between the close-up patch and the bounding box of the RoI (i.e. the largest square for patch extraction). For example, if n multi-scale patches are extracted, $n-2$ squares are evenly placed between the close-up patch and the bounding box. All extracted square areas are resized to fit the first layer of the CNN, resulting in a set of multi-scale patches used to train and test the classification model. Since each patch covers a square region in the image, the extracted areas are not stretched in any way by the resizing. In this guided multi-scale patch extraction process, the size of the close-up patch and the number of patches to be extracted can be customized depending on the resolution of the input image to the classification CNN.

Figure 2 illustrates the guided multi-scale patch extraction process described above. The binary mask (Figure 2(b)) generated by the segmentation process defines the RoI, which is used to guide the extraction of representative patches in various scales. Figure 2(d) shows the resulting patches extracted for 10 different scales. The MIT Scene Parsing works very well to indicate where the plant is located in a natural scene. Therefore, these extracted multi-scale patches provide well-represented samples for the fine-grained categorization of plants, discarding noisy backgrounds and focusing on highly informative regions of the image. This guided approach is initially set to extract 10 representative multi-scale patches but it can be programmed to extract more if necessary. Close-up patches around the centroids are extracted with the minimum resolution (or maximum zoom-in) and do not need to be resized. They are extracted with an area of 224x224 pixels for residual networks [6, 21] and 299x299 pixels for inception module networks [3, 22, 24] as recommended in the literature. As described above, all other extracted patches are resized to match these common sizes according to the input layer of the classification CNN. Furthermore, the guided multi-scale process allows the extraction of patches of arbitrary size in various scales, so it can be used by most of the CNN models. After this preprocessing stage of segmentation and multi-scale sample collection, extracted patches are now ready to be fed into the CNNs. The detailed steps of the proposed guided multi-scale method are shown in Algorithm 1.

4. Experimental Analysis

We implemented the proposed guided multi-scale approach using *Python 3.6* and *Keras 2.2.4 API*. Our testbed uses the *Ubuntu 16.04* operating system and a *NVIDIA GeForce GTX 1080 Ti* GPU to train the CNNs. Two datasets with annotated plant images are used for these experiments: the BJFU100 [21] and the UHManoa100 [10, 11]. Each of them contains 100 different plant species from the Beijing

Figure 2. (a) Input image (*Koelreuteria formosana*), (b) Mask produced by the MIT Scene Parsing [30], (c) Bounding box of the RoI (red square enclosing green), centroid (red dot), and close-up (red square around the centroid), and (d) Multi-scale extracted patches.

Algorithm 1: PATCHEXTRACTION(I,m,n,p)

Input: Image I, Mask m, # of patches n, Patch size p
Output: A set $\mathcal{M}$ of multi-scale patches each of size p

$r_{largest} \leftarrow$ Find largest RoI from m;
$mbr \leftarrow$ Find min. square bounding box from $r_{largest}$;
$c \leftarrow$ Calculate centroid of $r_{largest}$;
$q \leftarrow$ Find coordinates of square of size p around c;
$\delta \leftarrow [Area(mbr) - Area(q)] \div (n-1)$;
$\mathcal{M} \leftarrow \emptyset$;
$k \leftarrow q$ `// square coordinates`;
for $j \leftarrow 1$ **to** n **do**
  $i \leftarrow$ Crop I using k;
  $patch \leftarrow$ Resize crop image i to size p;
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{patch\}$;
  $k \leftarrow$ Increase the size of k by δ;

Forestry University campus (BJFU100) and the University of Hawai'i at Manoa campus (UHManoa100). All images used to train and test the CNNs are natural images, presenting complex backgrounds, varying illumination, occlusions, shadows, and a rich local covariance structure.

4.1. Training and Evaluating the CNNs

We implement the training process of our CNNs by splitting the datasets into training and testing sets. The training set is balanced with respect to the plant species to prevent the learning from being biased toward specific species. Patches extracted from the training images are randomly divided by selecting 80% for training and 20% for validation to perform cross-validation to assess the performance of each model during training. After each epoch of training, the predictive accuracy of the model is calculated on the validation set. The final trained model is the one with the minimum validation error rate after training is done for a pre-defined number of epochs.

4.2. Testing Process and Metrics

The trained CNN models are evaluated on the test set, which contains at least one image of each plant species that are unseen to the network during training. Multi-scale patches are extracted from each test image to perform the classification for evaluation. The CNN models make predictions for all patches in different scales extracted from a test image. These predictions are then averaged to classify the plant present in the image. This final averaging process

helps the models make a more robust prediction when categorizing plants in natural images since the plant is analyzed multiple times in different scales. For performance evaluation, we use the prediction accuracy, i.e. the percentage of test images correctly categorized versus the total number of images in the test set, as a metric. An image is considered correctly categorized when the top-1 output prediction matches the annotated species of the plant in the image.

4.3. Identifying Species with BJFU100 Dataset

Recently, a collection of annotated high-resolution images called BJFU100 is presented by Sun *et al.* [21]. This is one of the few available datasets containing plants in natural images. The BJFU100 dataset has 100 images per plant species, totalizing 10,000 natural images of ornamental plant species present on the campus of the Beijing Forestry University. Sun *et al.* used this dataset to train and test residual networks (*ResNets*) with different depths.

To explore the capability of *WTPlant*'s preprocessing stage to handle scale changes, we implement the guided multi-scale approach described in Section 3.2 using the BJFU100 dataset. This new approach differs from previous ones [10, 11] in three aspects: *i)* it uses 10 multi-scale patches guided by segmentation, *ii)* it limits the extracted area to the minimum square bounding the segmented plant, and *iii)* it converges to the center of mass or centroid of the plant region instead of the geometric center. In this way, unlike the previous approaches, which extract patches from multiple locations, it focuses exclusively on multi-scale representative patches of the plants extracted from a single location. The resulting system, called *WTPlant*, is used to train and test *ResNets* with 18, 34, and 50 layers similar to the work performed by Sun *et al.*. For comparison with our multi-scale approach, the *ResNets* are also trained with resized training images, as well as using the random and central crop methods.

As done by Sun *et al.*, 80% of the dataset is used for training and the rest for testing. Random crop extracts the same number of patches as used in *WTPlant* and the extracted patch size is 224x224 pixels as indicated in [6]. The *ResNets* are trained for 100 epochs in a two-fold cross-validation process. Table 1 presents the resulting prediction accuracy. It is clear that, regardless of the number of layers of *ResNets*, our proposed multi-scale approach guided by segmentation greatly improves the performance of the networks, significantly outperforming all other approaches. Sun *et al.* also proposed a customized *ResNet* architecture with 26 layers, which resulted in their best accuracy of **91.78%**. However, it is still far below the performance of the *ResNet* trained and tested using the guided multi-scale approach, which yielded the best accuracy of **97.80%**.

Although BJFU100 dataset provides a fair amount of annotated plant images of high quality, images in the dataset are all in the same size (3120x4208 pixels) and show relatively small variations of plant location (mostly centered), lighting condition (captured around the same time of the day) and, in particular, the scale across samples within each species. These aspects make the dataset relatively easy to classify, which explains the high prediction accuracy presented in Table 1. Even so, deeper *ResNets* present underfitting problems requiring more epochs to be fully trained. In addition, these CNNs may not have learned scale-invariant features because of the small intra-species scale variation in the dataset. Therefore, this dataset may not be the best choice for testing a model that aims to recognize plants in natural images showing a wide range of scales.

4.4. Identifying Species with UHManoa100 Dataset

Focusing on capturing variations in scale and appearance of plants in nature, we constructed the UHManoa100 dataset by collecting 4,500 natural images of plants, 45 images per each of the 100 plant species [11]. For each plant species, a set of test images at different scales and of different parts (leaf, flower, bush, and tree) is set aside for performance evaluation, which comprises a test set of around 300 images unseen by trained models. The annotation of the plant species in this dataset indicates the dominant plant present in the image. Different plants may appear in the background or even in front of the dominant plant, but the largest areas of these images are covered by the annotated plant species. Another important characteristic of this dataset is that images have different resolutions (ranging from 300x300 to 6000x4000 pixels) with varying orientations and locations of plants. Therefore, approaches that focus only on the center of the image such as the Central Crop and [7, 8, 16, 25] are unlikely to perform well on this complex dataset.

UHManoa100 dataset was first used in the work of Krause *et al.* [10, 11], where multi-location and multi-scale extractions of representative patches were proposed and used to train and test CNNs. Analyzing the individual patches extracted in *WTPlant* as well as in the work by Krause *et al.*, we noticed that zoomed in areas (close-up patches) are not helping the CNNs. As an alternative, *WTPlant v2.0* is implemented using only the five larger scales and their respective mirrored images balancing the training data for a fair comparison. The set of multi-scale patches extracted in this work is available online[1].

4.4.1 Comparison with Various CNN Models

Using this guided multi-scale approach, state-of-the-art CNN models including those with inception modules are trained to evaluate how helpful the proposed method is for the training of these CNNs. The six CNN models

[1] https://github.com/jonaskrause/UHManoa100

Table 1. Top-1 prediction accuracy of *ResNets* for the BJFU100 dataset.

CNN Model	Resizing	Random Crop	Sun *et al.* [21]	Central Crop	*WTPlant*
ResNet18	74.33%	87.78%	89.27%	90.05%	**97.80%**
ResNet34	71.38%	85.53%	88.28%	83.85%	97.58%
ResNet50	53.73%	73.73%	86.15%	75.25%	95.30%

incorporating residual blocks [6] and inception modules [3, 22, 23, 24] are listed in Table 2. In the experiment, four different ways of data preparation have been compared: *i)* Resizing images to fit to the input layer of CNNs, *ii)* extracting patches based on Random Crop, *iii)* extracting patches from the largest central square area (Central Crop), and *iv)* extracting patches based on our proposed method with selected scales and their mirrored images (*WTPlant v2.0*). Resulting performance for each case is shown in columns 2 through 5 of Table 2.

In this experiment, all CNNs are trained from scratch, initializing weights randomly for learning. The training is conducted for 100 epochs using 20% of the training set for validation. Only top-1 accuracy results are considered, meaning that presented percentages show the ratio of correctly categorized plants among the 100 species for all test images. Results presented in Table 2 show that *WTPlant v2.0* performs the best, indicating that the guided multi-scale approach improves the performance of all tested CNNs significantly. The results also show that inception models perform worse, potentially due to the overfitting problem. In that case, the use of pre-trained models and larger augmented data may improve the accuracy even further. As an initial conclusion, the *WTPlant* preprocessing step generally improves the performance of CNNs, and a guided multi-scale process that extracts more representative patches can further enhance the predictive power compared to other approaches that have been used commonly in the literature.

4.4.2 Fine-Tuning Pre-trained CNN Models

Although there are techniques that can help to deal with the limited data problem to some degree, it may not be enough to provide a good-sized dataset for training deep networks to obtain the best performance. CNN models such as the *Inception-v3* [24], *Inception-ResNet-v2* [22], and the *Xception* [3] have a large number of parameters (up to 54 million), and the lack of training data generally leads to overfitting and poor generalization. Consequently, these CNN models are commonly implemented using pre-trained weights [4]. To fully explore the capability of the six CNN models evaluated in the previous section, a similar experiment has been performed using the UHManoa100 dataset but applying fine-tuning to the pre-trained networks. We use CNN models pre-trained on the ImageNet dataset [18] and fine-tune them by initiating a training process using the learned weights as initial parameter values. In this way, filters learned from a general dataset such as the ImageNet can be adapted to the classification of plant species. Using data prepared the same ways as in the experiment described in the previous section, the pre-trained CNNs are fine-tuned for 50 epochs to classify 100 plant species.

The results presented in Table 3 show that the use of pre-trained models and fine-tuning improves the performance more significantly for CNNs with inception modules than the *ResNets*. This phenomenon is natural since the number of parameters in CNNs with inception modules is almost twice larger than the number of parameters of the *ResNets* trained in this experiment (up to 26 million), therefore the pre-learned weights have more impact on the CNNs with inception modules making them perform well on a relatively small dataset. The consistently superior performance of *WTPlant v2.0* across all CNN models reinforces the hypothesis that the guided multi-scale approach is helpful for training CNNs and even when fine-tuning pre-trained models. Also, the use of pre-trained models dramatically improves the performance of large CNNs, resulting in prediction accuracy of **89.21%** by the *Inception-ResNet-v2* model which is far better than the best performance (**65.11%**) of *ResNet18* trained from scratch.

5. Observations and Discussions

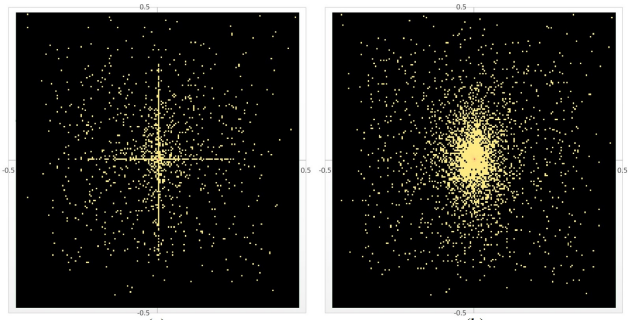
To better understand the complexity of the UHManoa100 dataset, we investigated the distribution of geometric centers of the extracted plant regions used by Krause *et al* [10, 11] and the centroids used in *WTPlant v2.0*. Point locations are estimated by normalizing all the image sizes and calculating the relative position to the center of the image as shown in Figure 3. Figure 3(a) shows the distribution of geometric centers of UHManoa100 dataset used for patch extraction in [10, 11]. It tightly clusters around the center as well as along the major axes, which is reasonable given that the object of interest can often be off from the center horizontally and/or vertically when the picture is taken. The centroids of the segmented areas from UHManoa100 dataset are shown in Figure 3(b). They are also clustered around the center mostly but a fair amount of the points are scattered all over the images, which means that plants in many images appear off-centered in this dataset. These centroids are the reference points for multi-scale patch extraction in the guiding process described in Section 3.2, ensuring that each extracted patch covers the plant area.

Table 2. Top-1 prediction accuracy of CNNs trained from scratch on the UHManoa100 dataset.

CNN Model	Resizing	Random Crop	Central Crop	*WTPlant v2.0*
ResNet18	39.21%	43.89%	44.24%	**65.11%**
ResNet34	40.29%	44.60%	42.81%	59.71%
ResNet50	28.42%	43.53%	39.21%	57.19%
Inception-v3	30.58%	40.65%	30.94%	49.28%
Inception-ResNet-v2	35.25%	59.35%	37.41%	62.23%
Xception	28.78%	39.57%	29.50%	53.24%

Table 3. Top-1 prediction accuracy of CNNs fine-tuned using the UHManoa100 dataset.

CNN Model	Resizing	Random Crop	Central Crop	*WTPlant v2.0*
ResNet18	60.79%	53.60%	60.07%	61.51%
ResNet34	56.83%	51.80%	56.83%	57.91%
ResNet50	53.60%	49.28%	54.68%	56.83%
Inception-v3	71.94%	79.50%	78.42%	85.61%
Inception-ResNet-v2	76.26%	83.81%	79.14%	**89.21%**
Xception	75.90%	82.37%	83.09%	87.05%



(a) (b)

Figure 3. Heatmap of point locations, relative to the image center. (a) Geometric centers of UHManoa100 dataset, and (b) centroids of UHManoa100 dataset.

Regarding the proposed guided multi-scale method, the *WTPlant* system demonstrates consistent performance improvement across almost all experiments conducted for the fine-grained categorization of plants in natural images. As shown in Table 1, over 97% accuracy is achieved when the proposed approach is applied to classify plants in the BJFU100 dataset. For the UHManoa100 dataset, *WTPlant v2.0* shows over 89% accuracy when the pre-trained and fine-tuned *Inception-ResNet-v2* model is used for classification. These results show that *WTPlant* may help addressing image-based phenotyping of plants in wild by observing properties of each species appearing at various scales in natural images.

6. Conclusion

In this paper, we present a multi-scale approach guided by segmentation for the training of CNNs designed for fine-grained plant classification tasks. Building on previous works on addressing scale issues for object classification, our approach uses a CNN [30] to parse a natural scene and segment the plant in the image from the complex background. The resulting segmentation information drives the extraction of representative patches at various scales. These patches are then fed into state-of-the-art CNNs allowing them to learn features that can address the large scale variation occurring in natural images of plant species. The proposed approach extends the previous work [10, 11] and uses segmentation information more efficiently, extracting patches that are limited to the minimum square bounding the segmented plant and using the center of mass to guide the extraction of multi-scale patches. As a direct result, extracting multi-scale samples using better guidance from the centroids help to select more scale representative patches.

A series of comparative experiments are conducted with two datasets of plants in natural images testing several CNN models. Our experimental analysis shows that (1) the proposed approach is effective in dealing with large scale variations within the natural images, (2) it is also robust to a varying location of the plant in the scene, and (3) it consistently improves classification accuracy of CNN models compared to other approaches.

We also notice that the scales of patches used for training can have a great influence on the performance. Hence, estimating an appropriate set of scales to extract patches from each image would be an important problem to investigate. In future work, we plan to explore whether the use of fractal dimensions can help us find this adequate range of scales for the extraction of more representative patches.

References

[1] Anelia Angelova and Shenghuo Zhu. Efficient Object Detection and Segmentation for Fine-Grained Recognition. In *CVPR*, pages 811–818. IEEE Computer Society, 2013. 3

[2] Steve Branson, Grant Van Horn, Serge J Belongie, and Pietro Perona. Bird Species Categorization Using Pose Normalized

Deep Convolutional Nets. *CoRR*, abs/1406.2952, 2014. 2

[3] Franois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. *CoRR*, abs/1610.02357, 2016. 4, 7

[4] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge J Belongie. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. In *CVPR*, pages 4109–4118. IEEE Computer Society, 2018. 7

[5] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge J Belongie. Fine-grained Categorization and Dataset Bootstrapping using Deep Metric Learning with Humans in the Loop. *CoRR*, abs/1512.05227, 2015. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015. 4, 6, 7

[7] Jing Hu, Zhibo Chen, Meng Yang, Rongguo Zhang, and Yaji Cui. A Multiscale Fusion Convolutional Neural Network for Plant Leaf Recognition. *IEEE Signal Processing Letters*, 25(6):853–857, 2018. 2, 6

[8] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2, 6

[9] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5546–5555, 6 2015. 3

[10] Jonas Krause, Gavin Sugita, Kyungim Baek, and Lipyeow Lim. What's That Plant? WTPlant is a Deep Learning System to Identify Plants in Natural Images. In *BMVC*, page 330. BMVA Press, 2018. 3, 4, 6, 7, 8

[11] Jonas Krause, Gavin Sugita, Kyungim Baek, and Lipyeow Lim. WTPlant (What's That Plant?): A Deep Learning System for Identifying Plants in Natural Images. In *ICMR*, pages 517–520. ACM, 2018. 3, 4, 6, 7, 8

[12] Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, and Joo V. B. Soares. Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV (2)*, pages 502–516, 2012. 1, 2

[13] Sue Han Lee, Chee Seng Chan, Simon Mayo, and Paolo Remagnino. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, 71:1–13, 2017. 2

[14] Sue Han Lee, Chee Seng Chan, Paul Wilkin, and Paolo Remagnino. Deep-plant: Plant identification with convolutional neural networks. *2015 IEEE International Conference on Image Processing (ICIP)*, pages 452–456, 2015. 1

[15] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition. *CoRR*, abs/1603.06765, 2016. 2

[16] Jeff Mo, Eibe Frank, and Vetrova Vetrova. Large-scale automatic species identification. In D. Alahakoon W. Peng and X. Li, editors, *Proceedings of 30th Australasian Joint Conference on Advances in Artificial Intelligence*, page 301312. Springer, 2017. 2, 6

[17] Michael P. Pound, Jonathan A. Atkinson, Darren M. Wells, Tony P. Pridmore, and Andrew P. French. Deep learning for multi-task plant phenotyping. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2055–2063, Oct 2017. 1

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7

[19] Asma R. Sfar, Nozha Boujemaa, and Donald Geman. Vantage Feature Frames for Fine-Grained Categorization. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 835–842, 2013. 3

[20] Wen Shi, Fanman Meng, and Qingbo Wu. Segmentation quality evaluation based on multi-scale convolutional neural networks. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 12 2017. 3

[21] Yu Sun, Yuan Liu, Wang Guan, and Haiyan Zhang. Deep Learning for Plant Identification in Natural Environment. *Computational Intelligence and Neuroscience*, 2017(7361042):6 pages, 2017. 4, 6, 7

[22] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *CoRR*, abs/1602.07261, 2016. 4, 7

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *CoRR*, abs/1409.4842, 2014. 7

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CoRR*, abs/1512.00567, 2015. 4, 7

[25] Ahmad P. Tafti, Fereshteh S. Bashiri, Eric LaRose, and Peggy Peissig. Diagnostic Classification of Lung CT Images Using Deep 3D Multi-Scale Convolutional Neural Network. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 412–414, 6 2018. 6

[26] Catherine Wah, Steve Branson, Pietro Perona, and Serge Belongie. Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*, pages 2524–2531, 11 2011. 2

[27] Jana Wäldchen and Patrick Mäder. Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*, 25(2):507–543, Apr 2018. 1

[28] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 2, 3

[29] Songfan Yang and Deva Ramanan. Multi-scale recognition with DAG-CNNs. *CoRR*, abs/1505.05232, 2015. 2

[30] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 5, 8