Fine-Grained Visual Classification of Plant Species In The Wild: Object Detection as A Reinforced Means of Attention

Matthew R. Keaton Ram J. Zaveri Meghana Kovur Cole Henderson Donald A. Adjeroh Gianfranco Doretto*
West Virginia University, Morgantown, WV 26506
{mrkeaton, rz0012, mk0174, cwh0015, daadjeroh, gidoretto}@mix.wvu.edu

Abstract

Plant species identification in the wild is a difficult problem in part due to the high variability of the input data, but also because of complications induced by the long-tail effects of the datasets distribution. Inspired by the most recent fine-grained visual classification approaches which are based on attention to mitigate the effects of data variability, we explore the idea of using object detection as a form of attention. We introduce a bottom-up approach based on detecting plant organs and fusing the predictions of a variable number of organ-based species classifiers. We also curate a new dataset with a long-tail distribution for evaluating plant organ detection and organ-based species identification, which is publicly available[1].

1. Introduction

Automated plant image analysis in its many forms has become an increasingly relevant research topic, impacting several related areas of research and application [27]. Herbarium specimens have proven useful tools for phenological research [33, 3, 4], while detection-based techniques are being deployed in precision agriculture [18, 12]. Furthermore, the collection of large crowdsourced datasets generated by "citizen scientists" [10] is expanding available research opportunities since image data is acquired "in the wild," where factors including the quality, viewpoint, and illumination of images as well as the shape and scale of the plant subjects are fully unconstrained. While simultaneously providing larger pools of data and generating a more faithful representation of real-world scenarios, this type of data poses additional challenges in contrast with former applications where different degrees of control could be imposed on the way data is collected.

Settings in the wild increase the variability of the data, leading to a larger intra-class variance. The most recent approaches to fine-grained visual classification cope with that variability by leveraging an attention mechanism, whereby they focus on a subset of the available feature space with the hope of decreasing the sources of nuisance factors causing variation. Several successful approaches exist, including the use of specialized network layers [17, 28] and part-based attention [38, 23, 37, 39]. On the other hand, when used in applications in the wild, where data often takes the form of a long-tailed distribution, these types of approaches tend to lose their effectiveness much more quickly as the number of samples per class decreases [31].

The contribution of this work is twofold. First, we introduce a bottom-up approach to plant species identification in the wild that uses object detection as a means of attention to localize plant organs. This allows us to decrease the effects of data variability by basing identification on important regions of interest while ruling out unwanted background noise. At the same time, the supervised nature of the detection task (as opposed to unsupervised attention) allows for better mitigation of the long-tail effects still induced by settings in the wild.

The second contribution is the introduction of a long-tail dataset that allows for training a plant organ detector as well as plant organ-based species classifiers. To the best of our knowledge this is the first dataset of its kind, and it is publicly available. In the experiment section we evaluate the proposed approach on this new dataset.

2. Related work

Plant species identification. Classically, plant image analysis was constrained to species identification using leaves or flowers as image subjects, often consisting of a single leaf or cluster of leaves laid across a white background [34, 21, 26]. Other organs of interest have been analyzed, with best results stemming from the identification of flowers [1, 25]. More recently, plant species identification has spread to more difficult and sizable datasets, increasing the number of available tasks in the field. Com-

*This material is based upon work supported in part by the National Science Foundation under Grants No. OAC-1761792, IIS-1657179.

[1] https://github.com/wvuvl/DARMA

petitions, such as those hosted by ImageCLEF [7, 8, 9] and Google AI [29, 20], produce valuable datasets for various analysis tasks including species identification, expanding to broader open challenges such as fine-grained visual classification "in the wild." Furthermore, crowdsourced datasets derived from images taken by "citizen scientists," such as iNaturalist [30] and LeafSnap [15], have gained popularity as their sheer size provides an advantage for deep neural networks.

Object detection. Object detection has mostly had its use in plant analysis constrained to invasive species detection [6] and phenotyping applications [2, 19]. This means related datasets either focus on detecting entire plants or individual organs, such as leaves, from a singular plant species. Recently, a few small-scale plant organ detection approaches and datasets using more than one species have been developed, each using at most a few hundred herbarium sheets with the intended use of phenological information gathering [36, 22, 32]. In the context of species identification, these approaches are unable to collect data on present biodiversity and the geographical distribution of different plant species, an advantage that crowdsourcing initiatives possess.

3. Proposed approach

Our species identification approach contains three main components. First, an object detector identifies and localizes *plant organs*, including leaves, flowers, fruit, stems, and regions with a high volume of leaves, termed "high-density leaves" (HDL). These regions of interest (ROIs) are then individually passed into an organ-based species classifier. Finally, the predictions for the given image are then aggregated by an information fusion step, which generates the final species prediction. See Figure 1.

Using an object detector to localize plant organs is advantageous for coping with the high variability of the data because it will select the information-rich regions to be individually classified while rejecting a large portion of the inherent background noise. Moreover, the downstream organ-based species classification and fusion steps will be able to handle a variable number of organs and to constructively aggregate predictions while being robust to false detections, as explained further below.

Organ detection. Given an input image I, we obtain n regions of interest for our downstream classifier by deploying an object detector, trained to localize and predict the classes of various plant organs. For the i-th ROI, the detector predicts the organ class o_i, where $o_i \in \{\textit{leaf, flower, fruit, stem, HDL}\}$. For this, we utilize the feature pyramid network-based Faster-RCNN object detector [24], built on the ResNet-101 backbone, which is shown to attain strong baseline results even with a large variability of object sizes, as discussed in [16]. Our model, pretrained

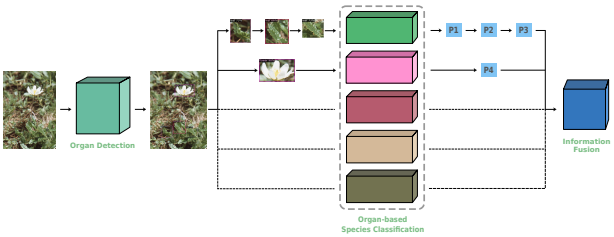

Figure 1: **Species identification in-the-wild overview.** The input image is processed by an organ detector producing the ROIs. Each ROI is fed into the corresponding organ-based species classifier. Then, the per-organ species classification probabilities are aggregated by an information fusion step into the final species prediction.

on the COCO detection dataset, is built using the Detectron2 library [35]. Using our dataset described in Section 4, we trained the model over 100k iterations using an SGD optimizer with a base learning rate of 5×10^{-5} and momentum of 0.9. Additionally, a non-maximum suppression threshold of 0.1 is used during both training and testing. Due to memory constraints, a "mini-batch" of 1 is used; training takes approximately 10 hours when run on a single NVIDIA GTX 1660 GPU.

Organ based species classification. Given the i-th ROI labeled by the organ detector as o_i, we compute the probability of the ROI to depict the species s, $p(s|o_i)$. We do so with an organ-based species classifier implemented with a convolutional neural network ending with a softmax layer. We use a ResNet-18 [11] as the backbone network, which we train with a cross-entropy loss. Each ROI was resized to 224×224, which is the average ROI size of the dataset. Additionally, each classifier was fine-tuned from a model pre-trained on ImageNet, for just enough iterations to observe the validation loss reaching the plateau. We used the default Adam optimizer, and learning rate of 1×10^{-4}, with a minibatch of size 32.

Information fusion. The species prediction entails finding the species s that maximizes the probability $p(s)$. We can express $p(s)$ as $p(s) = \sum_i p(s|o_i)p(o_i)$, where $p(o_i)$ is the prior on the organ o_i, and each of the organs in the input image are assumed to be independent. In lack of prior knowledge, the natural choice is to assume a uniform prior, which means that $p(s)$ becomes the average of the probabilities $p(s|o_i)$. This is equivalent to the so-called *sum rule* in information fusion [14].

Another way to pose the species prediction problem is to find the species s that maximizes the posterior $p(s|o_1, \cdots, o_n)$. Under the assumption that every species is equally likely, and that the organs are independent conditioned on the species, this is equivalent to finding s that maximizes $\prod_i p(o_i|s)$, which is the naive Bayesian fusion. If we further assume a uniform prior on the organs (like above), this is equivalent to finding s that maximizes

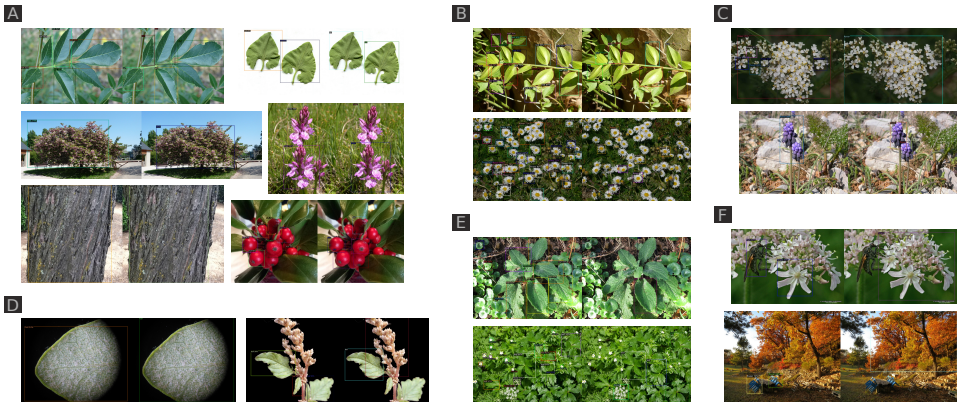

Figure 2: **Sample annotations from the DARMA dataset.** In (a), clockwise from the top-left, are samples of leaves, "scan-like" leaves that make up approximately an eighth of the dataset, flowers, fruit, stems and HDL. Highlighted in (b) through (f) are various challenges that potentially impact organ detection or classification performance. (b) and (c) depict artificial losses in average precision due to only a subset of leaves being annotated in (b) and random grouping/non-grouping of flowers due to annotator preferences in (c). (d)-(f) show various difficulties from the dataset and approach that affect classification accuracy. (d) shows random image effects added to photographs by submitters of the images, (e) provides examples where non-target species are included in an image and then extracted by the organ detector, and (f) shows that vibrant or large non-plant items found in images can confuse the organ detector.

Train	Test	Validation
62959	17995	9016

Table 1: **Statistics on the number of samples per split.**

$\prod_i p(s|o_i)$. This is known as the *product rule* in information fusion [5].

Finally, when $p(s|o_i)$ is approximated by a one-hot vector with the 1 corresponding to the species that maximizes $p(s|o_i)$ the sum rule becomes the *voting rule* [14]. In the experiments we compared both the sum, the product, and the voting rule.

4. DARMA dataset

In order to demonstrate the effectiveness of our approach, we first cultivated a benchmark dataset due to the lack of one for organ detection in the wild, which we named DARMA (i.e., short for Detection as A Reinforced Means of Attention). We derive our images from the Pl@ntView dataset used in the PlantCLEF 2015 challenge [13]. The images were gathered as part of a citizen scientist initiative, meaning that a large portion of them were taken by amateur photographers (see Figure 2).

Our dataset separates itself from the PlantCLEF 2015 dataset with the addition of bounding box annotations for four different plant organs - leaves, flowers, fruit, and stems - and regions we call high-density leaves (HDL). Additionally, the PlantCLEF 2015 challenge was centered around multi-observation queries, where one to five images are provided as part of a singular instance for prediction. Instead, we split these queries and always provide only one image for each prediction.

For each species, the images were split into 70% for training, 10% for validation, and 20% for testing. When

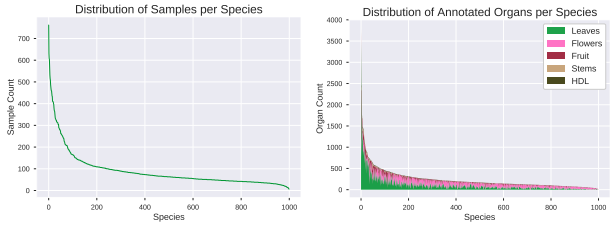

Figure 3: **Distribution of samples (left) and annotated organs (right) across each species.** Species are sorted in descending order of sample and total annotated organ count; both the sample number and organ count follow a long-tailed distribution.

Organs	Average	Standard Deviation
Leaf	184×199	145×172
Flower	210×220	188×190
Fruit	141×151	142×146
Stem	516×610	239×220
HDL	634×565	167×155

Table 2: **Statistics on the scale of bounding boxes per organ.**

fewer than 10 images were included for a species, at least one sample was placed into the validation and test set. Annotations were generated following a specific protocol. For leaves, only those that were normal to the image plane with less than 25% of their surface obstructed were to be annotated. Flowers following the same guideline for obstruction were annotated as well, including flower buds. The fruit category included fruits as well as pine cones and seeds. Stems included both upright (mostly vertical) plant stems and tree bark. Finally, we introduce the HDL category. These are regions where leaves are difficult to differentiate due to the distance the photo was taken from, and they include information in the form of a texture rather than a shape. Tables 1, 2, 3, and 4 illustrate the basic statistics of the dataset.

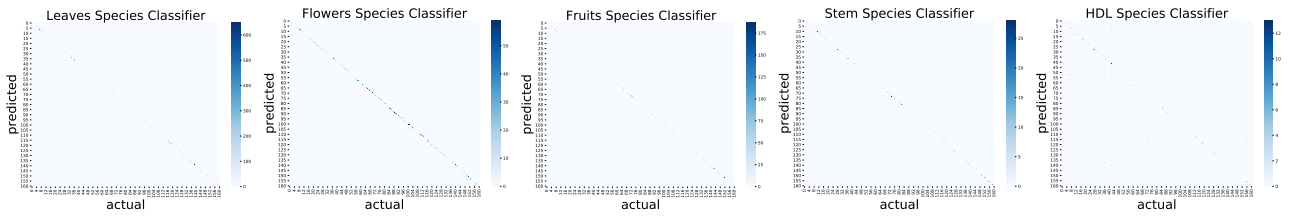

Figure 4: **Organ-based species classification.** Confusion matrices for each of the organ-based classifiers.

Mean	Standard Deviation	Minimum	Maximum
90.0	85.6	6	762

Table 3: **Statistics on the number of samples per species.**

Organ	Mean	Standard Deviation	Maximum
Leaf	119.0	219.7	3568
Flower	82.7	67.0	606
Fruit	30.7	66.1	1122
Stem	4.4	14.6	153
HDL	5.2	7.8	94

Table 4: **Statistics on the number of organs per species.** Minimum values are 0 for all organs.

Besides the images being taken in the wild, the dataset presents additional challenges. Figure 3 shows that the distribution of both images and organ annotations across species exhibit long tails where a large proportion of samples and organs is possessed by a few classes. In terms of each image, there are often other plants that appear within the frame, making proper classification more difficult. In addition, not all organs are annotated within each image - annotators were instructed to include as many as three leaves when at least that many were present. Although this should have a small impact on the efficacy of the object detector, AP results may be artificially lowered because of this, and efforts to improve annotations will be made in the future. Figure 2 portrays several samples with annotations and associated challenges.

5. Experiments

Organ detection. Our initial results are outlined in Tables 5 and 6. The organ detector tends to perform lower on our dataset than on other object detection datasets; this is due to a number of reasons, as discussed in Section 4. However, it is important to note that the overall approach is somewhat robust with respect to the calculated performance of the organ detector. This is because a few false positives are likely filtered out by the fusion stage of the final prediction. Moreover, several unannotated organs are also picked up by the detector, reinforcing a correct final species prediction while decreasing the average precision of the detector.

Organ-based species classification. We down-selected 161 species out of the original 1000 to retain those that had ROI data for every organ, and that had at least 130 leaf samples. Table 7 reports the accuracy of each of the organ-based classifiers. There is a significant accuracy drop for stem and HDL, which was to be expected given the much lower availability of data for these organs. In addition, Figure 4 shows the confusion matrices.

AP	AP_{50}	AP_{75}
42.9	66.6	46.2

Table 5: **Organ detection average precision.** We use the COCO AP evaluation metrics. The first AP is averaged across threshold values ranging from 0.50 to 0.95 with increments of 0.05.

Leaf	Flower	Fruit	Stem	HDL
40.9	36.0	25.7	74.0	37.6

Table 6: **Average precision for each organ.** AP is calculated in the same manner as in Table 5.

Leaf	Flower	Fruit	Stem	HDL
68.24	75.24	63.39	58.24	34.21

Table 7: **Organ-based classification accuracy.**

Rule	Sum	Product	Voting	ResNet-18
Accuracy	79.36	78.16	77.65	69.65

Table 8: **Species identification accuracy.** Comparison among three fusion techniques: sum rule, product rule, and voting rule.

Fusion-based species identification. Table 8 shows the species identification accuracy for the three fusion approaches, where the predictions were made on a per-image basis. The sum rule appears to outperform the others. In particular, we observe that the fusion accuracy outperforms the organ-based accuracies, proving the efficacy of combining multiple observations of discriminatory features from a single image. Finally, the fusion accuracy outperforms also the baseline of 69.65 obtained by training ResNet-18 on the whole input image for 30 epochs.

6. Conclusions

We introduced an approach for plant species identification in the wild and a new dataset for evaluating organ detection and organ-based species identification. We have shown that the dataset is long-tail distributed. Based on our initial evaluation, the approach exhibits robustness against false detections by fusing multiple species predictions, achieving accuracy values comparable with top-performing entries from related challenges in the wild.

References

[1] S. Arwatchananukul, K. Kirimasthong, and N. Aunsri. A New Paphiopedilum Orchid Database and Its Recognition Using Convolutional Neural Network. *Wireless Personal Communications*, 115(4):3275–3289, Dec. 2020. 1

[2] M. Buzzy, V. Thesma, M. Davoodi, and J. M. Velni. Real-Time Plant Leaf Counting Using Deep Object Detection Networks. *Sensors*, 20(23):6896, Jan. 2020. 2

[3] J. Carranza-Rojas, H. Goëau, P. Bonnet, E. Mata-Montero, and A. Joly. Going deeper in the automated identification of herbarium specimens. *BMC Evol. Biol.*, 17(1):181, Aug. 2017. 1

[4] S. D. Choudhury, A. Samal, and T. Awada. Leveraging image analysis for High-Throughput plant phenotyping. *Front. Plant Sci.*, 10:508, Apr. 2019. 1

[5] D. Dubois, W. Liu, J. Ma, and H. Prade. The basic principles of uncertain information fusion. an organised review of merging rules in different representation frameworks. *Inf. Fusion*, 32:12–39, Nov. 2016. 3

[6] H. Goëau, P. Bonnet, and A. Joly. Plant identification in an open-world (LifeCLEF 2016). In *CLEF: Conference and Labs of the Evaluation Forum*, pages 428–439, 2016. 2

[7] H. Goëau, P. Bonnet, and A. Joly. Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). In *CLEF: Conference and Labs of the Evaluation Forum*. hal.archives-ouvertes.fr, 2017. 2

[8] H. Goëau, P. Bonnet, and A. Joly. Overview of lifeclef plant identification task 2019: diving into data deficient tropical countries. In *CLEF 2019-Conference and Labs of the Evaluation Forum*, volume 2380, pages 1–13, 2019. 2

[9] H. Goëau, P. Bonnet, and A. Joly. Overview of lifeclef plant identification task 2020. In *CLEF 2020-Conference and labs of the Evaluation Forum*, 2020. 2

[10] T. Gura. Citizen science: amateur experts. *Nature*, 496(7444):259–261, Apr. 2013. 1

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[12] Y. Jiang, C. Li, A. H. Paterson, and J. S. Robertson. DeepSeedling: deep convolutional network and kalman filter for plant seedling detection and counting in the field. *Plant Methods*, 15:141, Nov. 2019. 1

[13] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, R. Planqué, A. Rauber, S. Palazzo, B. Fisher, and H. Müller. LifeCLEF 2015: Multimedia life species identification challenges. In *Lecture Notes in Computer Science*, pages 462–483. Cham, 2015. 3

[14] J. Kittler and F. M. Alkoot. Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):110–115, 2003. 2, 3

[15] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C Lopez, and J. V. B. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *The 12th European Conference on Computer Vision (ECCV)*, October 2012. 2

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 2

[17] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, J. Li, J. Yang, and S.-N. Lim. Cross-X Learning for Fine-Grained Visual Categorization. *arXiv:1909.04412 [cs]*, Sept. 2019. arXiv: 1909.04412. 1

[18] X. Mai, H. Zhang, X. Jia, and M. Q.-H. Meng. Faster R-CNN with classifier fusion for automatic detection of small fruits. *IEEE Trans. Autom. Sci. Eng.*, 17(3):1555–1569, July 2020. 1

[19] C. A. Manacorda and S. Asurmendi. Arabidopsis phenotyping through geometric morphometrics. *Gigascience*, 7(7), July 2018. 2

[20] E. Mwebaze, T. Gebru, A. Frome, S. Nsumba, and J. Tusubira. iCassava 2019 Fine-Grained Visual Categorization Challenge. *arXiv:1908.02900 [cs]*, Dec. 2019. arXiv: 1908.02900. 2

[21] P. Novotný and T. Suk. Leaf recognition of woody species in Central Europe. *Biosystems Engineering*, 115(4):444–452, Aug. 2013. 1

[22] T. Ott, C. Palm, R. Vogt, and C. Oberprieler. GinJinn: An object-detection pipeline for automated feature extraction from herbarium specimens. *Appl. Plant Sci.*, 8(6):e11351, June 2020. 2

[23] Y. Peng, X. He, and J. Zhao. Object-Part Attention Model for Fine-grained Image Classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, Mar. 2018. arXiv: 1704.01740. 1

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017. 2

[25] A. R. Sfar, N. Boujemaa, and D. Geman. Vantage Feature Frames for Fine-Grained Categorization. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 835–842, Portland, OR, USA, June 2013. IEEE. 1

[26] O. Söderkvist. Computer vision classification of leaves from swedish trees, 2001. 1

[27] P. S. Soltis, G. Nelson, A. Zare, and E. K. Meineke. Plants meet machines: Prospects in machine learning for plant biology. *Appl. Plant Sci.*, 8(6), June 2020. 1

[28] M. Sun, Y. Yuan, F. Zhou, and E. Ding. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. In *Computer Vision – ECCV 2018*, volume 11220, pages 834–850. Cham, 2018. 1

[29] K. C. Tan, Y. Liu, B. Ambrose, M. Tulig, and S. Belongie. The Herbarium Challenge 2019 Dataset. *arXiv:1906.05372 [cs, eess]*, June 2019. arXiv: 1906.05372. 2

[30] G. Van Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The INaturalist Species Classification and Detection Dataset. pages 8769–8778, 2018. 2

[31] G. Van Horn and P. Perona. The devil is in the tails: Fine-grained classification in the wild. Sept. 2017. 1

[32] W. N. Weaver, J. Ng, and R. G. Laport. LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Appl. Plant Sci.*, 8(6):e11367, June 2020. 2

[33] C. G. Willis, E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, G. Nelson, S. J. Mazer, N. L. Rossington, T. H. Sparks, and P. S. Soltis. Old plants, new tricks: Phenological research using herbarium specimens. *Trends Ecol. Evol.*, 32(7):531–546, July 2017. 1

[34] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang. A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. *arXiv:0707.4289 [cs]*, July 2007. arXiv: 0707.4289. 1

[35] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[36] S. Younis, M. Schmidt, C. Weiland, S. Dressler, B. Seeger, and T. Hickler. Detection and annotation of plant organs from digitised herbarium scans using deep learning. *Biodivers Data J*, 8:e57090, Dec. 2020. 2

[37] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. In *Computer Vision – ECCV 2018*, volume 11220, pages 595–610. Cham, 2018. 1

[38] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5219–5227, Oct. 2017. ISSN: 2380-7504. 1

[39] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5007–5016, Long Beach, CA, USA, June 2019. IEEE. 1