$S^2CL - LeafNet$: Recognizing Leaf Images Like Human Botanists

CONG ZOU and RUI WANG, State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences; School of Cyber Security, University of Chinese Academy of Sciences, China
CHENG JIN, School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, China
SANYI ZHANG, State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, China
XIN WANG, Department of Computer Science and Technology, Tsinghua University, China

Automatically classifying plant leaves is a challenging fine-grained classification task because of the diversity in leaf morphology, including size, texture, shape, and venation. Although powerful deep learning-based methods have achieved great improvement in leaf classification, these methods still require a large number of well-labeled samples for supervised training, which is difficult to get. In contrast, relying on the specific coarse-to-fine classification strategy, human botanists only require a small number of samples for accurate leaf recognition. Inspired by the classification strategy of human botanists, we propose a novel $S^2CL - LeafNet$, which exploits multi-granularity clues with a hierarchical attention mechanism and boosts the learning ability with the supervised sampling contrastive learning with limited training samples to classify plant leaves as human botanists do. Specifically, to fully explore and exploit the subtle details of the leaves, a novel sampling transformation mechanism is combined with the supervised contrastive learning to enhance the network's perception of details by amplifying the discriminative regions with a weighted sampling of different regions. Furthermore, we construct the hierarchical attention mechanism to produce attention maps of different granularity, which helps to discover details in leaves that are important for classification. Experiments are conducted on the open-access leaf datasets, including Flavia, Swedish, and LeafSnap, which prove the effectiveness of the proposed $S^2CL - LeafNet$.

CCS Concepts: • **Computing methodologies** → **Object recognition**; *Image representations*; • **Computer systems organization** → *Neural networks;*

Additional Key Words and Phrases: Fine-grained image classification, few-shot learning, leaf recognition

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. U20B2066 and 62176253.

Authors' addresses: C. Zou and R. Wang (corresponding author), State Key Laboratory of Information Security, Institute of Information Engineering, CAS; School of Cyber Security, University of Chinese Academy of Sciences, 19 Shucun Road, Haidian District, Beijing, Beijing, China; emails: {zoucong, wangrui}@iie.ac.cn; C. Jin, School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, No. 2005, Songhu Road, Yangpu District, Shanghai, Shanghai, China; email: jc@fudan.edu.cn; S. Zhang, State Key Laboratory of Information Security, Institute of Information Engineering, CAS, 19 Shucun Road, Haidian District, Beijing, Beijing, China; email: zhangsanyi@iie.ac.cn; X. Wang, Department of Computer Science and Technology, Tsinghua University, Qinghuayuan, Haidian District, Beijing, Beijing, China; email: xin_wang@tsinghua.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1551-6857/2023/09-ART30 $15.00
https://doi.org/10.1145/3615659

ACM Reference format:
Cong Zou, Rui Wang, Cheng Jin, Sanyi Zhang, and Xin Wang. 2023. $S^2CL - LeafNet$: Recognizing Leaf Images Like Human Botanists. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 1, Article 30 (September 2023), 20 pages.
https://doi.org/10.1145/3615659

1 INTRODUCTION

Given the image of the entire plant or just parts of it, such as flowers, fruits, leaves, and stems [16, 42], automatic plant classification refers to identifying plant images into botanical species with algorithms. Compared with organs such as flowers or fruits, which only appear at some specific stages of maturity, the leaf images of the plant are easier to collect, so it is more convenient to identify categories of plants with their leaves. In this article, we focus on classifying leaf images for plant identification. There are three challenges for leaf image recognition. First, closely related plants may be very similar in appearance, since they have the same ancestor in the evolution tree, as shown in Figure 1. Second, differences in the poses of objects and illumination, especially in the maturity level of leaves, can lead to large variances in images of the same species. Finally, accurate labeling of plant images needs expert domain knowledge, which makes it difficult to obtain sufficient well-labeled training samples for data-driven deep learning methods, greatly limiting its performance and practical use.

In the field of leaf image recognition, early works generally utilize hand-crafted features, such as textural features and shape-defining features, combined with SVM or nearest neighbor classifier for plant identification [10, 29, 39, 41]. Most of the recent works are based on **Convolutional Neural Networks (CNNs)**, which outperform hand-crafted based methods by a large margin. Such as approaches based on CNNs [3, 21, 47, 50], which utilize the data augmentation and stacked convolutional layers on the leaf image recognition.

Furthermore, leaf recognition is a typical fine-grained image classification task, which has attracted a lot of attention due to its wide application in practice. Both leaf recognition and fine-grained image classification have the same goal, intending to classify subcategories, such as birds, cars, and so on. General methods in the field of fine-grained classification can be grouped in two ways. One is discriminative feature learning [23, 34, 36, 52], which usually refers to enhancing the representation capability of features by end-to-end training. Reference [36] constructs a bilinear structure to extract the pairwise features by two parallel CNNs and use the pooled outer product of features to represent an image. After that, many methods [7, 12, 18, 30, 62] have further developed in this way and made great progress. The other is discriminative region discovery, which often contains two subnetworks: (1) localization subnetwork, localizing discriminative regions, and (2) classification subnetwork, combining those regions to produce the final prediction. Some previous methods [31, 33, 35, 58, 59] utilize additional part annotations to train their localization subnetworks. Recent methods [15, 19, 40, 48, 55, 61] use the attention capability of classification networks to localize discriminative regions with only class labels, removing the requirement of expensive part annotations. However, these fine-grained classification methods rely on an amount of well-labeled training data, which is expensive to get in the leaf recognition task.

The ability of humans to recognize objects has a specific property, i.e., even babies do not require many training samples to achieve high recognition accuracy. It is because humans can find the discriminative regions of objects and then abstract them into specific concepts for recognition. This is especially the case for experienced botanists, who have a specialized coarse-to-fine classification strategy to accurately distinguish leaves [4, 11, 17]. First, they analyze leaf appearance to get features of integral leaf shape, including edges and overall shapes. With this step, human botanists

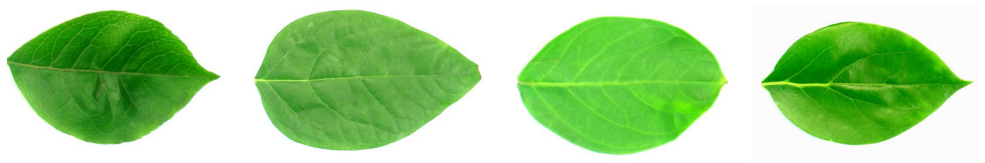

Fig. 1. Leaf images of the four plants, which are quite similar in appearance. From left to right, they are big-fruited holly, wintersweet, crape myrtle, and camphortree, respectively.

can classify the leaves into general categories, such as coniferous and broad-leaved, to make primary judgments. Second, botanists turn to analyze the texture and the main veins of leaves to get more detailed features for more fine-grained classification. Third, they put leaves under a magnifying glass or even a microscope to find the most detailed discriminative information such as minor veins and the serrated shape of the edge. Finally, they integrate all features obtained in the aforementioned three steps to get the final classification result. With this coarse-to-fine classification strategy, human botanists can obtain excellent classification performance, only utilizing a small number of training samples.

This strategy inspires us to design a novel hierarchical attention module to extract features of different granularity, which is helpful for the correct classification of leaf images. It is well-noted that there is an inherent hierarchical structure in convolutional neural networks, in which features from low-level to high-level are extracted. The low-level features respond to edge and color conjunctions, middle-level features respond to the similar texture (such as mesh pattern), and high-level features are more class-specific (such as dog faces and bird legs) [57]. This nature of CNNs is the same as the classification strategy of botanists mentioned above, and these features of different levels can be utilized to get the attention maps of different granularity to guide the classification network where to focus. However, due to shallow and mid-level maps that may respond to background regions, directly using the response maps of shallow and mid-level features may introduce noise. In our proposed method, these attention maps are obtained by aggregated class active maps, which can protect the final attention maps from noise, especially in low-level attention maps.

Additionally, aiming at reducing the number of required training data to make machines simulate the recognition ability like humans, we propose **Supervised Sampling Contrastive Learning (SSCL)**. Our proposed SSCL utilizes the sampling transformation to generate strong contrastive pairs and greatly improve the performance compared with the vanilla supervised contrastive learning [27]. Specifically, our SSCL densely samples the original image where the attention maps have high responses, while sparsely samples where the attention maps have low responses. As a result, under the guidance of attention maps, the mechanism of the sampling transformation is very close to that of magnifying glasses, by which the discriminative regions are magnified and the background regions are depressed. Then, the strong contrastive pairs are fed into a two-branch network, i.e., a discriminative branch focuses on the high-response regions and a complementary branch focuses on medium-response regions. By doing this, the SSCL can help improve the generalization of classification models, which is important in leaf recognition with limited training samples.

With the help of the fusion attention map produced by the hierarchical attention module and our supervised sampling contrastive learning, rich and detailed information is extracted for correctly recognizing the leaf images with a small number of training samples like human botanists do. Our contributions can be summarized into three aspects:

— Different from previous works that localize discriminative regions with a single deep feature, we are motivated by the coarse-to-fine classification strategy of plant botanists and propose

the hierarchical attention mechanism, which can localize the discriminative regions with different granularity.
— To solve the challenge of limited well-labeled data, we propose supervised sampling contrastive learning to handle the leaf recognition task in the few-shot learning scenario, by which small and subtle details of important discriminative regions are enlarged for accurate recognition.
— We evaluate our methods on three common-used public leaf datasets, outperforming the state-of-the-art methods.

2 RELATED WORK

2.1 Leaf Image Recognition

Recently, plant image recognition based on image classification methods has attracted a lot of attention in the field of computer vision. And many public plant datasets, such as Flavia [54], Swedish [49], and LeafSnap [32], are proposed, which provide a fair comparison of plant recognition methods. Leaf image recognition is a challenging task because of the difficulties in recognizing similar species, such as large intra-class variance, small inter-class distance, long-tail distribution, and noisy images.

Early work [39] proposes a symbolic method for leaf image recognition, which utilizes a nearest neighbor classifier and relies on the leaf textual features. Specifically, they propose a local binary pattern as the leaf textural features. Then, they utilize the clustering algorithm to decide the number of textual patterns, in which they use a threshold to cluster samples to decrease the variation of the samples belonging to the same class. Finally, the nearest neighbor classifier is utilized to produce the final results. However, this method relies on a multi-view of the leaf samples, which limits the usability of this method due to the difficulty of obtaining fine-grained leaf images. Similarly, Reference [2] also designs hand-crafted features, including shape descriptors and Fourier descriptors, for leaf recognition. Then, these features are fed to a multi-layer perceptron to produce the recognition results.

Despite proving their effectiveness in several public datasets, these hand-crafted features-based methods are soon outperformed by the deep learning methods. Reference [41] proposes a CNNs-based method for plant classification, which employs data augmentation based on low-level transformations applied to the leaf images such as shifting, scaling, and rotation. However, deep learning-based methods rely on a huge amount of well-labeled training data, which is not well applicable in leaf recognition due to the difficulty of collecting samples. Reference [5] points out that the CNNs need a large number of training samples to achieve high accuracy and provide solid results. Therefore, it is very important to use powerful deep learning methods and fewer data to achieve good leaf recognition results. Taking the problems above into account, Reference [51] proposes a few-shot learning method for leaf recognition. Specifically, they employ the **Siamese Convolutional Neural Network (S-CNN)** to extract the leaf features of pre-defined leaf image pairs and then apply metric learning to learn discriminative features. In this article, we also conduct research in the few-shot learning scenario.

2.2 Fine-grained Image Classification

In the past few years, great progress [20, 22, 24, 46] has been made in the field of image classification due to the rapid development of deep learning technology and the availability of large-scale image datasets [13]. Compared with generic image recognition[28, 37], methods in fine-grained visual classification should have the ability to localize discriminative regions or learn the discriminative features due to the high inter-class similarity and large intra-class variance. According to whether

there is an explicit localization process, methods can be categorized into localization-classification networks and discriminative feature learning.

One is discriminative feature learning [23, 34, 36, 52], which usually refers to enhancing the representation capability of features by end-to-end training. Reference [36] constructs a bilinear structure to extract the pairwise features by two parallel CNNs and uses the pooled outer product of features to represent an image. After that, many methods [7, 12, 18, 30, 62] have further developed in this way and made great progress. Discriminative feature learning methods are simple and straightforward but struggle with human interpretation and performance consistency.

The other is discriminative region discovery, which often contains two subnetworks: localization subnetwork and classification subnetwork combining those regions to produce the final prediction. Some previous methods [31, 33, 35, 58, 59] utilize additional part annotations to train their localization subnetworks. Recent methods [15, 19, 40, 48, 55, 61] use the attention capability of classification networks to localize discriminative regions with only class labels, removing the requirement of expensive part annotations.

3 METHODOLOGY

Like the human plant taxonomists, the proposed $S^2CL - LeafNet$ first utilizes the hierarchical attention mechanism to learn attention maps with different granularity, which localizes the discriminative regions of input images in a coarse to fine manner. Then, these attention maps are fed into the sampling transformation to selectively sample the input image into a pair of positive contrastive learning images, which should be similar in the feature space. Finally, we apply the supervised contrastive loss along with cross-entropy loss to train the whole framework, which minimizes the distance of images from the same species while maximizing that from different species. The network utilizes the CNN backbone to extract features and can be trained end-to-end, as illustrated in Figure 2.

3.1 Hierarchical Attention Mechanism

With professional consideration, human botanists utilize a three-step coarse-to-fine strategy to find discriminative details of leaf samples, as shown in Figure 3. First, they analyze the appearances of leaves to get features of integral leaf shapes, including their edges and shapes. Then, they record these general features and turn to further analyze the leaves by scrutinizing their colors, shapes, and textures. Third, they perform microscopic and laboratory tests on the leaves, concentrating on leaf veins to get the most discriminative details. After this three-step strategy, all discriminative information obtained above is taken into integrated consideration to get the recognition results.

Convolutional neural networks are inherently hierarchical, and input images are progressively downsampled to obtain features from low level to high level. It is well-noted that in this hierarchical structure, the low-level features correspond to primary patterns, the mid-level features correspond to some generalized patterns, and the high-level features correspond to high-level semantics.

This well inspired us to propose a novel hierarchical attention mechanism to simulate the classification strategies of human botanists. Given an input image X, we feed it into the backbone CNN to produce a series of features with different granularity $\{f_1, f_2, \ldots, f_l, \ldots, f_L\}$. And $f_l \in \mathbb{R}^{K_l \times H_l \times W_l}$ is extracted feature of the lth layer, where K_l denotes the number of feature channels and H_l and W_l denote the width and height of the feature map, respectively. Then, global average pooling is performed on the f_l followed by a fully connected layer to obtain classification score $s \in \mathbb{R}^N$, where N is the number of categories to be recognized. Denote the matrix of the fully connected layer with $W_l \in \mathbb{R}^{K_l \times N}$, the class active map of ith category in the lth layer $\mathbf{CAM}_{l,i}$ can be

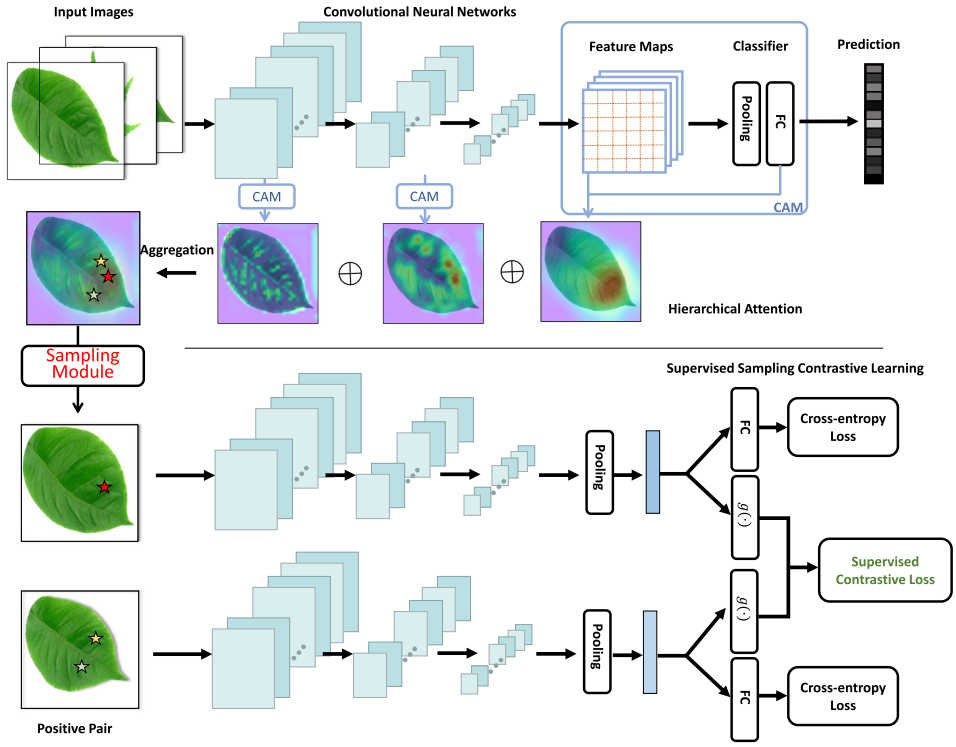

Fig. 2. An overview of the proposed $S^2CL - LeafNet$. First, the input leaf images are fed into the backbone CNN to produce a coarse prediction. Then, attention maps with different granularity are extracted by the hierarchical attention mechanism, which localizes the discriminative regions from coarse to fine. Second, sampling transformation is applied to the original leaf images with aggregated attention maps for producing a positive contrastive learning pair. Finally, supervised contrastive loss combined with cross-entropy loss is applied to train the whole network in an end-to-end way.

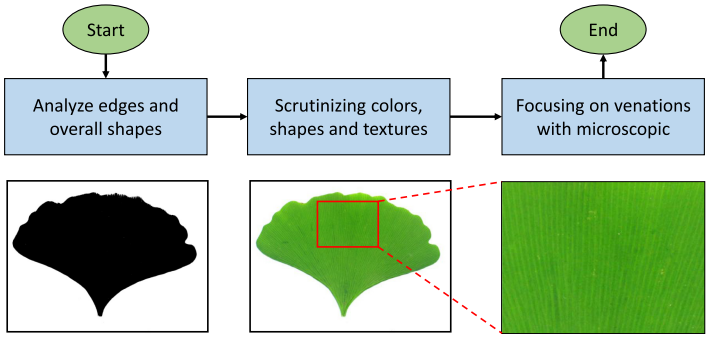

Fig. 3. The strategy of human botanists for recognizing leaves. First, they will analyze the margins and overall shapes of leaves. Second, the analysis is performed by scrutinizing both the shapes and colors of leaves. Finally, the microscopic observation is conducted to focus on the venations. The final decision is made based on the observation of these three steps.

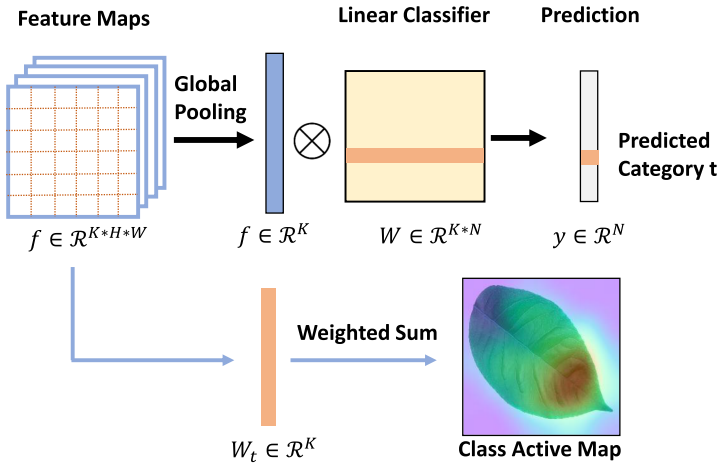

Fig. 4. CAM utilizes the classification weights in the linear classifier of the predicted category to weight the features of different channels, whose values represent the importance of different regions for the predicted category.

computed as

$$\mathbf{CAM}_{l,i} = \sum_{k=1}^{K_l} W_{l,i,k} \cdot \mathbf{f}_{i,k}. \tag{1}$$

For a better understanding, the process of producing CAM is shown in Figure 4.

Response values of the class active map indicate the importance of regions for correct classification, i.e., important regions have larger attention values, while background regions have smaller attention values. To take full advantage of the rich deep features of different layers, our hierarchical attention module aggregates class active maps of the predicted category by:

$$A = \sum_{l} \alpha_l \cdot \mathbf{b}(\mathbf{CAM}_l), \tag{2}$$

where α_l denotes aggregation weight, which is a trainable parameter. $\mathbf{b}$ is bilinear interpolation to make the class active maps the same size for aggregation. After obtaining the aggregated class active map, normalization is performed to scale the attention values in the same range:

$$A_{x,y} = \frac{A_{x,y}}{max(A_{x,y})}. \tag{3}$$

As shown in Figure 5, our hierarchical attention module well simulates the strategy of human botanists. In the shallow layer of the CNNs, the hierarchical attention mechanism focuses on the edges of leaves, which achieves the same goal as the first step of botanists' classification strategy. In the middle layer of the CNNs, the hierarchical attention mechanism turns to focus on the general leaf patterns, such as colors and shapes. Finally, in the high layer of the CNNs, the hierarchical attention mechanism focuses on the venations of leaves, which is the same with human botanists, because the venations are very important for recognizing categories of leaves.

3.2 Sampling Transformation

Like References [43, 63] and [14], our method also applies the sampling transformation to make the network zoom in the discriminative regions. First, we apply a sliding window size of r on the

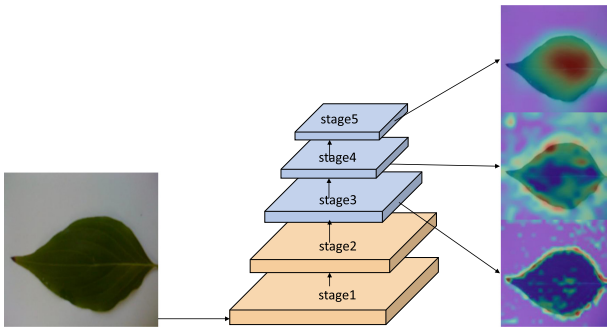

Fig. 5. The Hierarchical Attention Mechanism. There is an internal hierarchical structure in convolutional neural networks, where different layers focus on different information. In the shallow layers, the focus is on the edges and overall shapes of leaves. In the middle layers, the focus is on some general patterns, such as colors and textures. And in the high layers, the focus is on some high-level semantics that is closely related to recognizing. It well simulates the coarse-to-fine classification strategy of human botanists.

aggregated class active map to obtain the peak response points $P = \{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, where n is the number of valid peak point. Second, we use a threshold to partition the peak points into two sets P_d and P_c, according to their attention values,

$$\begin{aligned} P_d &= \{(x, y) | (x, y) \in P \quad , if A_{x,y} \geq \delta\} \\ P_c &= \{(x, y) | (x, y) \in P \quad , if A_{x,y} < \delta\}, \end{aligned} \tag{4}$$

where δ is the partition threshold.

Then, learnable Gaussian kernels are utilized to generate the sparse attention $S \in \mathbb{R}^{n \times H \times W}$ with

$$S_{i,x,y} = \begin{cases} A_{x_i,y_i} e^{\frac{(x-x_i)^2+(y-y_i)^2}{A_{x_i,y_i}\beta_1^2}}, & if (x_i, y_i) \in P_d, \\ \frac{1}{A_{x_i,y_i}} e^{\frac{(x-x_i)^2+(y-y_i)^2}{A_{x_i,y_i}\beta_2^2}}, & if (x_i, y_i) \in P_c, \end{cases} \tag{5}$$

where β_1 and β_2 is the parameter of Gaussian kernels, which are learnable parameters determined during the training process.

With the sparse attention defined in Equation (5), we perform image re-sampling to highlight fine-grained details from informative local regions while preserving surrounding context information. We construct two sampling maps S_d and S_c for the discriminative branch and complementary branch of feature extraction, respectively.

$$\begin{aligned} S_d &= \sum S_i, \quad if (x_i, y_i) \in P_d, \\ S_c &= \sum S_i, \quad if (x_i, y_i) \in P_c. \end{aligned} \tag{6}$$

Next, the sampling module g takes as input the saliency map S along with the full resolution image X to produce the re-sampled image X_{new} with

$$X_{new} = g(X, S). \tag{7}$$

As presented in Reference [43], we can compute a mapping between the sampled image and the original image and then use the grid sampler introduced in Reference [26]. This mapping can be written in the standard form as two functions $u(x, y)$ and $v(x, y)$ such that

$$X_{new} = X(u(x, y), v(x, y)). \tag{8}$$

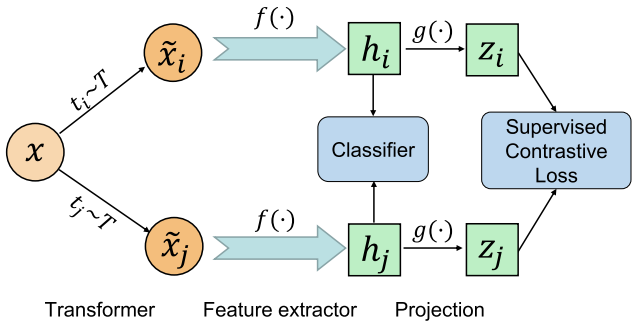

Fig. 6. The overview of supervised contrastive learning, which is a simple yet effective few-shot learning method. In supervised contrastive learning, samples of the same class and their transformations are positive examples. While samples of other classes and their transformations are negative examples.

The two mapping functions are used to map pixels proportionally to the normalized weight assigned to them by the saliency map. Assuming that $u(x, y)$, $v(x, y)$, x and y range from 0 to 1, an exact approximation to this problem would be to find u and v such that

$$\int_0^{u(x,y)} \int_0^{v(x,y)} S(x', y')dx'dy' = xy. \tag{9}$$

According to Reference [43], the solution can be described as

$$u(x, y) = \frac{\sum_{x',y'} S(x', y')k(x', y')x'}{\sum_{x',y'} S(x', y')k(x', y')}, \tag{10}$$

$$v(x, y) = \frac{\sum_{x',y'} S(x', y')k(x', y')y'}{\sum_{x',y'} S(x', y')k(x', y')}, \tag{11}$$

where k is a distance kernel that acts as a regularizer to avoid corner cases where all the pixels converge to the same value. By this, re-sampled images with the same dimensions as X have been produced from X. This sampling transformation makes regions with high attention values sampled more densely, since those regions have larger sampling weights. While regions with low attention values are sampled sparsely to preserve the context regions. By replacing S in Equations (10) and (11) with S_d and S_c, respectively, we can get two different groups of sampling functions. Thus, two different sampled images can be obtained with Equation (8).

On one hand, due to the partition of valid peak points, regions that are magnified in the two sampled images are different, which can provide rich but diverse discriminative information for classification. On the other hand, these two sampled images constitute a positive contrastive learning pair, which can be optimized with supervised contrastive learning, which is described in the next section.

3.3 Supervised Contrastive Learning for Leaf Recognition

Aiming at recognizing leaf images with a small number of training samples, we introduce supervised contrastive learning, which is a simple yet effective few-shot learning method. As shown in Figure 6, given an input image X, the supervised contrastive learning [27] first applies two different data transformations $t \sim T$ and $t' \sim T$ to obtain two transformed images. Both transformed images are fed into the encoder network $f(\cdot)$ to obtain a normalized embedding h. During training, this representation h is further propagated through a projection network $g(\cdot)$ that is discarded

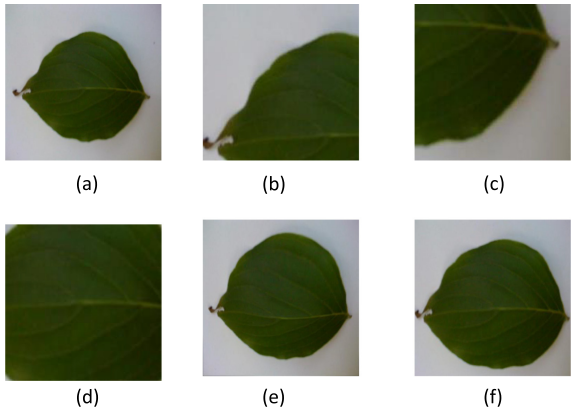

Fig. 7. Comparison of different transformed images. (a) is the original input image. (b) and (c) are images that are randomly cropped and resized, which is applied in vanilla supervised contrastive learning. (d) is the image that is center-cropped, which is applied in previous leaf recognition methods. (e) and (f) are a pair of the sampled image produced by the sampling module, which not only magnifies the discriminative regions but also preserves the other details.

at inference time. The supervised contrastive loss is computed on the outputs z of the projection network, bringing together representations from the same category while keeping representations from different categories away from each other. Further, a linear classifier is trained on top of the frozen representations equipped with a cross-entropy loss.

In vanilla supervised contrastive learning, the transformations of data augmentation include horizontal flip, random rotation, random crop, random resize, and so on. Though these transformations are effective for data augmentation and are commonly used in image classification, these transformations can not automatically attend to the discriminative regions of input images. However, localizing discriminative regions is an important step for image classification, especially for fine-grained classification tasks. As shown in Figure 7, compared with transformations used in vanilla supervised contrastive learning, our sampling module not only attends the discriminative regions but also reserves context regions, which is very helpful for recognizing.

Recapping the overall procedure, the input images are first fed into the hierarchical attention module to produce multi-level attention maps, which are later aggregated together and fed into the sampling transformation. Under the guidance of the attention maps, the sampling transformation performs dense sampling on the input image in areas with high attention values, while sparse sampling is performed in areas with low attention values. And the sampling transformation produces two sampled images with different attention regions, which constitutes a positive contrastive pair that can be used in supervised contrastive learning.

For a mini-batch of N training samplers, denote the index of augmented samples with $i \in I \equiv \{1, 2, \ldots, 2N\}$, and let $A(i) \equiv I \backslash i$ be the set of images of this mini-batch except the ith image. The supervised contrastive loss is defined as

$$\begin{aligned} L_{scl} &= \sum_{i \in I} L_{scl}^{i} \\ &= -\sum_{i \in I} log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{exp(z_i \cdot z_p / \tau)}{\sum_{j \in A(i)} exp(z_i \cdot z_j / \tau)} \right\}, \end{aligned} \tag{12}$$

Table 1. The Statistics of Datasets in this Article

Datasets	Number of classes	Number of training images	Number of test images
Flavia [54]	32	1,526	381
Swedish [49]	15	844	281
LeafSnap [32]	184	23,147	2,760

where $P(i) \equiv \{p \in A(i) : y_p = y_i\}$ is the set of all positives in this mini-batch, including the other augmented samples originating from the same source sample or samples from the same category. y_p is the label of the pth image and $|P(i)|$ is its cardinality.

The features extracted from the original image and the two sampled images are denoted with $F = \{F_o, F_d, F_c\}$. We utilize the concatenation of F_o, F_d, and F_c to make the final classification. So, there are four linear classifiers to be trained with

$$L_{cls} = \sum_{i \in I} L_{ce}(y_i, y^*) + L_{ce}(y_c, y^*), \tag{13}$$

where $I = \{I_o, I_d, I_c\}$. L_{ce} denotes the cross-entropy loss, and y^* denotes the ground-truth label. y_c is the predicted label of the concatenation feature. The whole model is trained end-to-end with

$$L = L_{cls} + \beta L_{scl}, \tag{14}$$

where β denotes the balance weight. L_{scl} is defined with Equation (12) and L_{cls} is defined with Equation (13).

4 EXPERIMENTS AND RESULTS

In this section, we conduct experiments on three public leaf datasets, i.e., LeafSnap, Flavia, and Swedish, to evaluate the effectiveness of our proposed method $S^2CL-LeafNet$.

4.1 Datasets

In this article, we evaluate the proposed approach on three fine-grained leaf datasets: LeafSnap, Flavia, and Swedish. The detailed statistics of these three datasets are shown in Table 1, including the number of training and test samples. These three datasets cover a wide variety of plant species, which makes these datasets challenging. Specifically, the LeafSnap dataset is imbalanced, and there are only a few samples available for some species. While in Flavia and Swedish, they are balanced. Noting that the goal of this article is to automatically recognize which species the leaf image belongs to with less labeled training data, we have designed different few-shot settings to evaluate our method. That is, we use 5 to 20 samples per class for training the models, which is a smaller amount of training data compared with existing leaf recognition methods and other fine-grained recognition methods.

Following the few-shot learning setting in S-CNN [51], the number of samples per class during training is set as 5, 10, 15, and 20, respectively, which are randomly sampled from the original training sets. For the test phase, all remaining images in the original training set and the original test images are used to evaluate the performance of our proposed method. Except for the few-shot learning setting, we also conduct experiments under the fully supervised setting, in which all original training sets are used.

To describe these three leaf image datasets more clearly, the samples are shown in Figure 8. In these three datasets, there is a high similarity between the different categories of leaves, which

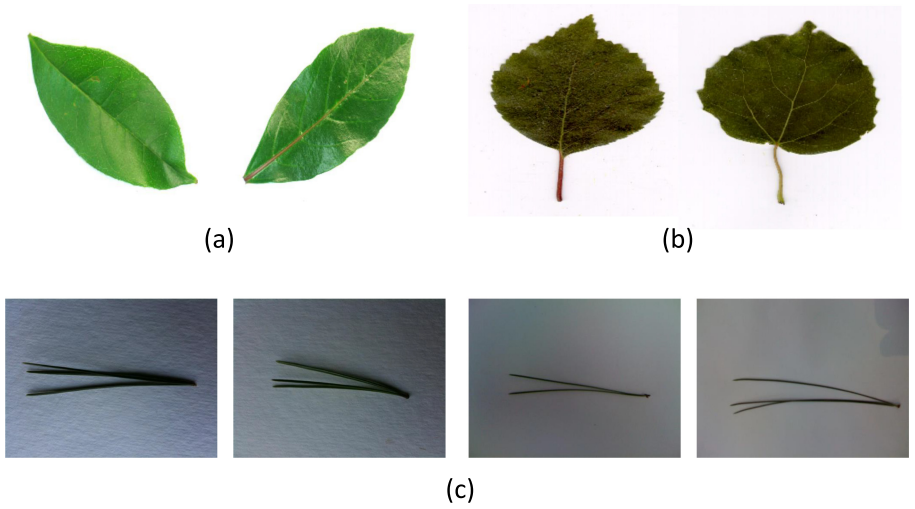

Fig. 8. Samples from three common-used leaf image datasets. (a) are from Flavia, (b) are from Swedish, and (c) are from LeafSnap.

makes it challenging to recognize their class correctly. Furthermore, compared with Flavia and Swedish, LeafSnap is a more complex dataset. It contains two subsets, lab subset and field subset, where the lab subset is captured in a laboratory environment, and the field subset is taken in a natural environment.

4.2 Implementation Details

We implement our proposed method with PyTorch and train our models on an NVIDIA Titan X (Pascal) GPU, whose memory is 12 GB. For a fair comparison, all images are resized to 224×224 during both the training and test phases. For data augmentation, randomly cropped, random resize, and random horizontal flip are deployed for all training samples. While for the test samples, only random horizontal flip is deployed. For training the whole model, we deploy the SGD optimizer. Specifically, the momentum is set to 0.9 and 0.0005 for the weight decay. During training, the size of a single batch is set to 30. The max training epoch is set to 50. The initial learning rate is 0.0005, with exponential decay of 0.95 every 4 epochs. Specifically, we train the parameters out of the backbone CNN with learning rate times 10 for weight parameters and times 20 for bias parameters.

4.3 Experiments and Analysis

4.3.1 Few-shot Learning Scenario. To verify the few-shot learning performance of our method, we conduct experiments in Flavia, Swedish, and LeafSnap datasets. Following the experiment setting of S-CNN [51], we set the number of training samples per category to 5, 10, 15, and 20, respectively, and all remaining images to compose the test set.

Note that this is different from the normal N-way-K-shot setting in few-shot learning; it can be seen as an all-way setting, which is more difficult than the commonly used 5-way and 10-way settings. As shown in Table 2, the experiment result on the Flavia dataset shows that we achieve a new state-of-the-art performance. Thanks to the hierarchical attention mechanism and the strong optimization strategy [20], our method without supervised sampling contrastive learning outperforms the previous methods. Our hierarchical attention mechanism helps the network find details of different granularity, which are useful for recognizing the categories leaves belong to. Furthermore,

Table 2. Overall Accuracy (%) of the Different Methods for the Flavia Dataset

Method	n = 5	n = 10	n = 15	n = 20
SSLDP [1]	32.2	44.1	58.7	74.6
SFFD [38]	42.8	77.8	83.2	85.8
SS-HCNN [9]	41.8	69.1	87.1	93.5
S-ResNet [51]	57.4	81.2	89.7	93.8
S-Inception [51]	59.2	85.2	92.3	95.3
PMG [15]	67.4	87.6	94.4	96.8
AE-Net [25]	68.5	88.3	95.5	97.9
Ours w/o SSCL	70.1	90.1	96.9	97.7
Ours	**75.3**	**93.8**	**97.7**	**98.3**

The SSCL denotes supervised sampling contrastive learning.

Table 3. Overall Accuracy (%) of the Different Methods for the Swedish Dataset

Method	n = 5	n = 10	n = 15	n = 20
SSLDP [1]	31.7	42.5	55.7	73.8
SFFD [38]	39.6	73.6	80.9	83.1
SS-HCNN [9]	38.5	66.5	85.0	92.0
S-ResNet [51]	52.8	77.2	88.1	91.7
S-Inception [51]	49.6	72.5	85.1	88.8
PMG [15]	71.3	87.0	94.9	98.1
AE-Net [25]	72.8	88.3	96.5	97.9
Ours w/o SSCL	73.9	88.9	96.8	99.0
Ours	**79.8**	**90.7**	**90.3**	**99.3**

The SSCL denotes the supervised sampling contrastive learning.

on this basis, our sampling transformation amplifies these useful image regions and effectively extends the data augmentation approaches used in supervised contrastive learning, which further boosts the performance of our approach.

Specifically, with a relatively small number of training samples such as the number of training samples per category equal to 5 and 10, our supervised sampling contrastive learning achieves a great improvement (75.3% vs. 70.1% and 93.8% vs. 90.1%) than without the SSCL. The good performance implies that our methods could be very useful in practice where training images are hard to get and annotate. And compared with the vanilla S-CNN [51], when the number of training samples increases, our proposed supervised sampling contrastive learning still improves the performance of the model (97.7% vs. 96.9% and 97.7% vs. 98.3%). This is because, by sampling transformation, the informative regions are effectively enlarged, which is highly important in fine-grained recognition.

As shown in Tables 3 and 4, our proposed method achieves consistent improvement in Swedish and LeafSnap. This demonstrates the scalability of our approach to other datasets.

Table 4. Overall Accuracy (%) of the Different Methods for the LeafSnap Dataset

Method	n = 5	n = 10	n = 15	n = 20
S-ResNet [51]	59.4	80.4	92.6	96.8
S-Inception [51]	61.0	80.0	90.0	93.2
PMG [15]	74.5	86.0	91.1	97.5
AE-Net [25]	74.8	85.5	91.5	97.9
Ours w/o SSCL	80.9	88.9	91.5	98.1
Ours	**82.4**	**89.6**	**92.8**	**98.7**

The SSCL denotes the supervised sampling contrastive learning.

Table 5. Repeated Experiment Results in Flavia Dataset

	1	2	3	4	5
n = 5	75.3	75.0	75.2	74.8	75.2
n = 20	98.3	98.3	97.9	98.1	98.0

Due to the training images being randomly sampled from the original datasets, we repeated the experiments to exclude the effect of randomness when k = 5 and k = 20 in Flavia dataset. The experiment results are shown in Table 5, which demonstrates the stability of the proposed $S^2CL - LeafNet$.

4.3.2 Experiments Results under Fully Supervised Setting. The proposed $S^2CL - LeafNet$ not only achieves state-of-the-art performance in the few-shot learning scenario but also shows well adaption under the fully supervised setting. Under the fully supervised setting, all training images are used in the training phase, which is the same as the fine-grained image recognition task. That is, we use all the 1,526 training samples in the Flavia dataset under this setting. And the computational costs are not much more than that in the case of few-shot learning. Thanks to the sampling transformation, the informative regions where details of the leaf images are well enlarged, which is important for fine-grained recognition [14, 63]. And as shown in Table 6, our proposed method also achieves state-of-the-art performance under the fully supervised setting.

4.3.3 Ablation Study and Visualization. To better illustrate the effect of our proposed method, ablation studies are conducted. Our proposed hierarchical attention mechanism could effectively find discriminative regions of different granularity that are highly related to correctly recognizing the leaf categories. In comparison, we design a simple module to fuse the attention maps. More specifically, the class active maps are not used to produce the raw attention maps, and the normalized convolutional response maps are replaced to generate the attention maps. Then, we average these raw attention maps to obtain the final attention map. Experimental results are shown in Table 7. Using the simple fusion method not only fails to improve the recognition accuracy but even harms the performance of the model when the training images are not sufficient. We think that it is because the middle-level features are sensitive to the changes in input images. Thus, the middle-level features are not robust like the high-level features, and even are noisy for recognition. Using these noisy features to produce attention maps that guide models to enlarge the attention regions may be harmful. Thanks to the class active maps focusing on the active regions that are

Table 6. Overall Accuracy (%) of the Different Methods on the Flavia Dataset under the Fully Supervised Setting

Method	Accuracy
PCNN [53]	96.9
LeafNet CNN [41]	97.9
Shape & Statistical & Vein Features, PCA + KNN [45]	98.8
DeepPlant+MLP [44]	99.4
SWP-Leaf Net [6]	99.7
PMG [15]	99.6
AE-Net [25]	99.7
Ours w/o SSCL	99.5
Ours	**99.8**

The SSCL denotes the supervised sampling contrastive learning.

Table 7. Effect of Hierarchical Attention Mechanism

Method	n = 5	n = 10	n = 15	n = 20	fully supervised
Ours w/o HAM	81.2	88.9	91.7	96.9	99.1
Ours w/simple fusion	80.1	87.8	91.5	96.3	98.5
Ours w/HAM	82.4	88.6	92.8	98.1	99.7

Experiments are conducted on the LeafSnap dataset.

related to the predicted category, our proposed hierarchical attention mechanism could help the procedure of recognizing.

As shown in Figure 9, we visualize the attention maps of the hierarchical attention mechanism. As shown in the first row in Figure 9, the hierarchical attention mechanism mainly focuses on the edge and veins of leaves in the shallow layer of CNN, which provides important details for the final recognition. In the middle layer of CNN, the hierarchical attention mechanism turns to focus on general leaf shapes, such as shape and color. Finally, in the high layer of CNN, the hierarchical attention mechanism focuses on the discriminative regions that are important for classification.

More importantly, we find that these hierarchical attentions could complement each other. For very narrow leaves such as coniferous leaves, as shown in the second row in Figure 9, it is difficult for the high-level attention module to accurately localize the discrimination regions. And the attention regions will diffuse to the surrounding background, which is unfavorable to the final classification procedure. However, just as mentioned before, attention maps produced by shallow layers can be very noisy. Just as shown in the third row in Figure 9, there are many active points scattered in the background. In this case, the attention maps from the middle and high levels help to correct the errors caused by these noises.

Furthermore, the sampled leaf images are presented in Figure 10. And for the discriminative sampled images, the main region of leaves is enlarged, just like using a magnifying glass magnifies important regions in the image. As the supplement of the discriminative sampled images, the complementary sampled images have little distortion compared to the original image and can be seen as a compromise between the discriminative images and original images. Compared with data augmentation methods, such as random crop and center crop, which are used in vanilla supervised

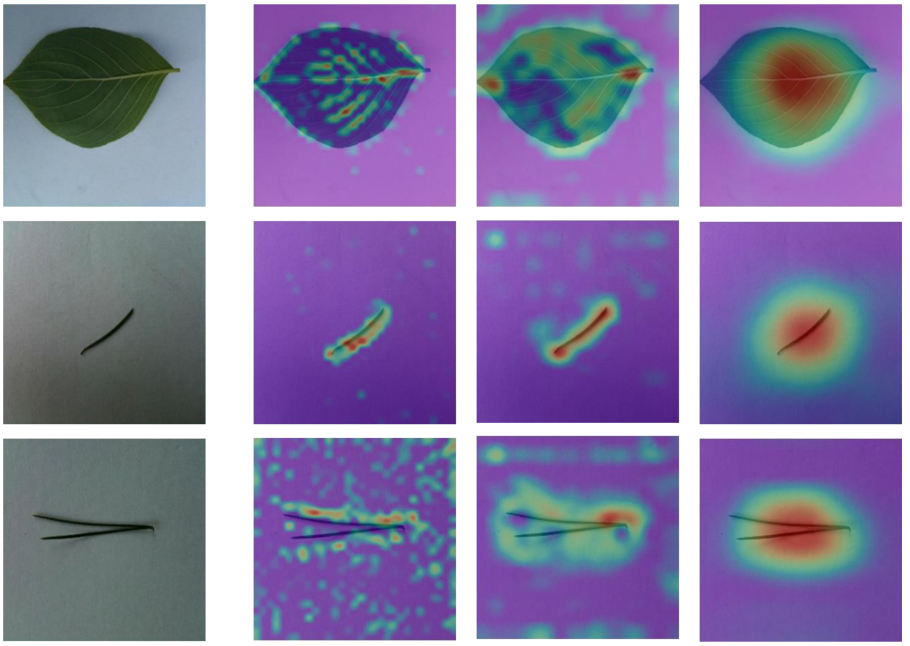

Fig. 9. Visualization of hierarchical attention. The first column is the input leaf image. And, from left to right, there are three attention maps of network layers from shallow to deep.

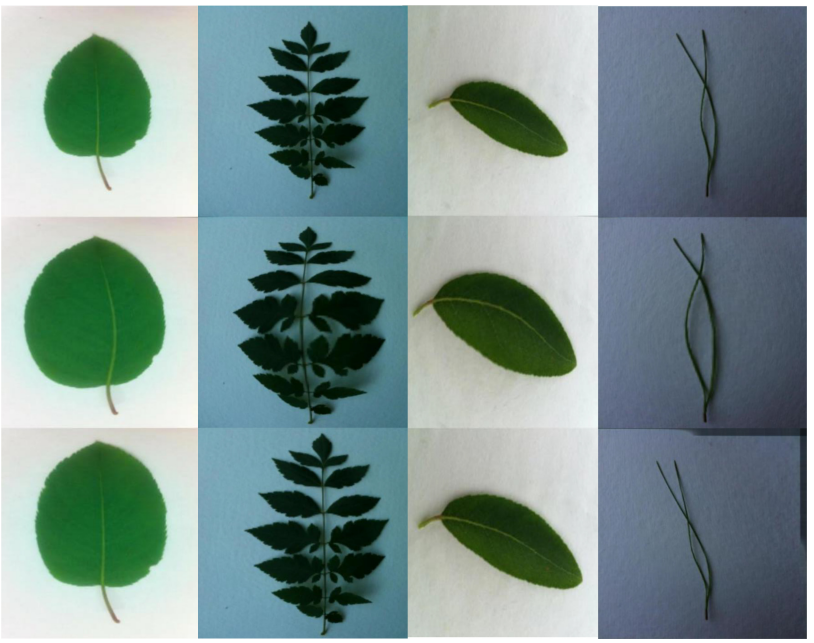

Fig. 10. Sampled images after sampling transformation. The first row is the original leaf image. Discriminative sampled images and complementary sampled images are presented in the second and third rows, respectively.

Table 8. Comparison with State-of-the-art Fine-grained Recognition Methods on CUB-200-2011 Dataset

Method	Accuracy
FT ResNet	84.1
DFL [52]	87.4
NTS [56]	87.5
TASN [63]	87.9
MAMC [48]	87.3
S3N [14]	88.5
PMG [15]	89.6
PMG with multi-head [8]	89.9
PART [60]	89.6
Ours	90.1

contrastive learning, using sampling transformation can magnify important details of leaves and preserve the original shape of the main object. This distortion of these important regions is a helpful data augmentation method, especially when used with supervised contrastive learning.

4.3.4 Evaluation under FGVC Setting. Though this is not the point of our proposed method, our method still can be utilized in traditional FGVC task. To prove the effectiveness of our proposed method, we conduct experiment on the common-used FGVC dataset, CUB-200-2011.

Due to our proposed method being based on the CNN backbone, we compare our proposed method with SOTA CNN-based FGVC method. As shown in Table 8, though this is not the point of our article, our method still achieves comparable performance to state-of-the-art CNN-based fine-grained methods under the traditional FGVC setting. We think this is because learning discriminative features is important in fine-grained recognition, which is the same in both leaf images and bird images. This proves that our proposed HAM and SSCL are also effective in more general fine-grained image classification.

5 CONCLUSION

To make the classification networks recognize the leaf images like human botanists, we propose the $S^2CL - Leaf Net$, which can correctly classify leaf images with a few training samples. This comes from two aspects, i.e., the hierarchical attention mechanism and supervised sampling contrastive learning. Inspired by the classification strategy of human botanists, our hierarchical attention mechanism localizes the discriminative regions of different granularity with the inherent hierarchical architecture of CNNs. And by using the supervised sampling contrastive learning, the original leaf images are augmented without damaging the discriminative regions, helping the recognition model achieve high accuracy with a few training samples. Furthermore, we conduct experiments on three commonly used leaf datasets and achieve new state-of-the-art performance. Visualization results are provided for a better understanding of our methods.

REFERENCES

[1] Shanwen Zhang, Ying-Ke Lei, and Yan-Hua Wu. 2011. Semi-supervised locally discriminant projection for classification and recognition. *Knowl.-based Syst.* 24, 2 (2011), 341–346.

[2] Aimen Aakif and Muhammad Faisal Khan. 2015. Automatic classification of plants based on their leaves. *Biosyst. Eng.* 139 (2015), 66–75.
[3] S. Anubha Pearline, V. Sathiesh Kumar, and S. Harini. 2019. A study on plant recognition using conventional image processing and deep learning approaches. *J. Intell. Fuzz. Syst.* 36, 3 (2019), 1997–2004.
[4] Amanda Ash. 1999. *Manual of Leaf Architecture: Morphological Description and Categorization of Dicotyledonous and Net-veined Monocotyledonous Angiosperms*. Smithsonian Institution.
[5] Jayme Garcia Arnal Barbedo. 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput. Electron. Agric.* 153 (2018), 46–53.
[6] Ali Beikmohammadi, Karim Faez, and Ali Motallebi. 2020. SWP-Leaf NET: A novel multistage approach for plant leaf identification based on deep learning. *arXiv preprint arXiv:2009.05139* (2020).
[7] Sijia Cai, Wangmeng Zuo, and Lei Zhang. 2017. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*. 511–520.
[8] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. 2021. Your "flamingo" is my "bird": Fine-grained, or not. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11476–11485.
[9] Tao Chen, Shijian Lu, and Jiayuan Fan. 2018. SS-HCNN: Semi-supervised hierarchical convolutional neural network for image classification. *IEEE Trans. Image Process.* 28, 5 (2018), 2389–2398.
[10] L. M. Chu, Cong Zhao, Wai-Kuen Cham, and Sharon S. F. Chan. 2015. Plant identification using leaf shapes—A pattern counting approach. *Pattern Recog.* 48, 10 (2015), 3203–3215.
[11] Allen J. Coombes. 2014. *The Book of Leaves: A Leaf-by-leaf Guide to Six Hundred of the World's Great Trees*. University of Chicago Press.
[12] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. 2017. Kernel pooling for convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2930.
[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.
[14] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. 2019. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6599–6608.
[15] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *Proceedings of the European Conference on Computer Vision*. Springer, 153–168.
[16] Esraa Elhariri, Nashwa El-Bendary, and Aboul Ella Hassanien. 2014. Plant classification system based on leaf features. In *Proceedings of the 9th International Conference on Computer Engineering & Systems*. IEEE, 271–276.
[17] Beth Ellis, Douglas Daly, Leo J. Hickey, Kirk R. Johnson, and Scott L. Wing. 2009. *Manual of Leaf Architecture*.
[18] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. 2016. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 317–326.
[19] Xiang Guan, Guoqing Wang, Xing Xu, and Yi Bin. 2021. Learning hierarchal channel attention for fine-grained visual classification. In *Proceedings of the ACM International Conference on Multimedia*. 5011–5019.
[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[21] J. Hu, Z. Chen, M. Yang, R. Zhang, and Y. Cui. 2018. A multi-scale fusion convolutional neural network for plant leaf recognition. *IEEE Sig. Process. Lett.* 25, 6 (2018), 853–857.
[22] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.
[23] Yunqing Hu, Xuan Jin, Yin Zhang, Haiwen Hong, Jingfeng Zhang, Yuan He, and Hui Xue. 2021. RAMS-Trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the ACM International Conference on Multimedia*. 4239–4248.
[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.
[25] Hu Yutao, Liu Xuhui, Zhang Baochang, Han Jungong, and Cao Xianbin. 2021. Alignment enhancement network for fine-grained visual categorization. *ACM Trans. Multim. Comput. Commun. Appl.* 17, 1s (2021), 12:1–12:20.
[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial transformer network. *Adv. Neural Inf. Process. Syst.* 28 (2015), 2017–2025.
[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362* (2020).
[28] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *Proceedings of the European Conference on Computer Vision*. 491–507.

[29] Hoshang Kolivand, Bong Mei Fern, Tanzila Saba, Mohd Shafry Mohd Rahim, and Amjad Rehman. 2019. A new leaf venation detection technique for plant species classification. *Arab. J. Sci. Eng.* 44, 4 (2019), 3315–3327.
[30] Shu Kong and Charless Fowlkes. 2017. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 365–374.
[31] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. 2015. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5546–5555.
[32] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, and J. V. B. Soares. 2012. LeafSnap: A computer vision system for automatic plant species identification. In *Proceedings of the European Conference on Computer Vision*.
[33] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. 2017. Fine-grained recognition as HSnet search for informative image parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2520–2529.
[34] Guangjun Li, Yongxiong Wang, and Fengting Zhu. 2021. Multi-branch channel-wise enhancement network for fine-grained visual recognition. In *Proceedings of the ACM International Conference on Multimedia*. 5273–5280.
[35] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. 2015. Deep LAC: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1666–1674.
[36] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhransu Maji. 2015. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 1449–1457.
[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2021. Swin transformer V2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883* (2021).
[38] Li Longlong, Jonathan M. Garibaldi, and He Dongjian. 2015. Leaf classification using multiple feature analysis based on semi-supervised clustering. *J. Intell. Fuzz. Syst.* 29, 4 (2015), 1465–1477.
[39] Y. G. Naresh and H. S. Nagendraswamy. 2016. Classification of medicinal plants: An approach using modified LBP with symbolic representation. *Neurocomputing* 173 (2016), 1789–1797.
[40] Meike Nauta, Ron van Bree, and Christin Seifert. 2021. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 14933–14943.
[41] Barré Pierre, Ben C. Stöver, Kai F. Müller, and Steinhage Volker. 2017. LeafNet: A computer vision system for automatic plant species identification. *Ecol. Inform.* 40 (2017), 50–56.
[42] C. Arun Priya, T. Balasaravanan, and Antony Selvadoss Thanamani. 2012. An efficient leaf recognition algorithm for plant classification using support vector machine. In *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering*. IEEE, 428–432.
[43] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2018. Learning to zoom: A saliency-based sampling layer for neural networks. In *Proceedings of the European Conference on Computer Vision*. 51–66.
[44] Sue Han Lee, Chee Seng Chan, Simon Mayo, and Paolo Remagnino. 2017. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognit.* 17 (2017), 1–13.
[45] Gulshan Saleem, M. Akhtar, Nisar Ahmed, and W. S. Qureshi. 2019. Automated analysis of visual leaf shape features for plant classification. *Comput. Electron. Agric.* (2019).
[46] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[47] Uday Pratap Singh, Siddharth Singh Chouhan, Sukirty Jain, and Sanjeev Jain. 2019. Multilayer convolution neural network for the classification of mango leaves infected by anthracnose disease. *IEEE Access* 7 (2019), 43721–43729.
[48] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision*. 805–821.
[49] Oskar J. O. Söderkvist. 2001. Computer vision classifcation of leaves from swedish trees. Master's Thesis, Linkoping University, 2001.
[50] Tan Kiet Nguyen Thanh, Quoc Bao Truong, Quoc Dinh Truong, and Hiep Huynh Xuan. 2018. Depth learning with convolutional neural network for leaves classifier based on shape of leaf vein. In *Proceedings of the Asian Conference on Intelligent Information and Database Systems*. Springer, 565–575.
[51] Bin Wang and Dian Wang. 2019. Plant leaves classification: A few-shot learning method based on siamese network. *IEEE Access* 7 (2019), 151754–151763.
[52] Yaming Wang, Vlad I. Morariu, and Larry S. Davis. 2018. Learning a discriminative filter bank within a CNN for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4148–4157.
[53] Zhaobin Wang, Xiaoguang Sun, Yaonan Zhang, Zhu Ying, and Yide Ma. 2016. Leaf recognition based on PCNN. *Neural Comput. Applic.* 27, 4 (2016), 899–908.
[54] S. G. Wu, F. S. Bao, E. Y. Xu, Y. X. Wang, and Q. L. Xiang. 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. In *Proceedings of the IEEE Symposium on Signal Processing and Information Technology*.

[55] Shaokang Yang, Shuai Liu, Cheng Yang, and Changhu Wang. 2021. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. *arXiv preprint arXiv:2102.09875* (2021).
[56] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision*.
[57] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*. Springer, 818–833.
[58] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. 2016. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1143–1152.
[59] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. 2014. Part-based R-CNNs for fine-grained category detection. In *Proceedings of the European Conference on Computer Vision*. Springer, 834–849.
[60] Yifan Zhao, Jia Li, Xiaowu Chen, and Yonghong Tian. 2022. Part-guided relational transformers for fine-grained visual recognition. *CoRR* abs/2212.13685 (2022).
[61] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 5209–5217.
[62] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Learning deep bilinear transformation for fine-grained image representation. *arXiv preprint arXiv:1911.03621* (2019).
[63] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5012–5021.

Received 27 November 2022; revised 10 July 2023; accepted 28 July 2023