Benchmark Platform for Ultra-Fine-Grained Visual Categorization Beyond Human Performance

Xiaohan Yu[1] Yang Zhao[1,2] Yongsheng Gao[1,*] Xiaohui Yuan[3] Shengwu Xiong[3]
[1]Griffith University [2]The University of Adelaide [3]Wuhan University of Technology
xiaohan.yu@griffith.edu.au yang.zhao01@adelaide.edu.au yongsheng.gao@griffith.edu.au
yuanxiaohui@whut.edu.cn xiongsw@whut.edu.cn

Abstract

Deep learning methods have achieved remarkable success in fine-grained visual categorization. Such successful categorization at sub-ordinate level, e.g., different animal or plant species, however relies heavily on the visual differences that human can observe and the ground-truths are labelled on the basis of such human visual observation. In contrast, few research has been done for visual categorization at the ultra-fine-grained level, i.e., a granularity where even human experts can hardly identify the visual differences or are not yet able to give affirmative labels by inferring observed pattern differences. This paper reports our efforts towards mitigating this research gap. We introduce the ultra-fine-grained (UFG) image dataset, a large collection of 47,114 images from 3,526 categories. All the images in the proposed UFG image dataset are grouped into categories with different confirmed cultivar names. In addition, we perform an extensive evaluation of state-of-the-art fine-grained classification methods on the proposed UFG image dataset as comparative baselines. The proposed UFG image dataset and evaluation protocols is intended to serve as a benchmark platform that can advance research of visual classification from approaching human performance to beyond human ability, via facilitating benchmark data of artificial intelligence (AI) not to be limited by the labels of human intelligence (HI). The dataset is available online at https://github.com/XiaohanYu-GU/Ultra-FGVC.

1. Introduction

The increasing popularity of deep learning methods has led to tremendous success in the research field of fine-grained visual categorization (FGVC). Human-observed prior knowledge has played a key role in developing current state-of-the-art fine-grained classification methods, *e.g.*, part-based methods [24, 61, 59, 68] driven by the human-observed facts that visual differences among categories might be subtle and reside in the unique properties of object parts. Yet it remains unclear how these state-of-the-art classification methods perform on what we call ultra-fine-grained visual categorization (ultra-FGVC) tasks. The ultra-FGVC is to classify images at an ultra-fine granularity where even human experts may fail to identify or depict the visual difference (see Fig.1 as an example).

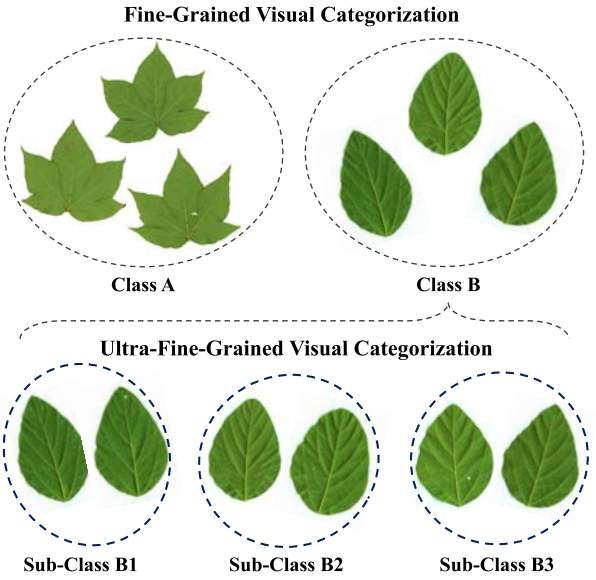

Figure 1. A visual comparison of the ultra-fine-grained visual categorization vs. the fine-grained visual categorization.

Despite the significant importance of ultra-FGVC in computer vision and artificial intelligence agriculture [65, 36], very few research has been reported to tackle the ultra-FGVC tasks. This is mainly due to the absence of large-scale ultra-fine-grained image datasets. To date, only a few attempts have been made on the collection of ultra-fine-grained image datasets [63, 35, 15]. Among these datasets, the largest as well as the only publicly available dataset [63]

*Corresponding author

contains 600 images from 100 categories, making it insufficient to provide current deep learning methods a practical and comprehensive test bed for ultra-FGVC evaluation. Nevertheless, these pioneering works have reported encouraging similar to human performance on the ultra-FGVC tasks, confirming the possibility of ultra-FGVC via machine learning methods.

To develop effective deep learning models for such a challenging scenario we require suitable and large-scale ultra-fine-grained image datasets. Data collection in the ultra-FGVC task, however, suffers from two limitations. First, it is beyond the capability of human experts or volunteers to accurately annotate ultra-fine-grained images via visual observation, due to the very high inter-class similarity (see Fig. 1). Second, as the spectrum of granularity moves down from a fine level to an ultra-fine level, the number of available samples in each category becomes smaller [53]. These limitations impede the progress towards building a desirable dataset for ultra-FGVC tasks.

To the best of our knowledge, no large-scale dataset exists for the task of ultra-FGVC. As a first step to fill this gap and to enable further research in the development of methods for ultra-FGVC, we introduce the ultra-fine-grained (UFG) image dataset that facilitates a thorough evaluation of such methods. The UFG image dataset contains in total 47,114 images from 3,526 ultra-fine-grained categories. Despite the difficulty in the collection of large-scale ultra-fine-grained images, our dataset is significantly larger than existing ultra-fine-grained image dataset, making it a desirable test bed for current deep learning methods to perform algorithm evaluations.

In contrast to existing fine-grained image datasets [54, 26, 53, 58, 60, 46, 32], the UFG image dataset differs in the following aspects. **Firstly**, the label of each image is determined by the cultivar name of its plant seed from the genetic resource bank. This solves one of the inherent problems in current benchmark databases for visual categorization, that is the ground truths are labelled by human, making the AI systems learned from these databases in theory unable to be more accurate than human. The proposed UFG image dataset provides a benchmark platform and task for developing machine learning approaches whose performances are not limited by human-observed labeling accuracy. **Secondly**, all the images are scanned in a laboratory environment, minimizing extra possible noises from complex backgrounds. Given that ultra-FGVC remains a pioneering and extremely challenging research problem, the major goal is to stay focused on classifying the objects themselves in this very early stage. **Thirdly**, more challenging than the current fine-grained image datasets [54, 58, 32] that enable the classification at a species level, the UFG image dataset further moves down the granularity of category to the subclass of species, *i.e.*, a cultivar level. Fig. 1 shows a visual comparison of objects in the ultra-fine-grained visual categorization vs. the fine-grained visual categorization. **Finally**, compared with the fine-grained image datasets, the UFG image dataset covers much smaller inter-class variations, making the ultra-FGVC a much more challenging task for current fine-grained classification methods. A more detailed discussion about the challenge of the UFG image dataset is given in Section 3.3.

To explore the performance of current state-of-the-art methods and motivate further research, we present an extensive evaluation on the proposed UFG image dataset. Given the extraordinary or even "superhuman" [44, 21] performance of deep learning methods in classification tasks, we evaluate 13 state-of-the-art classification methods as standard baselines for the UFG image dataset. The experimental results show both the critical need and challenges of developing new methods for the ultra-FGVC tasks.

The main contributions of this paper are:

- Further advancing the fine-grained visual categorization, we propose a challenging ultra-fine-grained visual categorization task that classifies objects at a much finer granularity to the level that human may fail to identify or provide affirmative description of their visual differences. To facilitate this research, a benchmark platform (dataset and evaluation baselines) is created.
- A large scale ultra-fine-grained (UFG) image dataset, which contains 47,114 images from 3,526 categories, is structurally developed to enable the exploration and evaluation for the ultra-FGVC tasks that is intended for developing algorithms to achieve accuracies beyond the capability of human.
- An extensive evaluation of recent state-of-the-art classification methods is performed, which serves as a baseline platform together with the dataset for the future advances towards addressing the challenging ultra-FGVC problem.

The paper is organized as follows. Section 2 reviews prior research on the ultra-fine-grained visual categorization and related techniques. Section 3 describes the proposed UFG image dataset. An extensive evaluation of current classification methods on the UFG image dataset is presented as baselines of the dataset in Section 4. Finally Section 5 concludes the paper.

2. Related Work

Ultra-Fine-Grained Visual Categorization. Ultra-fine-grained visual categorization aims for classifying objects at a very fine granularity where even human experts may fail to identify or describe the visual difference. This challenging research topic has recently gained traction for its better than human categorization capability and significant potential in artificial intelligence agriculture and smart farming,

Table 1. Comparison with existing ultra-fine-grained leaf datasets. The "Object Only" indicates if the image contains nothing but the object

Dataset	# of Categories	# of Images	Object Only	Publicly Available	Year
Soybean Leaf dataset [35]	3	422	Yes	No	2014
Grapevine Leaf dataset [15]	16	282	Yes	No	2018
SoyCultivarVein dataset [63]	100	600	Yes	Yes	2019
UFG image dataset (Ours)	3,526	47,114	Yes	Yes	2021

e.g., identifying the plant cultivars (a sub-class of species) via leaf images [36, 35, 63, 65]. As a classification task, the ultra-fine-grained visual categorization shares the same goal with the popular fine-grained visual categorization, *i.e.*, discriminating between large numbers of ultra-fine-grained categories with small numbers of training samples in those categories. However, very few research has been done on the ultra-fine-grained visual categorization, mainly due to the absence of ultra-fine-grained image datasets. There are only a few ultra-fine-grained image datasets as listed in Table 1. The first exploration of the ultra-fine-grained visual categorization was reported in [35, 36], where 422 leaf images from three different soybean cultivars were collected for classification. Their proposed vein-trait based classification method delivered an average classification accuracy ranging from 55.04% to 58.76%, significantly outperforming the performance obtained by human experts (41.56%). Despite the small number of cultivars included in the evaluation, their work showed that the ultra-fine-grained visual categorization (cultivar-level leaf classification) algorithms achieved better accuracies beyond the capability of human. Another exploration was performed on a grapevine leaf dataset [15], which contains 282 leaf images from 16 different grapevine cultivars. Both the two datasets, however, are not made publicly available. A SoyCultivarVein dataset [63] covering 600 leaf images from 100 different soybean cultivars was recently released. The classification on the SoyCultivarVein dataset was then formally defined as an ultra-fine-grained visual categorization task [65]. Classifying objects at such a fine granularity, inevitably brings a significant challenge to current methods.

Despite these significant pioneering efforts in the ultra-fine-grained visual categorization, existing datasets comprise only a few categories, making them insufficient to provide a practical and comprehensive test bed for baseline evaluations. The tremendous success of recent deep convolutional neural network methods in various areas of computer vision heavily relies on the development of large-scale image datasets. Consequently, recently published dataset papers [25, 3, 52, 37, 23, 19, 18, 27, 58, 47, 1, 45, 16, 72] have devoted to the dataset characterization, baseline establishment, and future inspiration, rather than delivering a specific method that is particular (or only) suitable for the proposed dataset. As such, to enable further research on the ultra-fine-grained visual categorization, we establish a large-scale ultra-fine-grained image dataset and evaluation baselines.

Fine-Grained Classification Methods. Fine-grained visual categorization has attracted extensive attention in the recent decade [20, 11, 13, 62, 7, 61, 57, 60, 34, 9, 53, 69, 4, 29, 10, 42, 28, 67, 43, 2]. The release of various datasets [54, 26, 53, 58, 60, 46] has played a key role in driving such a great progress. The taxonomy system of these datasets enables the classification down to a fine granularity, *e.g.*, species level. Consequently, various methods have been developed to address the species-level classification tasks, such as bird species classification [54], insect pest recognition [58], and plant species leaf classification [32], making species classification a rich and mature research field.

Ultra-fine-grained visual categorization (ultra-FGVC) originates from the popular fine-grained visual categorization (FGVC) but differs significantly from FGVC regarding the granularity of taxonomy system. While the FGVC deals with images mostly relying on the human-observed prior knowledge, the ultra-FGVC covers the images that even human experts may fail to identify or depict their differences. Given the wealth of fine-grained classification methods and the absence of the ultra-fine-grained research, we naturally follow and investigate several popular algorithms that have been proven effective in fine-grained species classification tasks. To evaluate a newly proposed fine-grained image dataset, a common strategy is to establish baselines using the deep learning methods, *e.g.*, Alexnet [31], VGG-16 [48] and ResNet-50 [22] for their practicality and effectiveness [58].

Since the subtle inter-class differences often reside in the unique properties of object parts, localizing informative sub-regions or parts is therefore considered as a promising solution [42, 24, 14, 70, 61]. By introducing extra part landmark annotations [42, 24] or additional network structure [14, 70], these methods have obtained significant performance improvements in fine-grained classification tasks. To enable a more practical solution, Yang *et al.* [61] introduced a self-supervision mechanism, which can effectively localize informative regions without requiring extra annotations of parts or key areas.

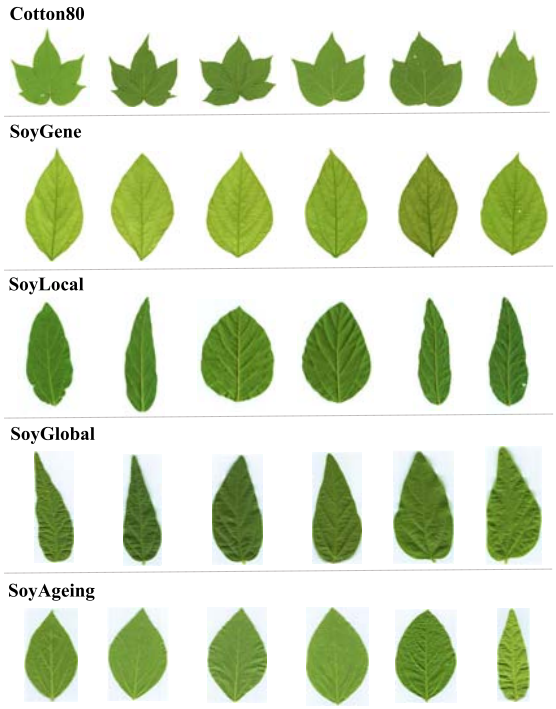

Figure 2. An overview of the proposed UFG image dataset. Each row shows images from a subset of the UFG image dataset. Each image represents a category in its associated subset.

Another line of research focuses on improving feature representations by employing second order statistics and covariance based representation [56, 39, 40]. Lin and Maji [41] introduced statistics normalization methods such that an architecture is able to capture second-order statistics of convolutional features in a translationally invariant manner. Li *et al.* [38] proposed an iterative matrix square root normalization method for fast end-to-end training of global covariance pooling networks. Despite the increase of feature dimension, these methods embedded the second order statistics in the feature level, and achieved state-of-the-art performances in various fine-grained classification tasks.

However, the number of images per category could be very small in the ultra-fine-grained visual categorization, leading to another challenge to the current classification methods, *i.e.*, how to avoid the overfitting problem. Many attempts have been made to mitigate this problem [50, 55, 51]. Among them, a data augmentation method [33] was introduced to randomly erase patches of given images and has been proven very effective in avoiding the overfitting. More recently, Chen *et al.* [7] introduced an effective Destruction and Construction (DCL) method for the fine-grained classification tasks. In their design, the input images are first partitioned into local regions and then shuffled via a region confusion mechanism. Their method effectively drives the learning within those discriminative regions for spotting the subtle inter-class differences.

Despite the tremendous success made by the classification methods in various fine-grained visual categorization tasks, it remains unclear how these methods perform on the ultra-fine-grained visual categorization tasks. This also motivates us to introduce the UFG image dataset that facilitates a thorough evaluation of current fine-grained methods on the ultra-fine-grained visual categorization tasks.

3. Dataset Collection and Design

In this section, we describe the image collection, design consideration, taxonomic system and insight analysis of the proposed UFG image dataset. An overview of the UFG image dataset is shown in Fig. 2.

3.1. Image Collection

Previous explorations have already shown the extreme difficulty of ultra-fine-grained visual categorization for both human and machine learning methods. However, only a small number of categories are involved (no more than 100 categories) in these works, impeding further progress towards a more practical and large-scale ultra-fine-grained visual categorization. The absence of large-scale datasets is mainly due to the fact that data collection for the ultra-fine-grained visual categorization is extremely difficult. Unlike general or fine-grained image datasets, which can be collected via online resources and annotated by human experts or volunteers, large-scale ultra-fine-grained images are difficult to be collected and accurately annotated via online resources. This is mostly because the annotation via human (experts or volunteers) observation is almost impossible for the ultra-fine-grained images (see Fig. 1).

To that end, we consider two main objectives in the image collection process. The first is to cover as many categories as possible. The second is to collect images in a clear and complete manner, as well as to avoid including noisy and complex backgrounds. This is because the ultra-FGVC remains an early stage and extremely challenging research problem, and thus should avoid extra noises in the current stage. In the collection, each of the plants has been provided with a confirmed cultivar name attached to the seed obtained from the genetic resource bank. Using such a breeding based label ensures the annotation is accurate, and can avoid the human observation bias or errors. And more importantly, it opens the door for researchers to develop visual categorization algorithms that can be more accurate than human. These picked leaves are then sent to be scanned in a laboratory environment, using an EPSON perfection V850 Pro scanner with reflective scanning mode, a resolution of 600 DPI and 48 bit true color setting. While we can not completely avoid any possible noise, such as the bug holes or dirty dots on the leaves, the collection process to the fullest extent ensured that the images meet the requirement of the ultra-fine-grained visual categorization.

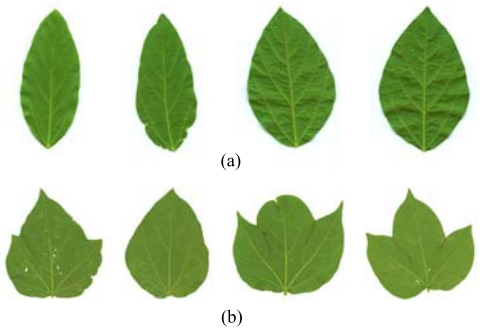

Figure 3. An example of illustrating the large intra-class variations in the UFG image dataset. (a) Four leaf images from the same category in SoyLocal subset. (b) Four leaf images from the same category in Cotton80 subset.

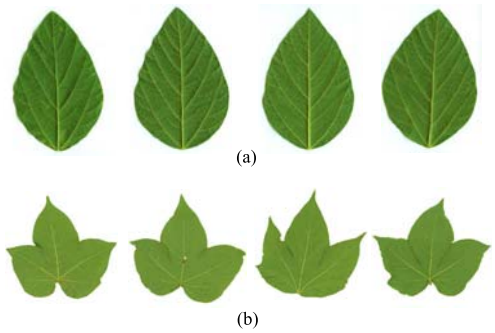

Figure 4. An example of illustrating the small inter-class variations in the UFG image dataset. (a) Four leaf images from four different categories in SoyLocal subset. (b) Four leaf images from four different categories in Cotton80 subset.

3.2. Dataset Structure

In this section, we describe the composition of the proposed UFG image dataset. According to the number of samples per category, we split the UFG image dataset into large-sample subsets and small-sample subsets. The first group covers the subsets that have more than 20 images per category, *i.e.*, SoyAgeing subset and SoyGene subset. The second group, on the contrary, is composed of subsets that have less than 20 images per category, *i.e.*, SoyLocal subset, SoyGlobal subset and Cotton80 subset.

Large-sample subsets: (1) SoyAgeing Subset contains the soybean leaves of five reproductive stages. There are 198 different categories in this subset. Each category contains leaf images from the above 5 reproductive stages. In each reproductive stage, there are 10 leaf images. Thus, each category consists of $5 \times 10 = 50$ leaf images. The total number of the leaf images is then $198 \times 50 = 9,900$. A more detailed introduction of this dataset is discussed in Supplementary File (Section 1). **(2)** SoyGene Subset covers 23,906 leaf images in total, from 1,110 different soybean categories. The number of images in each category ranges from 6 to 27.

Small-sample Subsets: (3) SoyGlobal Subset contains 1,938 different soybean categories with 6 leaf images per category, resulting in a total number of 11,628 images. **(4)** SoyLocal Subset consists of 1,200 images of 200 categories, with 6 images per category. **(5)** Cotton80 Subset includes 480 images of 80 categories, with 6 images per category. The source leaves of images in this subset are all from one plant species, cotton.

3.3. Challenges

We present an insightful analysis of the challenges brought by the proposed UFG image dataset.

Large-sample Subsets. The large-sample subsets improve existing ultra-fine-grained image datasets regarding the number of images per category and the number of categories. However, it remains a challenging dataset for current classification methods, due to the fact that very limited prior knowledge about ultra-fine-grained visual categorization is available to drive the learning.

Small-sample Subsets. The categories in the SoyCultivaLocal subset share closely related genotypes which may lead to similar phenotypes, *e.g.*, leaf shapes, among different categories. The main challenge of Cotton80 subset lies in the fact that cotton leaves commonly have the palm-like shapes, which may easily have a self-overlapping problem. The SoyGlobal subset covers a huge number of categories, making the classifier training a difficult task. Moreover, these subsets contain a small number of samples per category, which is more likely to suffer from overfitting problem.

In addition, a significant challenge exists in all the subsets of the UFG image dataset, *i.e.*, the large intra-class variations as well as the small inter-class variations. An example of illustrating the large intra-class variations and small inter-class variations is given in Fig. 3 and Fig. 4, respectively. In Section 4, we experimentally show that the UFG image dataset poses a new challenging problem for the state-of-the-art deep learning methods and the fine-grained classification techniques.

Relation to CIFAR-10 and iNaturalist Datasets. General object classification tasks such as CIFAR-10 [30] aim to categorize objects where inter-class variances are visually obvious and relatively large to distinguish. The recently developed iNaturalist2019 [53] includes 1,010 species from 72 genera (super-class of species) with 10-38 species per genus. It is a dataset for species or genera classification that belongs to Fine-Grained Visual Categorization (see top row in Fig. 1). In contrast, the proposed UFG image dataset further moves down the taxonomy granularity from species to cultivar (see bottom row in Fig. 1) and covers to date the largest number of cultivars (*e.g.* 3,446 within a single species). Fig 5 shows the difference and relation among

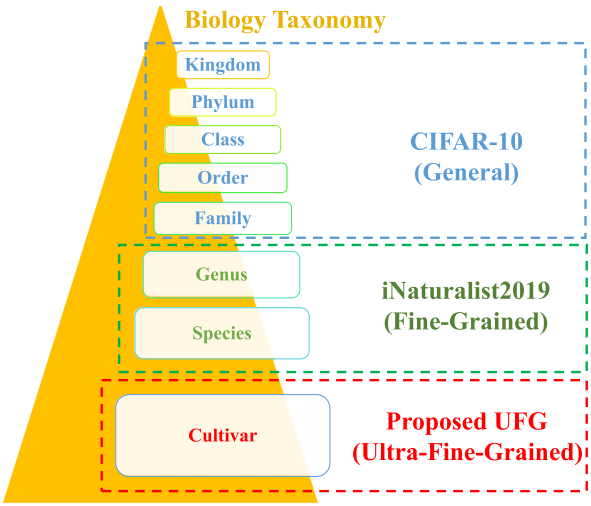

Figure 5. An example of illustrating the difference and relation among CIFAR-10 dataset, iNaturalist2019 dataset and the proposed UFG dataset.

CIFAR-10 dataset (for general object classification), iNaturalist2019 (for fine-grained object classification) and the proposed UFG dataset (for ultra-fine-grained object classification).

4. Experimental Analysis

We perform an extensive evaluation of multiple state-of-the-art methods for the ultra-fine-grained visual categorization on the proposed UFG image dataset. These methods serve as the associated baselines of this dataset for future work. Specifically, we first introduce the adopted baselines and their implementation details. Then we report an evaluation on the large-sample subsets and small-sample subsets. Finally, we present a discussion of the evaluation.

Table 2. The evaluation statistics of the UFG image dataset.

Datasets	# of Categories	Training	Test
SoyAgeing	198	4,950	4,950
SoyGene	1,110	12,763	11,143
Cotton80	80	240	240
SoyLocal	200	600	600
SoyGlobal	1,938	5,814	5,814

4.1. Baselines and Implementation

A statistics overview of the five subsets from the UFG image dataset is listed in Table 2. For the SoyAgeing subset and small-sample subsets, we use a setup where half of the images in each category are randomly selected as the training set, and the remaining images form the test set. As the SoyGene subset may contain odd-number images in some categories, we equally split the images in each category into training set and test set, and assign the extra images (if existed) to the training set. An extra experiment of validating the effect of increasing training samples is discussed in Section 4.4.

Table 3. The classification accuracy on the Cotton80 and SoyLocal subsets.

Method	Top 1 Accuracy (%)					
	SoyLocal			Cotton80		
	#3	#10	#50	#3	#10	#50
Human Expert	63.89	32.50	12.84	66.67	63.33	14.00
Alexnet [31]	88.89	53.33	33.33	88.89	66.67	24.67
VGG-16 [48]	77.78	50.00	56.00	77.78	56.67	53.33
ResNet-50 [22]	88.89	63.33	62.00	88.89	66.67	52.67
DCL [7]	88.89	63.33	60.67	77.78	73.33	54.67

We include in total 13 state-of-the-art classification methods as the baselines of the proposed UFG image dataset. The baselines are broadly categorized into two groups. One group covers the state-of-the-art deep learning methods including Alexnet [31], VGG-16 [48], ResNet-50 [22]. The other group is composed of 10 state-of-the-art fine-grained classification methods: SimCLR [5], MoCo v2 [6], BYOL [17], Cutout [12], Hide and Seek [49], ADL [8], Cutmix [66], fast-MPN-COV [38], DCL [7], and MaskCOV [64].

The baselines are implemented in the Pytorch framework. To keep the aspect ratio of the original object shapes, the training images are padded to square before being resized to the size of 440×440, and then randomly cropped to the size of 384×384. In the inference stage, the images are directly resized to 384×384.

The deep learning baselines are trained for 160 epochs using SGD with a batch size of 16. The learning rate is 0.001 initially and then decreases by a factor of 10 every 60 epochs. A more detailed implementation is given in Supplementary File (Section 2). Top-1 accuracy is used as the metric for classification evaluation.

4.2. Human Experts v.s. Deep Learning

Table 3 gives the classification accuracies from human experts and deep learning methods. Following [35, 36], we select the first three cultivars in SoyLocal and Cotton80 subsets to perform an evaluation of deep learning methods and human experts. The number of cultivars included is then increased to 10 and 50 respectively for a comprehensive evaluation. The human expert results are average results contributed from four soybean breeding experts and one cotton breeding experts.

Table 4. The classification accuracy on the Cotton80 (Cotton.), SoyLocal (Soy.Loc.), SoyGene (Soy.Gene), SoyGlobal (Soy.Glo.), SoyAgeing (Soy.Age.) datasets. For Soy.Age. dataset, the average classification accuracy on its five subsets (R1-R6) is reported.

Method	Backbone	Top 1 Accuracy (%)				
		Cotton.	Soy.Loc.	Soy.Gene	Soy.Glo.	Soy.Age.
Alexnet [31]	Alexnet	22.92	19.50	13.12	13.21	44.93
VGG-16 [48]	VGG-16	50.83	39.33	63.54	45.17	70.44
ResNet-50 [22]	ResNet-50	52.50	38.83	70.21	25.59	67.15
SimCLR [5]	ResNet-50	51.67	37.33	62.68	42.54	64.73
MoCo v2 [6]	ResNet-50	45.00	32.67	56.49	29.26	59.13
BYOL [17]	ResNet-50	52.92	33.17	60.65	41.35	64.75
Cutout [12]	ResNet-50	54.58	37.67	61.12	47.06	65.70
Hide and Seek [49]	ResNet-50	48.33	28.00	61.27	23.74	60.48
ADL [8]	ResNet-50	43.75	34.67	55.19	39.35	61.70
Cutmix [66]	ResNet-50	45.00	26.33	66.39	30.31	62.68
fast-MPN-COV [38]	ResNet-50	50.00	38.17	45.26	11.39	63.66
DCL [7]	ResNet-50	53.75	45.33	71.41	42.21	73.19
MaskCOV [64]	ResNet-50	58.75	46.17	73.57	50.28	75.86

4.3. Evaluation on Large-Sample Subsets

Table 4 shows the performance of the baseline methods on the SoyAgeing and SoyGene subsets. In addition to the average accuracies of the five stages, we also report the classification accuracy obtained on each stage in Supplementary File (Section 3 and Table 2). We observe that VGG-16 outperforms the other two deep learning methods, and the DCL achieves the best results among all the baselines. For the SoyGene subset, 11 out of 13 competing baselines achieve classification accuracies higher than 50%. Given that the SoyGene subset contains more than 1,000 categories, the baseline results seems to be very promising, which also confirms the possibility of addressing the ultra-fine-grained visual categorization.

The overall classification accuracies obtained on the large-sample subsets are very encouraging comparing to the accuracy of leading classification methods or human experts on previous ultra-fine-grained image datasets. This is possibly due to (1) a larger number of samples covered in the large-sample subsets which enables a more comprehensive learning process, and (2) the advancement of current data-driven deep learning techniques, especially the fine-grained classification methods.

4.4. Evaluation on Small-Sample Subsets

Table 4 shows the classification accuracy results of the baselines on the SoyGlobal subset. The Cutout achieves the best performance, outperforming the other competing baselines including the fine-grained classification methods. With such a large number (1,938) of categories and small number (6) of images per category, it makes sense that all the baselines obtain low classification accuracies (lower than 45.17%) on this challenging subset.

On the SoyLocal subset, MaskCOV and DCL obtain the best and second best performances respectively among all the baselines. Recall that the major challenge of SoyLocal subset is the very small inter-class variations (as stated in Section 3). This superior performance seems reasonable, given that both methods are designed to focus more on those local and subtle inter-class differences (as discussed in Section 2). For the Cotton80 subset, despite that the number of categories in this subset is only 80, the baselines achieve classification accuracies lower than 60%. This is partly due to the large intra-class variations caused by the self-overlapping of cotton leaf images.

4.5. Discussion

A more detailed evaluation is provided in Supplementary File (Section 3 and Tables 1&2). Throughout the extensive evaluation, we observe several interesting phenomenons that are discussed as follows. First, there is a large performance gap between the large-sample subsets and the small-sample subsets. Comparing the SoyGene subset and the SoyLocal subset, the baselines consistently achieve a higher classification accuracy on the SoyGene subset, which covers more training samples per category. For example, DCL [7] delivers 71.41% average classification accuracy on the SoyGene subset, which is 26.08% higher than that obtained on the SoyLocal subset. Given that the two subsets may

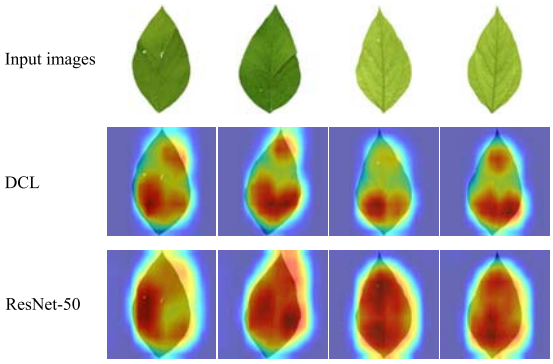

Figure 6. Examples of the Classification Activation Map (CAM) of DCL and ResNet-50 on four images from one category.

have different characteristics, the number of samples per category may not be the only factor to produce such a performance gap. Therefore, we evaluate the Cotton80 subset with two different protocols, *i.e.*, adjusting the ratio of training samples and test samples to 3:1 and 5:1, respectively. We report the results of VGG-16, ResNet-50, and DCL under the two new settings in Table 5. The performances of the baselines consistently improve as the training number increases. Therefore, it might be reasonable to consider improving data augmentation or other techniques that can increase training samples to partly address the ultra-FGVC task. Besides, we also report fine-grained classification results on the proposed UFG dataset by identifying objects at a species level, *i.e.*, differentiating cotton and soybean. We observe that both ResNet-50 and DCL can achieve 100% of classification accuracy on this fine-grained classification task. More details are provided in Section 4 in the Supplementary File.

Second, the fine-grained DCL method achieves competitive performance over the other competing baselines on all the five subsets. To provide a more comprehensive comparison, we adopt the Classification Activation Map (CAM) [71] to visualize the associated feature maps. Examples of the CAMs are given in Fig. 6. We observe that, compared with the ResNet-50, the DCL focuses in a more consistent and compact manner on local regions among the leaf images from the same category. This indicates that localizing local discriminative parts or regions might be a promising solution to the ultra-FGVC task.

5. Conclusion

Ultra-fine-grained visual categorization (ultra-FGVC) classifies objects further down the granularity of category to the sub-class of those in current fine-grained visual categorization (FGVC) tasks. Despite the promising and even saturated performance obtained by current state-of-the-art deep learning methods in the FGVC tasks, it remains unclear how these methods perform on the ultra-FGVC tasks where the human-observed prior knowledge is no longer available. The research on the ultra-FGVC has been heavily impeded due to the absence of large-scale ultra-fine-grained image datasets.

Table 5. The classification accuracy on the Cotton80 subset with different ratio of training samples and test samples.

Method	Ratio (Training : Test)	
	3:1	5:1
VGG-16 [48]	46.25	58.75
ResNet-50 [22]	46.25	57.50
DCL [7]	53.75	63.75

To fill this gap, we introduced the UFG image dataset, which has covered so far the largest scale data in the ultra-FGVC field. In contrast to existing fine-grained image datasets, UFG image dataset has a unique annotation system, *i.e.*, categorizing images with breeding based labels of the seeds from the genetic resource bank, making it a benchmark platform for developing machine learning methods whose performances are not limited by human-observed labeling accuracy. As such, the UFG image dataset opens up possibilities to advance the research of visual categorization from approaching human performance to beyond human capability, rather than serving merely as a challenging benchmark platform for visual categorization.

The UFG image dataset has offered two large-sample subsets and three small-sample subsets, depending on the number of samples per category. We have included in total 13 state-of-the-art classification methods as the baselines of the UFG image dataset. The evaluation on the large-sample subsets have shown very encouraging performances compared to those of leading methods or human experts on previous ultra-fine-grained image datasets, indicating the possibility of addressing the challenging ultra-FGVC tasks. In contrast, the performances obtained on the small-sample subsets are far from saturated, highlighting the difficulty in addressing the overfitting problem often accompanied with the ultra-FGVC. It is observed that localizing subtle discriminative parts or regions has shown signs of importance in addressing the ultra-FGVC. Given the extensive data and accompanying annotations available for analysis and benchmarking, the UFG image dataset may act as a cornerstone in research into ultra-FGVC. Evidently, it would also serve as a precious resource for potential applications in computer vision, pattern analysis, and artificial intelligence agriculture.

References

[1] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1707–1720, 2016. 3

[2] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Int. Conf. Comput. Vis.*, 2019. 3

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9592–9600, 2019. 3

[4] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Int. Conf. Comput. Vis.*, pages 511–520, 2017. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn.*, pages 1597–1607, 2020. 6, 7

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6, 7

[7] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5157–5166, 2019. 3, 4, 6, 7, 8

[8] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2219–2228, 2019. 6, 7

[9] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4109–4118, 2018. 3

[10] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2930, 2017. 3

[11] J. Deng, J. Krause, M. Stark, and L. Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):666–676, 2016. 3

[12] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6, 7

[13] Melih Engin, Lei Wang, Luping Zhou, and Xinwang Liu. Deepkspd: Learning kernel-matrix-based spd representation for fine-grained image recognition. In *Eur. Conf. Comput. Vis.*, pages 612–627, 2018. 3

[14] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4438–4446, 2017. 3

[15] Sigfredo Fuentes, Esther Hernández-Montes, JM Escalona, Josefina Bota, C Gonzalez Viejo, Carlos Poblete-Echeverría, E Tongson, and H Medrano. Automated grapevine cultivar classification based on machine learning using leaf morpho-colorimetry, fractal dimension and near-infrared spectroscopy parameters. *Computers and Electronics in Agriculture*, 151:311–318, 2018. 1, 3

[16] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(3):523–536, 2019. 3

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 6, 7

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5356–5364, 2019. 3

[19] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 939–948, 2019. 3

[20] Junwei Han, Xiwen Yao, Gong Cheng, Xiaoxu Feng, and Dong Xu. P-cnn: Part-based convolutional neural networks for fine-grained visual categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Int. Conf. Comput. Vis.*, pages 1026–1034, 2015. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 3, 6, 7, 8

[23] Mahdi S Hosseini, Lyndon Chan, Gabriel Tse, Michael Tang, Jun Deng, Sajad Norouzi, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11747–11756, 2019. 3

[24] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1173–1182, 2016. 1, 3

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6700–6709, 2019. 3

[26] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshop on Fine-Grained Visual Categorization*, volume 2, 2011. 2, 3

[27] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa,

Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9601–9611, 2019. 3

[28] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 365–374, 2017. 3

[29] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5546–5555, 2015. 3

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, pages 1097–1105, 2012. 3, 6, 7

[32] Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, and João V. B. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Eur. Conf. Comput. Vis.*, pages 502–516, 2012. 2, 3

[33] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3524–3533, 2017. 4

[34] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2520–2529, 2017. 3

[35] Mónica G Larese, Ariel E Bayá, Roque M Craviotto, Miriam R Arango, Carina Gallo, and Pablo M Granitto. Multiscale recognition of legume varieties based on leaf venation images. *Expert Systems with Applications*, 41(10):4638–4647, 2014. 1, 3, 6

[36] Mónica G Larese, Rafael Namías, Roque M Craviotto, Miriam R Arango, Carina Gallo, and Pablo M Granitto. Automatic classification of legumes using leaf vein image features. *Pattern Recognition*, 47(1):158–168, 2014. 1, 3, 6

[37] Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. A cross-season correspondence dataset for robust semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9532–9542, 2019. 3

[38] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 947–955, 2018. 4, 6, 7

[39] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *Int. Conf. Comput. Vis.*, pages 2070–2078, 2017. 4

[40] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. Distribution consistency based covariance metric networks for few-shot learning. In *AAAI*, 2019. 4

[41] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. In *Brit. Mach. Vis. Conf.*, 2017. 4

[42] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Int. Conf. Comput. Vis.*, pages 1449–1457, 2015. 3

[43] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S. Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *Int. Conf. Comput. Vis.*, 2019. 3

[44] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. 2

[45] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2019. 3

[46] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 1447–1454. IEEE, 2006. 2, 3

[47] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):271–284, 2019. 3

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 6, 7, 8

[49] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Int. Conf. Comput. Vis.*, pages 3544–3553. IEEE, 2017. 6, 7

[50] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 4

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–9, 2015. 4

[52] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1207–1216, 2019. 3

[53] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8769–8778, 2018. 2, 3, 5

[54] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 3

[55] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using drop-

connect. In *Proc. Int. Conf. Mach. Learn.*, pages 1058–1066, 2013. 4

[56] Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, and Wanqing Li. Beyond covariance: Feature representation with nonlinear kernel matrices. In *Int. Conf. Comput. Vis.*, pages 4570–4578, 2015. 4

[57] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 76:704–714, 2018. 3

[58] Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, and Jufeng Yang. Ip102: A large-scale benchmark dataset for insect pest recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 3

[59] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(5):1100–1113, 2016. 1

[60] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3973–3981, 2015. 2, 3

[61] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Eur. Conf. Comput. Vis.*, pages 420–435, 2018. 1, 3

[62] Jun Yu, Min Tan, Hongyuan Zhang, Dacheng Tao, and Yong Rui. Hierarchical deep click feature prediction for fine-grained image recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 3

[63] Xiaohan Yu, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Multiscale contour steered region integral and its application for cultivar classification. *IEEE Access*, 7:69087–69100, 2019. 1, 3

[64] Xiaohan Yu, Yang Zhao, Yongsheng Gao, and Shengwu Xiong. Maskcov: A random mask covariance network for ultra-fine-grained visual categorization. *Pattern Recognition*, 119:108067, 2021. 6, 7

[65] Xiaohan Yu, Yang Zhao, Yongsheng Gao, Shengwu Xiong, and Xiaohui Yuan. Patchy image structure classification using multi-orientation region transform. In *AAAI*, 2020. 1, 3

[66] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Int. Conf. Comput. Vis.*, pages 6023–6032, 2019. 6, 7

[67] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Int. Conf. Comput. Vis.*, 2019. 3

[68] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Eur. Conf. Comput. Vis.*, pages 834–849. Springer, 2014. 1

[69] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimedia*, 19(6):1245–1256, 2017. 3

[70] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5012–5021, 2019. 3

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2929, 2016. 8

[72] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. 3