Supplementary File for “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

1. SoyAgeing Dataset

The SoyCultivarAgeing subset contains the soybean cultivar leaves of five Reproductive Stages. The five stages are defined by [6] as **R1**: beginning bloom; open flower at any node on the main stem; **R3**: beginning pod; any pod that is 3/16 inch long and is on one of the four uppermost nodes of the main stem; **R4**: full pod; a 3/4 inch pod at one of the four uppermost nodes on the main stem; **R5**: beginning seed; seed is 1/8 inch long in a pod at one of four uppermost nodes on the main stem; **R6**: full seed; a pod containing a green seed that fills the pod capacity is located at one of the four uppermost main stem nodes.

There are 198 different soybean cultivars in this subset. Each cultivar contains leaf images from the above 5 reproductive stages. In each reproductive stage, there are 10 leaf images. Thus, each cultivar consists of $5 \times 10 = 50$ leaf images. The total number of the leaf images is then $198 \times 50 = 9,900$.

An overview of the UFG image dataset is shown in Fig. 1.

2. Implementation Details

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. The fine-grained baselines are implemented as reported in their papers with carefully fine-tuning. Specifically, for Alexnet [10], VGG-16 [13], ResNet-50 [8] and DCL [3], the batch size is set to 16, the learning rate is 0.001 with a learning rate decay of 10 for each 60 epochs, and SGD is used as the optimiser. Fast-MPN-COV [11] is trained using SGD with a learning rate of 0.0012, a weight decay of 0.001 and a batch size of 10. For MaskCOV [16], the batch size is 8 for Cotton80 subset and 16 for the remaining subsets, respectively. The learning rate is set to 0.001 with a learning rate decay of 10 for each 60 epochs and SGD is used as the optimiser. ADL [4] is evaluated with the drop threshold of 0.9, a learning rate of 0.01 and a batch size of 32. As ADL has a hyper-parameter (termed drop rate) that determines the drop rate of image content, we report two settings of ADL with drop rates of 0.25 and 0.5 respectively for a more comprehensive comparison. For Cutout [5], the learning rate is set to 0.01 with a batch size of 16. Similar to ADL, Cutout has a hyper-parameter (termed length) that controls the size of removed region, we therefore report the results with the best two settings adopted in their paper, *i.e.*, length = 8 and 16 respectively. For Hide and Seek [14], the learning rate is set to 0.001, with a learning rate decay of 10 for each 60 epochs and SGD as the optimiser. The hyper-parameter that controls the possibility of region removal is set to 0.5, with a training batch size of 16. For Cutmix [17], we adopt a training batch size of 16 and a learning rate of 0.01. Following [7], the performances of the self-supervised methods are evaluated using both linear evaluation (*i.e.*, only the classifier is optimized) and fine-tuning (*i.e.*, all the model parameters are fine-tuned). For SimCLR (fine-tuning) [1], the batch size is 32 with a learning rate of 0.1 which decreases by 10 times at 60, 80 and 100 epoch respectively. For SimCLR (linear), the initial learning rate is set to 1 and other settings remain the same. For MoCo v2 (linear) [2], the batch size is set to 16 and the learning rate is 30, which decreases by 10 times at 60, 80 and 100 epoch respectively. For MoCo v2 (fine-tuning), the learning rate ratio between the backbone and the classifier is set to 10^{-6} while the other settings remain the same. For BYOL (linear) [7], the batch size is set to 16 with a learning rate of 1, which decreases by 10 times at 60, 80 and 100 epoch respectively. For BYOL (fine-tuning), the learning rate is set to 0.1 and the other settings remain the same.

[*] Corresponding author

3. Ultra-Fine-Grained Visual Categorization

Table 1 lists the performance of all the competing methods on the four subsets of the proposed UFG dataset, *i.e.*, SoyLocal, SoyGlobal, Cotton80 and SoyGene subsets. Table 2 shows the performance of all the competing methods on the SoyAgeing subset. We observe that there remains a large room for performance improvement in the ultra-fine-grained visual categorization.

4. Fine-Grained Visual Categorization

Recall that the images from small-sample subsets can be grouped into two categories when the species-level taxonomic system is adopted (as stated in Section 3.3). Following other popular fine-grained species classification tasks [15, 9], we refer to our fine-grained visual categorization as the identification of the category at a species level, *i.e.*, differentiating cotton and soybean. The small-sample subsets are thus considered as an integrated two-category dataset, with one category containing 480 cotton leaf images and the other category covering 12,828 soybean leaf images.

We evaluate the ResNet-50 and DCL on this integrated two-category dataset for the fine-grained visual categorization. The classification accuracies of the two methods both achieve 100%, confirming the effectiveness of the baseline methods in the fine-grained visual categorization. This saturated performance is also in accordance with that obtained in other fine-grained species classification tasks, *e.g.*, bird species classification [15] (87.8%), flower species classification [12] (99.7%), and dog species classification [9] (92.1%).

In contrast, the performances of the baselines obtained in the ultra-fine-grained visual categorization are far from saturated. This indicates that the ultra-fine-grained visual categorization remains an unsolved task for current fine-grained classification methods. The proposed UFG image dataset may serve as a new challenge to develop new methods that scale to the level of ultra-fine-grained visual categorization.

References

[1] 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. 1, 4, 5

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1, 4, 5

[3] 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. 1, 4, 5

[4] 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. 1, 4, 5

[5] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1, 4, 5

[6] Walter R Fehr and Charles E Caviness. Stages of soybean development. 1977. 1

[7] 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. 1, 4, 5

[8] 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. 1, 4, 5

[9] 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

[10] 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. 1, 4, 5

[11] 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. 1, 4, 5

[12] 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

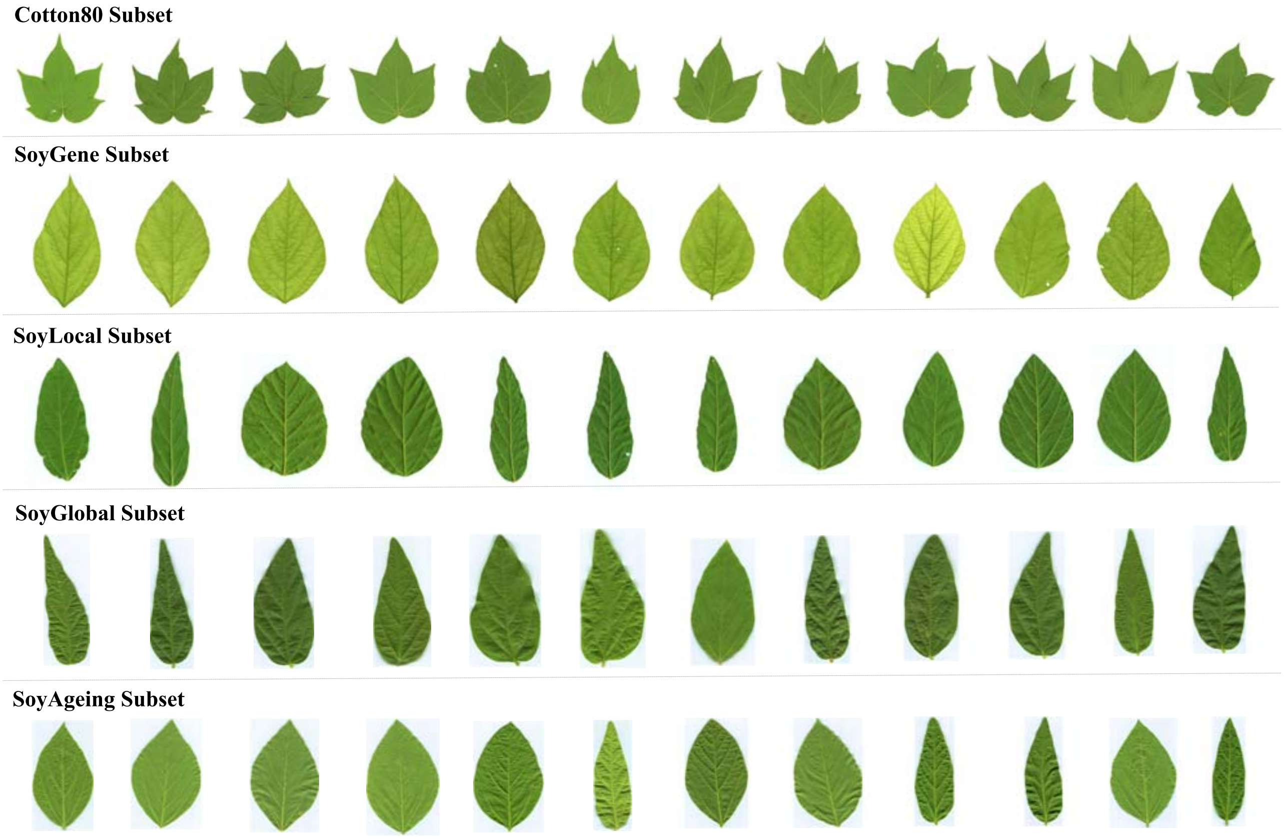
[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4, 5

[14] 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. 1, 4, 5

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

[16] 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. 1, 4, 5

[17] 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. 1, 4, 5



Cotton80 Subset

SoyGene Subset

SoyLocal Subset

SoyGlobal Subset

SoyAgeing Subset

Figure 1. An overview of the proposed UFG image dataset. Each row shows images from a subset of the UFG image dataset. Each image represents a unique cultivar (category) in its associated subset.

Table 1. The classification accuracy on the CottonCultivar80 (Cotton.), SoyCultivarLocal (Soy.Loc.), SoyCultivarGene (Soy.Gene), SoyCultivarGlobal (Soy.Glo.), SoyAgeing (Soy.Age.) datasets.

Method	Backbone	Top 1 Accuracy (%)				
		Cotton.	Soy.Loc.	Soy.Gene	Soy.Glo.	Soy.Age.
Alexnet [10]	Alexnet	22.92	19.50	13.12	13.21	44.93
VGG-16 [13]	VGG-16	50.83	39.33	63.54	45.17	70.44
ResNet-50 [8]	ResNet-50	52.50	38.83	70.21	25.59	67.15
SimCLR (linear) [1]	ResNet-50	41.25	29.17	29.62	13.48	46.18
SimCLR (fine-tuning) [1]	ResNet-50	51.67	37.33	62.68	42.54	64.73
MoCo v2 (linear) [2]	ResNet-50	30.42	27.67	26.58	12.99	38.26
MoCo v2 (fine-tuning) [2]	ResNet-50	45.00	32.67	56.49	29.26	59.13
BYOL (linear) [7]	ResNet-50	47.92	25.50	35.13	18.44	49.53
BYOL (fine-tuning) [7]	ResNet-50	52.92	33.17	60.65	41.35	64.75
Cutout (16) [5]	ResNet-50	55.83	31.67	62.46	44.65	63.68
Cutout (8) [5]	ResNet-50	54.58	37.67	61.12	47.06	65.70
Hide and Seek [14]	ResNet-50	48.33	28.00	61.27	23.74	60.48
ADL (0.25) [4]	ResNet-50	40.83	28.00	52.18	29.50	51.56
ADL (0.5) [4]	ResNet-50	43.75	34.67	55.19	39.35	61.70
Cutmix [17]	ResNet-50	45.00	26.33	66.39	30.31	62.68
fast-MPN-COV [11]	ResNet-50	50.00	38.17	45.26	11.39	63.66
DCL [3]	ResNet-50	53.75	45.33	71.41	42.21	73.19
MaskCOV [16]	ResNet-50	58.75	46.17	73.57	50.28	75.86

Table 2. The classification accuracy of the baselines on the five stages of the SoyAgeing subset. “Avg” denotes the average classification accuracy of the five subsets.

Method	Backbone	Top 1 Accuracy (%)					
		R1	R3	R4	R5	R6	Avg
Alexnet [10]	Alexnet	49.90	44.65	45.15	47.47	37.47	44.93
VGG-16 [13]	VGG-16	72.32	72.53	74.95	71.11	61.31	70.44
ResNet-50 [8]	ResNet-50	70.00	64.24	74.04	72.63	54.85	67.15
SimCLR (linear) [1]	ResNet-50	53.64	45.66	45.35	50.40	35.86	46.18
SimCLR (fine-tuning) [1]	ResNet-50	70.00	66.57	64.24	68.38	54.44	64.73
MoCo v2 (linear) [2]	ResNet-50	42.93	38.59	38.99	38.99	31.82	38.26
MoCo v2 (fine-tuning) [2]	ResNet-50	62.73	56.16	61.31	65.96	49.49	59.13
BYOL (linear) [7]	ResNet-50	55.35	48.38	50.40	49.60	43.94	49.53
BYOL (fine-tuning) [7]	ResNet-50	71.11	66.16	65.76	64.65	56.06	64.75
Cutout (16) [5]	ResNet-50	70.20	61.92	62.32	69.70	54.24	63.68
Cutout (8) [5]	ResNet-50	66.87	64.04	67.78	73.43	56.36	65.70
Hide and Seek [14]	ResNet-50	64.04	58.99	61.31	64.75	53.33	60.48
ADL (0.25) [4]	ResNet-50	53.54	54.34	55.15	52.83	41.92	51.56
ADL (0.5) [4]	ResNet-50	66.67	58.89	64.75	68.48	49.70	61.70
Cutmix [17]	ResNet-50	65.56	59.19	64.24	68.79	53.64	62.28
fast-MPN-COV [11]	ResNet-50	67.68	64.55	66.87	68.49	50.71	63.66
DCL [3]	ResNet-50	76.87	73.84	76.16	76.16	62.93	73.19
MaskCOV [16]	ResNet-50	79.80	74.65	79.60	78.28	66.97	75.86