Proceedings Paper:
Su, J., Anderson, S. and Mihaylova, L. orcid.org/0000-0001-5856-2223 (2023) Holistic self-distillation with the squeeze and excitation network for fine-grained plant pathology classification. In: 2023 26th International Conference on Information Fusion Proceedings. 2023 26th International Conference on Information Fusion (FUSION 2023), 27-30 Jun 2023, Charleston, SC, USA. Institute of Electrical and Electronics Engineers (IEEE) . ISBN 979-8-3503-1320-8

https://doi.org/10.23919/FUSION52260.2023.10224184

Holistic Self-Distillation with the Squeeze and Excitation Network for Fine-grained Plant Pathology Classification

Jingxuan Su, Sean Anderson and Lyudmila Mihaylova
Department of Automatic Control & Systems Engineering, University of Sheffield, S1 3JD, UK
Email: jsu14@sheffield.ac.uk, s.anderson@sheffield.ac.uk, l.s.mihaylova@sheffield.ac.uk

***Abstract*—Fine-grained plant pathology classification is an important task for precision agriculture, but at the same time, it is challenging due to the subtle difference in plant categories. Variances in the lighting conditions, position, and stages of disease symptoms usually lead to degradation of classification accuracy. Knowledge distillation is a popular method to improve the model performance to deal with the indistinguishable image classification problem. It aims to have a well-optimised small student network guided by a large teacher network. Existing knowledge distillation methods mainly consider training a teacher network that needs a high storage space and considerable computing resources. Self-knowledge distillation methods have been proposed to distil knowledge from the same network. Although self-knowledge distillation saves time and space compared with knowledge distillation, it only learns label knowledge. In this paper, we propose a novel self-distillation method to recognize the fine-grained plant category, which considers holistic knowledge based on the Squeeze and Excitation Network. We label this new method as holistic self-distillation because it captures knowledge through spatial features and labels. The performance validation of the proposed approach is performed on two public fine-grained plant datasets: Plant Pathology 2021 and Plant Pathology 2020 with the accuracy of 98.22% and 90.72% respectively. We also present experiments on the state-of-the-art algorithm (ResNet-50). The classification results demonstrate the effectiveness of the proposed approach with respect to accuracy.**

***Index Terms*—Knowledge distillation, Self-knowledge distillation, Fine-grained plant classification, Feature fusion, Plant pathology, Precision agriculture**

I. Introduction

Precision agriculture seeks to improve the production of plants and control environmental variations such as diseases that impact the production and quality of plant [1], [2]. Plant classification is an important technological challenge in precision agriculture [3] that aims to classify different subordinate categories under coarse large categories, e.g. plant diseases [4]. Plant classification tasks can be subdivided into coarse-grained and fine-grained [5] images. While coarse-grained image classification is interested in representing generic categories characterised with a large degree of dissimilarities, fine-grained image classification is a sub-field of object recognition that aims at representing categories with a large degree of similarity and is concerned with the problem of distinguishing between images of closely related entities, for instance, different species of plants from the same class. The focus of this paper is on fine-grained plant category classification [6] for plants. Intuitively, the fine-grained plant categories look very similar and are hard to distinguish, as shown in Fig. 1. Specifically, the inter-class variance is much smaller than the intra-class variance. Apparently, the fine-grained plant dataset increases the difficulty of classification. Moreover, the classification performance could directly affect society communities such as the farmers. The misclassification of plant diseases can lead to improper use of chemicals, to decreased yield, and potentially harming the

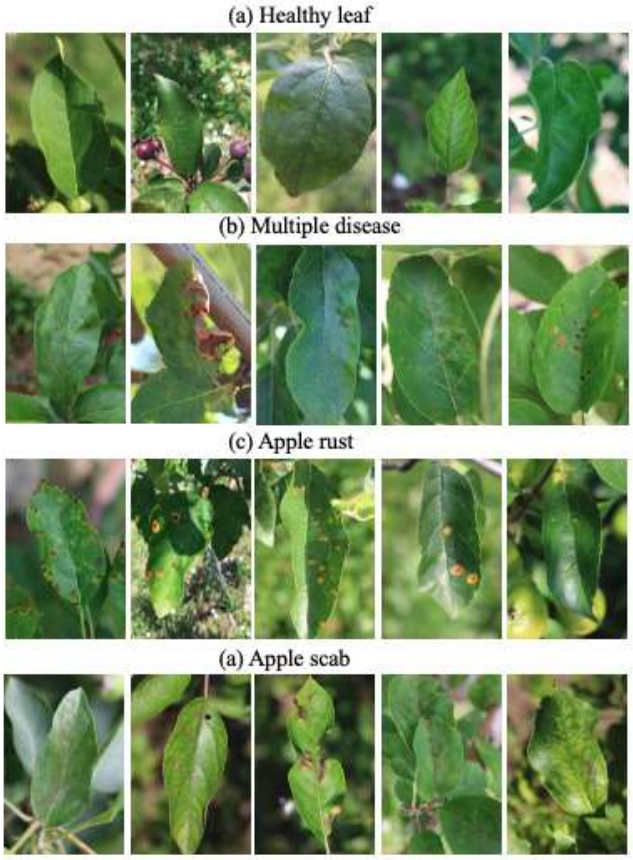

Fig. 1. Sample images from the fine-grain plant pathology datasets [4] showing the different symptoms (a) healthy leaf, (b)multiple diseases, (c) apple rust, (d) apple scab. Images show environment variation, e.g. lighting conditions, and capturing method.

entire farm [7], [8]. Currently, manual scouting based disease classification is time–consuming and expensive. While many deep learning methods have achieved remarkable success in classification [9]–[12], their application to fine-grained plant classification is still less satisfactory. This situation is even worse for great pathology variances due to genetic variations, and light conditions.

As a result, the difficulty of fine-grained plant classification comes from identifying subtle feature differences in particular regions. Residual network [13] as a state-of-the-art algorithm provides an effective architecture in general image classification. Squeeze and Excitation (SE) networks [14] have been proposed to focus on the feature details of specific regions, which won first place at the ILSVRC 2017 classification [15]. The main contribution of SE networks consists in the introduced Squeeze and Excitation (SE) block that finds the interdependencies between channels and adaptively pays attention to important features. The SE block can be stacked with any convolutional neural network, such as SE-ResNet-50, SE-Inception and others [14]. The SE network trains the binary assigned data (named hard label [16]). However, the performance of SE network may be restricted, since hard labels cannot provide sufficient feature information and the spatial features are lost in the SE block.

Knowledge distillation methods [17] aim at providing a well-optimised small student network guided by a large teacher network. The KD guides the student to learn the probability of each class (named soft labels [18], [19]) generated by the teacher network. Existing KD methods mainly consider training a teacher network that needs a high storage space and considerable computing resources. Self-KD methods [20] have been proposed to distil their own knowledge without a pretrained teacher network. These approaches help the network to enhance classification performance. However, these methods often rely on extra networks and soft labels to capture additional knowledge, which loses the spatial features.

To address these challenges in existing classification methods, we propose a novel self-distillation approach, named Holistic Self-Distillation (HSD). The proposed approach is designed to extract spatial feature information before the SE block. We demonstrate that HSD is superior to state-of-the-art (SOTA) method and other SE network approaches on plant image classification tasks. Extensive experiments on two public datasets further show the superiority of HSD in learning knowledge comprehensively from spatial feature information and soft labels. The main contributions of this work are as follows:

1) We propose Holistic Self-Distillation (HSD), a novel method to learn holistic knowledge from the teacher network through distilling feature maps and soft labels.
2) The proposed HSD method employs the Squeeze and Excitation (SE) network to integrate feature information and soft labels. It can be applied on all SE networks due to similar constructions, e.g. SE-Residual networks.
3) Extensive experiments are conducted on fine-grained publicly available plant pathology benchmark datasets to evaluate the performance of the HSD method. The efficiency of the HSD framework in providing a new direction of self-knowledge distillation is demonstrated.

The paper is organised as follows. Section II provides a brief overview of the knowledge distillation methods. Section III describes the proposed holistic self-distillation method with the Squeeze and Excitation network. The following Section IV presents the experimental results and analysis. Section V summarises the results and discusses future work.

II. Theoretical Background Knowledge

A. Knowledge Distillation

Knowledge distillation (KD) is a powerful method for network compression that includes a complex pre-trained teacher network that provides a supervisory signal to train the light student network [21], [22]. There are two main types of KD methods, logits distillation (also known as soft-label distillation or target distillation) and feature distillation. The most popular and classic work on logits distillation uses the softmax output as the soft label [17]. Specifically, the optimization of the student network is to minimize the Kullback-Leibler (KL) divergence between the soft and hard labels from teacher and student networks respectively. Unlike the logits method, feature distillation learns the feature of the middle layer [23]. Despite its compression performance for the student network, the training of the teacher network still requires time and computation resources [24].

B. Self-knowledge Distillation

Instead of training extra networks, Self-knowledge Distillation (Self KD) utilizes self-knowledge to enhance effectiveness and improve performance [18], [20]. The student network is trained by the mixed soft and hard labels. When the network uses hard labels only, the model actually loses the information of the original data. It makes the model prone to overfitting and results in a decrease in the generalization ability. Usually, soft labels will alleviate the degeneration of model generalization by providing extra knowledge, e.g. the similarity and difference between two close labels [25], [26]. There are several works to develop Self KD. A self-attention distillation method [27] uses attention maps as soft targets to rich the learning knowledge for lane detection. Snapshot distillation [28] effectively prevents under-fitting problems by increasing the difference between teacher and student. A novel Self KD was proposed that redefines the probabilities of the soft label through the training network [29]. These variant self-distillation methods [30]–[32]are all around soft labels or regularization. However, the knowledge of soft and hard labels is not enough when the teacher network goes deeper.

This paper proposes a holistic self-distillation framework with the Squeeze and Excitation network [14] for learning features and soft label knowledge. The methodology details are given in the following Section III.

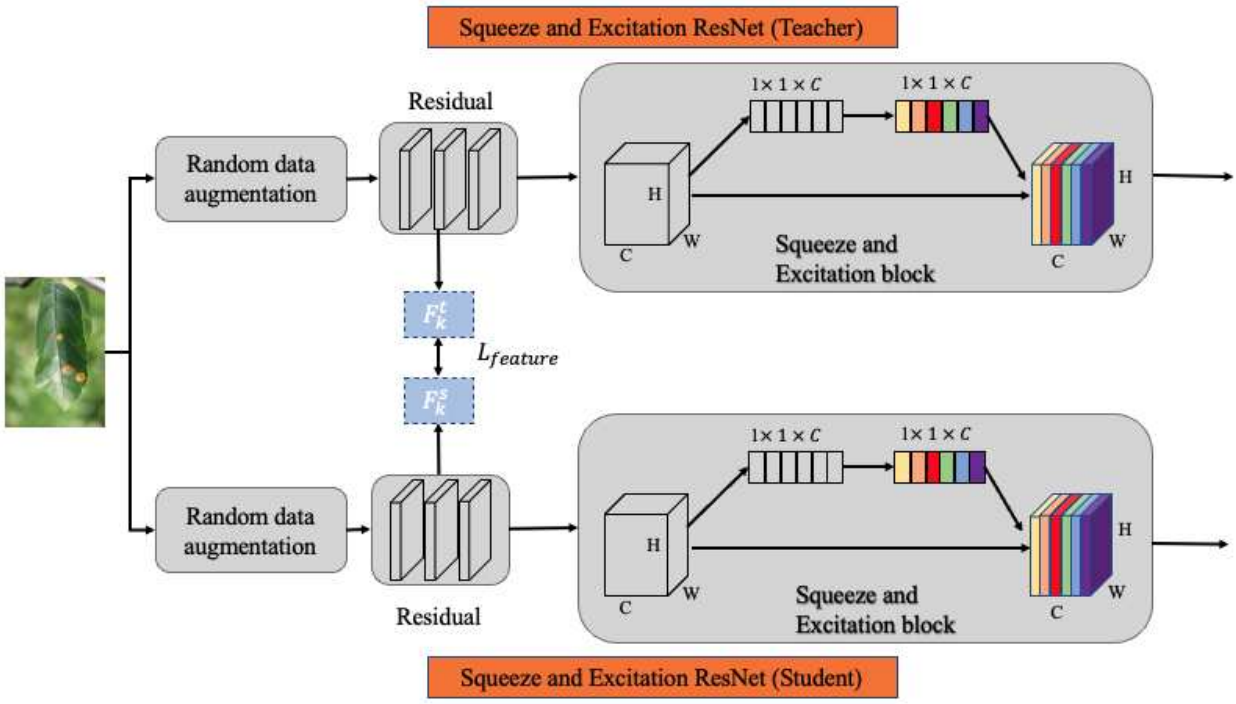

Fig. 2. The architecture of the fusion features based on Squeeze and Excitation Residual network.

III. Methodology

This section introduces briefly the classification network construction [14]. As we mentioned in Section II, the main challenge of fine-grained image classification consists in the difficulty to discriminate the subtle and local traits within the same basic level category. To handle the slight difference in the image features, many researchers developed extended strategies of knowledge distillation [17]. The holistic self-distillation is a new type of distillation method proposed in this paper, which captures knowledge through spatial features and soft labels.

A. The Structure of the Classification Network

Inspired by the significant improvements of the Squeeze and Excitation network in feature spatial encoding and classification tasks, we apply the Squeeze and Excitation Residual network 50 (SE-ResNet-50) [14], [33], [34] on the fine-grained plant classification to extract the holistic feature knowledge. It consists of two main parts, Residual framework [13] and Squeeze and Excitation block. This network captures the interdependencies between feature channels that obtain the importance of each feature channel through learning. The core idea is that useful features are promoted and the other features are suppressed. Fig. 2 shows the schematic of SE-ResNet-50 with feature maps computation. Formally, the Squeeze operation $\mathbf{F}_{sq}(\mathbf{u})$ transforms the size of feature map $H \times W \times C$ to the size of feature map $1 \times 1 \times C$, which is calculated by:

$$\mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u(i,j), \tag{1}$$

where H denotes the height, W is the width and C is the channel dimension of the feature map and u is the shrink operate.

However, the classical self-knowledge distillation focuses only on soft label knowledge distillation [17]. The student network could ignore spatial feature information. Therefore, we propose holistic self-distillation to learn the knowledge of the teacher network from both soft labels and spatial features.

B. Preliminary

Consider a batch of the K-category labelled dataset $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where N represents the number of training instance in the dataset, $\mathbf{x}_i$ is the input data and y_i is the corresponding label of $\mathbf{x}_i$.

The hard labels are fed into the Squeeze and Excitation network $H(y_i, \mathbf{p}_i)$. The cross-entropy loss function is defined as follows

$$\mathbf{L}_{CE} = \frac{1}{n} \sum_{i=1}^{n} H(y_i, \mathbf{p}_i). \tag{2}$$

The predictive distribution $\mathbf{p}_i$ is computed through the softmax layer that compares the logit $f_k(\mathbf{x}_i)$ with other logits. It is formulated as

$$\mathbf{p}_i(k) = \frac{\exp(f_k(\mathbf{x}_i)/\tau)}{\sum_{j=1}^{K} \exp(f_j(\mathbf{x}_i)/\tau)}, \tag{3}$$

where $f_k(\mathbf{x}_i)$ represents the corresponding logit of the k and the temperature constant τ is normally set to 1. Using the Kullback-Leibler (KL) divergence, it optimizes the student

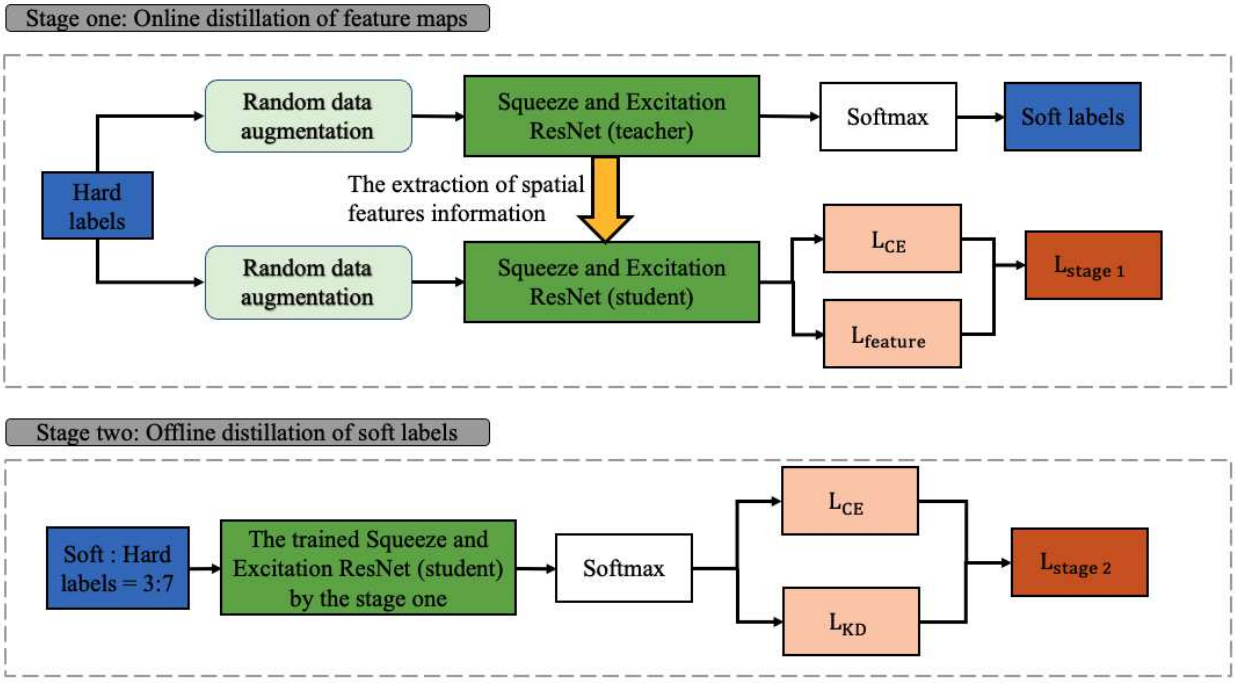

Fig. 3. The diagram of the holistic self-distillation method

network [17], which minimizes the loss between soft label $\mathbf{p}_i^t$ and $\mathbf{p}_i^s$ generated by student and teacher respectively:

$$\mathbf{L}_{KD} = \frac{1}{n}\sum_{i=1}^{n} \tau^2 \cdot D_{KL}\left(\mathbf{p}_i^s \| \mathbf{p}_i^t\right). \tag{4}$$

The next subsection presents the components of holistic self-distillation.

C. Holistic Self-Distillation

Feature maps often contain the context and spatial information of images. Instead of training mixed soft and hard labels alone, our proposed method utilizes feature map information. We encourage the student network to learn discriminative features between soft labels and hard labels. Motivated by the hint loss from FitNet [23], we consider employing the squared l_2-norm for teacher feature maps $\left\{\mathbf{F}_k^t\left(\boldsymbol{x}\right) \in \mathbb{R}^{H\times W\times C}\right\}_{k=1}^{K}$ and student feature maps $\left\{\mathbf{F}_k^s\left(\boldsymbol{x}\right) \in \mathbb{R}^{H\times W\times C}\right\}_{k=1}^{K}$. Then, we train the model by minimizing the loss function (5). Meantime, the loss function introduces feature fusion, which is defined as:

$$\mathbf{L}_{feature} = \sum_{k=1}^{K} \frac{1}{HWC}\left\|\mathbf{F}_k^t\left(\boldsymbol{x}\right) - \mathbf{F}_k^s\left(\boldsymbol{x}\right)\right\|^2. \tag{5}$$

Benefiting from the above equation, $\mathbf{L}_{feature}$ would learn the meaningful spatial feature from the different between teacher and student network. A good student network is able to learn holistic knowledge from feature fusion and probabilities of soft labels. The student network is trained to optimize two stages of loss:

$$\begin{aligned} \mathbf{L}_{stage1} &= \mathbf{L}_{CE} + \mathbf{L}_{KD}, \\ \mathbf{L}_{stage2} &= \mathbf{L}_{CE} + \mathbf{L}_{feature}. \end{aligned} \tag{6}$$

The $\mathbf{L}_{CE}$ is the cross-entropy (CE) loss between hard labels and results. In short, the Squeeze and Excitation network distil soft labels and feature maps. The Squeeze and Excitation network is trained by a new training dataset with mixed soft and hard labels. Meanwhile, the distilled feature map is involved in the loss function. The whole training process is the holistic distillation visualised in Fig. 3.

IV. Experiments and Analysis

This section presents performance evaluations and analysis of Holistic Self-Distillation (HSD) on two plant pathology fine-grained datasets.

A. Datasets and Implementation Details

The plant pathology datasets [4] are available at the Kaggle community and are a part of the Computer Vision and Pattern Recognition (CVPR) Fine-Grained Visual Categorization (FGVC) workshop 2020 and 2021. The Plant Pathology 2020 dataset contains 3,651 high-quality RGB images of four apple foliar categories: healthy, scab, rust and multiple diseases. These images are captured under different illumination, angle, surface and noise conditions (Fig. 1). The plant pathology of FGVC 2021 increased the images to the number of 23,249 and added two categories of disease powdery mildew and frog eye leaf spot.

We validate the proposed method over these two datasets. The 3,651 images of Plant Pathology 2020 are used to train the

model. The model performance is tested on the hidden dataset of the Kaggle leaderboard. The Plant Pathology 2021 dataset is divided into train and test data with a ratio of 6:4. The teacher network is essentially the same as the SE-ResNet-50. The network is used in all experiments and is pre-trained by ImageNet [35].

In the first stage, the networks of teacher and student are trained simultaneously through the same dataset with random data augmentation. We randomly apply 12 types of data augmentation, such as compose, resize, random brightness, different blur and flip etc. The networks will generate different feature maps for the same image. We train the student network by minimising feature loss. Meanwhile, the teacher network generates soft labels. In the second stage, we adopt 30% soft label and 70% hard label to train the student network that is pre-trained from stage one. The whole stage is named holistic self-distillation.

B. Performance Validation Results and Analysis

This section presents testing results for the performance evaluation of the holistic self-distillation on Plant Pathology 2020 and 2021 datasets with the SE-ResNet-50 network. All these experiments were run under the PyTorch framework over two NVIDIA Tesla K80 GPUs.

We have shown the performance of the method with different datasets in Table II. The holistic self-distillation (HSD) with the SE-ResNet-50 gives an accuracy of 98.22% in Plant Pathology 2020 and 90.72% in Plant Pathology 2021, which is a large improvement compared with the teacher model and the SOTA algorithm. The classic self-distillation reaches the same level of performance as the SE-ResNet-50. It is reasonable to assume that the improvement in results comes from the learned spatial knowledge.

Table I shows the experimental results of the HSD method for each category in Plant Pathology 2021. Three metrics [36] are applied to each category, which is computed by True Positive (TP), False Positive (TP), True Negative (TN), and False Negative (FN) as formalised in the following equations:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP,} \\ Recall &= \frac{TP}{TP+FN,} \\ F1\,score &= \frac{2 \times Precision \times Recall}{Precision + Recall}. \end{aligned} \tag{7}$$

The precision [37] indicates the predicted positive is the true positive. The recall [38] represents the correct prediction in positive samples. The F1 score finds a balance between both precision and recall. Combining the above metrics, the macro average computes the arithmetic mean of the metrics of each category. The weighted average takes into account the weight of each category [39]. Among them, the HSD achieves brilliant performance in all the categories. We also observe that the healthy category gets the best results within three metrics over 1,950 test images. The multiple diseases category is prone to be misclassified.

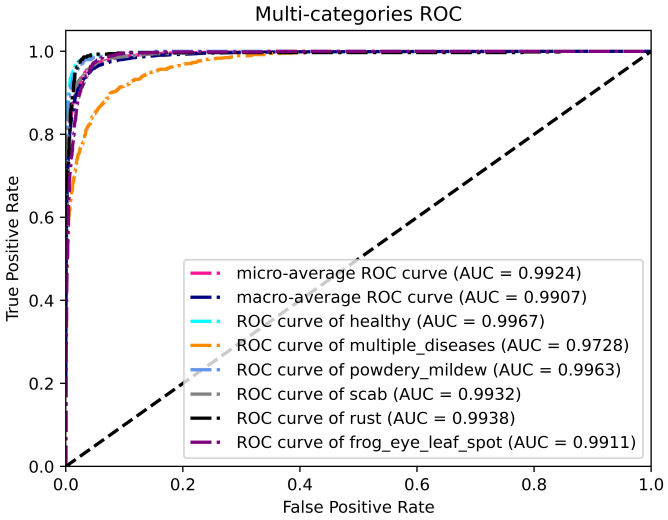

Fig. 4. The ROC curves of the Plant Pathology 2021. The AUC (Area Under Curve) is defined as the area under the ROC curve [40].

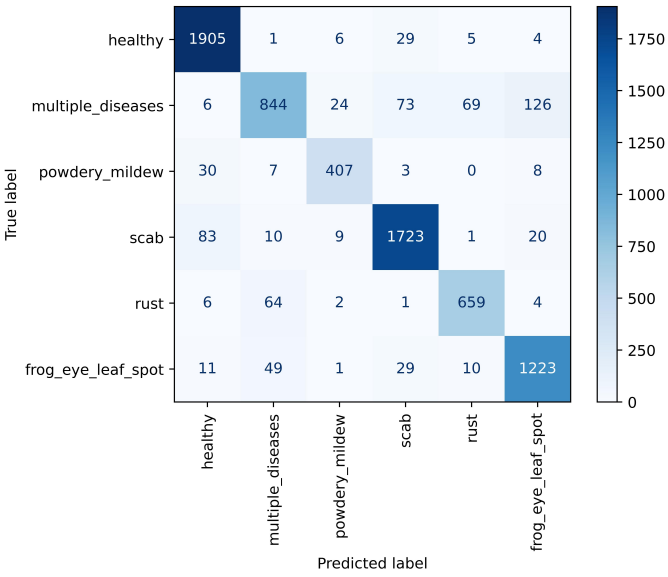

Fig. 5. The confusion matrix on the Plant Pathology 2021 dataset.

We further visualize the performance of the HSD method in Fig. 4. The Receiver Operating Characteristics (ROC) curve is usually used to measure the performance of a model by True Positive (TP) rate and False Positive (FP) rate [41]. The ROC curve has robustness even though the imbalanced positive and false samples hardly change the shape of curves [40]. It is calculated as:

$$\begin{aligned} FP\_rate &= \frac{FP}{FP+TN}, \\ TP\_rate &= \frac{TP}{TP+FN}. \end{aligned} \tag{8}$$

TABLE I
A PERFORMANCE OF DIFFERENT CATEGORIES ON PLANT PATHOLOGY 2021.

Categories	Precision (%)	Recall (%)	F1 score (%)	Number of testing images
Healthy	93.34	97.69	95.46	1950
Multiple diseases	86.56	73.91	79.74	1142
Powdery mildew	90.65	89.45	90.04	455
Scab	92.73	93.34	93.03	1846
Rust	88.58	89.54	89.05	736
Frog eye leaf spot	88.30	92.44	90.32	1323
Macro avg	90.03	89.39	89.61	7452
Weighted avg	90.62	90.73	90.58	7452

TABLE II
A PERFORMANCE COMPARISON ON PLANT PATHOLOGY 2020 AND 2021 IN TERMS OF ACCURACY (%).

Method	Plant Pathology 2020	Plant Pathology 2021
ResNet-50 [13] (SOTA)	97.34	89.98
SE (teacher)	97.96	90.48
SE + KD	97.97	90.51
SE + HSD	**98.22**	**90.72**

Apparently, if the ROC curve closes to the upper left corner with a high value of TP and a low value of FP, it represents the high performance of the classifier. As shown in curves of Fig. 4, the HSD method can effectively classify the diseases with robust ability. The multiple diseases ROC curve is obvious fluctuations that match the class accuracy in Table I. We also calculate the macro-average and micro-average ROC curves to evaluate the overall characteristics.

Additionally, we use the confusion matrix to visualize the performance of the proposed method. Each row of the confusion matrix indicates the true label and each column indicates the predicted label [42]. As seen in Fig. 5, the confusion matrix illustrates the correlation of categories in the Plant Pathology 2021 dataset. The diagonal presents the true prediction, and the rest of the same column is the misclassified diseases. For instance, the rust disease has 1,723 images correctly classified, and 73 images misclassified as multiple diseases. In addition to the misclassified rust disease, there are also mispredictions between powdery mildew and scab, healthy and frog eye leaf spot etc. It is intuitive that all diseases are prone to misidentification as multiple diseases and the HSD method achieves good performance for all categories.

V. CONCLUSIONS

In this paper, a holistic self-distillation method based on the squeeze and excitation network is proposed to solve the bottleneck of the fine-grained classification in plant pathology. An advantage of the method consists in its ability to capture both spatial image information and label knowledge. This is due to the joint work of a teacher and a student network working collaboratively in the online distillation stage, generating soft labels (probabilities) and extracting the spatial features. We use the feature fusion in the squeeze and excitation block to make use of the features training the student network. Such operations can improve the student network performance and prepare soft labels for the next stage of distillation. In the offline distillation stage, the mixed soft and hard labels are fed into the trained student network. Meantime, the squeeze and excitation block explicitly models interdependencies between channels that adaptively recalibrate channel-wise feature responses. This structure improves the ability of the student network further. Our proposed method focuses on learning holistic knowledge from both spatial features and label information. The proposed method achieves 98.22% and 90.72% on the Plant Pathology 2020 and 2021 test datasets respectively. Future work will focus on the development of other deep learning methods for fine-grained classification with out-of-distribution data and derive uncertainty bounds for the proposed solution.

ACKNOWLEDGMENT

We acknowledge the support of the UK EPSRC project EP/V026747/1 (Trustworthy Autonomous Systems Node in Resilience). We are grateful to the UK EPSRC Council via the project EP/V026747/1 (Trustworthy Autonomous Systems Node in Resilience. We are also grateful to the UK EPSRC for funding this work through the EP/T013265/1 project NSF-EPSRC: "ShiRAS. Towards Safe and Reliable Autonomy in Sensor Driven" and the support for ShiRAS by the National Science Foundation under Grant USA NSF ECCS 1903466. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, T.-L. Su, and M. Zuo, "Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, p. 1058, 2019.

[3] H. S. Abdullahi, R. Sheriff, and F. Mahieddine, "Convolution neural network in precision agriculture for plant image recognition and classification," in *Proc. of the 2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, vol. 10, pp. 256–272, Ieee, 2017.

[4] R. Thapa, K. Zhang, N. Snavely, S. Belongie, and A. Khan, "The plant pathology challenge 2020 data set to classify foliar disease of apples," *Applications in Plant Sciences*, vol. 8, no. 9, p. e11390, 2020.

[5] M. Palmer, H. T. Dang, and C. Fellbaum, "Making fine-grained and coarse-grained sense distinctions, both manually and automatically," *Natural Language Engineering*, vol. 13, no. 2, pp. 137–163, 2007.

[6] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 420–435, 2018.

[7] T. B. Sutton, H. S. Aldwinckle, A. M. Agnello, and J. F. Walgenbach, *Compendium of apple and pear diseases and pests*. Am Phytopath Society, 2014.

[8] A. Peil, V. G. Bus, K. Geider, K. Richter, H. Flachowsky, and M.-V. Hanke, "Improvement of fire blight resistance in apple and pear," *Int J Plant Breed*, vol. 3, no. 1, pp. 1–27, 2009.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[10] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449–1457, 2015.

[11] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," *arXiv preprint arXiv:1603.06765*, 2016.

[12] J. Su, S. Anderson, and L. S. Mihaylova, "A deep learning method with cross dropout focal loss function for imbalanced semantic segmentation," in *2022 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–6, IEEE, 2022.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[16] J. Snedeker, L. Gleitman, *et al.*, "Why it is hard to label our concepts," *Weaving a Lexicon*, vol. 257294, 2004.

[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[18] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722, 2019.

[19] Z. Zhang and M. Sabuncu, "Self-distillation as instance-specific label smoothing," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2184–2195, 2020.

[20] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[21] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541, 2006.

[22] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13876–13885, 2020.

[23] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[24] Y. Shen, L. Xu, Y. Yang, Y. Li, and Y. Guo, "Self-distillation from the last mini-batch for consistency regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11943–11952, 2022.

[25] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3048–3068, 2021.

[26] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5008–5017, 2021.

[27] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self attention distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1013–1021, 2019.

[28] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2859–2868, 2019.

[29] S. Hahn and H. Choi, "Self-knowledge distillation in natural language processing," *arXiv preprint arXiv:1908.01851*, 2019.

[30] D. Sun, A. Yao, A. Zhou, and H. Zhao, "Deeply-supervised knowledge synergy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6997–7006, 2019.

[31] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

[32] X. Zhu, S. Gong, *et al.*, "Knowledge distillation by on-the-fly native ensemble," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[33] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module," *PloS One*, vol. 14, no. 3, p. e0214587, 2019.

[34] R. Deng, M. Tao, H. Xing, X. Yang, C. Liu, K. Liao, and L. Qi, "Automatic diagnosis of rice diseases using deep learning," *Frontiers in Plant Science*, vol. 12, p. 701038, 2021.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[36] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756*, 2020.

[37] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 79–91, 2020.

[38] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, pp. 345–359, Springer, 2005.

[39] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[40] J. Muschelli III, "Roc and auc with a binary predictor: a potentially misleading metric," *Journal of Classification*, vol. 37, no. 3, pp. 696–708, 2020.

[41] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[42] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection.," *Maics*, vol. 710, no. 1, pp. 120–127, 2011.