Key Area Acquisition Training for Practical Image-based Plant Disease Diagnosis

Kaito Odagiri, Shogo Shibuya, Quan Huu Cap, and Hitoshi Iyatomi
Applied Informatics, Graduate School of Science and Engineering
Hosei University, Tokyo, Japan
Email: kaitodagiri@gmail.com, shogo.shibuya.5u@gmail.com, huuquan.cap.75@hosei.ac.jp, iyatomi@hosei.ac.jp

***Abstract*—Automatic diagnosis of plant diseases using images is a fine-grained task, and disease symptoms are often ambiguous and highly variable. Pre-extraction of the region of interest (ROI) exhibiting disease symptoms (such as one or more leaves) is known to have a certain effect on improving accuracy. However, the ROI extraction at runtime is time-consuming, resulting in issues of system usability. This paper proposes a new training method called key area acquisition training (KAAT). KAAT reduces the variation in prediction results between images before and after the extraction of the ROI. By directing the model's attention to the ROI through learning, KAAT contributes to improved diagnostic performance without sacrificing execution time during diagnosis. In the evaluation, we conducted nine class diagnosis task (eight diseases and healthy) using 77K and 9K images of cucumber leaves (collected from different fields) for training and testing, respectively. The proposed KAAT improved diagnostic accuracy by 3.8% in macro-F1 and 2.0% in micro accuracy without increasing execution time.**

***Index Terms*—automated plant diagnosis, stability training, convolutional neural networks, segmentation, ROI**

I. Introduction

According to the Food and Agriculture Organization of the United Nations (FAO), it is estimated that up to 40% of food crops are lost each year due to plant pests and diseases [1]. Early detection of pests and appropriate action are therefore important, but diagnosis requires expert knowledge and experience and, in some cases of genetic testing. These pose challenges such as difficulties of availability and prohibitive time and financial costs. With the recent development of machine learning technology, reports of research on automatic diagnosis of plant diseases by means of images using deep learning began with the analysis of data from actual fields with a three-layer convolutional neural networks (CNN) [2]. Since then, the PlantVillage dataset, a large-scale open dataset, has become widely available on the web, and deep learning-based systems using this dataset have been proposed one after another. Studies using this dataset have reported numerically excellent discriminative performance [3]–[7].

However, the training and evaluation data they used from this dataset reportedly yielded significantly lower discriminative power under real-world conditions because the leaves used were pre-cropped images that were analyzed and placed on a plain background. For example, it has been reported that when evaluated in a real-world environment different from the training data, the diagnostic performance, which was over 99%, drops to about 30% [3], [7].

Unlike these methods that analyze data from laboratory environments, many methods have been proposed in recent years that are built on originally constructed datasets based on images collected from actual fields. Most of these systems use CNNs and assume that the analysis target, such as a leaf or a fruit, is in the center of a single image, and similarly high discrimination accuracy has been reported [8]–[15]. However, there are serious concerns about this assessment, and it has been pointed out in recent years that the accuracy is far overestimated compared to the true performance, which is much lower [16]–[20]. This is because in most of these reports, the evaluations are conducted by arbitrary partitioning or cross-validation of the datasets created, which leads to high potential similarity in the datasets and thus padding of diagnostic performance.

While plant disease symptoms are often small and ambiguous in terms of the number of features on the image, there is also a great deal of variability in appearance within the same disease category, as well as environment-dependent diversity unrelated to disease symptoms [20]. In situations where the intrinsic diversity and number of training images is limited, deep discriminators, such as CNNs, tend to capture features that include background, composition, and image features (brightness, tint) that are usually larger in area than disease features. It is clear that the accuracy of the evaluation data is improved in cases where highly similar images, such as a series of images of the same subject taken in a short period of time, are divided into training and evaluation data. Evaluations using a sample much larger than in any other existing study, with over 220K actual datasets from four crops and discrimination models more sophisticated than previously applied, reported nearly 100% diagnostic accuracy for the same test data, but this dropped to 40-70% when properly evaluated [20]. This discrepancy highlights the need to obtain a clear distinction between training and evaluation data in order to properly evaluate the model. It should be noted that CNNs are not robust to scaling, and are vulnerable to the difference in the distance between the camera and the object to be identified, i.e., the difference in composition. The bounding box (BB)-type discriminator [21]–[23], which can simultaneously identify the location of the target of interest and its content, is more robust to variations in distance to the target (scaling in the image) than conventional CNN discriminators, and can also present the region of interest (ROI), making it easier to interpret. This is a promising technology. However, in the task of plant disease diagnosis [10], [16], [24], the biggest challenge is the extremely high cost of creating training labels for labeling each disease symptom; therefore, in pracice, the number of images that can be used for training is much smaller than for CNN models. In addition, the training and evaluation of small lesions in a relatively large image area requires high-resolution images, which requires a large amount of computational resources and, especially, execution time. Therefore, it is important to improve the performance of conventional CNN-based discriminators in practical system construction.

On the other hand, in image-recognition problems, pre-segmentation of ROIs, or regions that are important in realizing the goal, has been routinely performed in various tasks. Segmentation used to be a difficult procedure that was task-dependent; however, great progress has been made with the development of deep learning techniques, especially with the generative adversarial networks (GAN) [25] technique. Saikawa et al. [17] proposed anti-overfitting pre-treatment (AOP), which detects the region to be diagnosed (leaves, fruits, etc.) based on pix2pix [26] and calibrates brightness and color, as background is one of the factors that cause overlearning in automatic plant disease diagnosis tasks. AOP found that the VGG network, trained on 35,694 cucumber disease images in eight classes, improved diagnostic accuracy by 12.2% across 9,115 test data, including many early disease symptoms collected over different fields. The effectiveness and importance of directing the discriminator's attention to the region of interest has been confirmed not only intuitively, but also by the attention mechanism [27] and the vision transformer [28], which applies this technique to image recognition.

Recent studies using our large-scale data [20] suggest that the cause of overlearning is not merely in the background of the image but also in the ROI in a way that is difficult for us to perceive, indicating that the problem cannot be solved only by pre-extraction of these regions. Although this large-scale study showed that training with a large number of high-resolution images can suppress the effects of background-based overlearning to some extent, there is no doubt that prior extraction of ROIs is important for diagnosis. However, when segmentation is performed as a preprocessing step during diagnosis, the execution time increases compared to the case without segmentation and becomes a burden during actual use.

Therefore, this paper proposes a new learning method, key area acquisition training (KAAT), which focuses on the ROI by means of the discriminator itself during training and aims to improve the intrinsic discriminative power. The proposed KAAT utilizes stability training [29], which builds a model robust to disturbances, and learns such that the diagnostic results of the original image x and the image x' from which the key ROIs are pre-extracted for diagnosis are the same. This allows a model trained with KAAT to expect the same results as if a time-consuming ROI pre-extraction had been performed at the time of diagnostic execution. This is not only a great benefit in terms of reduced runtime in practical terms but also a method that leads to the construction of an essential and robust system since the discriminator's attention is directed to the region containing the lesion features.

II. Key area acquisition training (KAAT)

This paper proposes key area acquisition training (KAAT), a learning method to improve the accuracy of essential discriminators and reduce the time required for execution in order to realize a practical automatic plant disease diagnosis system. The proposed KAAT is a method inspired by stability training [29], which increases the robustness of a model by adding constraints such that the output is the same even when the input is disturbed.

A. Stability training

Stability training [29] is an effective method for building discriminators that are robust to input disturbances. Let $p(y|x)$ (hereafter abbreviated as y, and similarly for other symbols of probability distribution) be the output probability distribution of the model for the input x, y' be the output distribution for the disturbed input $x' = x + \xi$, and t be the correct label distribution. Then the stability training minimizes the Kullbuck-Leibler (KL) divergence between y and y' in addition to the original classification task error L_{org}. This can be expressed as

$$L_{\mathrm{STAB}} = L_{\mathrm{org}} + \lambda L_{\mathrm{stab}} = D(y, t) + \lambda D_{\mathrm{KL}}(y, y'). \quad (1)$$

Here λ is a hyperparameter to balance the loss term, and D is a general error measure, such as cross entropy. The difference between this stability training and general data augmentation is that the former explicitly adds a constraint such that x and x' have the same result so that the discriminator itself is robust to the explicitly specified disturbances. The original paper used random perturbation, which is reported to be superior in terms of generalizability improvement. Learning methods that actively use disturbances in learning include adversarial training [30] and its extension to semi-supervised learning, virtual adversarial training (VAT) [31]. These methods have been widely used, with excellent accuracy gains reportedly achieved by calculating and including small amplitude perturbation with adversarial directions in the learning process. Stability training, on the other hand, has a major advantage in that it allows the designer to design the perturbation according to the objective task.

B. Definition and Implementation of KAAT

Fig. 1 shows a schematic diagram of the proposed KAAT. By applying stability training, KAAT learns to ensure that the output of the model for the original input image x and the image from which only ROIs (leaves, fruits, etc.) containing symptoms are extracted x' is the same. In other words, we consider that x' contains a large disturbance ξ that removes parts of the image other than the ROI.

There is one difference here from the stability training that KAAT is based on. Stability training implicitly assumes that the disturbance ξ is small and the difference between y and y'

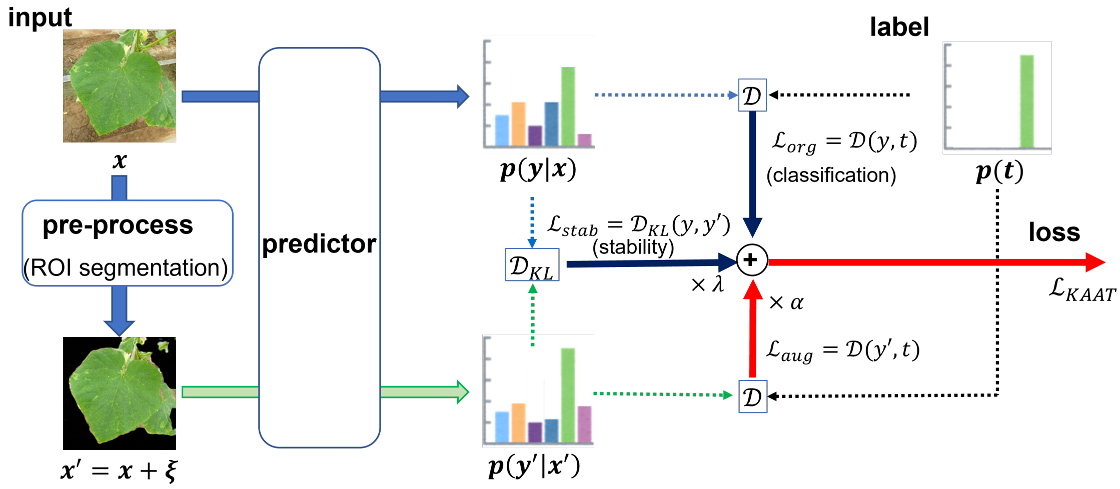


Fig. 1. Schematics of the key area acquisition training (KAAT)

is not so large during training, i.e., it is a learning algorithm that does not assume that the class of output results will change due to disturbances. However, the disturbance ξ applied by KAAT this time is an extremely large signal that removes all the background of the original image x, so y' is often essentially very different from y when learning x', making it difficult to achieve the expected stabilization effect. Therefore, a new loss, L_{aug}, is added to the KAAT learning function to perform supervised learning to obtain the same output for $x' = x + \xi$, which includes large changes to the original input x. Finally, the loss function of KAAT, L_{KAAT}, is written as follows, modifying equation (1):

$$\begin{aligned} L_{\text{KAAT}} &= L_{\text{STAB}} + \alpha L_{\text{aug}} \\ &= L_{\text{org}} + \lambda L_{\text{stab}} + \alpha L_{\text{aug}} \\ &= D(y,t) + \lambda D_{\text{KL}}(y,y') + \alpha D(y',t). \end{aligned} \tag{2}$$

The newly introduced L_{aug} can also be regarded as an additional data augmentation of ROI extraction added to the original image. The α is a hyperparameter.

C. Detection of ROI

For the extraction of the ROI (i.e., a leaf region) in this study, we used AOP [17]. AOP is a highly accurate method that achieves precision and recall rates of 98.6% and 97.5%, respectively, for the segmentation of cucumber leaf regions. However, we do not limit ourselves to this method as long as foreground extraction is possible. Although AOP can generate segmentation images with automatic correction of image brightness and color, in this experiment, in order to directly evaluate the effect of the proposed KAAT, the segmentation results were obtained by turning the generated image into a binary mask and applying it to the original image. The same 8,000 cucumber leaf images as in [17] were used to train the AOP network. These were included in the training images for this experiment, and no test images were included.

III. Experiments

A. Dataset

As noted above, many previous studies have pointed to the effects of overlearning associated with the splitting of potentially similar images into training and evaluation images. In this experiment, we eliminated such effects by clearly separating the training and evaluation image sets. The statistics for the dataset used in the experiment are shown in Table I, and examples of images of typical symptoms of each disease are shown in Fig. 2. The experiment was conducted using a total of 76,964 cucumber leaf images for training and 9,338 test images collected in different fields than the training set. Each case is either infected with one of the following eight diseases or a healthy leaf. The types of diseases include five fungal diseases ((a) powdery mildew (PM), (b) downy mildew (DM), (c) corynespora leaf spot (CLS), (d) gummy stem blight (GSB), and (e) bacterial spot (BS)), as well as three viral diseases ((f) cucumber mosaic virus (CMV), (g) cucumber chlorosis (CCYV), and (h) melon yellow spot virus (MYSV)).

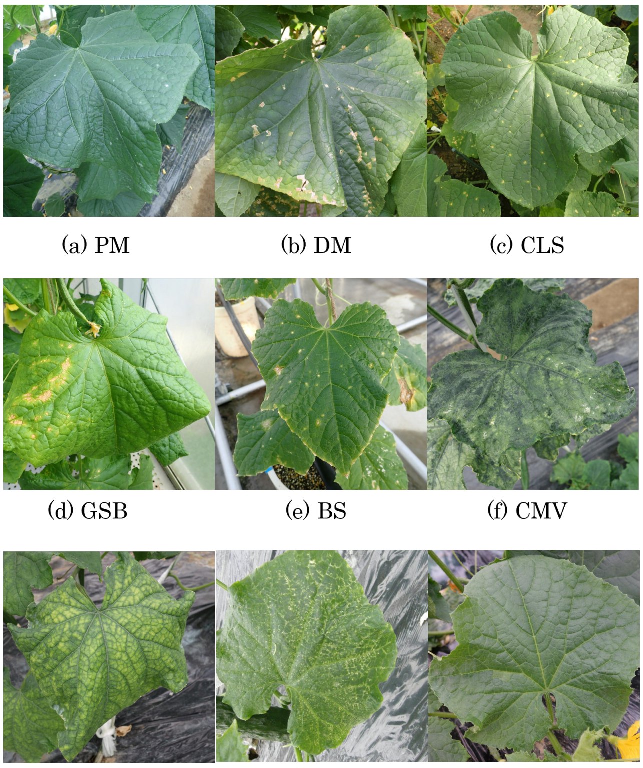
B. Classification model and evaluation

In this experiment, EfficientNet-B4 [32], a CNN model with reported superior classification performance, was used as the plant disease discriminator. For data augmentation, we used RandAugment [33], which has been reported to be equally effective. The resolution of the discriminator input was set to 380×380, the optimization method was Adam, the learning coefficient was 0.001, and the batch size was 32. The parameters of RandAugment were set to $n = 4$ and $m = 5$, based on the results of preliminary experiments. In this experiment, these conditions were used as a baseline to compare the performance. The hyperparameters for stability training and KAAT were set to $\lambda = 1.0$ and $\alpha = 0.85$, based on the results of preliminary experiments.

Four types of comparison of disease diagnosis performance were conducted: the baseline, diagnosis on images of pre-

TABLE I
DATASET USED IN OUR STUDY

	fungal diseases					viral diseases			(i) Healthy	Total
	(a) PM	(b) DM	(c) CLS	(d) GSB	(e) BS	(f) CMV	(g) CCYV	(h) MYSV		
# train	6,390	6,803	6,675	1,475	1,071	22,042	4,721	10,670	17,117	76,964
# test	1,135	117	491	100	946	1,588	1,248	1,468	2,245	9,338


(a) PM (b) DM (c) CLS

(d) GSB (e) BS (f) CMV

(g) CCYV (h) MYSV (i) Healthy

Fig. 2. Examples of infected and healthy cucumber leaves

extracted leaf areas by AOP [17] (+ AOP), baseline with stability training [29] (+ stability training), and baseline with the proposed KAAT (+ KAAT).

When constructing a practical diagnostic system, it is not enough to evaluate the diagnostic performance for each disease based on the percentage of correct answers (i.e., number of correct answers/number of data = $recall$); it is also important to evaluate the conflicting indicator $precision$ (i.e., number of correct answers/number of detected instances). Therefore, in this study, we used the F1 score (i.e., $2 \times (precision \times recall)/(precision + recall)$), which is the harmonic mean of those conflicting indicators. On the other hand, micro accuracy, which is widely used for performance evaluation in literature, is also included for reference. It should be noted that the number of training and test data varies widely by disease category.

IV. Results

Although we did not quantitatively evaluate the accuracy of leaf area extraction by AOP for the test images, qualitative evaluation confirmed that the extraction was reasonable and natural.

Table II shows a comparison of diagnostic performance based on F1 scores for each disease. Table III summarizes how much of an advantage KATT had over the baseline. The proposed KAAT improved the diagnostic performance for six of the nine disease categories, although the degree of improvement in identification for individual diseases varied from small to large. The largest improvement (14.4%) was seen in the F1 score for GSB, which originally had very low diagnostic performance. Fig. 3 shows an example of GSB images and the leaf areas extracted by AOP for them. GSB is symptomatically similar to CLS and PM, with 21% false positive identification each in the baseline, but with the introduction of KAAT, the false positive identification rate for PM was reduced to 5%. Conversely, BS showed a 5.6% decrease. As it turned out, KAAT improved the overall diagnostic performance by 3.77% for macro-F1 and 2.01% for (micro) accuracy. These results are better than those obtained with AOP pre-processing in addition to the baseline.

On the other hand, in this experiment, the classifier that applied naive stability training had lower diagnostic performance than the baseline that did not. Various changes in the hyperparameter of the stability training did not change this trend.

The speed of the diagnosis was the same at 25.12 images/sec for the baseline and the proposed KAAT, whereas it dropped to 1.13 images/sec when diagnostics were conducted after ROI extraction by AOP was performed as preprocessing. This is because the ROI extraction process takes much more time than the diagnosis process.

V. Discussion

We have confirmed that the proposed KAAT steadily improves diagnostic performance under the practical conditions. In addition, KAAT has the great advantage of not increasing the execution time of the diagnostic run by virtue of a learning method that directs the discriminator's attention to the ROI. The images used in this evaluation were collected in a realistic situation and included images from completely different fields, unlike the simple evaluation in which the evaluation data are similar to the training data, which is often the case in conventional plant disease research. Therefore, the absolute values of diagnostic performance appear lower than those in

TABLE II
Classification performance for each disease in F1 metric (%)

	fungal diseases					viral diseases					
	(a) PM	(b) DM	(c) CLS	(d) GSB	(e) BS	(f) CMV	(g) CCYV	(h) MYSV	(i) Healthy	macro F1	accuracy
Baseline †	76.8	50.4	49.6	14.3	74.4	52.5	84.9	56.2	75.0	59.35	67.00
+ AOP [17] ‡	81.6	48.3	49.0	36.1	64.2	51.8	83.3	62.2	77.8	61.57	68.16
+ stability training [29]	80.4	52.4	48.2	18.0	46.7	51.8	82.6	54.3	71.9	56.25	63.56
+ KAAT (proposed)	83.6	62.9	49.2	28.7	68.8	52.7	83.9	62.0	76.3	**63.13**	**69.01**

†: EfficientNet-B4 [32] + RandAugment [33] $(n = 4, m = 5)$.
‡: Classification for images with prior ROI (leaf area) extraction by AOP.

TABLE III
Performance improvement with KATT relative to the baseline (%)

	recall	precision	F1
(a) PM	4.5	10.4	6.8
(b) DM	-0.9	26.8	12.4
(c) CLS	-2.8	1.0	-0.4
(d) GSB	2.0	16.6	14.4
(e) BS	-4.4	-7.3	-5.6
(f) CMV	4.9	-6.0	0.2
(g) CCYV	-1.8	0.6	-1.0
(h) MYSV	5.6	6.0	5.8
(i) Healthy	2.3	0.8	1.3
Total	1.06	5.41	3.77

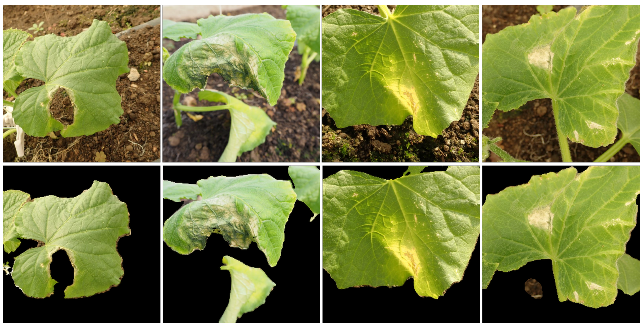


Fig. 3. Example of GSB images and their leaf regions determined with AOP

the literature to date. However, this experiment included tens of thousands of training images, and only current cutting-edge CNN model + data augmentation techniques are able to achieve this level of diagnostic performance. The difficulty of diagnosis differs for each disease, and the diagnostic performance varies depending on the similarity of the composition of the training image and the evaluation image.

Although stability training has been reported to be highly effective, it did not produce the desired results in the experiment with the large background-removing disturbance ξ in this case. This is presumably because the estimated y' for the input containing extremely large disturbances $x + \xi$ is essentially very different from y, as mentioned above, and it was difficult to reduce the difference between these two by learning. Since KAAT performs supervised learning even for images with extensive background removal thanks to the newly introduced loss L_{aug}, we believe that the stabilization effect is enhanced even for such large disturbances.

Since DM and GSB have very few test images (117 and 100 images, respectively), the precision in question is greatly reduced due to misidentification among categories. Therefore, the precision that constitutes the F1 score is inevitably lower, resulting in a lower F1 score. For these difficult categories, KAAT also achieves performance comparable to that achieved when AOP is introduced as preprocessing.

In this experiment, the improvement with respect to GSB was significant. The typical symptom of PM, which was originally frequently misidentified, was white blotchiness, but there were also many symptoms, such as wilting of leaves, that were similar to those of GSB. The KAAT allows the discriminator to focus on the leaf area, which may have facilitated improvement in performance, as it allows the discriminator to focus on differences that are essentially symptoms.

For BS, scores decreased by 5.6%, which was almost entirely due to increased misidentification as healthy cases. A closer look reveals that for BS, there was a significant decrease in performance (-10.3%) when AOP was applied to the baseline. BS is typically characterized by small spots caused by fungal growth, and it appears that AOP increased misidentification to the healthy category by excluding symptomatic or pronounced leaves as background. In contrast, KAAT does not completely remove the background information, allowing it to take the original information into account, which may have resulted in better results than the baseline +AOP.

VI. Conclusion

In this paper, we proposed key area acquisition training (KAAT) for the automatic diagnosis of plant diseases using images, which enables the classifier model to focus on the ROI including lesions. KAAT is a learning method in which the diagnostic results are constrained to be the same before and after the ROI is extracted. To achieve this, explicit constraints are added to obtain correct results even after ROI extraction. Using a total of approximately 86,000 images collected in several actual fields, classification experiments of nine categories found in cucumber leaf images showed that KAAT did not require ROI (i.e., leaf area) extraction at runtime but nevertheless provided diagnostic performance equal to or better than that of the extracted images.

Acknowledgment

This work was supported by the Japanese Ministry of Agriculture, Forestry and Fisheries (MAFF) commissioned project study on "Development of pest diagnosis technology using AI" (JP17935051).

References

[1] FAO, "Protecting plants, protecting life," *Food and Agriculture Organization of the United Nations*, 2020. [Online]. Available: http://www.fao.org/plant-health-2020/about/en/

[2] Y. Kawasaki, H. Uga, S. Kagiwada, and H. Iyatomi, "Basic study of automated diagnosis of viral plant diseases using convolutional neural networks," in *Proceedings of the International Symposium on Visual Computing*, 2015, pp. 638–645.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016.

[4] H. Durmuş, E. O. Güneş, and M. Kırcı, "Disease detection on the leaves of the tomato plants by using deep learning," in *Proceedings of the 6th International Conference on Agro-Geoinformatics*, 2017, pp. 1–5.

[5] G. Wang, Y. Sun, and J. Wang, "Automatic image-based plant disease severity estimation using deep learning," *Computational Intelligence and Neuroscience*, vol. 2017, p. 2917536, 2017.

[6] M. Brahimi, M. Arsenovic, S. Laraba, S. Sladojevic, K. Boukhalfa, and A. Moussaoui, "Deep learning for plant diseases: Detection and saliency map visualisation," in *Human and Machine Learning*. Springer, 2018, pp. 93–117.

[7] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.

[8] E. Fujita, Y. Kawasaki, H. Uga, S. Kagiwada, and H. Iyatomi, "Basic investigation on a robust and practical plant diagnostic system," in *Proceedings of the IEEE International Conference on Machine Learning and Applications*, 2016, pp. 989–992.

[9] T. Hiroki, R. Kotani, S. Kagiwada, U. Hiroyuki, and H. Iyatomi, "Diagnosis of multiple cucumber infections with convolutional neural networks," in *Proceedings of the Applied Imagery Pattern Recognition Workshop*, 2018, pp. 1–4.

[10] A. F. Fuentes, S. Yoon, S. Kim, and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," *Sensors*, vol. 17, no. 9, p. 2022, 2017.

[11] J. Boulent, S. Foucher, J. Théau, and P.-L. St-Charles, "Convolutional neural networks for the automatic identification of plant diseases," *Frontiers in Plant Science*, vol. 10, p. 941, 2019.

[12] Y. Toda and F. Okura, "How convolutional neural networks diagnose plant disease," *Plant Phenomics*, vol. 2019, p. 9237136, 2019.

[13] H.-J. Yu and C.-H. Son, "Leaf spot attention network for apple leaf disease identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 229–237.

[14] M. Zekiwos, A. Bruck *et al.*, "Deep learning-based image processing for cotton leaf disease and pest diagnosis," *Journal of Electrical and Computer Engineering*, vol. 2021, p. 9981437, 2021.

[15] M. A. Mithu, S. I. Momo, M. Hasan, K. M. Rahman, A. Sattar *et al.*, "Pumpkin leaf disease detection: Convenience of cnn over traditional machine learning in terms of image classification," in *Smart Systems: Innovations in Computing*. Springer, 2022, pp. 347–357.

[16] K. Suwa, Q. H. Cap, R. Kotani, H. Uga, S. Kagiwada, and H. Iyatomi, "A comparable study: Intrinsic difficulties of practical plant diagnosis from wide-angle images," in *Proceedings of the IEEE International Conference on Big Data Workshops*, 2019, pp. 5195–5201.

[17] T. Saikawa, Q. H. Cap, S. Kagiwada, H. Uga, and H. Iyatomi, "AOP: An anti-overfitting pretreatment for practical image-based plant diagnosis," in *Proceedings of the IEEE International Conference on Big Data Workshops*, 2019, pp. 5177–5182.

[18] Q. H. Cap, H. Uga, S. Kagiwada, and H. Iyatomi, "LeafGAN: An effective data augmentation method for practical plant disease diagnosis," *IEEE Transactions on Automation Science and Engineering*, 2020.

[19] S. Kanno, S. Nagasawa, Q. H. Cap, S. Shibuya, H. Uga, S. Kagiwada, and H. Iyatomi, "PPIG: Productive and pathogenic image generation for plant disease diagnosis," in *Proceedings of the IEEE-EMBS Conference on Biomedical Engineering and Sciences*, March 2021, pp. 554–559.

[20] S. Shibuya, Q. H. Cap, S. Nagasawa, S. Kagiwada, H. Uga, and H. Iyatomi, "Validation of prerequisites for correct performance evaluation of image-based plant disease diagnosis using reliable 221K images collected from actual fields," in *Proceedings of the AAAI Conference on Artificial Intelligence Workshops*, 2021.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2016.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 21–37.

[23] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2778–2788.

[24] J. Liu and X. Wang, "Early recognition of tomato gray leaf spot disease based on MobileNetv2-YOLOv3 model," *Plant Methods*, vol. 16, pp. 1–16, 2020.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative adversarial nets," in *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–21.

[29] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4480–4488.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations*, 2015, pp. 1–11.

[31] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[32] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, May 2019, pp. 6105–6114.

[33] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the Advances in Neural Information Processing Systems*, 2020, pp. 18 613–18 624.