Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-32 (2022) 89–94

Investigation of the added value of CycleGAN on the plant pathology dataset

Bart M. van Marrewijk*, Gerrit Polder* and Gert Kootstra*

* *Wageningen University and Research, 6700 AA Wageningen, the Netherlands, e-mail: bart.vanmarrewijk@wur.nl*

Abstract: A typical problem when using deep neural networks in the domain of agriculture is the limited availability of labelled training data. Generative adversarial networks (GAN) hold the potential that they can learn to create realistic fake images, which might be used to enrich the training set. In this research, the added value of image augmentation using CycleGAN is evaluated for image classification on a plant-pathology dataset. CycleGAN was trained to generate new images of infected leaves by converting healthy apple leaves to leaves with scab symptoms. This resulted in two new datasets, which were compared with a benchmark dataset. Each dataset was trained with a ResNet-18 classifier. The trained classifiers were tested on the complete independent plant pathology 2021 dataset. The datasets with the new images generated by CycleGAN had an accuracy of 69.49 and 66.54, which was not significant compared with the benchmark (68.43). An additional dataset was made in which the benchmark dataset was extended with real images. This dataset without generated images obtained a classification accuracy of 77.75%, which was a significant improvement compared to the benchmark. This does not mean that GANs are useless, but it shows that in the evaluation of GANs it can be interesting to determine how many real images you would need to add to obtain a similar performance gain as the dataset with the synthetic images.

Keywords: Disease classification, deep learning, generative adversarial networks, domain adaptation

1. INTRODUCTION

To cater a growing world population while reducing the negative effects on the environment, agriculture is focussing on producing more with less, increasing production while minimizing the use of fertilizers and pesticides. At the moment, global losses due to pests and diseases can be up to 30.0% (Savary et al., 2019). Early detection of these infections can reduce the use of pesticides, reduce food waste, and improve food quality.

In recent years, the development of deep learning has resulted in a new era for image-based disease detection in agriculture (Kamilaris and Prenafeta-Boldú, 2018). Deep neural networks outperformed classical methods that are based on hand-crafted features, since they can learn complex features, which is needed to deal with the variations present in the agricultural environment (Kamilaris and Prenafeta-Boldú, 2018). The downside, however, is that deep learning requires many training images to deal with all variations in climate, light, and crop diversity (Cap et al., 2020b, Kuznichov et al., 2019). For agricultural applications, it is often challenging to acquire high-quality and -quantity labelled datasets (Kuznichov et al., 2019, Nazki et al., 2020). Research is consequently focused on maximising the performance with minimal data.

The ability of a network to deal with new unseen data is also known as generalisation (Goodfellow et al., 2014, Shorten and Khoshgoftaar, 2019). Several papers have been published to improve generalisation of networks in agriculture (Espejo-Garcia et al., 2020, Blok et al., 2020). In these papers, data augmentation was used to reduce overfitting and improve generalisation (Shorten and Khoshgoftaar, 2019). Many different types of image augmentation exist like geometric (rotation, cropping, and scaling) and photometric (light, colour, and texture) transformations. Slightly more advanced methods are CutOut (DeVries and Taylor, 2017), Cutmix (Yun et al., 2019), and Mixup (Zhang et al., 2017) in which random parts of the images are removed and mixed. The mentioned augmentation techniques are basic image transformations that modify the original images. A more advanced method is to generate data using generative deep learning. One approach is the use of generative adversarial networks (GANs) (Goodfellow et al., 2014), which can learn to generate realistically looking fake images. They hold great promise as a data-augmentation approach to enrich the training set (Bowles et al., 2018, Shorten and Khoshgoftaar, 2019). In this paper, we investigate the added value of using a GAN to enrich the training set.

The next subsections give some background on GANs and their use in agricultural applications.

1.1 GANs for domain adaptation

Most GAN architectures have a generator that is creating images and a discriminator that determines if the input image is real or fake (Goodfellow et al., 2014). These GANs can be extended with conditions (cGANS) to force the generator to create an image of a specific class/conditions, which results in more realistic images (Mirza and Osindero, 2014, Isola et al., 2017). Without conditioning, the network only learns if the output looks realistic, it does not include the mismatch between domain A and domain B, which results in a reduced performance. cGANs can be used for image-to-image translation, which is a process that converts an image of domain A to domain B. For example, converting horses to zebras (Zhu et al.,

10.1016/j.ifacol.2022.11.120

2017). Image-to-image translation can be divided into paired and unpaired image translation. In the first, conditioning is applied since both the source and target image have to be similar. In unpaired image-to-image translation, no 1-to-1 correspondence between source and target images is needed. The algorithm should learn to translate an image from the source domain to the target domain. To improve stability, cycle consistency is used (Park et al., 2020). Here, the transformation of image x to domain B using generator $G(x)$ is learned, as well as the inverse mapping with generator F to transform the transformed image back to domain A so that $F(G(x)) \approx x$ (Zhu et al., 2017).

1.2 GANs for agricultural applications

Several studies have applied GANs to agricultural applications to deal with limited data. Nazki et al. (2020) applied an improved version of CycleGAN named AR-GAN to enlarge and balance a tomato-disease dataset. The dataset consisted of 9 different plant diseases of which two occupied more than 50% of the dataset. AR-GAN was used to balance the dataset by creating new disease images using domain adaptation on healthy images. The new dataset was tested on a classifier of which the accuracy improved by 5.2 percentage points (pp), which was better than using classical augmentation methods (0.8 pp). Arsenovic et al. (2019) applied StyleGAN to generate images for plant-disease classification. The existing training dataset of ±14500 images was increased with 5000 synthetic images. Multiple classifiers were trained to explore the added value of the generated images. For each classifier, the classification accuracy improved using the synthetic images, but the difference was minimal with an average improvement of the classification accuracy of 0.8pp (88.0 to 88.8). Madsen et al. (2019) created a custom version of WGAN-GP and ACGAN to generate images of plant seedlings. Their GAN was able to obtain a classification accuracy of 58.9%±9.2%., which was the same as without the additional data, meaning that the additional value of the GAN was minimal. The only advantage was that it converged earlier.

1.3 Contribution of the paper

The above suggests that the increase in performance when enriching the dataset using GANs is minimal. Furthermore, GANs are hard to train since convergence is difficult and sometimes mode-collapse occurs in which the generator tends to learn to map different input samples to the same output (Pang et al., 2021). As a result, increasing a dataset using GANs is a time-consuming process, and it is uncertain if the quality of the transferred images is sufficient. From a practical point of view, it is therefore unknown if it is better to use a GAN or to collect more images to improve performance. Moreover, related work often used additional data to train the GANs, which does not make for a completely fair comparison.

In this paper, we applied CycleGAN on the plant-pathology 2020 dataset (also known as fgvc7) to generate additional apple disease images[1]. The added value of the GAN was investigated by training a classifier with and without the additional data and comparing the performance of classifying health and infected leaves on the independent plant pathology 2021 (fgvc8) dataset[2].

2. MATERIAL AND METHODS

2.1 Original dataset

Two public datasets were used; plant pathology 2020 also known as fgvc7 (Thapa et al., 2020) and plant pathology 2021 dataset (fgvc8). In both datasets, only the images with healthy leaves and scab-infested leaves were selected. In Figure 1, an example of both the fgvc7 and fgvc8 datasets is shown. The right column shows leaves with the scab symptoms.

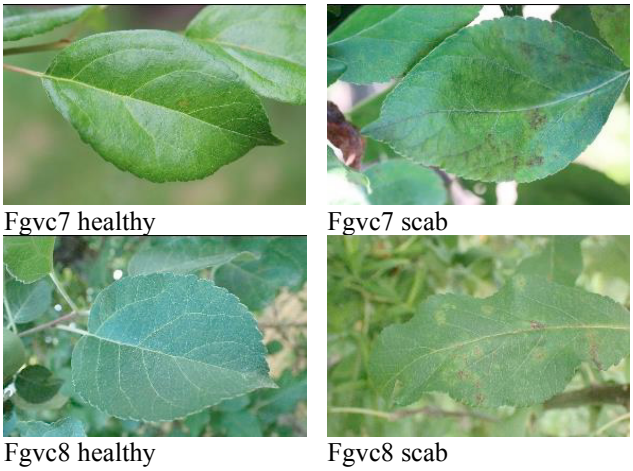

Figure 1. Example images of the fgvc7 and fgvc8 datasets.

The fgvc7 was split into a train, validation, and test set with a 70/10/20 division. The fgvc7 was used to train and validate the various CycleGANs (Section 2.2) and to test the within-domain generalization. The out-of-domain generalisation was tested on the completely independent dataset of fgvc8. The datasets are summarized in Table 1. We refer to "A" for images with healthy and "B" for images with scab-infected leaves.

Table 1. Overview of fgvc7 and fgvc8 dataset.

	fgvc7			**fgvc8**
	train	valid	test	
A: healthy	364	54	98	4623
B: scab	412	65	115	4825
Total	776	119	213	9448
Healthy/scab	0.88	0.83	0.85	0.95

2.2 CycleGAN

The popular CycleGAN was used for unpaired image-to-image translation (Zhu et al., 2017). CycleGAN uses two generators, G and F, to translate images from one domain to another and back. The discriminator (D_A, D_B) judges whether the images belong to domain A or domain B. Generator G tries to convert image from domain A to domain B, $G: A \rightarrow B$, while

[1] https://www.kaggle.com/c/plant-pathology-2020-fgvc7

[2] https://www.kaggle.com/c/plant-pathology-2021-fgvc8/overview/cvpr-2021

discriminator D_B aims to distinguish between transformed images G(A) and real samples from B. Therefore, G aims to minimize the difference between G(A) and B, while the discriminator tries to maximise the difference. This process is also known as the adversarial loss. To prevent a transformation of each input image to the same output image cycle-consistency is used, which includes a second generator, F, that learns the reverse mapping, $F: B \rightarrow A$. Cycle consistency means that $F(G(A)) \approx A$, in other words, an input image from A transformed first by G and then by F should be similar to the original input image (Zhu et al., 2017).

To preserve color composition between the input image and the generated image the identity loss is included. In the identity loss function, the input of the generator is a real image of the other domain. Instead of G(A), the input of the network is a real image of domain B: G(B). The idea behind this mechanism is that if an image has a high similarity with the target domain, then the generator should not do any mapping at all. For more detailed information about CycleGAN, we refer to the original paper of Zhu et al. (2017).

CycleGAN was trained on the fgvc7_train dataset to translate the images of healthy leaves (A) to an image of scab-infected leaves (B). The network was trained for 200 epochs. At that point, the network was converged. The initial learning rate was set to $2 \cdot 10^{-4}$, with λ_{cyc} and $\lambda_{identity}$ set to 10 and 0.5 respectively. These parameters were based on the standard settings in the following repository (Cap et al., 2020a). The fake scab-infected images generated by this CycleGAN were used as additional training data for the scab classifier (see Section 2.5).

2.3 Classification of images of healthy and infected leaves

To classify the images of healthy and infected leaves, a ResNet-18 network was used. The network was trained for each experiment (see section 2.5) for 150 epochs with a batch size of 12, and an image size of 256 by 256 pixels. Early stopping was applied when the validation loss on the fgvc7 valid dataset did not improve for 25 epochs.

The performance of the classifier is dependent on the hyperparameters. The optimal settings of the classifier were found by testing 9 different combinations of learning rate, optimizer, and learning scheduler on the fgvc7 train dataset, see Table 2. The momentum and weight decay of SGD were fixed to 0.9 and $5 \cdot 10^{-4}$ respectively. The settings with the best accuracy on the fgvc7 test dataset were used to fix training settings for all other datasets. Normally the test set is not used to determine the optimal settings, however, in this research the actual test set is fgvc8, which was not used during training and hyperparameter optimization.

Table 2. Summary of different settings for training resnet18 classifier on fgvc7 dataset. The training settings with the highest accuracy on fgvc7 test were used to train all other datasets.

Training combination	Optimizer	Learning rate	Scheduler
1	SGD	0.01	CosineAnnealingLR
2	SGD	0.001	CosineAnnealingLR
3	SGD	0.0001	CosineAnnealingLR
4	SGD	0.01	stepLR, stepsize=15 ,λ=0.1
5	SGD	0.001	stepLR, stepsize =15, λ =0.1
6	SGD	0.0001	stepLR, stepsize =15, λ =0.1
7	Adam	0.01	CosineAnnealingLR
8	Adam	0.001	CosineAnnealingLR
9	Adam	0.0001	CosineAnnealingLR

2.4 Evaluation

Quality of fake images: The generated images of CycleGAN were evaluated using the Fréchet inception distance (FID), which is a metric that calculates the distance between real and fake images by representing each image by a feature vector of 2048 elements (Heusel et al., 2017). This vector is made using the Inceptionv3 classification network pre-trained on the ImageNet dataset. All vectors of both domains are summarized into two Gaussian distributions. The distance between both distributions is known as the Fréchet distance (Heusel et al., 2017). The higher the similarity between images the lower the distance and FID, with a perfect score of zero indicating that both groups are completely identical.

Classification performance: The classification performance was tested by calculating the precision, recall, F-score, and accuracy on the independent fgvc8 dataset.

2.5 Experiments

The goal of this research is to investigate if adding fake training images using domain adaptation with CycleGAN improves the performance of a classifier on an independent test set. To make a fair comparison, the CycleGAN was trained on the training set also used by the classifier, so no new data was used for training. Occasionally, in literature new data is used to train the CycleGAN (Cap et al., 2020b), but this gives an unfair advantage for the classifier using fake data.

Apart from the original fgvc7-train set, three new datasets were created to train the ResNet classifier. In the first dataset, named "ExtraTrain", 50% of the healthy images in fgvc7-train (also used to train CycleGAN) were converted to scab images using CycleGAN. In the second dataset, "ExtraValid", all healthy fgvc7-valid images (not used during training of the CycleGAN) were converted to scab images. The last set, "ExtraBenchmark", additional real data from the validation set was used to compare the added value of the fake images with respect to additional real images. The datasets to train the classifier are summarized in Table 3.

Significant differences in performances between the datasets were calculated using an unpaired two-sided Wilcoxon test.

Table 3. Summary of new composed datasets

Name	Classifier trainings dataset	# images healthy/scab
Benchmark	Fgvc7_train	364/412
ExtraTrain	Fgvc7_train + 50% healthy fgvc7_train -> scab	364/594
ExtraValid	Fgvc7_train + 100% healthy fgvc7_valid -> scab	364/466
ExtraBenchmark	Fgvc7_train + fgvc7_valid	418/412

3. RESULTS

3.1 Selection of optimal classifier settings

In Table 4, the accuracy and standard deviation of the classifier are given for all training setting combinations (Table 2). The combination with the highest performance on the fgvc7 test dataset is selected for all other experiments. According to Table 4 setting 2 is optimal, which is SGD with an initial learning rate of 0.001 and a cosine annealing scheduler. In this table, the settings with the highest initial learning rate also have the highest standard deviation, indicating that the steps were too large to converge properly. The settings with the lowest initial learning rate (3, 6, and 9) have the lowest standard deviation meaning that the model was precise but not accurate.

Table 4. Accuracy and standard deviation (std) of multiple training settings on the fgvc7 dataset

Training combination	fgvc7 val	std	fgvc7 test	std
1	93.11	8.23	91.55	6.75
2	97.14	1.01	96.57	0.76
3	91.09	0.41	90.61	0.21
4	87.06	12.29	88.17	8.82
5	95.13	2.05	95.68	0.98
6	93.61	0.41	96.53	0.23
7	72.94	6.43	75.07	6.31
8	90.92	5.61	89.48	3.23
9	97.82	1.20	94.98	1.21

3.2 Evaluation of CycleGAN

The FID score is calculated for all fake scab images in both the "ExtraValid" and "ExtraTrain" dataset. The results are summarized in Table 5. In this table, the fgvc7 test set is used as reference, because this is the only dataset that is not used in either training CycleGAN or creating the images. Also, the similarity between the fgvc7 test with the real valid and train subsets are included. There is a large difference between ExtraValid and ExtraTrain FID scores indicating that the images of the ExtraTrain dataset have a higher similarity with the fgvc7 test dataset than ExtraValid. This can be caused by the fact that CycleGAN is trained on the fgvc7 train datasets.

Table 5. FID similarity scores, the lower the score the higher the similarity between both datasets. Valid and train of the real dataset are included as reference.

Dataset 1	Dataset 2	FID score
fgvc7_testB	ExtraValid_fakeB	69.0
fgvc7_testB	ExtraTrain_fakeB	55.0
fgvc7_testB	fgvc7_valid_realB	51.0
fgvc7_testB	fgvc7_train_realB	35.8

The FID scores on the real scab images compared to ExtraTrain and ExtraValid indicates that the original healthy leaves in ExtraValid and ExtraTrain are not perfectly adapted. In Figure 2, two examples are shown with in the left column the real images and in the right column the fake images. Both fake leaves contain scab symptoms, but it should be noted that the images are slightly blurred. This is especially visible on the nerves of the leaf. Furthermore, it seems that not only the leaf but also the background is altered.

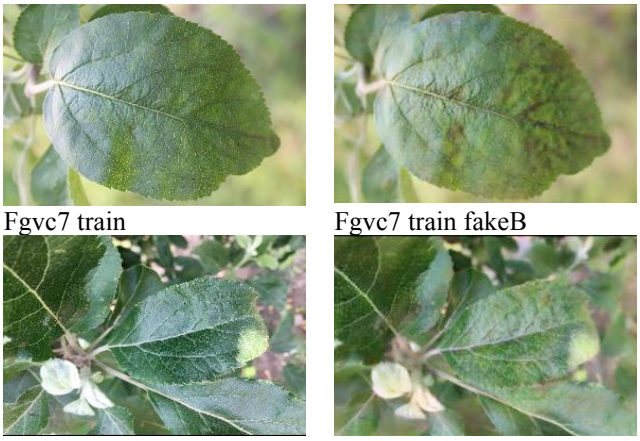

Figure 2. Example of real images (left column) and their corresponding generated fake image (right column).

3.3 Evaluation on fgvc8 dataset

The generated datasets described in section 2.4 are evaluated on the fgvc8 dataset by training the ResNet classifier. The performance of the classifier is summarized in Table 6. From this table, it can be concluded that each network has a relatively high precision compared to the recall. This means that many of the scab images that were classified as scab were correct, with only ±1% false positives. However, the recall is a poor 0.4 indicating that many scab images were missed (false negatives). The extra scab images in the ExtraTrain dataset resulted in higher accuracy than the benchmark, but both ExtraTrain ($p=0.24$) and ExtraValid ($p=0.19$) were non-significant. Only the ExtraBenchMark dataset with only non-adapted healthy images added had a significant difference with the benchmark.

Table 6. Result of classifier on fgvc8 dataset. Pr=precision, Re=recall, Fs=Fscore and Acc=accuracy. *significant difference with benchmark dataset $p<0.05$.

	Benchmark		ExtraTrain		ExtraValid		ExtraBench-Mark	
Pr	0.99	0	0.99	0	0.99	0	0.99	0
Re	0.38	0.05	0.41	0.05	0.35	0.05	0.53	0.03
Fs	0.55	0.05	0.58	0.05	0.51	0.06	0.69	0.02
Acc	68.43	2.55	69.49	2.45	66.54	2.72	75.75*	1.47

In section 3.1, we selected training option 2 for training the classifiers. However, given the high standard deviation in Table 4 other options could also lead to a reasonable result. Therefore, the datasets are trained again, but now with training option 5 (Table 2). The results are summarized in Table 7. Both Extravalid and ExtraBenchMark were significantly different, indicating that adding images with CycleGAN can result in lower performance compared to the benchmark. A possible explanation for this phenomena is shown in Figure 3. All sub-images in this figure should have scab symptoms, but none of these symptoms are clearly visible. These images may be perceived as mislabelled.

Table 7. Result of classifier on fgvc8 dataset training with training setting 6. *significant difference with benchmark dataset $p<0.05$.

	Benchmark		ExtraTrain		ExtraValid		ExtraBench-Mark	
Pr	0.99	0.01	0.97	0.01	0.99	0.01	0.98	0.01
Re	0.44	0.04	0.46	0.06	0.38	0.05	0.57	0.06
Fs	0.60	0.03	0.62	0.05	0.55	0.05	0.72	0.05
Acc	70.93	1.70	71.87	2.48	68.09*	2.27	77.49*	2.89

Figure 3. Some generated fake scab images in the Extravalid dataset do not contain any scab symptoms.

4. DISCUSSION

4.1 Classifier training settings

In section 3.1, the optimal settings for the ResNet classifier were determined. This was done by changing the learning rate/optimizer or scheduler on the benchmark dataset. Finding the optimal settings is therefore biased for this dataset. It is possible that the ExtraValid and ExtraTrain datasets have different optimal training settings. This results in an underestimation of the actual performance of the new datasets. Furthermore, only 9 different settings have been tested, which is minimal considering the many unexplored parameters; momentum, weight decay warming up etc. however, we should take into account, that the focus of this paper is determination of the added value of a GAN on a semi-optimal classifier, rather than optimising the performance.

4.2 CycleGAN

CycleGAN was trained with almost all standard parameters. Compared to the classifier no grid search has been done to find the optimal parameters. It is therefore likely that the settings of CycleGAN are suboptimal. On the other hand, there are other peer-reviewed papers available that applied GANs without mentioning any hyperparameter optimisation (Arsenovic et al., 2019, Madsen et al., 2019). This does not mean that you should not do any hyperparameter optimisation, but it does indicate that it is not always done in literature. More importantly, it is actually a question if you want to include all generated images. As can be seen from Figure 3 several images do not have any clear scab symptoms. Adding these images to the training data will confuse the classifier. In addition, CycleGAN also tends to alter the entire image instead of only the object of interest (Figure 2)(Cap et al., 2020b). There are algorithms available that remove the focus on the background like AR-GAN or leafGAN (Nazki et al., 2020, Cap et al., 2020b), both have their pros and cons. AR-GAN is only tested for simple backgrounds according to Cap et al. (2020b) and leafGAN requires a semi-supervised segmentation module. More research is needed to create images with clear symptoms without altering the background.

4.3 Added value of GANs

In both Table 6 and Table 7, the datasets with the extra GAN data did not show a significant improvement compared to the benchmark. This result was unexpected since GANs are often mentioned as the solution for limited datasets and time-consuming annotations. A critical literature review of GANs in agriculture shows that the added value of a GAN is not always properly researched. For example, Nazki et al. (2020) compared the classification accuracy of a custom tomato plant disease dataset with 9 identifiable diseases classes. The performance of ResNet-50 was compared when trained with no augmentation, standard augmentation (random distortion, rotation, and flipping) and synthetically image generation by transferring healthy images to one of the 9 diseases. The overall accuracy for baseline was 80.9%, classical augmentation 81.7%, and synthetically generated diseases 86.1%. The result between the synthetically and base line was significant. Despite the improved result, the question arises whether it is fair to compare both datasets. The synthetic data was generated by converting healthy images to any disease, while the healthy class was not included in the 9 diseases dataset. This means that images that are converted by the GAN were new images, resulting in a larger dataset compared to the original one. This makes the comparison with the baseline and classical approach biased. A similar approach is found in the paper of (Cap et al., 2020b), which applied leafGAN on the validation dataset. In the research of (Arsenovic et al., 2019), StyleGAN was applied to generate images for plant disease detection. Although the average accuracy improved with 0.9% (88.0 to 88.8) no numbers about significance are mentioned.

Instead of creating new artificial data, the addition of real healthy images to the training set was investigated. The results were summarized in the column “ExtraBenchMark” in both Table 6 and **Table 7**. The results shows a significant improvement, while adding GAN generated images did not show a significant improvement. This result could partly be expected since new data is always better than data augmentation. On the other hand, it strongly shows that sometimes a small investment in extra images can result in a major improvement. It is of course interesting to know whether a GAN can improve the accuracy on an independent dataset. However, if a similar improvement can be reached by only adding a few real images, then the added value of the GAN is arguable.

5. CONCLUSION

The added value of CycleGAN is investigated by training CycleGAN on a 2020 plant pathology dataset and testing the performance using a classifier on the independent 2021 plant pathology dataset. Two different sub-datasets have been used, but none of them had a significant improvement with respect to the benchmark. This is partly caused by the fact that a GAN mostly focuses on the conversion of the complete image instead of only the object of interest. As shown in this research

some images are not transferred properly, resulting in wrongly labelled images. Instead of using GAN's to create artificial images a better solution is to add real images to the dataset, even when only one class e.g. 'healthy' is available. In this study, expanding the dataset with only 'healthy' plant images resulted in a significant improvement over the benchmark dataset. This shows that the added value of the GAN was arguable. As a result, in the evaluation of a GAN it can be interesting to determine how many real images you would need to add to obtain a similar performance improvement.

ACKNOWLEDGEMENT

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 773718, project OPTIMA and the Dutch KB (knowledge base) program 38 "Advances in data-driven phenotyping" of Wageningen UR.

REFERENCES

ARSENOVIC, M., KARANOVIC, M., SLADOJEVIC, S., ANDERLA, A. & STEFANOVIC, D. 2019. Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry,* 11**,** 939.

BLOK, P. M., VAN EVERT, F. K., TIELEN, A. P., VAN HENTEN, E. J. & KOOTSTRA, G. 2020. The effect of data augmentation and network simplification on the image-based detection of broccoli heads with Mask R-CNN. *Journal of Field Robotics,* 38**,** 85-104.

BOWLES, C., CHEN, L., GUERRERO, R., BENTLEY, P., GUNN, R., HAMMERS, A., DICKIE, D. A., HERNÁNDEZ, M. V., WARDLAW, J. & RUECKERT, D. 2018. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

CAP, Q. H., UGA, H., KAGIWADA, S. & IYATOMI, H. 2020a. *Leafgan* [Online]. Available: https://github.com/IyatomiLab/LeafGAN [Accessed 03-02-2022].

CAP, Q. H., UGA, H., KAGIWADA, S. & IYATOMI, H. 2020b. Leafgan: An effective data augmentation method for practical plant disease diagnosis. *IEEE Transactions on Automation Science and Engineering*.

DEVRIES, T. & TAYLOR, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

ESPEJO-GARCIA, B., MYLONAS, N., ATHANASAKOS, L. & FOUNTAS, S. 2020. Improving weeds identification with a repository of agricultural pre-trained deep neural networks. *Computers and Electronics in Agriculture,* 175**,** 105593.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. & BENGIO, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems,* 27.

HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., NESSLER, B. & HOCHREITER, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems,* 30.

ISOLA, P., ZHU, J.-Y., ZHOU, T. & EFROS, A. A. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 1125-1134.

KAMILARIS, A. & PRENAFETA-BOLDÚ, F. X. 2018. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture,* 147**,** 70-90.

KUZNICHOV, D., ZVIRIN, A., HONEN, Y. & KIMMEL, R. Data augmentation for leaf segmentation and counting tasks in rosette plants. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019. 0-0.

MADSEN, S. L., DYRMANN, M., JØRGENSEN, R. N. & KARSTOFT, H. 2019. Generating artificial images of plant seedlings using generative adversarial networks. *Biosystems Engineering,* 187**,** 147-159.

MIRZA, M. & OSINDERO, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

NAZKI, H., YOON, S., FUENTES, A. & PARK, D. S. 2020. Unsupervised image translation using adversarial networks for improved plant disease recognition. *Computers and Electronics in Agriculture,* 168**,** 105117.

PANG, Y., LIN, J., QIN, T. & CHEN, Z. 2021. Image-to-Image Translation: Methods and Applications. *arXiv preprint arXiv:2101.08629*.

PARK, T., EFROS, A. A., ZHANG, R. & ZHU, J.-Y. Contrastive learning for unpaired image-to-image translation. European Conference on Computer Vision, 2020. Springer, 319-345.

SAVARY, S., WILLOCQUET, L., PETHYBRIDGE, S. J., ESKER, P., MCROBERTS, N. & NELSON, A. 2019. The global burden of pathogens and pests on major food crops. *Nature ecology & evolution,* 3**,** 430-439.

SHORTEN, C. & KHOSHGOFTAAR, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data,* 6**,** 1-48.

THAPA, R., ZHANG, K., SNAVELY, N., BELONGIE, S. & KHAN, A. 2020. The Plant Pathology Challenge 2020 data set to classify foliar disease of apples. *Applications in Plant Sciences,* 8**,** e11390.

YUN, S., HAN, D., OH, S. J., CHUN, S., CHOE, J. & YOO, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 6023-6032.

ZHANG, H., CISSE, M., DAUPHIN, Y. N. & LOPEZ-PAZ, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

ZHU, J.-Y., PARK, T., ISOLA, P. & EFROS, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017. 2223-2232.