IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON NEW TRENDS IN BRAIN SIGNAL PROCESSING AND ANALYSIS

Received January 23, 2019, accepted February 6, 2019, date of publication February 21, 2019, date of current version April 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2900327

Classification of Canker on Small Datasets Using Improved Deep Convolutional Generative Adversarial Networks

MIN ZHANG[1], SHUHENG LIU[2], FANGYUN YANG[3], AND JI LIU[1]

[1]Chongqing Key Laboratory of Software Theory and Technology, College of Computer Science, Chongqing University, Chongqing 400044, China
[2]Department of Automation, Chongqing University, Chongqing 400044, China
[3]Citrus Research Institute, Southwest University, Chongqing 400712, China

Corresponding author: Min Zhang (zm@cqu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61701051.

ABSTRACT This paper proposes a deep learning model for the classification of citrus canker that overcomes the shortcomings of traditional approaches, and the scarce number of available images for training, which have been subject to the overfitting limitation. To address the issues, we propose two approaches, namely, the feature magnification and objective breakdown optimization, to augment datasets, and emphasizes meaningful features and prevent the model from overfitting noise signals. The first approach mimics the distribution of positive samples with a generative model based on deep convolutional generative adversarial networks. The second approach updates different parts of the model with optimization objectives, whereby back-propagated error signals are no longer the only signal for updating parameters. To validate the proposed approaches, we present theoretical proofs to justify the correctness of our methods and conduct extensive case studies and experiments to show that the proposed approaches clearly outperform traditional approaches on the classification of accuracy and efficiency of small training sets. In this paper, a methodology is proposed to generate a general model. The model can be applied to other bio-medical applications, where the scarcity of visual samples makes it difficult for a normal deep learning model to work without overfitting.

INDEX TERMS Generative adversarial networks, Siamese learning, feature magnification, optimization objective breakdown.

I. INTRODUCTION

A. TRADITIONAL APPROACHES USED TO IDENTIFYING CITRUS CANKER

Citrus canker is an infectious and fatal disease for citrus plants, that could potentially lead to heavy crop losses. Given that one of the common measures enforced to prevent citrus canker infections is to decimate all citrus products near the site of infection, owners of citrus plants have very little motive to report citrus canker epidemics, or to provide visual samples of afflicted plants.

Before the mass application of computer vision, the diagnosis of citrus canker was difficult owing to the facts that a) the disease has very obscure visual traits in its early stages, and b) its identification requires considerable human experience.

Traditional computer vision techniques seek to program the model to identify a series of traits that empirically differentiate positive samples from negative ones.Yet, such approaches require complicated logic structures and it is thus difficult to achieve adequate accuracy.

Recently, machine learning and especially convolutional neural network models have exhibited considerable strengths for image identification applications. However, most deep learning algorithms are too complex in network structures and require a large training set. For citrus canker classification applications, the training instances and related data are scarce. Novel models and algorithms are in high need for the exploitation of the scarcity of training images for yielding a good identification accuracy.

B. PROPOSED APPROACHES

Our approach includes multiple steps:

1)We first augment the training set by feeding it to the convolutional network with both real-world and artificial samples. The artificial samples are generated from generative adversarial networks and are trained to match the distribution

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Hugo Albuquerque.

of authentic samples. This process is referred to as feature magnification because it essentially replicates and emphasizes meaningful features to prevent the model from overfitting to noise signals. With the expanded training set, the convolutional network is capturing characteristic visual traits based on which positive samples can be identified.

2)Furthermore, in order to speed up training and reduce the possibility of overfitting, we base our model on a lightweight AlexNet with implemented modifications of the optimization objective and the parameter updating mechanism where Siamese training is involved. Specifically, we split the model into two parts, separated by a chosen latent layer l. Preceding parameters are updated with the objective of minimizing the Siamese loss on l, and forcing latent representations to be linearly separable. Parameters in successive layers are updated to identify a decision boundary and allow the completion of the classification task. In view of the optimization objectives, we are able to mathematically prove and experimentally demonstrate that the model performance is enhanced.

II. RELATED STUDIES

Recently, considerable research efforts were paid to the problem of citrus canker identification. Traditional approaches used for the identification of citrus canker usually employed machine learning algorithms such as AdaBoost. Zhang *et. al* [1] proposed a method to detect citrus canker with the use of a comprehensive method based on global and local descriptors that achieved a similar classification accuracies as those achieved by human experts. Additionally, the authors of this study extended their methodology even further and proposed an approach that can adaptively detect citrus canker online. Lately, a pipeline was introduced by Sunny and Gandhi in [2] for the identification of canker with a support vector machine (SVM) classifier.

Models that apply deep learning techniques have not been introduced until recent years given the lack of computational power. Most of the latest approaches that have been proven to be efficient rely on neural networks to identify plant features that are highly intractable by usual methods. Reyes *et. al* [3] used convolutional neural networks to classify 1000 plant images, including stems or leaves, flowers, fruit, and entire trees. The results showed that the average accuracy rate was 0.486, and the classification accuracy for leaves and flowers was the highest compared to other types. Xiaolong *et al.* [4] applied deep convolutional neural networks to promote the capacity to recognize plant leaves in complex environments. Tan *et al.* [5] proposed a method to identify fruit–melon lesions with a convolutional neural network (CNN).

Although modern neural networks can be considerably deep and complex, many medical classification tasks can be pursued with a relatively shallower CNN classifier. Li *et al.* [6] customized a shallow CNN framework to classify interstitial lung disease (ILD). Another classification model for ILD with a seven-layer CNN was proposed by Hattikatti *et al.* [7] in 2017, which used a local binary pattern for feature extraction. Because of the rapid advancement of CNNs, such measures are easily transferred to other classification tasks in the field of biomedicine, including the recognition of citrus canker.

AlexNet, proposed by A. Krizhevsky, is one of the most prevalent CNNs with publicly available pretrained weights. The model comprises five convolutional layers and three fully connected layers. AlexNet achieved a state-of-the-art performance in the ImageNet Large Scale Visual Recognition Challenge in 2012, with a top five error of 15.3%, and laid the foundation for its successors. AlexNet was designed to allow the input of images with matrix sizes of 227*227 pixels, and performed a 1,000 label classification task on these images. During the training of the model, error signals were back propagated through each of the layers, and all the encountered parameters were updated. Tajbakhsh *et al.* [8] considered three distinct medical fields (radiology, cardiology, and gastroenterology) before they evaluated the effects of pretraining CNNs on large datasets. Their findings showed that models with pretrained initial weights can perform as well as—if not better than—models trained from the beginning. This finding allows the fitting of models with limited data for specific tasks if the model is pretrained on an extensive dataset.

In this study, however, we approach the image classification problem from another perspective and update different parts of the model with different optimization goals. Errors computed against ground truths are no longer the sole signal for tuning our model, and peer comparisons are conducted to update model parameters as proposed by Krizhevsky *et al.* [9].

Apart from the CNNs, another major development in deep learning is the development of generative adversarial networks (GAN) first proposed by Goodfellow *et al.* in [10]. A GAN comprises two major parts: a) a discriminator that aims to distinguish real samples from all other images, and b) a generator that aims to fool the discriminator by generating artificial images that closely resemble the real ones [11]. Successful GAN variants include the deep convolutional GAN (DCGAN) [12], InfoGAN [13], energy-based GAN [14], Wasserstein GAN (WGAN) [15], Wasserstein GAN with gradient penalty (WGAN–GP) [16], RCNN [17], and others.

In this study, we base our model on the DCGAN, whose discriminator and generator both contain convolution (deconvolution) layers, batch normalization layers, and periodically repeating ReLU activations.

III. MODEL STRUCTURE

The performance of an image classification model depends heavily on its ability to capture abstract features that are distinctive for each category. To ensure that a model learns these features, a sufficiently large training set with labels is usually required. However, in the context of citrus canker recognition, samples available for model training are often scarce, and the

visual traits of positive samples are too obscure such that only very few experts are able to correctly annotate the samples.

To address these problems, we come up with specific alternatives that prove to improve model performance.

A. FEATURE MAGNIFICATION

To remedy the inadequacy of training samples (positive samples in particular), we construct a generative model to produce realistic artificial samples that share a similar distribution to that for genuine samples, a process which we refer to as feature magnification.

The generative model we employ is a custom-architected variant of DCGAN, which proves to be effective in many applications. Yet, owing to limitations of the loss function [18] [15], the discriminator network tends to overfit, thus making it difficult to update the generator in the right direction. In light of this, we implemented the following modifications to the original model.

We performed data augmentation before the learning of the statistical distribution of the genuine samples. Specifically, for each image sample, we added to the dataset copies of the sample that were flipped horizontally and vertically, along with four other copies of the sample rotated by 0°, 90°, 180°, and 270°. In this manner, the size of the training data is six times larger, making it less likely for the discriminator to overfit.

We implemented our discriminator model using a custom architecture based on DCGAN by incorporating it between each convolutional and batch normalization layers, which we refer to as a two-dimensional mute layer. A mute layer works in a similar manner as the dropout layer except that the former leaves out a certain proportion of signals during both training and validation, while the latter only does so during training, and is skipped during validation.

The insertion of such a layer is an unorthodox practice because existent studies have shown that dropout layers do not work well where batch normalization (BN) is also employed (see [19]). Accordingly, we explain the reasoning behind this choice as follows. Normally, when dropout is applied in the training phase, the scaling factor $\frac{1}{p}$ causes variance shifts of the batch of samples, thus misleading the batch normalization. Therefore, in the validation phase, where the dropout layer is no longer active, the batch normalization layer does not yield the ideal input for the pursuant convolutional layer (which is trained to accept input normalized by a different variance).

For a mute layer, however, because we indiscriminately leave out a part of the signals during the training and validation processes, the influence of the variance shift is cancelled out. The mute layer merely checks for possible overfitting by the discriminator and does not enhance its performance.

After successfully training the generative part of the model, we mix the output of the artificial samples with genuine positive samples and feed them to the classification part of the model. Experimental results in Section 4 indicate that such procedure can increase model performance and speed up the training process.

B. OPTIMIZATION OBJECTIVE BREAKDOWN

In the context of citrus canker recognition, the task is to identify positive samples from negative ones, including the leaves that are unaffected and the leaves that suffer from other diseases.

Traditional image classification models deal with multi-label classification with hundreds or thousands of classes, which is excessive for our tasks, and can result in underfitting. Actually, for the purpose of recording and extracting relevant features in our model, a simple AlexNet model should suffice provided that the model knows the traits which it should look for.

Because the de novo training of AlexNet can be very slow, even with the use of the feature magnification procedure, we initialize the model with publicly available weights pre-trained on ImageNet and update the weights only on the three fully connected (FC) layers, thus leaving all other parameters unchanged.

A trickier problem arises from the fact that some samples of different class labels are so close to each other that visual traits differ in a very subtle manner. The model initialized with pretrained weights can hardly learn to capture these features without occasional overfitting. Therefore, we incorporated Siamese training into the network. Specifically, we split the model into two parts with the use different training objectives for its different parts [20] [21].

The two parts are separated by a latent layer l based on a chosen hyperparameter. The objective of all its preceding layers is to minimize the computed Siamese loss, while the objective of all its succeeding layers is to minimize the cross-entropy computed against ground truths on the output layer (the last layer).

The Siamese loss on a latent layer is defined to be a function of latent representation vectors. Let x be the input image tensor, $c \in \{0, 1\}$ be the class label, $y \in (0, 1)$ be the output prediction scalar, and z_i $(i = 0, 1, \ldots, n)$ be the latent representation in the ith layer, where $z_0 = x$ and $z_n = y$.

The Siamese loss function of latent representations of the two input images $x^{(1)}$ and $x^{(2)}$ in layer i is defined as,

$$\text{loss} = \begin{cases} \max\left(0, \|z^{(1)} - z^{(2)}\| - \alpha_0\right)^p & c^{(1)} = c^{(2)} = 0 \\ \max\left(0, \|z^{(1)} - z^{(2)}\| - \alpha_1\right)^p & c^{(1)} = c^{(2)} = 1 \\ \max\left(0, \alpha_2 - \|z^{(1)} - z^{(2)}\|\right)^p & \text{otherwise} \end{cases} \quad (1)$$

where the subscript i in z_i is omitted for brevity purposes. Intuitively, the loss function imposes penalties on the model when

1) latent representations of two images from the same class fall too far away from each other (beyond the margins of α_0 or α_1), or
2) latent representations of two images from different classes fall too close to each other (within the margin of α_2).

FIGURE 1. Sample augmentation.

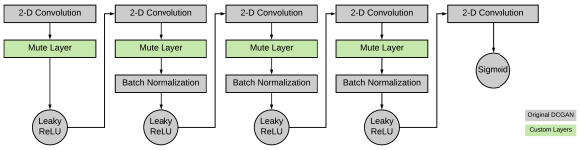

FIGURE 2. Architecture of our custom variant of DCGAN.

The total Siamese loss over a random batch of 2 m samples are given by,

$$L = \frac{1}{2m}\sum_{k=1}^{m} \text{loss}(z^{(2k-1)}, z^{(2k)}, c^{(2k-1)}, c^{(2k)}) \tag{2}$$

Ideally, L is optimized to its global minimum 0 for any random sample batch, thus leading to the following conclusions,

1) latent representation of any negative sample lies within a ball B_0 of radius $\frac{\beta_0}{2} \leq \frac{\alpha_0}{2}$
2) latent representation of any negative sample lies within a ball B_1 of radius $\frac{\beta_1}{2} \leq \frac{\alpha_1}{2}$
3) positive and negative latent representations are separable by uncountably many choices of a hyperplane if $\alpha_2 > \sqrt{\frac{\beta_0^2 + \beta_1^2}{2}}$ (see Appendix for proof)

In the scenario where latent representations are linearly separable, the objective of successive layers is to identify a hyperplane based on which different sample categories are separated. Therefore, a simple affine transformation followed by ReLU activation would suffice, and can be performed with a single layer of perceptrons.

Note that the Siamese loss imposes a much looser constraint than traditional loss functions, such as cross-entropy computed based on prediction logits against ground truths. Siamese loss is insensitive to the absolute positions of input vectors and focuses on their relative distances. Cross-entropy, however, is susceptible to any rotation or shift of the representations as a whole, even if their relative distances remain constant.

In practice, the Siamese loss cannot always be reduced to 0, yet a latent layer trained to separate representations still intuitively helps facilitate the training process. Empirically, even for cases where $\min(\beta_0, \beta_1) < \alpha_2 < \sqrt{\frac{\beta_0^2 + \beta_1^2}{2}}$, the latent layer still maintain some weak separability of the representations, although Siamese loss is therefore very helpful when dealing with classification tasks where we are only concerned about the relative distribution of sample representations instead of their absolute positions. In practice, the Siamese loss cannot always be reduced to 0, yet a latent layer trained to separate representations still intuitively helps facilitate the training process.

Empirically, even for cases where $\min(\beta_0, \beta_1) < \alpha_2 < \sqrt{\frac{\beta_0^2 + \beta_1^2}{2}}$, the latent layer still maintains some weak separability of the representations, although representations of the same class label may be scattered into several clusters that cannot be separated by a single hyperplane.

A low-dimensional example of such irregularity is shown in Appendix B. To remedy the irregularity encountered in practice, there should always be more than one layer of perceptrons following the latent layer.

IV. EXPERIMETNAL RESULT

A. FEATURE MAGNIFICATION

To expand the set of positive training samples with feature magnification, we augmented 800 samples images to 6 times as stated previously, and input the expanded images to the generative model.

In our case, the generative model is a DCGAN with inserted custom mute layers. The model was trained for 2,000 epochs, with the generator and discriminator updated once every epoch.

The batch size was set to 128 for an ideal batch normalization performance. We then split the 891 positive samples into two categories based on the diagnosis of human experts. Positive samples that have been correctly identified are used to form the training set, and positive samples that were falsely reported as negative were used to form the validation set. For the negative samples, we employed a 90%-10% training–validation split based on 2,244 negative samples randomly cropped from leaf images (which did not contain canker spots).

We then trained the AlexNet classifier for 100 epochs with a dropout rate of 0.5 and a batch size of 128. The learning curve for accuracy based on the validation set is illustrated in blue with its maximum set at 93.7%.

We then incorporated the generated positive samples into the training set, and performed the experiment described above once again. This yielded the red curve with a peak at 96.7%. Apparently, employing the artificial output to expand the training set can magnify some key features of the image samples and produce ideal results.

Our generated samples can boost the classification performance, as they represent a class of images that are similar to citrus canker. Therefore, they help train our classifier to find fine distinctions between the citrus canker and other diseases. We conducted an experiment where human experts were requested to identify generated images from real ones.

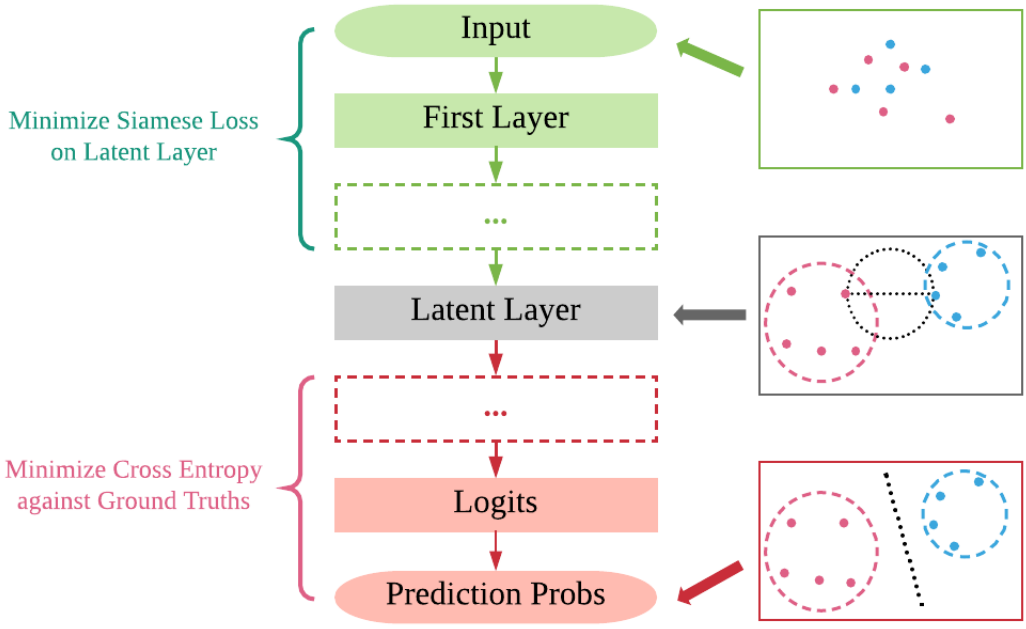

FIGURE 3. The model is divided into two parts, each with a different optimization objective. Error signals are not backpropagated beyond the latent layer.

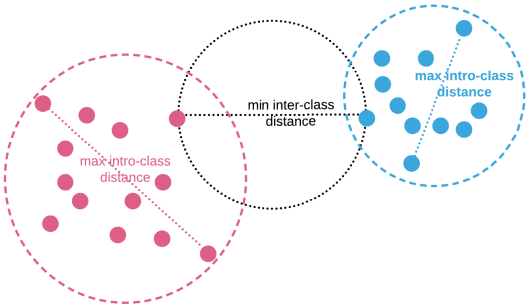

FIGURE 4. When Siamese loss is reduced to zero for any pair of samples, fake and real representations are respectively contained in two compact sets that are linearly separable.

TABLE 1. Best classification performances of human experts.

Human Experts (At Best)			
Precision	Recall	F1 Score	Accuracy
0.5514	0.4270	0.4813	0.6419

TABLE 2. Classification performance of human experts on average.

Human Experts (On Average)			
Precision	Recall	F1 Score	Accuracy
0.5224	0.3987	0.4523	0.6236

Given that experts fail to make a definite judgment on a few samples, we summarized their classification performances in three different charts.

Conversely, our classifier performed far better than human experts on this task. We trained our classifier for 100 epochs and obtained the best results, as follows.

The training curves of precision, recall, f-1 scores, accuracy, and cross-entropy, are respectively illustrated in Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9.

TABLE 3. Classification performance of human experts at worst.

Human Experts (At Worst)			
Precision	Recall	F1 Score	Accuracy
0.4715	0.3800	0.4208	0.5931

TABLE 4. Best classification performance of model trained for 100 epochs.

Classifier (Best of 100 epochs)			
Precision	Recall	F1 Score	Accuracy
0.8765	0.8658	0.8712	0.9009

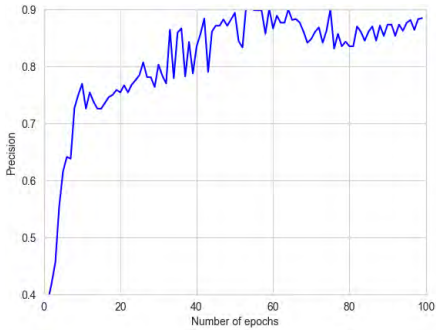

FIGURE 5. Training curve of precision.

B. MUTE LAYERS

Mute layers prove to be very important to model performance as indicated in the following experiment. We trained two models with exactly the same input as described in Section 4.1. These models share the same choice of hyper-parameters and only differ slightly in their discriminator network architectures. Specifically, one of them is implemented with an architecture which contained additional customized mute layers, and the other has no such layers. Both models were sufficiently trained (with 2,000 epochs) and yielded

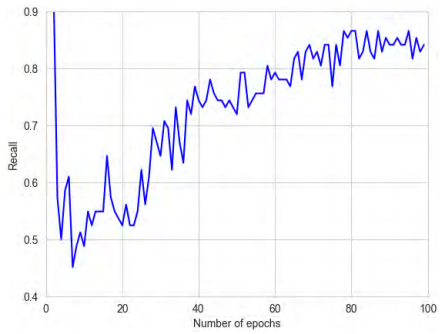

FIGURE 6. Training curve of recall.

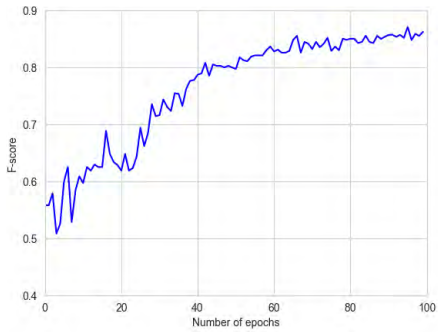

FIGURE 7. Training curve of F-score.

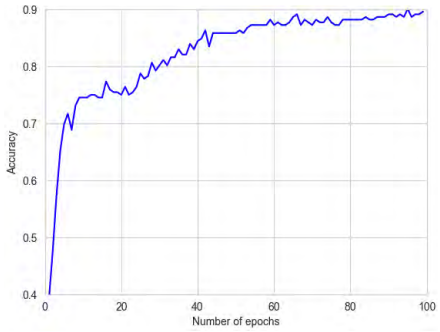

FIGURE 8. Training curve of accuracy.

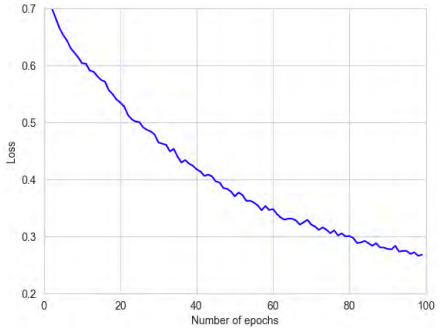

FIGURE 9. Training curve of loss.

meaningful results. Shown below are some image samples generated by both models as well as the real samples used to train them.

Closer examination of these photos empirically shows that samples generated by the first model (with mute layers) is less blurry in shape and more realistic in color (more samples are shown in the Appendix).

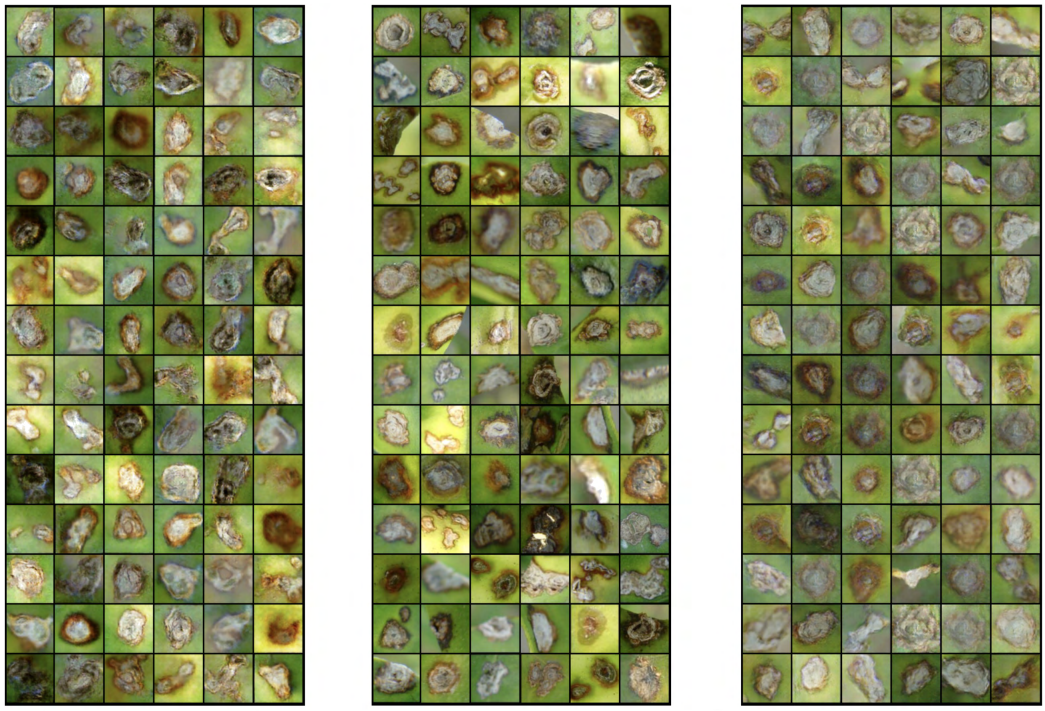

FIGURE 10. Real images and generated images.

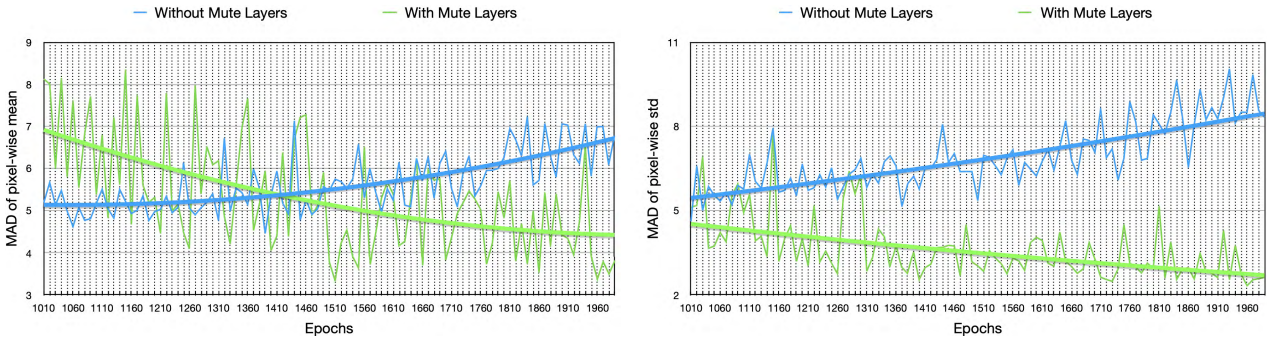

FIGURE 11. MAD curves of pixel-wise mean and standard deviation on a [0, 256) scale.

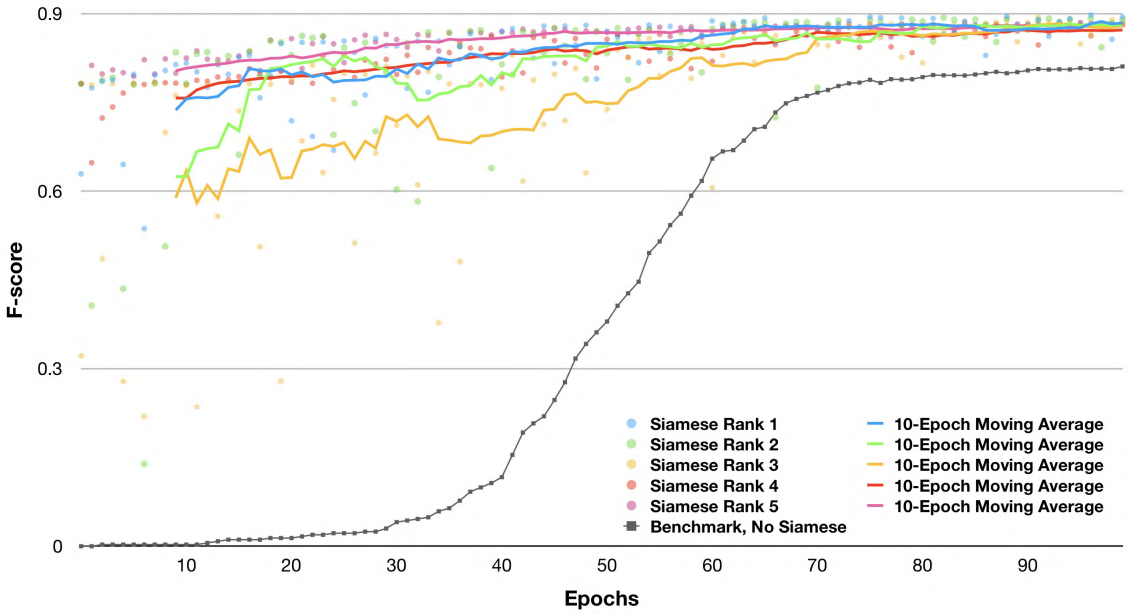

FIGURE 12. F1 score of five parallel experiments against the benchmark in association with the trendlines of a 10-epoch moving average.

We quantitatively evaluated how well the model mimicked the real distribution with the use of the mean and standard deviation. These two metrics were computed on a pixelwise basis across all images in a batch of 128 samples.

Obviously, without the mute layer the mean absolute differences (MAD) of the metrics of the real distribution indicated how poorly the model performed. We plotted the MAD curves because the model has been stabilized (after the 1,000th epoch), and fitted them to an exponential curve in the following figure.

It is immediately apparent from these curves that mute layers do help in the assessment of the overfitting of the discriminator network and thus improve the quality of generated samples. More importantly, the generative model does not even converge without the mute layers.

C. OPTIMIZATION OBJECTIVE BREAKDOWN

We now experimentally prove that the optimization objective breakdown helps improve the performance of the classification networks.

For illustration purposes, we still use the AlexNet in this experiment, and choose the first dropout layer (a.k.a. dropout–6) as the latent layer.

Hyperparameters α_0, α_1 and α_2 are set to 4, 3 and $\frac{5}{\sqrt{2}}$ respectively, which satisfies the constraint $\sqrt{\frac{\alpha_0^2+\alpha_1^2}{2}} \leq \alpha_2$.

We executed and compared our model against a benchmark. In this case, Siamese training was not applied, and the sole objective of all parameters was the minimization of cross-entropy computed on the last layer.

Given that Siamese loss was not monotonically decreasing during training, we took a snapshot of model parameters in conjunction with current Siamese loss computed on validation set in each epoch. We then sorted these losses in descending order. Subsequently, a total of five parallel experiments were conducted with sets of parameters which corresponded to the five lowest losses.

These five candidate models were tested against a benchmark. Siamese training was not involved, and all layers

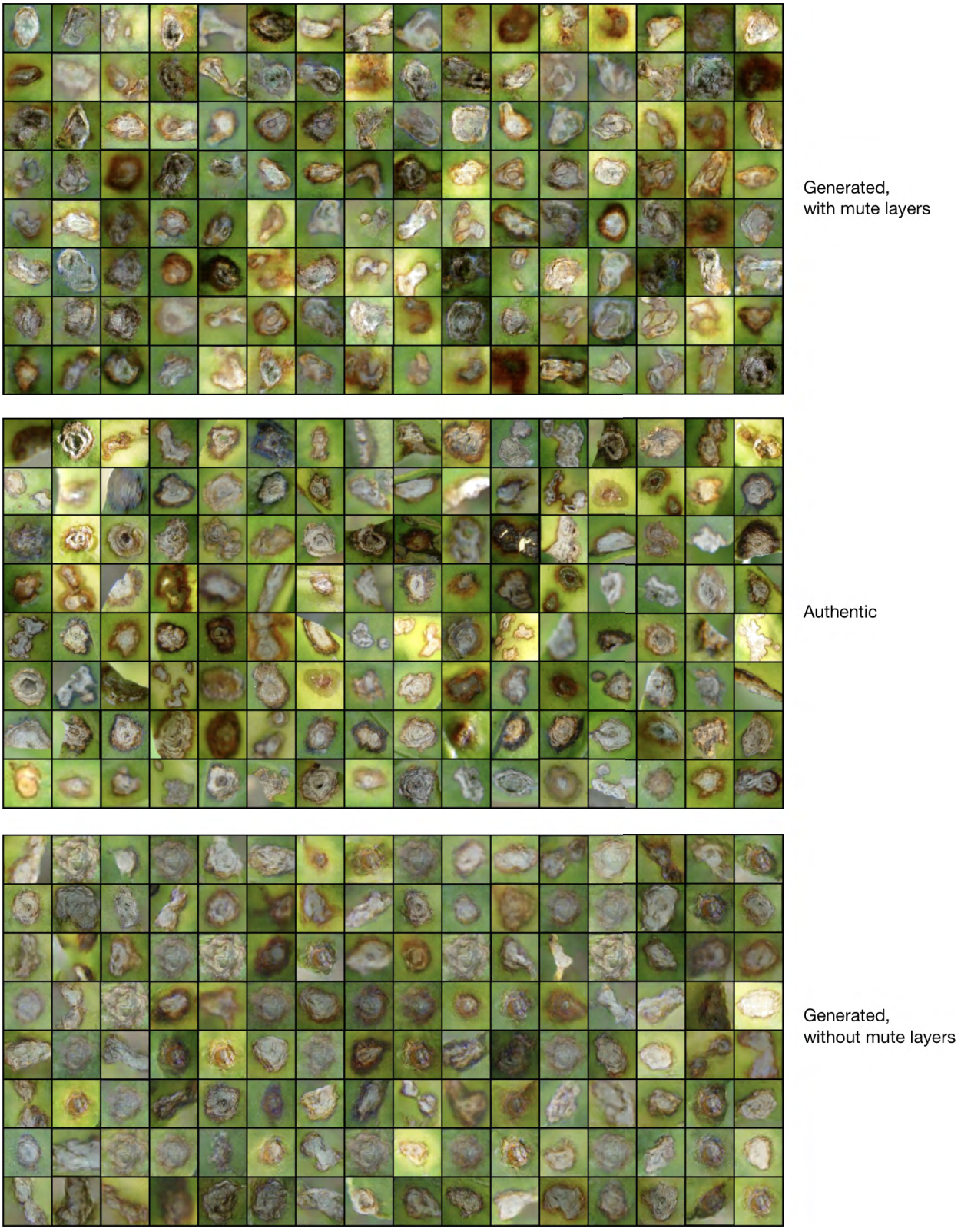

FIGURE 13.

were merely updated as the back-propagated error signal dictated.

The technique of optimization objective breakdown does not guarantee a high performance on a consistent basis.

Therefore, hypertuning is recommended. In our experiment, the F-score of the model without this technique reaches a value at approximately 70%, as shown in Fig. 12. This is probably because the model failed to update its weight in the right direction.

V. GENERALIZATION OF MODEL

Our model used for the identification of positive samples of citrus canker can also be applied to other biomedical applications where the scarcity of visual samples makes it difficult for normal deep learning models to capture relevant features without overfitting.

Herein, are some general ideas on how to build a good model. The first measure to be considered is performing feature magnification by selecting a generative model that captures the distribution of authentic samples. In the identification of citrus canker, our choice is a model with a custom-architecture based on DCGAN, yet a model from the GAN class is not always required. However, it is worth pointing out that if a GAN is chosen, the prevention of discriminator overfitting is crucial to the generator's performance. In our case, we employed data augmentation and mute layers to limit the performance of the discriminator that in turn served to enhance the quality of the generator's output.

When the generative model is sufficiently trained, a moderate amount of artificial samples can be input to the classifier model along with the genuine samples. In the case where only a small number of class labels are of concern, breaking down the optimization objective is helpful. We recommend employing a similar Siamese loss function on one or more latent layers because we are able to mathematically interpret the behavior of the model. Such methods help reduce model redundancy and prevent overfitting, provided that appropriate choices are made for the marginal parameters.

The Siamese training process on the latent layer has drawbacks. During Siamese training, a considerable number of possible pairs can be drawn from the training set. Therefore, the gradient descent on some arbitrary pairs of data points does not guarantee better latent representations. The only way to combat stochasticity in this procedure is to increase batch size during training. However, this requires a tremendously large hardware memory.

Nevertheless, there are some limitations associated with this study. First, we only chose mute layer owing to the scarcity of our data. Different ways of preventing GANs from overfitting can be used when the circumstances permit the use of larger datasets. Wasserstein GAN, for example, updates the generator network by minimizing the Wasserstein-1 distance between real and fake distributions. This approach also proves to be effective with large datasets. Second, for multilabel classification, the corollary in Appendix A indicates that the breakdown of the optimization objective with Siamese training is still an ideal tool for the linear separation of the latent representations. However, it should be noted that the number of required marginal parameters quadratically increases as a function of the label count. This increases the difficulty of hyperparameter grid searching. However, it must be noted that most biomedical tasks do not involve classifications of too many labels, and the technique should work as expected.

VI. CONCLUSION

In essence, feature magnification is a tool that generates more label samples to balance the dataset for classification. When sufficiently fine-tuned, the tool replicates the essential traits of the label and helps the subsequent classifier to capture the contour of each label in the sample space.

The breakdown of the optimization objective revolutionized the way model parameters are updated based on the addition of one or more internal optimization objectives for the model. These objectives collectively contributed to the refinement of the overall objective function, which is typically cross-entropy computed on ground truths and model outputs.

Possible defects include a slower training process if there are many available data points for sampling. However, for cases where training samples are scarce, the method is expected to outperform pure gradient-based training both in classification metrics and in the training speed.

To summarize, feature magnification and the optimization objective breakdown are effective auxiliary for the classification CNNs when the

- number of training samples is scarce
- dataset is unbalanced or even extremely skewed
- samples from different datasets share considerable similarities and differ only in minor traits

APPENDIX A
PROPOSITION AND COROLLARY

A. PROPOSITION

Let $X, Y \in \mathbb{R}^n$ be nonempty sets in the Euclidean space, and $\alpha_x, \alpha_y, \alpha_{xy}$ be some positive real numbers.

There are uncountably many choices of a hyperplane by which X and Y are separable if

1) $\forall x, x' \in X, \rho(x, x') < \alpha_x$,
2) $\forall y, y' \in Y, \rho(y, y') < \alpha_y$, and
3) $\forall x \in X, \forall y \in Y, \rho(x, y) > \alpha_{xy} > \sqrt{\frac{\alpha_x^2+\alpha_y^2}{2}}$.

B. PROOF

Let $x_0 \in \overline{X}$ and $y_0 \in \overline{Y}$ be two points such that $\rho(x_0, y_0) = \inf_{x\in X, y\in Y} \rho(x, y) = \rho_0$, where $\overline{S}$ denotes the closure of set S.

Let B_1 be a ball of radius α_x with a center at x_0, B_2 be a ball of radius α_y with a center at y_0, and B_3, B_4 denote two open balls with common radii α_{xy} with centers at x_0 and y_0, respectively.

It is obvious that $X \subset \overline{B_1}, Y \subset \overline{B_2}$ and that $X \cap B_4 = Y \cap B_3 = \emptyset$.

We now prove that X and Y are respectively contained in some disjoint convex sets that are both compact.

1) Closed balls $\overline{B_1}$ and $\overline{B_2}$ are disjoint.
   Accordingly, $\rho_0 > \alpha_x + \alpha_y$ is valid. Given that $\overline{B_1}$ and $\overline{B_2}$ are convex and compact, according to the hyperplane separation theorem, there exist two parallel hyperplanes in between $\overline{B_1}$ and $\overline{B_2}$ separated by a gap of $\rho_0 - \alpha_x - \alpha_y > 0$.
2) Closed balls $\overline{B_1}$ and $\overline{B_2}$ are not disjoint.
   Because $X \subset B_1 \subset \overline{B_1}$, $Y \subset B_2 \subset \overline{B_2}$, and $X \cap B_4 = Y \cap B_3 = \emptyset$, it follows that $X \subset \overline{B_1} \backslash B_4$ and $Y \subset \overline{B_2} \backslash B_3$, where both $\overline{B_1} \backslash B_4$ and $\overline{B_2} \backslash B_3$ are compact.
   Let C_1 and C_2 be respectively the convex hulls of $\overline{B_1} \backslash B_4$ and $\overline{B_2} \backslash B_3$.
   By construction, C_1 and C_2 are disjoint and compact. Based on the hyperplane separation theorem, there exist two parallel hyperplanes in between C_1 and C_2 separated by a gap of $\dfrac{2\alpha_{xy}^2 - \alpha_x^2 - \alpha_y^2}{2\rho_0} > 0$.

In both of these cases, there are uncountably many hyperplane choices that separate two convex sets in which X and Y are contained.

Q.E.D.

C. COROLLARY

Let $\mathcal{F} = \{X_i \subset \mathbb{R}^n : i = 1, 2, \ldots, m\}$ be a finite family of nonempty subsets of the Euclidean space and $A = (\alpha_{ij})_{m \times m}$ be a positive symmetric matrix such that

1) $\forall x_i, x_i' \in X_i$, $\rho(x_i, x_i') < \alpha_{ii}$, and
2) $\forall x_i \in X_i, \forall x_j \in X_j$, $\rho(x_i, x_j) > \alpha_{ij} > \sqrt{\dfrac{\alpha_{ii}^2 + \alpha_{jj}^2}{2}}$ for $i \neq j$.

There exist uncountably many choices of a partition of $\mathbb{R}^n$, denoted as $\{S_i\}_{i=1}^m$, such that

1) $X_i \subset S_i$, where $i = 1, 2, \ldots, m$, and
2) S_i and S_j $(i \neq j)$ are separable by some hyperplane H_{ij}.

APPENDIX B
GENERATED SAMPLES

In addition to the samples shown on page 49685, more samples are illustrated on page 49687. On the top, there are 128 images generated by a model with mute layers. In the middle, there are 128 images acquired from the real world. On the bottom, there are 128 images generated by a naive DCGAN. Both generative models are sufficiently trained based on 2,000 epochs.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their insightful comments and suggestions, which could significantly help in improving the quality of this paper.

REFERENCES

[1] M. Zhang and Q. Meng, "Automatic citrus canker detection from leaf images captured in field," *Pattern Recognit. Lett.*, vol. 32, no. 15, pp. 2036–2046, 2011.
[2] S. Sunny and M. P. I. Gandhi, "An efficient citrus canker detection method based on contrast limited adaptive histogram equalization enhancement," *Int. J. Appl. Eng. Res.*, vol. 13, no. 1, pp. 809–815, 2018.
[3] A. Reyes, J. Caicedoc, and J. Camargo, "Fine-tuning deep convolutional networks for plant recognition," in *Proc. CEUR Workshop*, 2015. [Online]. Available: http://CEUR-WS.org
[4] X. Zhu, M. Zhu, and H. Ren, "Method of plant leaf recognition based on improved deep convolutional neural network," *Congn. Syst. Res.*, vol. 52, pp. 223–233, Dec. 2018.
[5] W. Tan, C. Zhao, and H. Wu, "Intelligent alerting for fruit-melon lesion image based on momentum deep learning," *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 16741–16761, Dec. 2016.
[6] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proc. 13th Int. Conf. Control Automat. Robot. Vis.*, Dec. 2014, pp. 844–848.
[7] P. Hattikatti, "Texture based interstitial lung disease detection using convolutional neural network," in *Proc. Int. Conf. Big Data, IoT Data Sci.*, Dec. 2017, pp. 18–22.
[8] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[10] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[11] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.
[12] A. Radford, L. Metz, and S. Chintala. (Nov. 2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434
[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.
[14] J. Zhao, M. Mathieu, and Y. LeCunc. (Sep. 2016). "Energy-based generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.03126
[15] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein GAN." [Online]. Available: https://arxiv.org/abs/1701.07875
[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
[17] H. Li, Y. Huang, and Z. Zhang, "An improved faster R-CNN for same object retrieval," *IEEE Access*, vol. 5, pp. 13665–13676, 2017.
[18] M. Arjovsky and L. Bottou. (Jan. 2017). "Towards principled methods for training generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1701.04862
[19] X. Li, S. Chen, X. Hu, and J. Yang. (Jan. 2018). "Understanding the disharmony between dropout and batch normalization by variance shift." [Online]. Available: https://arxiv.org/abs/1801.05134
[20] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Trans. Graph.*, vol. 34, no. 4, Aug. 2015, Art. no. 98.
[21] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 850–865.

MIN ZHANG received the master's degree in computer software and theory and the Ph.D. degree in computer application from Chongqing University, China, in 2003 and 2008, respectively, where she is currently a Faculty Member of the College of Computer Science. Her research interests include machine learning algorithm, machine vision, and artificial intelligence applications.

SHUHENG LIU is currently pursuing the B.S. degree in automation with Chongqing University. He has received some awards in different competitions. He is expected to receive the B.S. degree this year, and he will pursue the M.S. degree in computer science thereafter. His research interest includes deep learning and its applications.

FANGYUN YANG received the master's degree from Southwest Agricultural University, in 2004, and the Ph.D. degree from Southwest University, China, in 2012, where he is currently a Faculty Member. He studies on virus indexing and virus exclusion for the superior citrus cultivars or bud lines selected in domestic or imported from abroad.

JI LIU received the master's and Ph.D. degrees in computer software and theory from Chongqing University, China, in 2005 and 2009, respectively, where he is currently a Faculty Member of the College of Computer Science. He has started working on tree modeling already during the Ph.D. thesis. His research interests include computer graphics, 3D reconstruction, and tree modeling.