in silico Plants, 2025, 7, diaf002
https://doi.org/10.1093/insilicoplants/diaf002
Advance access publication 24 February 2025
Original Research

in silico Plants
OXFORD

Original Research

Integrating phenotyping and modelling approaches StomaGAN: improving image-based analysis of stomata through generative adversarial networks

Jonathon A. Gibbs*, and Alexandra J. Gibbs

Agriculture and Environmental Sciences, School of Biosciences, University of Nottingham Sutton Bonington Campus, Loughborough LE12 5RD, United Kingdom

*Corresponding author. Agriculture and Environmental Sciences, School of Biosciences, University of Nottingham Sutton Bonington Campus, Loughborough LE12 5RD, United Kingdom. E-mail: jonathon.gibbs1@nottingham.ac.uk

Handling Editor: Przemyslaw Prusinkiewicz

Abstract Stomata regulate gas exchange between plants and the atmosphere, but analysing their morphology is challenging due to anatomical variability and artefacts during image acquisition. Deep learning (DL) can address these challenges but often requires large and diverse datasets, which are costly and error prone to produce. Generative adversarial networks (GANs) offer a solution by generating artificial data via unsupervised learning. However, GANs often suffer from problems including mode collapse, vanishing gradients, and network failure, particularly with small datasets. Here, we present StomaGAN, a deep convolutional GAN (DCGAN) with tailored modifications to address common GAN issues. We collected 559 stomatal impressions of field, or faba bean (*Vicia* faba) consisting of ~3000 stoma, 80% of which were used to train StomaGAN. Evaluation metrics, including generator and discriminator loss progression and a mean Fréchet Inception Distance (FID) score of 61.4 across eight experimental runs confirm successful training. To validate StomaGAN, we generated artificial images to train a deep convolutional neural network (DCNN) based on the DeepLabV3 framework for stomata detection from real, unseen images. The DCNN achieved a mean Interception over Union (IoU) of 0.95 on artificial training images and 0.91 on real, unseen, images across varying magnifications. Our results demonstrate that StomaGAN effectively generates high-quality synthetic datasets, enabling reliable stomatal detection and enhancing phenotypic analysis. This approach reduces the need for extensive manual data collection and simplifies complex morphological assessments.

KEYWORDS: artificial data; deep convolutional neural network; deep learning; generative adversarial network; plant phenotyping; stomata.

1. INTRODUCTION

Crop yield largely depends on the cumulative rate of photosynthesis as well as the availability of water, in which stomata play a fundamental role (Long et al. 2006; Franks et al. 2015; Furbank et al. 2015; Condon 2020). Stomata (singular 'stoma') refers to the complex consisting of a central pore surrounded by specialized cells, called guard cells, located on above- ground plant organs. These structures regulate pore aperture in response to internal and external signals, driven by changes in the turgor pressure of the guard cells, facilitating gas exchange between the plant and the atmosphere (Lawson and Blatt 2014). A comprehensive understanding of stomatal form and function can help enhance photosynthetic activity and water use efficiency (Franks and Farquhar 2007), ultimately increasing crop yield and stability across increasingly extreme environments.

Analysing stomata presents significant challenges, partly due to their diverse appearances across species (Peterson et al. 2010). Typically, stomatal anatomy is studied using microscope-based images, either captured directly from the plant surface or obtained from surface impressions made using dental resin, nail varnish or other means (Matthaeus et al. 2020; Pathoumthong et al. 2023). Following image collection, manually analysing stomatal traits such as counts, or morphology is time-consuming and error prone.

Deep learning (DL) and deep neural networks (DNNs), offer a fast and efficient solution to automating plant phenotyping tasks, including for the analysis of stomata (Thompson et al. 2017; Balacey et al. 2023; Gibbs and Burgess 2024). Using multiple artificial neural layers, DNNs can recognize, classify and describe data, making them particularly effective for image analysis-related tasks (Rawat and Wang 2017). However, the accuracy and precision at which stomata can be identified and characterized depends upon the provision of an initial training data set where the stomata, or other relevant phenotypic features, have been accurately annotated. Creating this training dataset is often time-consuming, tedious and requires some biological expertise to ensure accurate labelling of sufficient images. Combined with a lack of shared data resources, this image collection and annotation phase represents a critical bottleneck in the throughput of phenotyping tasks.

In well-established fields like object detection or handwriting recognition, existing datasets provide access to hundreds of thousands of annotated images (e.g. Krizhevsky 2009; IMAGENET 2012). Similarly, increased research into stomata has led to a growing number of publicly available datasets. However, these datasets predominantly include annotations in the form of bounding boxes (Gibbs and Burgess 2024), limiting the extraction of detailed morphometric data and the ability to perform more complex analyses (Gibbs et al. 2021; Wang et al. 2024). Improving access to high-quality datasets could significantly alleviate this bottleneck whilst supporting more in-depth analysis. One promising approach to expanding data availability is the application of Generative Adversarial Network (GAN).

GANs were first introduced by Goodfellow et al. (2014) and are a subclass of generative models. Their primary use is to generate artificial representations of real data via unsupervised learning. By identifying and learning patterns in input data, GANs can produce realistic and plausible outputs (Creswell et al. 2018; Goodfellow et al. 2020). The most successful use of GANs has been in image processing and computer vision, with applications including face generation, portrait creation, pose generation, imager super-resolution and medical applications (Creswell et al. 2018). Beyond these domains, GANs have also been applied to tasks involving natural language processing, music composition, speech synthesis and time series analysis (Aggarwal et al. 2021; Gui et al. 2023). GANs have also been applied to plant phenotyping tasks, such as the artificial generation of *Arabidopsis thaliana* rosettes to facilitate segmentation and counting tasks (Giuffrida et al. 2017).

GANs consist of two interconnected sub-models, a generator *G* and a discriminator *D.* The generator is tasked with producing new data, while the discriminator, typically a binary classifier, attempts to distinguish between real data (from the original input dataset) and fake data (generated by *G*; Figure 1) (Goodfellow et al. 2014). *Both G* and *D* are trained simultaneously in a minimax, or zero-sum game; referred to as adversarial learning. Here, *G* aims to maximize the likelihood of *D* misclassifying its generated data as real. In essence, *G* aims to produce data that closely resembles the training set to deceive the *D*, thereby driving *G* to generate increasingly realistic samples. Simultaneously, *D* learns to improve its ability to correctly classify data as real or fake, creating a dynamic balance between the two models. Many variants of GANs have been proposed, and whilst a full review of all GAN variants is out of the scope of this paper, the most recurring methods include CycleGAN, InfoGAN, Conditional GANs (cGAN), Deep Convolutional GAN (DCGAN), Wasserstein GAN (WGAN), Identity GAN (Fathallah et al. 2023) and Least Squares GAN (for a review see Gui et al. 2020).

Although GANs have relatively simple network architectures, they are notoriously difficult to train and evaluate. Even minor changes to hyperparameters or optimization randomness can lead to poor or incomplete results. For instance, adjustments to hyperparameters may cause mode collapse, where the *G* submodel produces limited data variations, or a diminished gradient, where the *D* becomes overly effective at distinguishing real from fake data, preventing the generator from learning. Moreover, there is no robust or consistent method for evaluating GANs, making it challenging to objectively determine the optimal network structure (Lucic et al. 2017; Borji 2022).

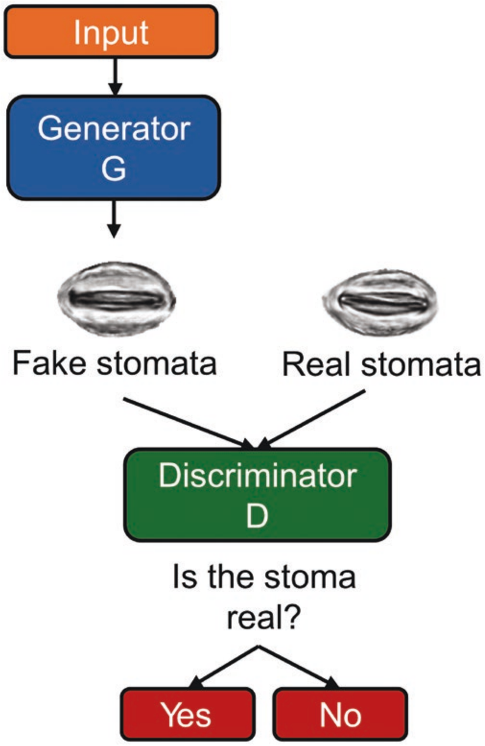

Figure 1. Basic GAN architecture applied to stomata. The generator (G) creates a synthetic stoma from a random seed, while the discriminator (D) evaluates the stoma to determine whether it can classify it as real or fake based on its training. The feedback from this evaluation is then used to iteratively refine both the generator and discriminator, improving their performance over time.

Here, we present a modified DCGAN to help alleviate common issues associated with GANs and novel evaluation methodology applied to a relatively small dataset of leaf surface impressions of field, or faba, bean (*Vicia faba*).

2 MATERIALS AND METHODS

Our approach consists of several key stages, as illustrated in Figure 2:(i) Data acquisition—the initial data stage in which data is collected and annotated manually. Notably, this is the only manual and labour-intensive component in our proposed approach. (ii) Preprocessing—to overcome key issues with data collected under various conditions. (iii) StomaGAN—The training of the proposed GAN, which incorporates modifications relative to the original DCGAN. (iv) Post processing—Application of a series of postprocessing steps to improve the quality of the output of StomaGAN. (v) Fake image generation—a series of tools to generate artificial images, including additional augmentations to increase the size and variety of the artificial dataset. This can further be used to validate the GAN method.

2.1 Data Acquisition

We acquired 559 images of nail varnish-based surface impressions taken of field bean using a Leica DM 5000 B microscope (Wetzlar, Germany) at a magnification of 10 × 40 (Figure 2). The total dataset consisted of around ~3000 individual stomata. Stomata were annotated using pixel-wise methods using the Pixel Annotation Tool (Bréhéret 2017). Whilst annotations performed as bounding boxes could be equally used within StomaGAN, semantic segmentation permits morphology and boundaries to be preserved (Gibbs and Burgess 2024). Furthermore, here we removed the background, prior to training the GAN so that the generated stoma presents only this complex with preserved boundaries.

2.2 PreProcessing

Microscope-based images of stomata often face challenges such as inconsistent lighting and varying environmental conditions due to the wide range of microscope configurations and data acquisition methods. StomaGAN aims to address these inconsistencies by providing a generalized tool for stomatal analysis, regardless of the data acquisition method. To achieve this, an automated preprocessing step is implemented, which is free from constraints and universally applicable to all annotated images (Figure 2). Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to the annotated images to highlight features and standardize the dataset by eliminating any colour biases. Individual stomata were identified and extracted using blob detection on the image mask, enabling the detection of each stoma and extraction of its contours. A bounding box was placed around the contour of the stoma, obtained using the minimum and maximum coordinates for a best-fit box. Each stoma, and associated mask, were cropped from the original image and saved as separate files. Finally, stoma alignment was performed, which is the process of rotating the stomata to horizontally align it with the y-axis. The angle of rotation was determined from the major and minor axis of the image mask.

The proposed GAN requires that training images be square (width == height) so additional padding was applied to resize images accordingly and prevent distortions (Fig. 2). During training, images are resized to the network default, as specified in the configuration file, and therefore the dimension of images does not have to be consistent across images. In most cases, variation in image size can improve training by reflecting variations in stoma size and quality, for example., scaling up may retain defects. To introduce further variability, random padding was added to further adjust the size of stomata. Typically, GANs are trained with tens of thousands of images, however, this study used a significantly smaller dataset. To address this limitation, a series of augmentations were applied, increasing the dataset fivefold through random transformations, flips and contrast enhancements. Degenerative augmentations, such as blur and random noise, were intentionally excluded to maximize the quality of generated synthetic images. Instead, these distortions can be introduced later when training on fake data, to improve the model's ability to detect unseen real stomata, as discussed in subsequent sections.

2.3 Modified DCGAN

The original DCGAN, proposed by Radford et al. (2015), modifies the traditional GAN architecture by replacing the perceptron layers with convolutional neural networks (CNN), while excluding pooling and sampling layers. Here, we incorporate additional modifications to help alleviate issues such as overfitting, overconfidence, mode collapse and vanishing gradients; all of which are more susceptible when training on smaller datasets. We discuss these below:

Replacement of ReLU: We replaced the rectified linear unit (ReLU) activation functions with parametric ReLU (PReLU), to mitigate the vanishing gradient problem. Vanishing gradients occur when gradients become too small, causing learning to slow down or cease altogether, while exploding gradients involve excessively large gradients (Liu et al. 2022). Both issues are known to contribute to the instability of GANs. PreLU not only addresses these issues but also offers additional advantages in terms of computational efficiency. Unlike ReLU and Leaky ReLU, PreLU offers a learnable slope parameter, which enhances model accuracy and convergence (He et al. 2015).

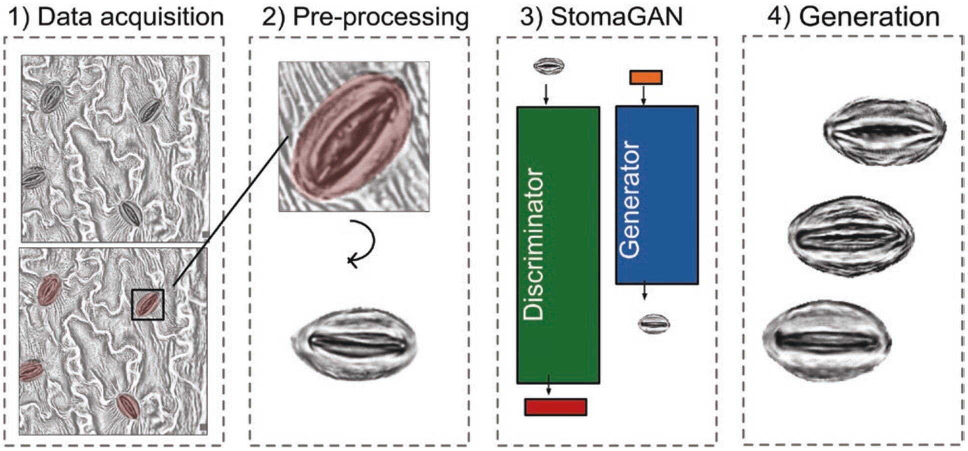

Figure 2. Overview of the pipeline from image acquisition of leaf impressions of field bean (*Vicia faba*) to the generation of synthetic stomata via StomaGAN. Microscope-based images of leaf impressions were taken at 10 × 40 magnification and annotated using pixel-wise segmentation. Stoma were extracted and rotated, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied, and resulting images were padded to create square images (width == height). StomaGAN used these preprocessed stoma as an input to generate synthetic stoma.

Noisy labels: We replaced the instance labels, traditionally 1 for a true (a real image) and 0 for false, with two-sided noisy labels. For real images, labels were randomly applyied in the range of 0.9 to 1.0, and for fake images, labels ranged from 0.0 to 0.1., These labels were dynamically adjusted per epoch. Applying dynamic noisy labels helps to stabilize training and prevent overconfidence. Overconfidence occurs when the discriminator focuses on minimal features to classify an image. Consequently, the generator exploits this behaviour by producing only the feature the discriminator uses for classification, undermining the training process (Wenzel 2023).

Dropout: Within GANs, the discriminator is known to be more dominant than the generator and tends to overfit the training data. Consequently, the discriminator tends to perform well for seen data but fails to adapt to new data. To alleviate this problem, dropout layers, a regularisation technique, were added to the discriminator with a probability of 0.5. These layers function by intentionally omitting random data points from the network during training, helping to reduce overfitting and improve generalization.

Spectral normalization is the process of normalizing the weights in the discriminator. This aids to stabilize training by mitigating the exploding and vanishing gradient problem as well as alleviating mode collapse (Miyato et al. 2018). By restricting the weight changes in each iteration, spectral normalization ensures that the discriminator is not over dependent on a small set of features in distinguishing images. We applied spectral normalization to the final block in our discriminator network (Figure 3).

Simulated Annealing with top_k: Research suggests that updating the generator and discriminator with more realistic weights improves the realism of the samples generated (Wu et al. 2019). Based on this theory, Sinha et al. (2020) proposed a simple approach leveraging the *top k* gradients. In this approach, during each update step where k decreases by a constant factor over time, lower weights are ignored. Whilst the proposed method works, it does not take into consideration the quality of the weights by instead selecting a random distribution.

We propose a simple change to introduce *adaptive top k* based on the quality of the weights. Early in training, the scoring function's ability to correctly classify weights as good or bad is unreliable due to a lack of knowledge. Discarding these weights at this stage would be equivalent to discarding random samples. To address this, we first applied an initial set of warmup epochs during which the temperature remained constant at the starting value. Following this, we applied an annealing process to gradually decrease the temperature over time, adjusting the base batch size and allowing lower-quality weights to be included in the initial stages of training. We adjusted the batch size based on the mean of the results of the weights (Eq. 1 and 2).

$$b_i = (1 - T_i) \cdot b_0 \tag{1}$$

$$k = \ b_i + \overline{x_j^i} \cdot (b_0 - b_i) \tag{2}$$

where T_i is the temperature at time at time i, b_i is the base batch size at time i, and $\overline{x_j^i}$ is mean of the output generated by the discriminator.

Generality: We have aimed to make the source code as general and applicable as possible through various approaches. The generator and discriminator are designed to be adaptive, automatically resizing the model based on the input, eliminating the need for manual adjustments or rewrites. Robust evaluation is facilitated by integrating Comet ML (https://www.comet.com/drjonog/stomagan/), which supports adjustable parameters specified in a configuration file that can be edited without requiring technical expertise. Additionally, the StomaGAN repository on GitHub includes a suite of helper functions for image preprocessing tasks.

2.4 Experimental setup

An overview of the StomaGAN architecture is presented in Figure 3. Both the discriminator and generator were initialized with random weights and a random seed. Since training is highly sensitive to minor changes in weights, we trained the model eight times to ensure fairness. Unless otherwise specified, the mean value of these eight experiments is presented when discussing results. The models used the Adam optimizer with a learning rate of 1×10^{-4} and a Binary Cross Entropy loss function. A batch size of 16 was used and maintained throughout all experiments, as a smaller or larger number can significantly impact the results (Brock et al. 2018). Each experiment ran for 250 epochs. Terminating the run after 20 un-improving FID evaluations could be more computationally efficient, however, for evaluation purposes and fairness, we completed all 250 epochs. All hyperparameters, except the random seed, remained consistent across all experiments. The experiments were performed on a 12GB Titan V graphics card, an Intel Core i9-9980XE CPU running at 3.00GHz, with a total of 112GB of RAM. During evaluation, we recorded both the total run time and the run time per epoch, excluding evaluation time. This was necessary because, for the purpose of this paper, we included additional evaluation metrics, such as the estimation of Fréchet Inception Distance (FID) at each step (see below), which significantly increased computational time.

StomaGAN proposes two architectures of similar size (Fig. 3). The generator contains 32 layers, primarily composed of blocks of 2D transposed convolutions, batch normalization, and the PreLU activation function, with a final layer applying the hyperbolic tangent (tanh) function. The discriminator comprises 33 layers made up of blocks of 2D convolutions, batch normalization, the PreLU activation function, and a dropout rate of 0.5. Its final layer incorporates spectral normalization and a sigmoid activation function. Examples real and generated (fake) stoma are shown in Figure 4.

2.5 Potential application proof of concept: the artificial dataset

This section aids in illustrating the relevance and potential application of StomaGAN. Consequently, we do not provide an in-depth analysis of the results, but instead propose this as a proof of concept.

An overview of the application of an artificial dataset generated using StomaGAN is presented in Figure 5. (i) We used the trained StomaGAN to produce a series of artificial stomata, which, due to the training set, are individual images of 128 × 128

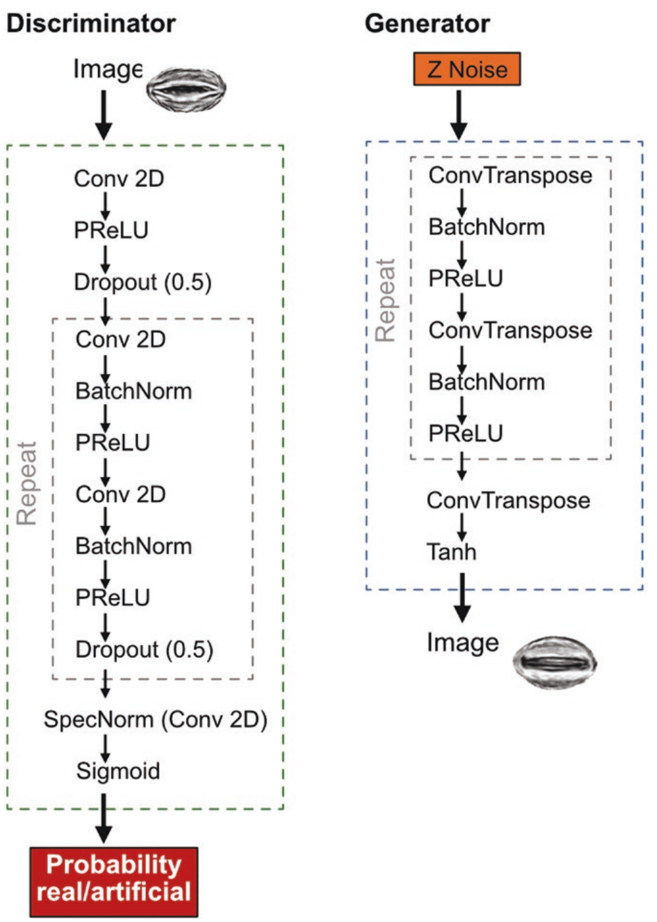

Figure 3. StomaGAN network structure. The generator (G) contains 32 layers comprised primarily of blocks featuring 2D transposed convolutions, batch normalization and the Parametric Rectified Linear Unit (PreLU) activation function. The final layer employs a hyperbolic tangent function, tanh. The discriminator comprises 33 layers made up of blocks of 2D convolutions, batch normalization, PReLU activation functions and a dropout rate of 0.5. The final layer integrates spectral normalization along with a sigmoid activation function.

pixels. (ii) The original microscope-based images of leaf surface impressions and their corresponding annotations were passed to a data manipulator tool (provided on GitHub). This extracts background segments of the images, comprising epidermal cells, vein structures, etc. but omits stomata. Each background segment is cropped to 128 × 128 pixels, and two simple augmentations are applied: namely, resizing- in which images are kept at the original size or resized to 64 × 64 or 32 × 32 pixels; and a random horizontal or vertical flip. (iii) The background to the artificial images was generated through random tiling of background segments (step 2) to create a base background of 512 × 512 pixels. (iv) Artificially generated stomata (step 1) were subsequently assigned random coordinates on top of this background, ensuring there was no overlap by using bounding box collision detection. During insertion, random augmentations were applied to each stoma including scaling, ranging from 0.2 to 1.2 of its original size; rotation; Gaussian blur and gamma adjustment. Each of these augmentations adjusts the stomata in a way that is commonly seen in higher magnification images and significantly increases the variability within the dataset. Corresponding masks were then generated using the known stomata locations and contours. (v) We trained DeepLabV3 with MobileNet, an advanced neural network for semantic segmentation (Sandler et al. 2018) from scratch. The training was conducted on artificial images using an NVIDIA Jetson Nano Orin (Santa Clara, USA). The dataset, comprising 10 000 images, was divided into training and testing sets in a 4:1 ratio, and the model was trained for 150 epochs. To further validate the usefulness of the GAN-generated dataset, we trained DeepLabV3 on solely real images as a benchmark, and again on a combination of real and artificial images.

3. RESULTS

Here, we present StomaGAN, a modified and enhanced GAN designed to generate artificial images of stomata, with applications across various plant phenotyping tasks. All data and tools associated with this project are publicly available. These include the StomaGAN source code, pretrained models, and a suite of image analysis and manipulation tools; available at https://github.com/DrJonoG/StomaGAN. Additionally, we provide three new datasets available at: https://www.stomatahub.com:

1) The original, semantically annotated images of field bean leaf impressions.
2) The original stomata extracted from their backgrounds, along with individual artificial stomata generated by StomaGAN.
3) A collection of artificial images created by combining original image backgrounds with the artificial stomata (Fig. 5).

In StomaGAN evaluation was conducted using multiple metrics (a full evaluation of can be found at: https://www.comet.com/drjonog/stomagan/). Whilst we generated loss functions for both the generator and discriminator independently, these values alone provide limited insight into assessing GAN quality. However, GANs are known to exhibit specific trends during successful training, which we observed here (Fig. 6A and B). Typically, the discriminator loss function starts high, around 0.8, indicating that it struggles to correctly distinguish between real and artificial data. Over time, this value gradually converges towards 0.5, reflecting a 50% probability of guessing whether data is real or artificial, as expected by chance (Fig. 6A). In comparison, the generator loss function increases as training progresses, starting around 0.75 and rising up to approximately 1.75 (Fig. 6B). Together, these trends suggest that the discriminator successfully learns key features of real images, while the generator improves its ability to produce realistic artificial images.

We employed the Fréchet Inception Distance (FID), as it has been shown to align closely with human judgement (Heusel et al. 2017). A lower FID score indicates greater similarity between real and generated (fake) data. Whilst a perfect FID score of 0 is theoretically achievable, it is often unrealistic without overfitting. However, a decreasing FID score over time indicates successful learning, which was observed for StomaGAN (Fig. 6C).

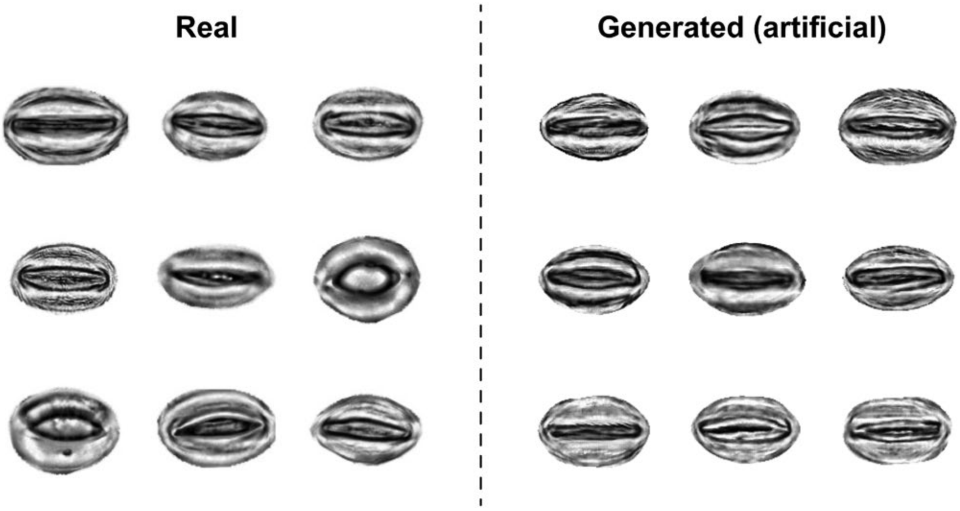

Figure 4. Example stoma where Real (left side) present original images of stoma following extraction, rotation and Contrast Limited Adaptive Histogram Equalization (CLAHE) and Generated (synthetic) stoma produced via StomaGAN.

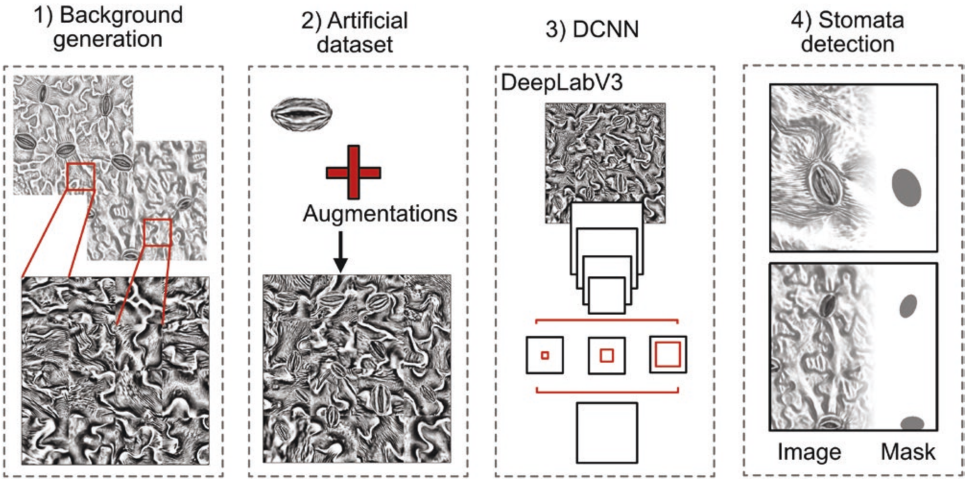

Figure 5. Evaluation pipeline for StomaGAN using an artificially generated dataset. Small sections of leaf impression background (i.e. areas in which stomata are not present) were cropped from the original dataset and tiled to create a base. Variability was increased by applying random augmentations to the StomaGAN-generated stoma before embedding them into the tiled background images. The resulting artificial dataset was split 4:1 for training and validation and used to train a deep convolutional neural network (DCNN), DeepLabV3. Once trained, the DCNN was applied to unseen original microscope-based images for the detection of real stomata.

Although three common metrics were applied, they do not address the issue of variability, realism, or prove a future application of StomaGAN. To demonstrate proof of concept, we utilized a DCNN for stomatal detection trained exclusively on artificial data. While synthetic data generation is not always strictly necessary, it can be particularly valuable in data-scarce environments. In addition, this provides an alternative and complementary approach to improving the performance of a DCNN, such as DeepLab, when real data is limited.

DeepLab achieved a mean Interception over Union (IoU) of 0.95 during training on the artificial dataset. When the trained model was applied to real, unseen images, a mean IoU of 0.91 was achieved. Notably, these real images were not only unseen during GAN training but were also at different resolutions. Specifically, whilst the GAN was trained on images at 40x magnification, the DeepLab model successfully processed images at 10x and 20x magnification, where stomata appear significantly smaller and often exhibit more defects.

To further evaluate the potential value of an artificial dataset generated through StomaGAN, the DeepLab v3 model was trained and validated using other combinations of data: (i) trained on real images and validated on real images, and (ii) trained on a combination of artificial and real images and validated on real images. Both models were evaluated on an independent test set to assess their performance. The model trained solely on real data achieved a detection accuracy of approximately 94.7%, whereas the model incorporating artificial data attained an accuracy of 99.7%, misclassifying only a single instance. This latter case is similar to the results of Giuffrida et al. (2017), we found that the inclusion of both real and artificial data led to an improvement in accuracy.

4. DISCUSSION

This study offers a novel GAN architecture, StomaGAN, and application. StomaGAN offers a proof of concept for using artificially generated images to train neural networks with high accuracy, we are however aware of limitations of this study, which are discussed here.

4.1 Evaluation of GANs

Despite significant advancements in improving the quality of GANs, the evaluation and comparison of methods remain underdeveloped (Borji 2018, 2022). Since GANs rely on the coordinated training of two models, the generator and the discriminator (Figure 1), there is no objective loss function to directly evaluate the generator's performance. Consequently, it is not possible to assess the progress of the training based solely on loss, requiring evaluation to be based on the quality of the generated synthetic images. Whilst various methods to evaluate GANs have been proposed, none have been universally adopted (Borji 2022). Even under ideal conditions, the training can be unstable and highly sensitive to hyperparameters (Wenzel 2023). Further difficulties arise because optimal weights correspond to saddle points rather than to a minimum or maximum loss function (Li et al. 2017). Furthermore, issues such as mode collapse, vanishing and exploding gradient exacerbate the difficulties in training and evaluating GANs (Wenzel 2023).

Focus on qualitative measures, such as visually comparing results, is often used when evaluating GANs (Zhou et al. 2019; Borji 2022). While improved frameworks have been proposed to improve human evaluation metrics (Zhou et al. 2019), this approach remains subjective, inconsistent, and potentially misleading (Le et al. 2010; Salimans et al. 2016). Moreover, humans process data differently from machines, limiting their ability to assess model outputs accurately (Denton et al. 2015; Olsson et al. 2018). Therefore, alternative, more quantitative evaluation measures have been proposed. Inception score (*IS*; Salimans et al. 2016) is an evaluation metric based on the comparison between generated data and an existing image library. Therefore, *IS* is appropriate for generated images of objects known to the model used to calculate the conditional class probabilities but is unsuitable for objects outside of these categories (Barratt and Sharma 2018). For example, the Inception v3 model recognizes 1000 object types as part of the ILSVRC 2012 dataset (IMAGENET 2012), whereas the CIFAR-10 and CIFAR-100 models recognize 10 and 100 object classes, respectively (Krizhevsky 2009). However, current published models lack object categories useful to biological analysis, making *IS* unsuitable for evaluating StomaGAN.

The pattern of change in loss functions of the generator and discriminator provides another means of evaluating GAN performance. Whilst the witnessed pattern within this study indicates successful training (Fig. 6), this is not always the case. For example, the discriminator could learn a specific feature which allows it to distinguish between real and generated data. Alternatively, the generator could be producing the same, or very similar images, which therefore have the same features. This would make it easier for the discriminator to distinguish and, consequently, results in the generation of artificial data with little variability.

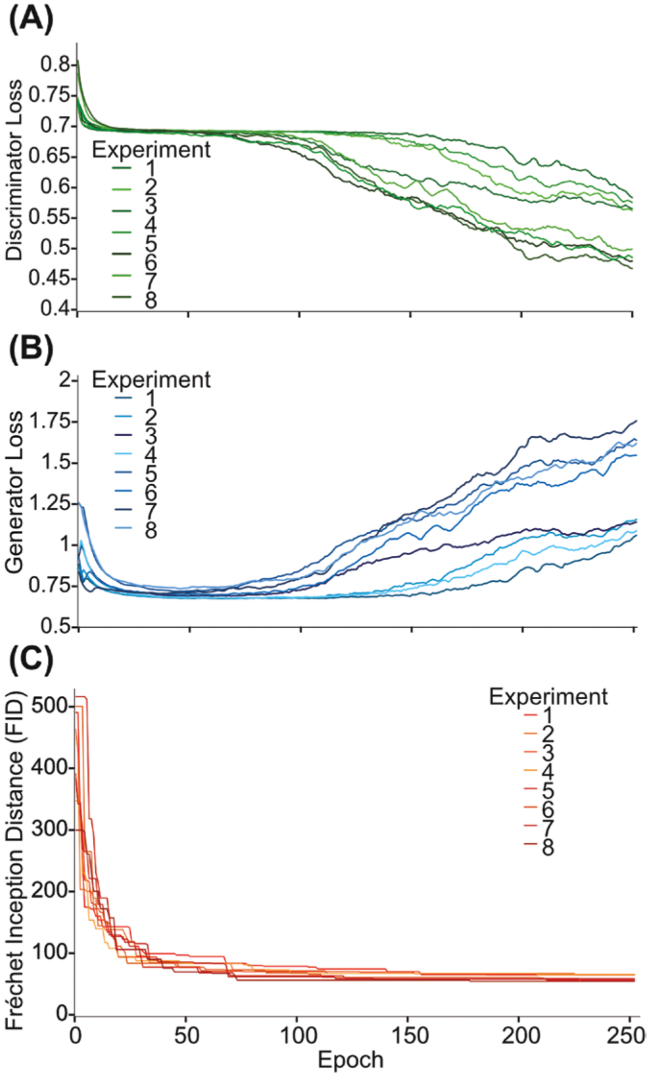

Figure 6. Evaluation of StomaGAN performance during training for 250 epochs indicating the moving average of the (A) Discriminator loss function and (B) Generator loss function. (C) Fréchet Inception Distance (FID).

Here, we introduced an additional metric to assess the accuracy of the GAN via the use of a DCNN trained solely on an artificial dataset. This test highlights the capabilities of StomaGAN; first by producing sufficiently plausible stomata to deceive the DCNN, and second, by demonstrating its applicability to more difficult phenotyping tasks. Microscope-based images taken at lower magnifications contain more stomata but often suffer from a higher degree of artefacts such as blur (Millstead et al. 2020). This makes annotation significantly more difficult, time-consuming and computationally expensive. Furthermore, accurately preserving the boundaries of small stoma using pixel-wise annotation is more difficult than those of larger sizes, dependent on the radius of the annotation tool and the resolution of the image. For this reason, the majority of deep learning approaches applied to stomatal analysis have utilized bounding box annotation, as opposed to the more informative semantic segmentation (Gibbs and Burgess 2024). StomaGAN provides a solution to this problem through the generation of artificial data that can be applied to any magnification. This is achieved via an initial image set captured and annotated at high magnification (40×), and through augmentations that represent many of the key artefacts seen in surface impressions. Furthermore, as the results are semantic, more in-depth analysis can be performed such as the estimation of G_{smax} (Gibbs et al. 2021), compared to the limiting bounding box approaches.

Although pixel-wise annotation is more time-consuming than bounding boxes, this process can be significantly accelerated through the application of a GAN. As demonstrated in this study, even a small dataset can enable GANs to produce sufficient data representations to permit detection via a DCNN. Here, we started with 559 images of surface impressions and were able to generate 10 000 artificial images, which can then be used for the preliminary classification of further unseen images. However, we expect that an even smaller dataset could be used initially. It is expected that this would work via an iterative process; where outputs of the DCNN provide additional training data that can be fed into the GAN. This would be particularly valuable for datasets whereby artifacts or errors during data collection result in distorted target objects (i.e. stomata); such that manual annotation is not feasible. Further augmentations, such as blur, could be applied to GAN-generated artificial datasets to aid the detection of these objects (e.g. Gibbs et al. 2019). Common errors in the DCNN output often relate to rough or incomplete boundaries, most of which can be repaired via additional blob detection and contour smoothing.

4.2 Future directions and current limitations

While DL offers a promising solution for enhancing the throughput of biological image analysis, there are several limitations that hinder its broader applicability. Variability in image capture and annotation pipelines, along with restricted access to datasets, pose significant bottlenecks for phenotypic analysis. Additionally, the lack of alignment between image analysis outputs and conductance measurements limits the ability to correlate findings with underlying physiological functions. For a more detailed discussion of these limitations, see Gibbs and Burgess (2024).

The use of GANs can help to increase dataset variability and size; however, there are areas for improvement. Hyperparameter tuning plays a critical role in GAN performance, and comprehensive optimization could enhance output quality. Whilst this was beyond the scope of our project, it presents an important next step. Existing evaluation metrics for StomaGAN have been discussed, but incorporating additional real-time metrics during training could provide deeper insights into GAN performance. Currently, there are limited options for such metrics in GANs, and robust methods for comparing models are needed to determine if hyperparameter changes lead to meaningful improvements.

In certain phenotyping tasks, it is essential to generate artificial data that represents multiple features or labels. For instance, StomaGAN could be trained separately on background leaf segments and stomata. Similarly, Park et al. (2019) introduced Spatially-Adaptive Normalization (SPADE), a normalization layer designed to retain semantic information within a network, thereby preventing the loss of spatial details. This is particularly relevant for phenotyping tasks that require segmentation maps of specific features. Biological images, such as those obtained through microscopy and other imaging techniques, often exhibit complex structures where preserving spatial and semantic information is crucial. Techniques like SPADE can therefore be adapted to maintain and control spatial features during data generation.

Alternative generative models present valuable opportunities to enhance dataset diversity and size. One example is Latent Diffusion Models (LDMs), which incrementally add noise to data and then apply a reverse denoising process to generate synthetic samples (Rombach et al. 2021). Although LDMs are generally more stable to train compared to GANs, they are computationally intensive, making them particularly well-suited for tasks involving high-dimensional data. However, for certain tasks, pretrained LDMs exist (e.g. Anagnostopoulou et al. 2023). Similarly, Variational Autoencoders (VAEs) encode images into a latent space before decoding them back into image form (Pu et al. 2016). Conditional VAEs, in particular, offer a powerful approach for generating synthetic representations of specific features, such as plant traits. The optimal method depends on the specific requirements of the application, such as the need for realism, control over features, computational efficiency, or the scale of the dataset.

Future work should focus on developing fair evaluation techniques for less common biological images, such as stomata, which are often overlooked compared to objects featuring in datasets such as ILSVRC 2012, CIFAR-10 or CIFAR-100 datasets (Krizhevsky 2009; IMAGE-NET 2012). Emphasis should also shift towards shared resources and datasets, including a unified GAN capable of generating stomata across various species to represent the full spectrum of structural diversity.

5. CONCLUSION

Here, we introduce StomaGAN, a novel method for generating artificial images of stomata to support automated analysis of stomatal traits or other plant phenotyping tasks using deep learning. Alongside StomataGAN, we provide publicly available, high-quality resources, including the StomaGAN source code, pretrained models, image analysis and manipulation tools, and three diverse datasets. These tools and datasets represent a significant advancement in stomatal analysis, offering enhanced throughput and expanded capabilities with applications in broader image analysis tasks. Notably, StomaGAN is optimized to work with relatively small datasets, enabling the generation of larger, more diverse training data for DCNNs. Additionally, StomaGAN simplifies complex phenotyping tasks, such as translating higher-resolution images to lower magnifications, making it a versatile tool for plant science and image-based research.

ACKNOWLEDGEMENTS

This work was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) International Partnership on AI for the Biosciences [grant number BB/Y513866/1]. In addition, this work was supported by a Rank Prize Nutrition New Lecturer Award, the Gatsby Grant for Exceptional Researchers and a Royal Society Research Grant awarded to A.J.G. We wish to thank all co-investigators and project partners on the H2YOLO project- we are thankful to be working with you all. We hope that through this paper we are able to encourage collaborative working and the sharing of datasets. If you would like to be involved or are willing to share your datasets on www.stomatahub.com, please get in touch.

AUTHOR CONTRIBUTIONS

J.G. gathered the data, annotated it, and processed the data. Additionally, J.G. designed, developed, tested and evaluated StomaGAN, the additional tools developed here, and modified the existing DeepLab code for further evaluation and testing. A.G. and J.G. wrote the manuscript. A.G. created figures based on images supplied by J.G.

CONFLICT OF INTEREST

The authors developed www.stomatahub.com, a nonprofit web resource to promote the sharing of datasets and best practice, as an accompaniment to this manuscript and as part of work on the acknowledged project 'H2YOLO'.

Data Availability

The dataset(s) supporting the conclusions of this article are available on the Stomata Hub, https://www.stomatahub.com/datasets.php and in the GitHub repository https://github.com/DrJonoG/StomaGAN. This latter resource include the source code for StomaGAN and corresponding trained models as well as image manipulation and image processing tools. The original DeepLab network can be found at https://github.com/VainF/DeepLabV3Plus-Pytorch. Additionally, a more comprehensive evaluation of StomaGAN can be found here https://www.comet.com/drjonog/stomagan/.

REFERENCES

Aggarwal A, Mittal M, Battineni G. Generative adversarial network: an overview of theory and applications. *Int J Inform Manage Data Insights* 2021;1:100004. https://doi.org/10.1016/j.jjimei.2020.100004

Anagnostopoulou D, Retsinas G, Efthymiou N, Filntisis P, Maragos P. A Realistic synthetic mushroom scenes dataset. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, 6282–9. https://doi.org/10.1109/cvprw59228.2023.00668

Balacey S, Capone D, Sullivan W, Tyerman S. 2023. Transpiration Responses to Potential Volatile Signals and Hydraulic Failure in Single Leaves of Vitis Vinifera (CV. Shiraz) and Arabidopsis Thaliana (Col 0) Utilising Sensitive Liquid Flow and Simultaneous Gas Exchange. *bioRxiv*: 2023.01.24.525440.

Barratt S, Sharma R. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Borji A. Pros and cons of GAN evaluation measures. *Comput Vis Image Underst* 2018;179:41–65. https://doi.org/10.1016/j.cviu.2018.10.009

Borji A. Pros and cons of GAN evaluation measures: new developments. *Comput Vis Image Underst* 2022;215:103329. https://doi.org/10.1016/j.cviu.2021.103329

Bréhéret A. 2017. Pixel Annotation Tool. 11 Nov. 2020.

Brock A, Donahue J, Simonyan K. 2018. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:*1809.11096*.

Condon A. Drying times: plant traits to improve crop water use efficiency and yield. *J Exp Bot* 2020;71:2239–52. https://doi.org/10.1093/jxb/eraa002

Creswell A, White T, Dumoulin V et al. Generative adversarial networks: an overview. *IEEE Signal Process Mag* 2018;35:53–65. https://doi.org/10.1109/MSP.2017.2765202

Denton E, Chintala S, Szlam A, Fergus R. Deep generative image models using a laplacian pyramid of adversarial networks In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc, 2015, 28.

Fathallah M, Sakr M, Eletriby S. Stabilizing and improving training of generative adversarial networks through identity blocks and modified loss function. *IEEE Access* 2023;11:43276–85. https://doi.org/10.1109/access.2023.3272032

Franks P, Doheny-Adams T W, Britton-Harper Z et al. Increasing water-use efficiency directly through genetic manipulation of stomatal density. *New Phytol* 2015;207:188–95. https://doi.org/10.1111/nph.13347

Franks P, Farquhar G. The mechanical diversity of stomata and its significance in gas-exchange control. *Plant Physiol* 2007;143:78–87. https://doi.org/10.1104/pp.106.089367

Furbank R, Quick W, Sirault X. Improving photosynthesis and yield potential in cereal crops by targeted genetic manipulation: Prospects, progress and challenges. *Field Crops Res* 2015;182:19–29. https://doi.org/10.1016/j.fcr.2015.04.009

Gibbs J, Burgess A, Pound M et al. Recovering wind-induced plant motion in dense field environments via deep learning and multiple object tracking. *Plant Physiol* 2019;181:28–42. https://doi.org/10.1104/pp.19.00141

Gibbs J, Mcausland L, Robles-Zazueta C et al. A deep learning method for fully automatic stomatal morphometry and maximal conductance estimation. *Front Plant Sci* 2021;12:2703. https://doi.org/10.3389/fpls.2021.780180

Gibbs JA, Burgess AJ. Application of deep learning for the analysis of stomata: a review of current methods and future directions. *J Exp Bot* 2024;75:6704–18. https://doi.org/10.1093/jxb/erae207

Giuffrida MV, Scharr H, Tsaftaris SA. 2017. ARIGAN: synthetic arabidopsis plants using generative adversarial network. *IEEE International Conference on Computer Vision Workshops (ICCVW)*. Venice, Italy, 2017, 2064–2071. https://doi.org/10.1109/ICCVW.2017.242

Goodfellow I, Pouget-Abadie J, Mirza M et al. Generative adversarial networks. *Commun ACM* 2020;63:139–44. https://doi.org/10.1145/3422622

Goodfellow IJ, Pouget-Abadie J, Mirza M et al. Generative adversarial networks. *Sci Rob* 2014;63:139–44. https://doi.org/10.1145/3422622

Gui J, Sun Z, Wen Y et al. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans Knowl Data Eng* 2020;35:3313–32. https://doi.org/10.1109/tkde.2021.3130191

Gui J, Sun Z, Wen Y et al. A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Trans Knowl Data Eng* 2023;35:3313–32. https://doi.org/10.1109/tkde.2021.3130191

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification In: Proceedings of the IEEE international conference on computer vision, 2015, 1026–34.

Heusel M, Ramsauer H, Unterthiner T et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Adv Neural Inform Process Syst* 2017;1:6627–38 https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

IMAGE-NET. 2012. ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). https://image-net.org/challenges/LSVRC/2012/ (23 May 2024, date last accessed).

Krizhevsky A. 2009. *Learning Multiple Layers of Features from Tiny Images*. Canada: Computer Science University of Toronto. https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

Lawson T, Blatt M. Stomatal size, speed, and responsiveness impact on photosynthesis and water use efficiency. *Plant Physiol* 2014;164:1556–70. https://doi.org/10.1104/pp.114.237107

Le J, Biewald L, Edmonds A, Hester V. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution In: Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010), 2010, 17–20.

Li Y, Schwing A, Wang K-C et al. Dualing GANs. *arxiv* 2017;1:1–11. preprint arXiv.1706.06216.

Liu B, Lv J, Fan X et al. Application of an Improved DCGAN for Image Generation. *Mobile Inform Syst* 20222022;2022:1–14. https://doi.org/10.1155/2022/9005552

Long S, Zhu X-G, Naidu S et al. Can improvement in photosynthesis increase crop yields? *Plant Cell Environ* 2006;29:315–30. https://doi.org/10.1111/j.1365-3040.2005.01493.x

Lucic M, Kurach K, Michalski M et al. Are GANs created equal? A large-scale study. *Adv Neural Inform Process Syst* 2018-December 2017:700–9. https://proceedings.neurips.cc/paper_files/paper/2018/file/e46de7e1bcaaced9a54f1e9d0d2f800d-Paper.pdf

Matthaeus W, Schmidt J, White J et al. Novel perspectives on stomatal impressions: rapid and non-invasive surface characterization of plant leaves by scanning electron microscopy. *PLoS One* 2020;15:e0238589. https://doi.org/10.1371/journal.pone.0238589

Millstead L, Jayakody H, Patel H et al. Accelerating automated stomata analysis through simplified sample collection and imaging techniques. *Front Plant Sci* 2020;11:1–14. https://doi.org/10.3389/fpls.2020.580389

Miyato T, Kataoka T, Koyama M, Yoshida Y. 2018. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

Olsson C, Bhupatiraju S, Brown T, Odena A, Goodfellow I. 2018. Skill rating for generative models. arXiv preprint.

Park T, Liu M-Y, Wang T-C, Zhu J-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. arXiv. *ACM SIGGRAPH 2019 Real-Time Live!*

Pathoumthong P, Zhang Z, Roy S et al. Rapid non-destructive method to phenotype stomatal traits. *Plant Methods* 2023;19:36. https://doi.org/10.1186/s13007-023-01016-y

Peterson K, Rychel A, Torii K. Out of the mouths of plants: the molecular basis of the evolution and diversity of stomatal development. *The Plant Cell* 2010;22:296–306. https://doi.org/10.1105/tpc.109.072777

Pu Y, Gan Z, Henao R, et al. 2016. Variational Autoencoder for Deep Learning of Images, Labels and Captions In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, (eds.), *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*. Curran Associates, Inc.

Radford A, Metz L, Chintala S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In: 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings.

Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput* 2017;29:2352–449. https://doi.org/10.1162/NECO_a_00990

Rombach R, Blattmann A, Lorenz D et al. High-resolution image synthesis with latent diffusion models 2021;2022:10674–85. https://doi.org/10.1109/cvpr52688.2022.01042

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved Techniques for Training GANs In: 30th Conference on Neural Information Processing Systems (NIPS 2016). 2016, 1–9.

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 4510–20. https://doi.org/10.1109/cvpr.2018.00474

Sinha S, Zhao Z, Goyal A et al. Top-k training of GANs: improving GAN performance by throwing away bad samples. In: *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*. 2020. https://proceedings.neurips.cc/paper/2020/file/a851bd0d418b13310dd1e5e3ac7318ab-Paper.pdf

Thompson A, Senin N, Giusca C et al. Topography of selectively laser melted surfaces: a comparison of different measurement methods. *CIRP Ann* 2017;66:543–6. https://doi.org/10.1016/j.cirp.2017.04.075

Wang J, Renninger HJ, Ma Q et al. Measuring stomatal and guard cell metrics for plant physiology and growth using StoManager1. *Plant Physiol* 2024;195:378–94. https://doi.org/10.1093/plphys/kiae049

Wenzel M. Generative adversarial networks and other generative models In: Colliot O (ed.), *Machine Learning for Brain Disorders*. New York, NY: Humana Press Inc, 2023, 139–92. https://doi.org/10.1007/978-1-0716-3195-9_5

Wu Y, Donahue J, Balduzzi D, Simonyan K, Lillicrap T, London D. 2019. LOGAN: Latent optimisation for generative adversarial networks.

Zhou S, Gordon M, Krishna R, Narcomey A, Fei-Fei L, Bernstein M. 2019. HYPE: a benchmark for human eye perceptual evaluation of generative models In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).