StomaGAN: Improving image-based analysis of stomata through Generative Adversarial Networks

Jonathon A. Gibbs[1*], Alexandra J. Gibbs[1]

[1]Agriculture and Environmental Sciences, School of Biosciences, University of Nottingham Sutton Bonington Campus, Loughborough, LE12 5RD, UK

ORCiD: JAG- 0000-0002-2772-2201

AJG- 0000-0002-1621-6821

Corresponding author: Jonathon A. Gibbs Jonathon.gibbs1@nottingham.ac.uk

Abstract

Stomata regulate gas exchange between plants and the atmosphere, but analysing their morphology is challenging due to anatomical variability and artifacts 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 ~3,000 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 confirms 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 a 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 (DCNN), Deep learning (DL), Generative Adversarial Network (GAN), 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; Furbank *et al.* 2015; Franks *et al.* 2015; Condon 2020). Stomata (singular 'stoma') refers to the complex consisting of a central pore surrounded by specialised 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 recognise, 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 characterised 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. IMAGE-NET, 2012; Krizhevsky, 2009). 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 maximise 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 of GAN variants is out 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 seeGui et al. 2020).

Although GANs have relatively simple network architectures, they are notoriously difficult to train and evaluate. Even minor changes to hyperparameters or optimisation randomness can lead to poor or incomplete results. For instance, adjustments to hyperparameters may cause mode collapse, where the *G* sub-model 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).

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 Method

Our approach consists of several key stages, as illustrated in Figure 2:1) 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. 2) Pre-processing – to overcome key issues with data collected under various conditions. 3) StomaGAN – The training of the proposed GAN, which incorporates modifications relative to the original DCGAN. 4) Post processing – Application of a series of post-processing steps to improve the quality of the output of StomaGAN. 5) 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 10x40 (Figure 2). The total dataset consisted of around ~3,000 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 generated stoma present only this complex with preserved boundaries.

2.2 Pre-Processing

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 generalised tool for stomatal analysis, regardless of the data acquisition method. To achieve this, an automated pre-processing step is implemented, which is free from constraints and universally applicable to all annotated images (Figure 2). Contrast Limited Adaptive Histogram Equalisation (CLAHE) was applied to the annotated images to highlight features and standardise 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 are square (width == height) so additional padding was applied to resize images accordingly and prevent distortions (Figure 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 maximise 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).

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 stabilise 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 to 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 normalisation is the process of normalising the weights in the discriminator. This aids to stabilise 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 normalisation ensures that the discriminator is not over dependant on a small set of features in distinguishing images. We applied spectral normalisation 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 as 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&2).

$$b_i = (1 - T_i) \cdot b_0 \quad (1)$$

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

where T_i is the temperature at time at time i, b_i is the base batch size at time i, and $\bar{x_j^i}$ 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 pre-processing tasks.

2.4 Experimental Setup

An overview of the StomaGAN architecture is presented in Figure 3. Both the discriminator and generator were initialised 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 optimiser with a learning rate of 1 x 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 (Figure 3). The generator contains 32 layers, primarily composed of blocks of 2D transposed convolutions, batch normalisation, 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 normalisation, the PreLU activation function, and a dropout rate of 0.5. Its final layer incorporates spectral normalisation and a sigmoid activation function. Example real and generated (fake) stoma are shown in Figure 4.

2.5 Potential application proof of concept: the Artificial Dataset

This section aids to illustrate 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. 1) We used the trained StomaGAN to produce a series of artificial stomata, which, due the training set, are individual images of 128x128 pixels. 2) 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 128x128 pixels, and two simple augmentations are applied: namely, resizing- in which images are kept at the original size or resized to 64x64 or 32x32 pixels; and a random horizontal or vertical flip. 3) The background to the artificial images was generated through random tiling of background segments (step 2) to create a base background of 512x512 pixels. 4) 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 adjust 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. 5) 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 task. All data and tools associated with this project are publicly available. These include the StomaGAN source code, pre-trained 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 (Figure 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 successfully training, which we observed here (Figure 6A&B). Typically, discriminator loss function starts high, around 0.8, indicating that it struggles to correctly distinguishing 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 (Figure 6A). In comparison, the generator loss function increases as training progresses, starting around 0.75 and rising up to approximately 1.75 (Figure 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 (Figure 6C).

Although three common metrics were applied, they do not address of the issue of variability, realism, or prove a future application of StomaGAN. To demonstrate proof of concept, we utilised 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 limiting.

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: 1) trained on real images and validated on real images, and 2) 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 advancement in improving the quality of GANs, the evaluation and comparison of methods remains underdeveloped (Borji 2018, 2022). Since GANs relay 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 to 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 recognises 1,000 object types as part of the ILSVRC 2012 dataset (IMAGE-NET 2012), whereas the CIFAR-10 and CIFAR-100 models recognise 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 (Figure 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.

Here, we introduced an additional metric to assess the accuracy of the GAN via 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 artifacts 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 resolution of the image. For this reason, the majority of deep learning approaches applied to stomatal analysis have utilised

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 which can be applied to any magnification. This is achieved via an initial image set captured and annotated at high magnification (40x), and through augmentations that represent many of the key artifacts 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 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 dataset to aid 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 several limitations 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 function. 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 optimisation 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, pre-trained 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, pre-trained 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 optimised 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.

Figure Legends

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.*

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 10x40 magnification and annotated using pixel wise segmentation. Stoma were extracted, rotated, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied, and resulting images were padded to create square images (width == height). StomaGAN used these pre-processed stoma as an input to generate synthetic stoma.*

Figure 3: *StomaGAN network structure. The generator (G) contains 32 layers comprised primarily of blocks featuring 2D transposed convolutions, batch normalisation and the Parametric Rectified Linear Unit (PreLU) activation function. The final layer employs a hyperbolic tangent function, tanh. The discriminator comprises33 layers made up for blocks of 2D convolutions, batch normalisation, PReLU activation functions and a dropout rate of 0.5. The final layer integrates spectral normalisation along with a sigmoid activation function.*

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 through 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.

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).*

Abbreviations

CNN	Convolutional Neural Network
D	Discriminator- sub model of the GAN
DCGAN	Deep Convolutional Generative Adversarial Network
DL	Deep Learning
DNN	Deep Neural Network
FID	Fréchet Inception Distance
G	Generator- sub model of the GAN
GAN	Generative Adversarial Network
IoU	Interception over Union
IS	Inception Score
PReLU	Parametric Rectified Linear Unit
ReLU	Rectified Linear Unit

Declarations

Model and 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/.

Authors' contributions

J.G. gathered the data, annotated, and processed the data. Additionally, J.G. designed, developed, tested and evaluated StomaGAN, the additional tools developed here, and modified 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 interests

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'.

Acknowledgements

This work was supported by the 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.

References

Aggarwal A, Mittal M, Battineni G. **2021**. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights* **1**: 100004.

Anagnostopoulou D, Retsinas G, Efthymiou N, Filntisis P, Maragos P. **2023**. A Realistic Synthetic Mushroom Scenes Dataset In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.6282–6289.

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. **2018**. Pros and Cons of GAN Evaluation Measures. *Computer Vision and Image Understanding* **179**: 41–65.

Borji A. **2022**. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding* **215**: 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. **2020**. Drying times: plant traits to improve crop water use efficiency and yield. *Journal of Experimental Botany* **71**: 2239–2252.

Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath A. **2018**. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* **35**: 53–65.

Denton E, Chintala S, Szlam A, Fergus R. **2015**. 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., 28.

Fathallah M, Sakr M, Eletriby S. **2023**. Stabilizing and Improving Training of Generative Adversarial Networks Through Identity Blocks and Modified Loss Function. *IEEE Access* **11**: 43276–43285.

Franks P, Farquhar G. **2007**. The mechanical diversity of stomata and its significance in gas-exchange control. *Plant Physiology* **143**: 78–87.

Franks P, W. Doheny-Adams T, Britton-Harper Z, Gray J. **2015**. Increasing water-use efficiency directly through genetic manipulation of stomatal density. *New Phytologist* **207**: 188–195.

Furbank R, Quick W, Sirault X. **2015**. Improving photosynthesis and yield potential in cereal crops by targeted genetic manipulation: Prospects, progress and challenges. *Field Crops Research* **182**: 19–29.

Gibbs JA, Burgess AJ. **2024**. Application of deep learning for the analysis of stomata: A review of current methods and future directions. *Journal of Experimental Botany*: erae207.

Gibbs J, Burgess A, Pound M, Pridmore T, Murchie E. **2019**. Recovering wind-induced plant motion in dense field environments via deep learning and multiple object tracking. *Plant Physiology* **181**: 28–42.

Gibbs J, Mcausland L, Robles-Zazueta C, Murchie E, Burgess A. **2021**. A Deep Learning Method for Fully Automatic Stomatal Morphometry and Maximal Conductance Estimation. *Frontiers in Plant Science* **12**: 2703.

Giuffrida MV, Scharr H, Tsaftaris SA. **2017**. ARIGAN: Synthetic Arabidopsis Plants using Generative Adversarial Network.

Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al.* 2014. Generative Adversarial Networks. *Science Robotics* **3**: 2672–2680.

Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* 2020. Generative adversarial networks. *Commun. ACM* **63**: 139–144.

Gui J, Sun Z, Wen Y, Tao D, Ye J. **2020**. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering* **35**: 3313–3332.

Gui J, Sun Z, Wen Y, Tao D, Ye J. **2023**. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering* **35**: 3313–3332.

He K, Zhang X, Ren S, Sun J. **2015**. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification In: *Proceedings of the IEEE international conference on computer vision*.1026–1034.

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. **2017**. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems* **2017-December**: 6627–6638.

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

Krizhevsky A. **2009**. *Learning Multiple Layers of Features from Tiny Images*.

Lawson T, Blatt M. **2014**. Stomatal Size, Speed, and Responsiveness Impact on Photosynthesis and Water Use Efficiency. *Plant Physiology* **164**: 1556–1570.

Le J, Biewald L, Edmonds A, Hester V. **2010**. 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)*.17–20.

Li Y, Schwing A, Wang K-C, Zemel R. **2017**. Dualing GANs. *arxiv*: 1–11.

Liu B, Lv J, Fan X, Luo J, Zou T. **2022**. Application of an Improved DCGAN for Image Generation. *Mobile Information Systems* **2022**.

Long S, Zhu X-G, Naidu S, Ort D. **2006**. Can improvement in photosynthesis increase crop yields? *Plant, cell & environment* **29**: 315–330.

Lucic M, Kurach K, Michalski M, Bousquet O, Gelly S. **2017**. Are GANs Created Equal? A Large-Scale Study. *Advances in Neural Information Processing Systems* **2018-December**: 700–709.

Matthaeus W, Schmidt J, White J, Zechmann B. **2020**. Novel perspectives on stomatal impressions: Rapid and non-invasive surface characterization of plant leaves by scanning electron microscopy. *PLOS ONE* **15**: e0238589-.

Millstead L, Jayakody H, Patel H, *et al.* 2020. Accelerating Automated Stomata Analysis Through Simplified Sample Collection and Imaging Techniques. *Frontiers in Plant Science* **11**: 1–14.

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*.

Pathoumthong P, Zhang Z, Roy S, El Habti A. **2023**. Rapid non-destructive method to phenotype stomatal traits. *Plant Methods* **19**: 36.

Peterson K, Rychel A, Torii K. **2010**. Out of the Mouths of Plants: The Molecular Basis of the Evolution and Diversity of Stomatal Development. *The Plant Cell* **22**: 296–306.

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. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., .

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

Rawat W, Wang Z. **2017**. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation* **29**: 2352–2449.

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. **2021**. High-Resolution Image Synthesis with Latent Diffusion Models.

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

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. **2018**. MobileNetV2: Inverted Residuals and Linear Bottlenecks In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.4510–4520.

Sinha S, Zhao Z, Goyal A, Raffel C, Odena A. **2020**. Top-k Training of GANs: Improving GAN Performance by Throwing Away Bad Samples. *Advances in Neural Information Processing Systems* **2020-December**.

Thompson A, Senin N, Giusca C, Leach R. **2017**. Topography of selectively laser melted surfaces: A comparison of different measurement methods. *CIRP Annals* **66**: 543–546.

Wang J, Renninger HJ, Ma Q, Jin S. **2024**. Measuring stomatal and guard cell metrics for plant physiology and growth using StoManager1. *Plant Physiology* **195**: 378–394.

Wenzel M. **2023**. Generative Adversarial Networks and Other Generative Models In: *Machine Learning for Brain Disorders*. Humana Press Inc., 139–192.

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)*.

Figure 1

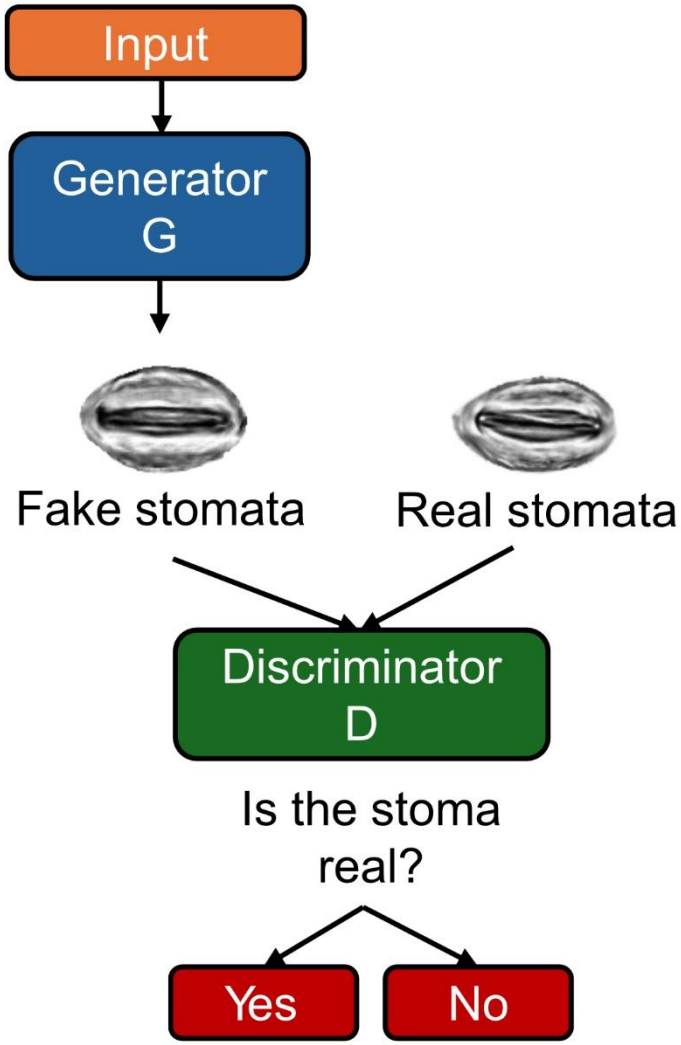

Figure 2

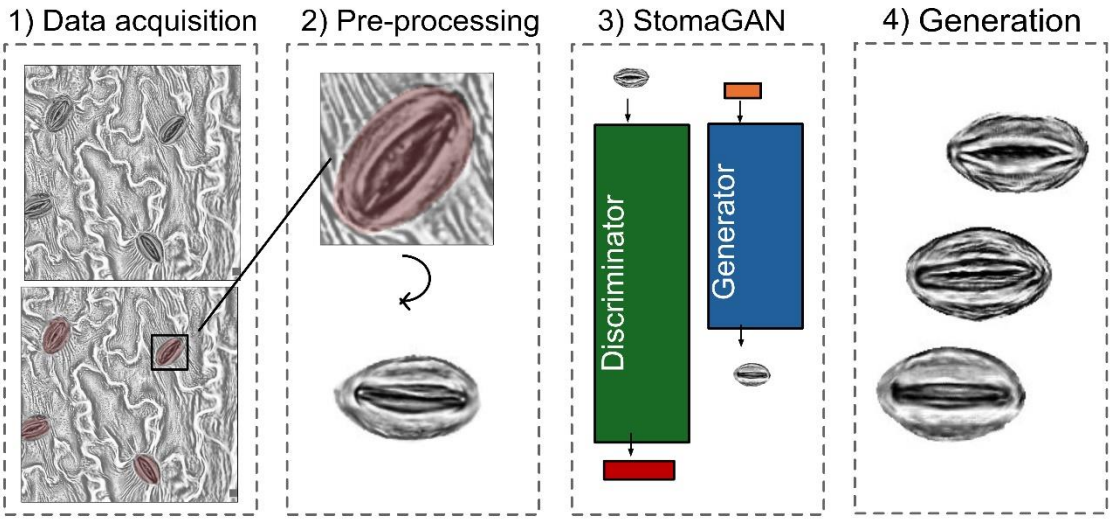

Figure 3

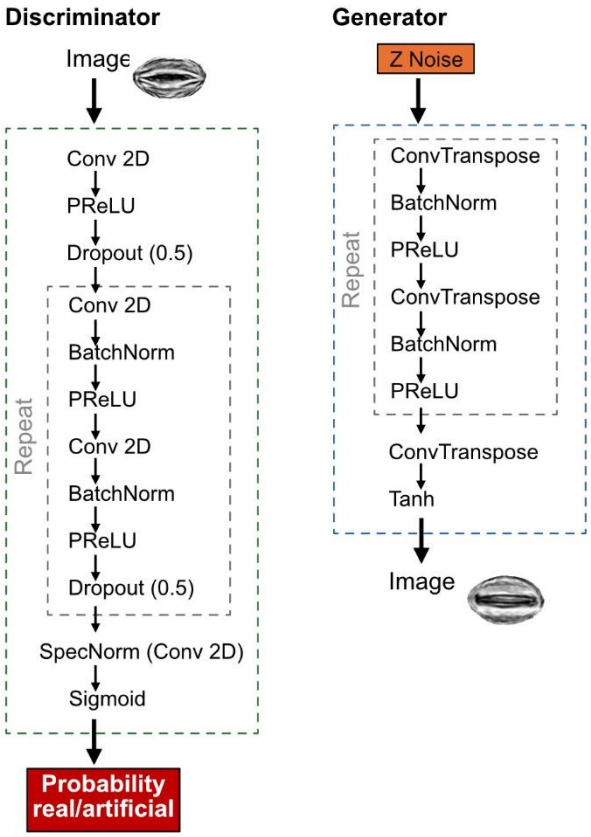

Figure 4

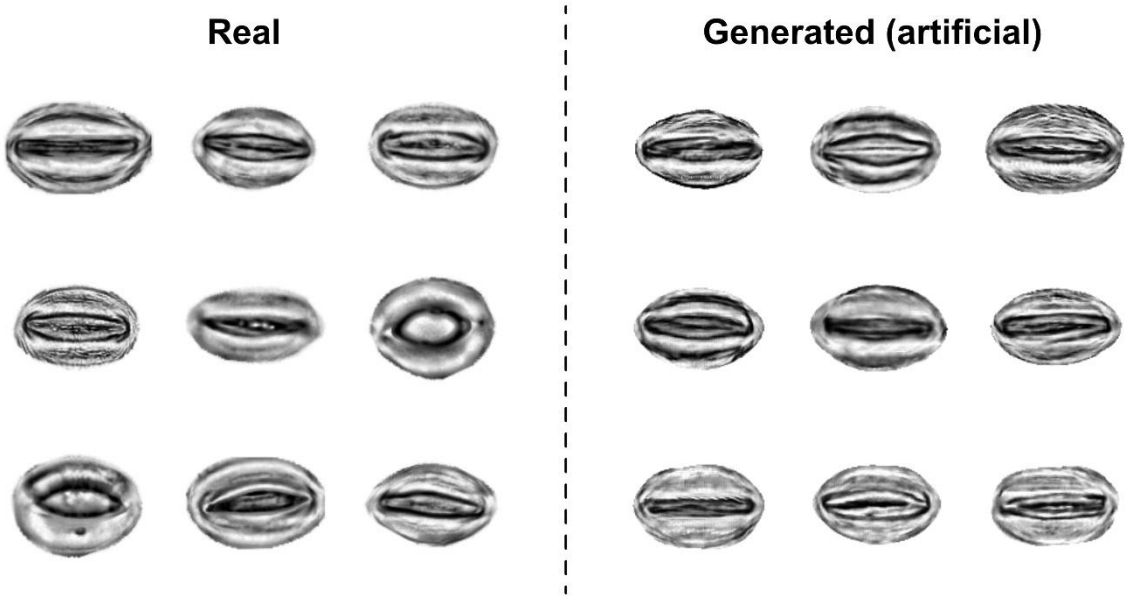

Figure 5

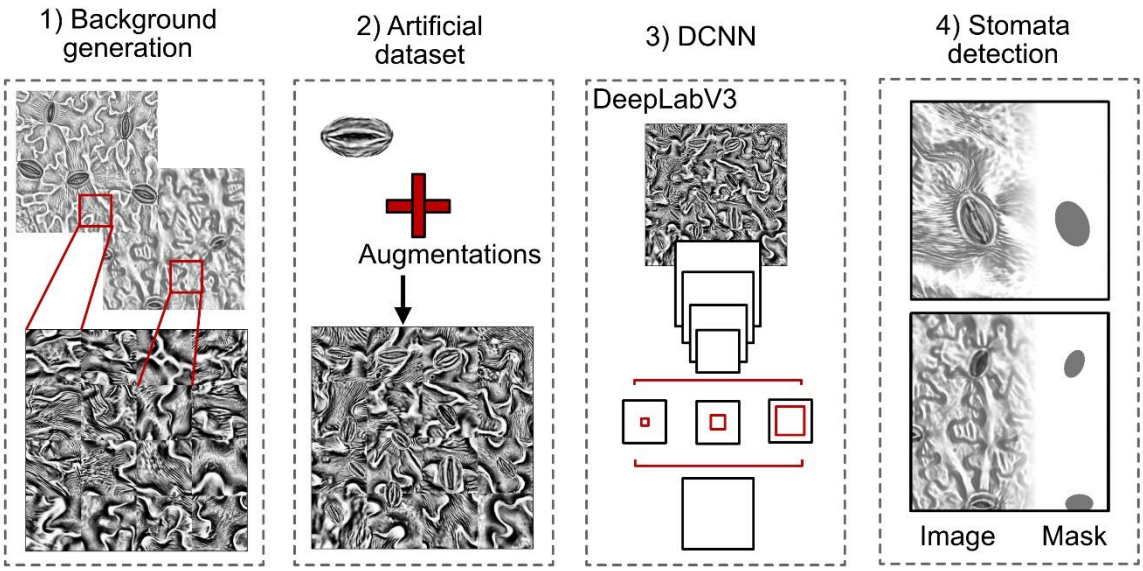

Figure 6