StomataCounter: a deep learning method applied to automatic stomatal identification and counting.

Karl C. Fetter[a,b,1,2],**Sven Eberhardt**[c,1],**Rich S. Barclay**[b],**Scott Wing**[b],**Stephen R. Keller**[a]

[a]Department of Plant Biology, University of Vermont
[b]Department of Paleobiology, Smithsonian Institution, National Museum of Natural History
[c]Amazon.com, Inc

Corresponding author: Karl C. Fetter (e-mail: kf@uvm.edu, tel: +1 802 656 2930).

[1]S.E. and K.F. contributed equally to this work

ABSTRACT

- Stomata fulfill an important physiological role and are often phenotyped by researchers in many fields. Currently, no fully automated method exists to perform this task. Researchers typically rely on manual counts of stomata, which is an error-prone method and difficult to reproduce.
- We introduce StomataCounter, an automated stomata counting system using a deep convolutional neural network to identify pores in a variety of different microscopic images. We used a human-in-the-loop approach to train and refine a neural network on a large variety of microscopic images, which helps us achieve robust detection among a number of datasets.
- Our network achieves 98.1% identification accuracy on Ginkgo SEM micrographs, and 94.2% transfer accuracy when tested on untrained species.
- To facilitate adoption of the method, we make a web tool available under http://www.stomata.science/.

INDEX TERMS Stomata, Computer Vision, Convolutional Neural Network, Deep Learning, Phenotyping

Total word count (excluding summary, references and legends)	3909
Summary	129
Introduction	788
Materials and Methods	1712
Results	506
Discussion	903
Acknowledgements	88
No. of Figures	7
Figures in color	1–4, 6, 7
No. of Tables	3
No. of Supporting Information files	1

I. INTRODUCTION

STOMATA (singular "stoma", Greek word for "mouth") play a critical role in plant physiology and are particularly important for photosynthesis. A plant's cuticle is impervious to gases and stomata evolved to permit exchange between internal and external sources of gases, most notably CO_2 and water vapor (Kim *et al.*, 2010). A stoma is composed of a pair of guard cells forming an aperture that, in some species, are flanked by a pair of subsidiary cells (Bergmann and Sack, 2007). Regulation of the aperture pore size, and hence, the opening and closing of the pore, is achieved through changing turgor pressure in the guard cells (Shimazaki *et al.*, 2007).

Because of their importance in regulating plant productivity and response to the environment, stomata have been one of the key functional traits of interest to researchers working across scales in plant biology. At the molecular level, regulation of stomata has been the subject of numerous genetic studies (see Shimazaki *et al.*, 2007; Kim *et al.*, 2010 for detailed reviews), and can be trait targeted in plant breeding and crop improvement projects (Fischer *et al.*, 1998). Stomata also mediate tradeoffs between carbon gain and pathogen exposure that are of interest to plant ecophysiologists and pathologists. For example, foliar pathogens frequently exploit the aperture pore as a site of entry. In *Populus*, some species, and even populations within species, evolve growth strategies that maximize carbon fixation through increased stomatal density and aperture pore size on the adaxial leaf surface. This adaptation results in a cost of increased infection rates from fungal pathogens that have more sites of entry into the leaf (McKown *et al.*, 2014). Stomata, as sites of water vapor output, are also implicated in driving environmental change across biomes (Hetherington and Woodward, 2003) and variation of stomatal density and aperture pore length are linked to changes of ecosystem productivity (Wang *et al.*, 2015). Stomatal traits are of particular interest to paleoecologists and paleoclimatologists due to the relationship between stomatal traits and gas exchange. Measurements of stomatal density from fossil plants have been proposed as an indicator of paleoatmospheric CO_2 concentration (Royer, 2001), and measuring stomatal traits to predict paleoclimates has become widely adopted (McElwain and Steinthorsdottir, 2017).

It is clear from their physiological importance that researchers across a wide variety of disciplines in plant biology will phenotype stomatal traits for decades to come. Stomatal traits that are often phenotyped include density, guard cell size, aperture pore size, and interstomatal length. A typical stomatal phenotyping pipeline consists of collecting plant tissue, creating a mounted tissue for imaging (typically a nail polish peel or a cleared and stained leaf), imaging of the specimen, and manual phenotyping of a trait of interest. This last step can be the most laborious, costly, and time-consuming task, reducing the efficiency of the data acquisition and analysis pipeline. This is especially important in large-scale plant breeding and genome-wide association studies, where phenotyping has been recognized as the new bottleneck in data collection the relative ease of generating large genome sequence datasets (Hudson, 2008).

Here, we seek to minimize the burden of high-throughput phenotyping of stomatal traits by introducing an automated method to recognize and count stomata from plant epidermal micrographs. Although automated methods using computer vision have been suggested (Higaki *et al.*, 2014; Laga *et al.*, 2014; Duarte *et al.*, 2017; Jayakody *et al.*, 2017), these highly specialized approaches require feature engineering specific to an image class. Such features typically transfer poorly to images from a new source, such as images recorded using different microscopy and illumination techniques, or different image processing protocols. Tuning hand-crafted methods to work on a general set of conditions is cumbersome and often impossible.

Deep convolutional neural networks (DCNN) circumvent this problem by training the feature detector along with the classifier (LeCun *et al.*, 2015). The method has been widely successful on a range of computer vision problems in biology such as medical imaging (See Shen *et al.* (2017) for a review) or macroscopic plant phenotyping (Ubbens and Stavness, 2017). The main caveat of deep learning methods is that a large number of parameters have to be trained in the feature detector. Although gradient descent learning methods prove to be surprisingly resilient against overfitting (Poggio *et al.*, 2017), and many small improvements in network structure (He *et al.*, 2016) and training procedure (Simonyan and Zisserman, 2014) have helped training of large networks. Nevertheless, a large number of correctly annotated training images is still required to allow the optimizer to converge to a correct feature representation. For classification between common feature classes (such as animals or cars), large labeled training like the ImageNet database exist (Jia Deng *et al.*, 2009), but for a highly specialized problems, such as stomata identification, publicly available datasets are not available at the scale required to train a typical DCNN.

We solve this problem by training a network with a human-in-the-loop approach using a publicly available web-based tool called StomataCounter. Our development of StomataCounter allows plant biologists to rapidly upload plant epidermal image datasets to pre-trained networks and then annotate stomata on cuticle images when desired. We apply this tool to a diverse set of plant cuticle images and achieve robust identification and counts of stomata on a variety of angiosperm and pteridosperm taxa.

II. MATERIALS AND METHODS

A. PLANT MATERIAL

Micrographs of plant cuticles were collected from four sources: the cuticle database (https://cuticledb.eesi.psu.edu/ Barclay *et al.*, 2016); SEM images of the Maidenhair tree (*Gingko biloba*) created from 2013 to 2016; a large intraspecific collection balsam poplar (*Populus balsamifera*) micrographs; and a new collection made in 2017 from fresh plant material from a variety of taxa. Dormant cuttings of *P. balsamifera* genotypes were collected across the eastern and central portions of its range in the United States and Canada by S.R. Keller and collaborators, planted in common gardens in Vermont and Saskatchewan, and fresh leaves sampled during June-July 2015. For the 2017 collection, fresh leaves were opportunistically sampled in June 2017 from the gardens around the National Museum of Natural History building (USNM) and from the collections of the United States Botanic Gardens (USBG). The taxonomic identity of each specimen was recorded according to the existing label next to each plant. Hereafter, we refer to this collection as USNM/USBG.

For freshly collected samples (*P. balsamifera* and USNM/USBG collections), leaves were clipped from plants and transferred into plastic bags for temporary storage (up to three hours) and transported back to the lab. Nail polish (Sally Hansen, big kwik dry top coat #42494) was painted onto a 1 cm^2 region of the adaxial and abaxial leaf surface and allowed to dry for approximately twenty minutes. The dried cast of the epidermal surface was lifted with clear tape and mounted onto a glass slide. The USNM/USBG collection specimens were imaged with an Olympus BX-63 microscope using differential intereference contrast (DIC). Some specimens had substantial relief and z-stacked images were created using cellSens image stacking software (Olympus Corporation). *P. balsamifera* material was imaged with an Olympus BX-60 using DIC.

Specimens in the cuticle database collection were previously prepared by clearing and staining leaf tissue and then imaged. The entire collection of the cuticle database was downloaded on November 16, 2017. Downloaded images contained both the abaxial and adaxial cuticles, and were automatically separated with a custom bash script. Abaxial cuticle micrographs were discarded if no stomata were visible or if the image quality was so poor that no stomata could be visually identified by a human. Micrographs from all four collections were imaged at 200x and/or 400x magnification. Mounted material for the USNM/USBG and Ginkgo collections are deposited in the Smithsonian Institution, National Museum of Natural History in the Department of Paleobiology. Material for the *P. balsamfiera* collection is deposited in the Keller laboratory in the Department of Plant Biology at the University of Vermont, USA.

B. DEEP CONVOLUTIONAL NEURAL NETWORK

We used a Deep Convolutional Neural Network (DCNN) to generate a stomata likelihood map for each input image, followed by thresholding and peak detection to localize and count stomata (Fig. 1). Because dense per-pixel annotations of stoma versus non-stoma are difficult to acquire in large quantity, we trained a simple image classification DCNN based on the AlexNet structure instead (Krizhevsky *et al.*, 2012), and copied the weights into a fully convolutional network to allow per-location assessment of stomata likelihood. Although this method does not provide dense per-pixel annotations, the resulting resolution proved to be high enough to differentiate and count individual stomata.

C. DCNN TRAINING

Because of the relatively small training dataset sampled from 4,767 images, we used pre-trained weights for the lowest five convolutional layers from conv1 to conv5. The weights were taken from the ILSVRC image classification tasks (Russakovsky *et al.*, 2015) made available in the caffenet distribution (Jia *et al.*, 2014). All other layers were initialized using Gaussian initialization with scale 0.01. Training was performed using a standard stochastic gradient descent solver as in Krizhevsky *et al.* (2012), with learning rate 0.001 for pre-trained layers and 0.01 for randomly initialized layers, momentum 0.9. Because the orientation of any individual stomata does not hold information for identification, we augmented data by rotating all training images into eight different orientations, applied random flipping and randomly positioned crop regions of the input size within the extracted 256x256 image patch. For distractors, we sampled patches from random image regions on human-annotated images that were at least 256 pixels distant from any labeled stoma. The trained network weights were transferred into a fully convolutional network (Long *et al.*, 2015), which replaces the final fully connected layers by convolutions. To increase the resolution of the detector slightly, we reduced the stride of layer $pool5$ from two to one, and added a dilation (Yu and Koltun, 2015) of two to layer $fc6_{conv}$ to compensate. Due to margins (96 pixels) and stride (32 over all layers), application of the fully convolutional network to an image of size s yields an output probability map p of size $s-(96*2)/32$ along each dimension.

D. STOMATA COUNTING

To avoid detecting low-probability stomata within the noise, the probability map p was thresholded and all values below the threshold p-thresh=0.98 were set to zero.

Local peak detection was run on a 3x3 pixel neighborhood on the thresholded map, and each peak, excluding those located on a border, was labeled as a stoma center. This intentionally excludes stomata for which the detection peak is found near the border within the model margin to match the instructions given to human annotators. Resulting stomata positions were projected back onto the original image (see figure 1).

E. WEBSERVER

We built a user-friendly web service, StomataCounter, freely available at http://stomata.science/ , to allow the scientific community easy access to automated stomata counting. We are using a flask/jquery/bootstrap stack. Source codes for network training, as well as the webserver are available at http://stomata.science/source. To use StomataCounter, users upload a zip file of their images containing leaf cuticles prepared and imaged using an appropriate protocol (as above). A new dataset is then created where the output of the automatic counts, image quality scores and image metadata are recorded and can easily be exported for further analysis.

In addition to automatic processing, the user can manually annotate stomata and determine the empirical error rate of the automatic counts through a straightforward and intuitive web interface. The annotations can then be reincorporated into the training dataset to improve future performance of StomataCounter by contacting the authors and requesting retraining of the DCNN.

F. EVALUATION DATASETS

We tested the performance of the DCNN by creating a test set of images from the four sources of stomata images described in the plant material section (i. e. cuticle database, Ginkgo, *P. balsamifera*, and USNM/USBG sets). Whenever possible, images from a given species were used in either the training or test set, but not both, resulting in only 37 out of 1,378 species included in both the training and test sets. After running the test set through the network, stomata were manually annotated. If the center of a stoma intersected the bounding box around the perimeter of the image, it was not counted.

For the evaluation, micrographs from 1,378 species totaling 6,530 images were collected. Training and testing datasets were constructed as shown in table 1.

G. IDENTIFICATION EVALUATION

Identification accuracy is tested by applying the DCNN to a small image patch either centered on a stoma (target), or taken from at least 256 pixels distance to any labeled stomata (distractor). This yields true positive (N_{TP}), true negative (N_{TN}), false positive (N_{FP}) and false negative (N_{FN}) samples. We define the classication accuracy A as:

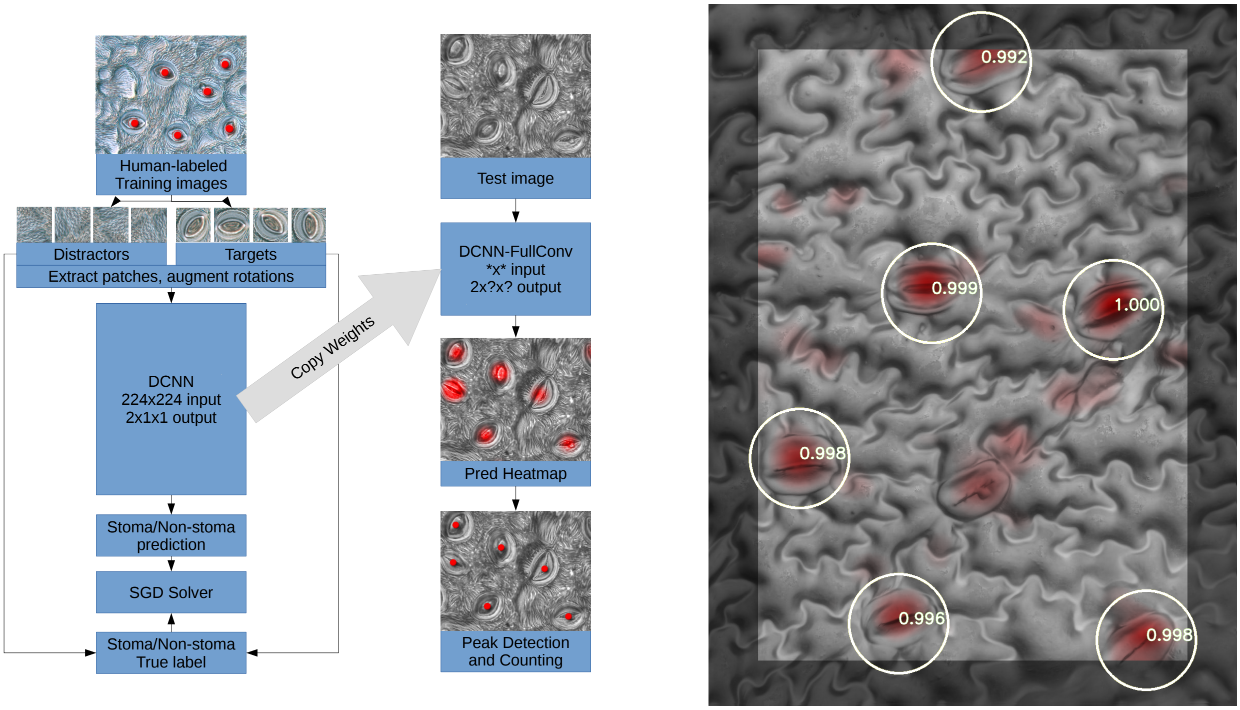

Human-labeled Training images
Distractors
Targets
Extract patches, augment rotations
DCNN 224x224 input 2x1x1 output
Copy Weights
Stoma/Non-stoma prediction
SGD Solver
Stoma/Non-stoma True label
Test image
DCNN-FullConv *x* input 2x?x? output
Pred Heatmap
Peak Detection and Counting
0.992
0.999
1.000
0.998
0.996
0.998

Figure 1. Left: Training and testing procedure. First column: Target patches were extracted and centered around human-labeled stomata center positions; distractor patches were extracted in all other regions. A binary image classification network was trained. Second column: The image classification network was applied fully convolutional to the test image to produce a prediction heatmap. On the thresholded heatmap, peaks were detected and counted.
Right: Example heatmap output of the model on a stomata image. Probability map overlaid in red channel. Detected stomata marked with circles; peak value given in white. One stoma in the bottom center was not detected because the probability was below threshold.

Table 1. Train (top) and Test (bottom) dataset sizes. μ: mean number of samples for each taxonomic level (family/genus/species). $\tilde{\mathbf{n}}$: median number of samples for each level. Poplar images from *P. balsamifera*.

Train datasets

Dataset	N_{images}	N_{species}	$\tilde{n}_{\text{family}}$ μ_{family}	$\tilde{n}_{\text{genus}}$ μ_{genus}	$\tilde{n}_{\text{species}}$ μ_{species}
Cuticle	784	578	2 11.5	1 2.5	1 1.2
USNM/USBG	431	128	4 6.3	3 3.9	3 3.3
Poplar	3,144	1	-	-	3,144
Ginkgo	408	1	-	-	408
Total	4,767	708	4 41.8	1 12.1	1 6.6

Test datasets

Dataset	N_{images}	N_{species}	$\tilde{n}_{\text{family}}$ μ_{family}	$\tilde{n}_{\text{genus}}$ μ_{genus}	$\tilde{n}_{\text{species}}$ μ_{species}
Cuticle	671	573	5 18.1	1 3.3	1 1.2
USNM/USBG	694	132	6 12.6	6 6.9	6 5.5
Poplar	198	1	-	-	198
Ginkgo	200	1	-	-	200
Total	1,763	707	6 23.2	3 6.3	1 2.7

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{TP}}$$

H. DETECTION EVALUATION

To evaluate the DCNN, we first determined if it could identify stomata when they are known to be present and fail to identify them when they are absent. To execute this test, a set of 25 randomly selected abaxial plant cuticle micrographs containing stomata were chosen from each of the four datasets for a total of 100 images. To create a set of test images known to lack stomata, 100 adaxial cuticle micrographs (adaxial surfaces typically lack stomata) were randomly sampled from the cuticle database. Visual inspection confirmed that none of the adaxial images contained stomata. We also included an additional 116 micrographs of human aorta tissue (provided by A. Chapman UVM) and 58 breast cancer tissue micrographs (Gelasca *et al.*, 2008) to assess if the DCNN is likely to detect structures within images that share no homology with the target trait.

Stomata counts between human and machine annotations were compared with the log precision of the counts, given that the manual counts contain true positives and the automatic counts contain true positives plus false positives defined as:

$$\log\left(\frac{\text{ManualCount}}{\text{AutomaticCount}}\right)$$

The log precision identifies over-counting errors as negative values and under-counting as positive values. Because this measure is undefined if either manual or automatic count is zero, such values were discarded from the evaluation. This affected 30 of 1793 samples and these samples were either out of focus, lacked stomata entirely, or too grainy for human detection of stomata.

We use linear regression to understand the relationship between human and automatically counted stomata. The images were partitioned into four groups defined by the characteristics of the data set. As the sample preparation and imaging of the cuticle database was quite different from the other images created with light microscopy, they were grouped separately. Samples from the USNM/USBG and *P. balsamifera* collections were prepared with the same method and imaged similarly; thus, these images were collectively partitioned into a group of micrographs imaged at 200x and 400x. The Ginkgo images were grouped together as the sample preparation and imaging (SEM) was unique to this collection.

To understand how different sources of variance contribute to precision variance, we collected data on the taxonomic family, magnification, imaging technique, and three measures of image quality. The taxonomic family of each image was determined using the rotl R package (Michonneau *et al.*, 2016). Two of the image quality measures were based on the properties of an image's power-frequency distribution tail and described the standard deviation (fSTD) and mean (fMean) of the tail. Low values of fSTD and fMean indicate a blurry image, while high values indicate non-blurry images. The third image quality measure, entropy, is a measure an image's information content. High entropy values indicate high contrast/noisy images while low values indicate lower contrast. These image quality measures were created with PyImq (Koho *et al.*, 2016). Random effects linear models were created with the R package lme4 (Bates *et al.*, 2014) by fitting log precision to taxonomic family, magnification, and imaging technique as factors. Linear models were fit for the scaled error and image quality scores. We used the root mean square error (RMSE) of the model residuals to understand how the factors and quality scores described the variance of log precision. Higher values of RMSE indicate larger residuals. Statistical analyses were conducted in python and R (R Team, 2016).

III. RESULTS

A. STOMATA DETECTION

StomataCounter was able to accurately identify and count stomata when they were present in an image; stomata were detected in all of the abaxial cuticle images. False positives were detected in the adaxial cuticle, aorta, and breast cancer cell image sets at a very low frequency (Fig. 2). The mean number of stomata detected in the adaxial, aorta, and breast cancer image sets was 1.5, 1.4, and 2.4, respectively, while the mean value of the abaxial set was 24.1. No stomata were detected in 105 out of the 274 (38.3%) non-stomata containing images.

Correspondence between automated and human stomata counting varied among the respective sample sets, with close agreement among all but the 200x samples where StomataCounter tended to under count relative to human observers (Fig. 3). The slopes of all models were close to 1 (Table 2), despite the total variation among the datasets and the large error present in the 200x set. The 400x, Ginkgo, and cuticle database sets all performed well at lower stomata counts, as indicated by their proximity to the expected one-to-one line.

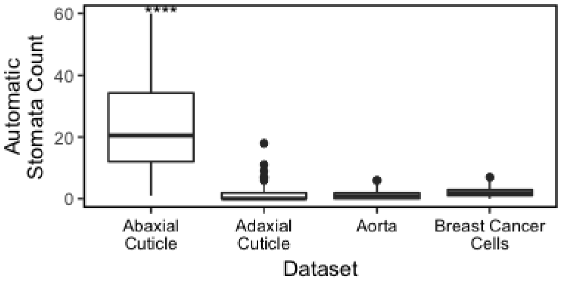


Figure 2. StomataCounter correctly identified stomata in all abaxial cuticle images containing stomata, while producing a low frequency of false positive counts in other image types. More false positives were identified in adaxial cuticle images than in images from human aorta or breast cancer cells. Results of post-hoc significance testing of pairwise means indicated by: *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

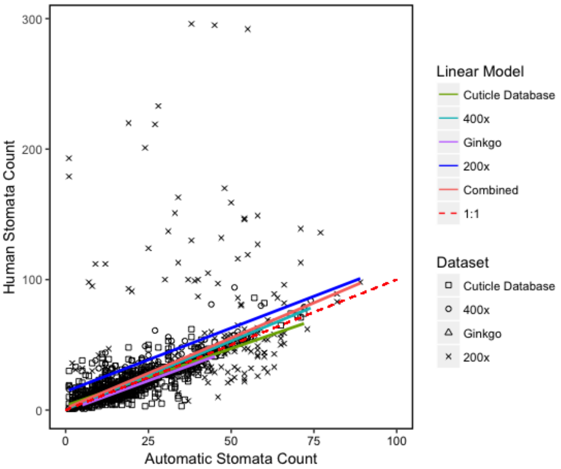


Figure 3. Human versus automatic stomata count. The one-to-one line (red dashed line) indicates a perfect count between the human and automatic count. StomataCounter performed best on the Ginkgo.

Comparison of human versus automated counting revealed patterns of scaled error that varied with sample preparation and imaging methods (Fig. 4). Scaled error was related to image quality (entropy, fMean, fSTD), while among the different imaging methods, SEM had the lowest error and brightfield the most, with DIC intermediate. There was little bias apparent, with scaled errors centered symmetrically around zero (Fig. 4). RMSE values were lowest for taxonomic family and the family:magnification interaction, suggesting these factors contributed less to deviations between human and automated stomata counts than image quality or imaging technique (Table 3).

Table 2. Summary of linear model fit parameters in Figure 3 for different test datasets. SEa: Intercept standard error. CI_s: Slope 95% confidence interval. R^2: Mean squared residual. N: Sample count in test dataset. Significance indicated by: *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

Dataset	Combined	Cuticle	400x	200x	Ginkgo
Intercept a	1.329	3.454***	0.499	14.236**	-0.795*
SEa	0.841	0.571	0.388	4.855	0.392
Slope s	1.078***	0.873***	1.050***	0.973***	0.933***
CI_s	$[1.01, 1.14]$	$[0.82, 0.93]$	$[1.02, 1.08]$	$[0.72, 1.22]$	$[0.89, 0.97]$
R^2	0.382	0.609	0.878	0.162	0.933
N	1738	671	572	295	200

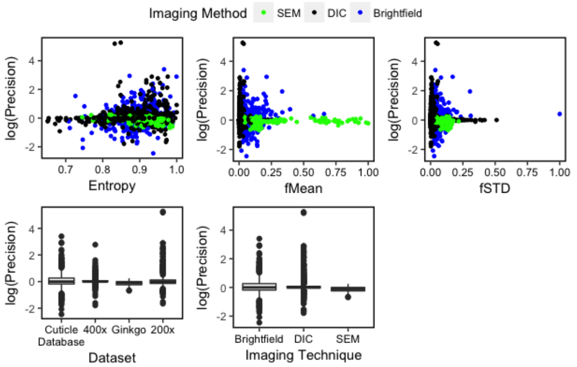


Figure 4. Log precision plotted against image quality scores and the dataset and imaging techniques. Positive values of scaled error indicate undercounting by the StomataCounter relative to human counts, negative values indicate overcounting.

Table 3. Influence of taxonomic family and imaging technique/quality on explaining the log precision of stomata human and automatic stomata counts. RMSE, root mean square error of the model residuals (lower RMSE suggests greater contribution of the predictor to the scaled error).

Model	RMSE
log(Precision) $\sim$ family	0.463
log(Precision) $\sim$ magnification	0.523
log(Precision) $\sim$ family:magnification	0.399
log(Precision) $\sim$ image_technique	0.519
log(Precision) $\sim$ fMean	0.523
log(Precision) $\sim$ fSTD	0.523
log(Precision) $\sim$ entropy	0.521
log(Precision) $\sim$ entropy:image_technique	0.515

B. STOMATA IDENTIFICATION

Results of the classification accuracy of the DCNN are shown in figure 5. Unsurprisingly, the peak accuracy (94.2%) on the combined test sets is achieved when all training sets are combined. The combined dataset performs best on all test subsets of the data; i. e. adding additional training data - even from different sets - is always beneficial for the generalization of the network. Accuracy from train to test within a single species is higher (e. g. Ginkgo training for Ginkgo test at 97.4%) than transfer within datasets with a large number species across families (400x training to 400x test: 85.5%).

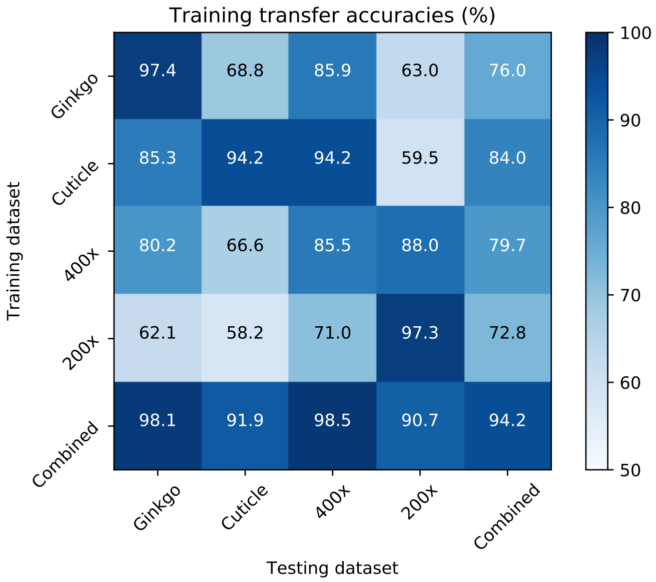


Figure 5. Accuracies for models trained on different training datasets (vertical), tested on different test datasets. *Combined* is a union of all training and test sets. For precision and recall values, see supplement 1.

The network does not generalize well between vastly different scales, i.e. the 200x-dataset, which contains images downsacaled to half the image width and height. In this case, only training within the same scale achieves high accuracy (97.3%), while adding additional samples from the larger scale reduces the performance (to 90.7%). We argue that the current architecture does not transfer well to different scales and images should be pre-processed to match the training scale of the network.

Precision values are generally higher than recall (0.99 precision on the combined training and test sets; 0.93 recall, see supplement 1), which shows that we mostly miss stomata rather than misidentifying non-existing stomata.

The effect of the training size is visualized in figure 6. Providing a large number of annotated images is beneficial, as it lifts training accuracy from 72.8% to 94.2% on the test dataset.

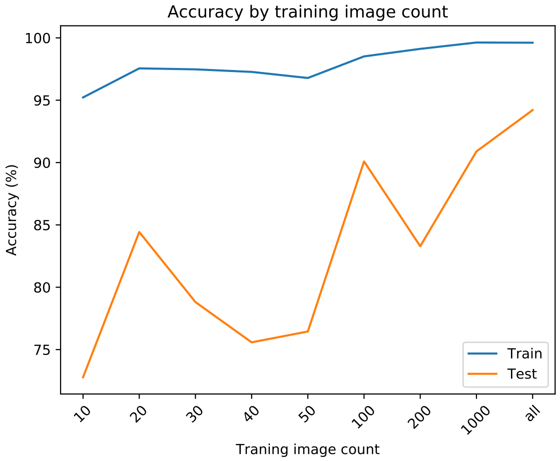


Figure 6. Effect of the training set size on classification accuracy of the training (blue) and test set (line). Training image count "*all*" includes all 4,566 annotated training images. Because this is a binary classification task, chance level is at 50%.

IV. DISCUSSION

Stomata are an important functional trait to many fields within plant biology, yet manual phenotyping of stomata counts is a laborious method that has few controls on human error and reproducibility. We created a fully automatic method for counting stomata that is both highly sensitive and reproducible, allows the user to quantity error in their counts, but is also entirely free of parameter optimization from the user. Furthermore, the DCNN can be iteratively retrained with new images to improve performance and adjust to the needs of the community. This is a particular advantage of this method for adjusting to new taxonomic sample sets.

Apart from the reduced workload, automated image processing provides better reproducibility than manual stomata annotations. For instance, if multiple experts count stomata, they may not agree, causing artificial differences between compared populations. This includes how stomata at the edge of an image are counted, and what to do with difficult to identify edge cases. Automatic counters will have an objective measure, and introduce no systematic bias between compared sets as long as the same model is used. As the complexity of processing pipelines in biological studies increases, repeatability of studies increasingly becomes a concern (Bruna, 2014, Open Science Foundation, 2015).

Our method is not the first to identify and count stomata. However, previous methods have not been widely adopted by the community and many researchers are likely to manually count stomata. Previous methods relied on substantial image pre-processing to generate images for thresholding to isolate stomata for counting (Oliviera *et al.*, 2014; Duarte *et al.*, 2017). Thresholding can perform well in a homogenous collection of images, but quickly fails when images collected by different preparation and micrscopy methods are provided to the thresholding method (K. Fetter, pers. obs.). Some methods also require the user to manually segment stomata and subsequently process those images to generate sample views to supply to template matching methods (Laga *et al.*, 2014). Object-oriented methods (Jian *et al.*, 2011) also require input from the user to define model parameters. These methods invariably requires the user to participate in the counting process to tune parameters and monitor the image processing, and are not fully autonomous. More recently, cascade classifier methods have been developed which perform well on small collections of test sets (Vialet-Chabrand and Brendel, 2014; Jayakody *et al.*, 2017). Additionally, most methods rely on a very small set of images (50 to 500) typically sampled from just a few species to create the training and/or test set.

Apart from generalization concerns, all published methods require the user to have some experience coding in python or C++, a requirement likely to reduce the potential pool of end users. Our method resolves these issues by being publically available, fully autonomous of the user, who is only required to upload a zip file of jpeg formatted images, free of any requirement for the user to code, and is trained on a relatively large and taxonomically diverse set of cuticle images.

We have demonstrated that this method is capable of accurately identifying stomata when they are present, but false positives may still be generated by shapes in images that approximate the size and shape of stomata guard cells. Conversely, false negatives are generated when a stomata is hidden by a feature of the cuticle or if poor sample preparation/imaging introduces blur. This issue is likely avoidable through increased sample size of the training image set and good sample preparation and microscopy techniques from the end user.

The importance of having a well-matched training and testing image set was apparent at 200x, where there was a subset of observations where transfer from the other 400x training sets was low (Fig. 5), and StomataCounter consistently under counted relative to human observation (Fig. 2). Our training set of images spanned 82 different families and was overrepresented by angiosperms. Stomata in gymnosperms are typically sunken into pores that make it difficult to obtain good nail polish casts. The resulting images from gymnosperms lead to a loss in accurate counting. Consistent with this, models tested to explain variation in scaled error revealed that taxonomic family and its interaction with magnification were the factors that had the best explanatory power for scaled error. Future collections of gymnosperm cuticles could be uploaded to the DCNN to retrain it in order to improve the performance of the method for gymnosperms.

More generally however, this highlights how users will need to be thoughtful about their matching of training and test samples for taxa that may deviate in stomata morphology from the existing reference database. We therefore

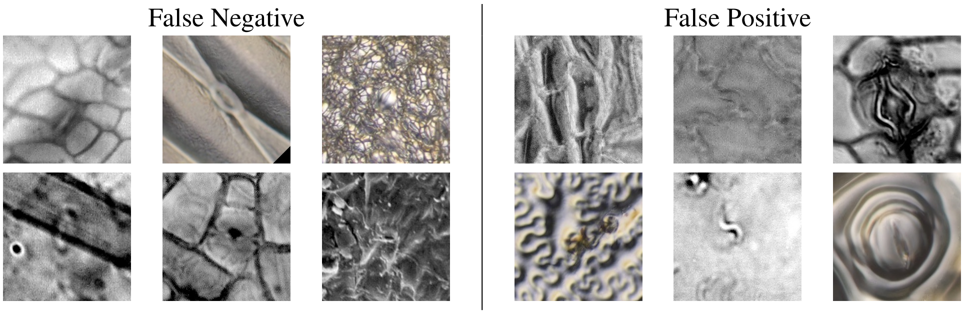


Figure 7. Samples that were mislabeled with high confidence. Left: Human-annotated as stoma; machine annotated as non-stoma with high confidence (missed targets).
Right: Human-annotated as non-stoma; machine annotated as stoma with high confidence (false positives). Typical errors are blurry images, artifacts in images, artifacts that look like stomata, stomata at a very small scale, or simple human annotation errors, where the model did a correct prediction (e.g. bottom right and top right images).

recommend that users working with new or morphologically divergent taxa first run several pilot tests with different magnification and sample preparation techniques to find optimal choices that minimize error for their particular study system. SEM micrographs had the least amount of error, followed by DIC, and finally bright field (Fig. 4). Lastly, image quality was strongly related to log precision; predictably, images that are too noisy (i.e. high entropy) and out-of-focus (low fMean, fSTD) will generate higher error. Obtaining high quality, in-focus images should be a priority during data acquisition.

Fast and accurate counting of stomata increases productivity of workers and decreases the time from collecting a tissue to analyzing the data. Until now, assessing measurement error required phenotyping a reduced set of images multiple times by, potentially, multiple counters. With StomataCounter, users can instantly phenotype their images and annotate them to create empirical error rates. The open source code and flexibility of using new and customized training sets will make StomataCounter and important resource for the plant biology community.

V. ACKNOWLEDGEMENTS

We thank Kyle Wallick at the United States Botanic Garden for facilitating access to the living collections of the Garden. Scott Whitaker aided in imaging cuticle specimens with SEM and DIC microscopy in the Laboratories of Analytical Biology at the Smithsonian Institution, National Museum of Natural History. Dr. Abbie Chapman Dr. Terry Delaney at the University of Vermont kindly allowed KCF to use his microscope for imaging. Funding to create Stomata Counter was provided by an NSF grant to SRK (IOS-1461868) and a Smithsonian Institution Fellowship to KCF.

VI. AUTHOR CONTRIBUTIONS

The research was conceived and performed by KCF and SE. The website and python scripts were written by SE. Data were collected and analyzed by KCF. Ginkgo Images were submitted by RSB and SW. All authors interpreted the results. The manuscript was written by KCF, SE, and SRK. All authors edited and approved the manuscript.

References

Barclay RS, Bush R, Baczynski AA, France C, and Wing S. 2016. Constraining the drivers of variance in leaf δ^{13}C in fossil Ginkgo adiantoides and extant Ginkgo biloba: canopy effect or diagenesis? In: *Geological Society of America Annual Meeting and Exposition Program*, p. 241.

Bates DM, Maechler M, Bolker BM, and Walker S. 2014. Fitting Linear Mixed-Effects Models Using lme4. In: *Journal of Statistical Software* arXiv, p. 1406.5823.

Bergmann DC and Sack FD. 2007. Stomatal Development. In: *Annual Review of Plant Biology* 58.1, pp. 163–181.

Bruna EM. 2014. Reproducibility {&} Repeatability in Tropical Biology: a call to repeat classic studies. In: *Biotropica.Org*.

Collaboration OS *et al.* **2015**. Estimating the reproducibility of psychological science. In: *Science* 349.6251, aac4716.

Duarte K, Carvalho M de, and Martins P. 2017. Segmenting High-quality Digital Images of Stomata using the Wavelet Spot Detection and the Watershed Transform. In: *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, (VISIGRAPP 2017)*, pp. 540–547.

Fischer RA, Rees D, Sayre KD, Lu ZM, Condon AG, and Larque Saavedra A. 1998. Wheat yield progress associated with higher stomatal conductance and photosynthetic rate, and cooler canopies. In: *Crop Science* 38.6, pp. 1467–1475.

Gelasca ED, Byun J, Obara B, and Manjunath BS. 2008. Evaluation and benchmark for biological image segmentation. In: *Proceedings - International Conference on Image Processing, ICIP*, pp. 1816–1819.

He K, Zhang X, Ren S, and Sun J. 2016. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. arXiv: 1512.03385.

Hetherington AM and Woodward FI. 2003. The role of stomata in sensing and driving environmental change. In: *Nature* 424.6951, pp. 901–908.

Higaki T, Kutsuna N, and Hasezawa S. 2014. CARTA-based semi-automatic detection of stomatal regions on an Arabidopsis cotyledon surface. In: 26, pp. 9–12.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. In: *Molecular Ecology Resources* 8.1, pp. 3–17.

Jayakody H, Liu S, Whitty M, and Petrie P. 2017. Microscope image based fully automated stomata detection and pore measurement method for grapevines. In: *Plant Methods* 13.1.

Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, and Darrell T. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In: *arXiv preprint arXiv:1408.5093*.

Jian S, Zhao C, and Zhao Y. 2011. Based on remote sensing processing technology estimating leaves stomatal density of Populus euphratica. In: *International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 547–550.

Kim TH, Bohmer M, Hu H, Nishimura N, and Schroeder JI. 2010. Guard Cell Signal Transduction Network: Advances in Understanding Abscisic Acid, CO 2 , and Ca 2+ Signaling. In: *Annual Review of Plant Biology* 61.1, pp. 561–591.

Koho S, Fazeli E, Eriksson JE, and Hänninen PE. 2016. Image Quality Ranking Method for Microscopy. In: *Scientific Reports* 6.

Krizhevsky A, Sutskever I, and Hinton GE. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances In Neural Information Processing Systems (NIPS)*, pp. 1–9. arXiv: 1102.0183.

Laga H, Shahinnia F, and Fleury D. 2014. Image-based plant stornata phenotyping. In: *2014 13th International Conference on Control Automation Robotics and Vision, ICARCV 2014*, pp. 217–222.

LeCun Y, Bengio Y, and Hinton G. 2015. Deep learning. In: *Deep learning.* arXiv: arXiv:1312.6184v5.

Long J, Shelhamer E, and Darrell T. 2015. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

McElwain JC and Steinthorsdottir M. 2017. Paleoecology, Ploidy, Paleoatmospheric Composition, and Developmental Biology: A Review of the Multiple Uses of Fossil Stomata. In: *Plant Physiology* 174.2, pp. 650–664.

McKown AD, Guy RD, Quamme L, Klápště J, La Mantia J, Constabel CP, El-Kassaby YA, Hamelin RC, Zifkin M, and Azam MS. 2014. Association genetics, geography and ecophysiology link stomatal patterning in Populus trichocarpa with carbon gain and disease resistance trade-offs. In: *Molecular Ecology* 23.23, pp. 5771–5790.

Michonneau F, Brown JW, and Winter DJ. 2016. rotl: an R package to interact with the Open Tree of Life data. In: *Methods in Ecology and Evolution* 7.12, pp. 1476–1481.

Oliviera MWdS, Silva NR da, Casanova D, Pinheiro LFS, Kolb RM, and Bruno OM. 2014. Automatic counting of stomata in epidermis microscopic images. In: *X Workshop de Visao Computacional* 3, pp. 253–257.

Poggio T, Kawaguchi K, Liao Q, Miranda B, Rosasco L, Boix X, Hidary J, and Mhaskar H. 2017. Theory of Deep Learning III: explaining the non-overfitting puzzle. In: arXiv: 1801.00173.

Royer DL. 2001. Stomatal density and stomatal index as indicators of paleoatmospheric CO2concentration. In: *Review of Palaeobotany and Palynology* 114.1-2, pp. 1–28.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A, and Fei-Fei L. 2015. ImageNet large scale visual recognition challenge. In: *International Journal of Computer Vision* 115.3, pp. 211–252.

Shen D, Wu G, and Suk Hi. 2017. Deep Learning in Medical Image Analysis. In: March, pp. 221–248.

Shimazaki Ki, Doi M, Assmann SM, and Kinoshita T. 2007. Light Regulation of Stomatal Movement. In: *Annual Review of Plant Biology* 58.1, pp. 219–247.

Simonyan K and Zisserman A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: *arXiv preprint*, pp. 1–10. arXiv: 1409.1556.

TEAM RDC and R DEVELOPMENT CORE TEAM. 2016. A Language and Environment for Statistical Computing. In:

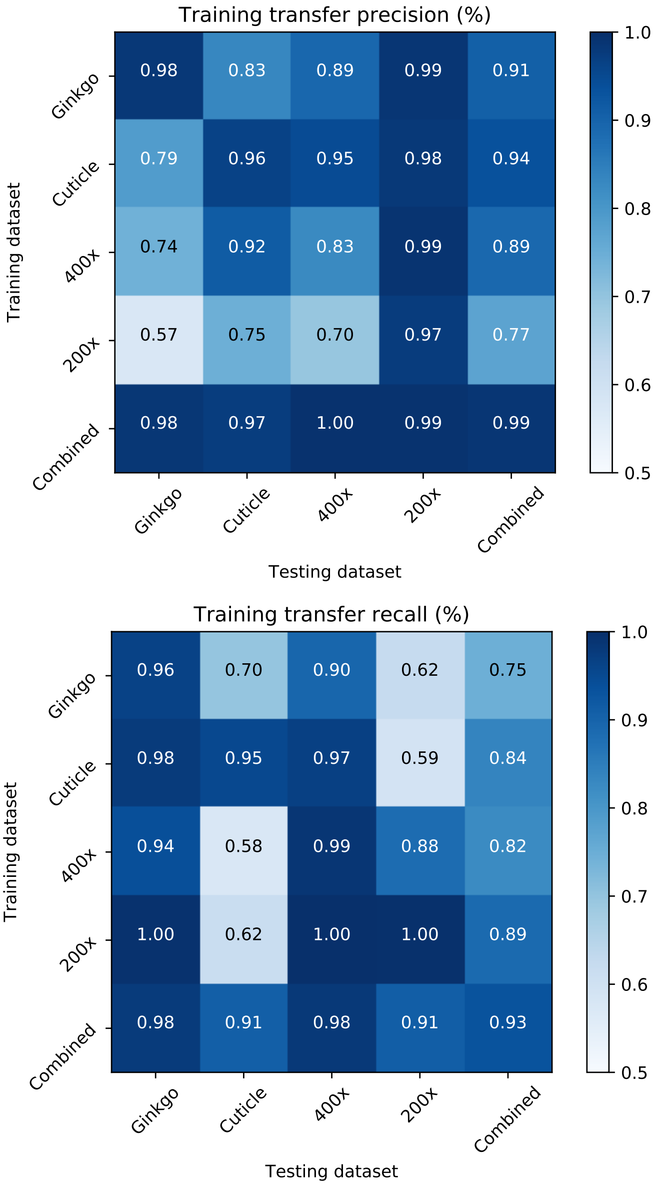
Ubbens JR and Stavness I. 2017. Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. In: *Frontiers in Plant Science* 8.

Vialet-Chabrand S and Brendel O. 2014. Automatic measurement of stomatal density from microphotographs. In: *Trees - Structure and Function* 28.6, pp. 1859–1865.

Wang R, Yu G, He N, Wang Q, Zhao N, and Xu Z. 2015. Latitudinal variation of leaf stomatal traits from species to community level in forests : linkage with ecosystem productivity. In: *Nature Scientific Reports* 5.14454, pp. 1–11.

Yu F and Koltun V. 2015. Multi-Scale Context Aggregation by Dilated Convolutions. In: arXiv: 1511.07122.

•••



Training transfer precision (%)

Training transfer recall (%)

SUPPLEMENT 1. Precision and Recall for models trained on different training datasets (vertical), tested on different test datasets. *Combined* is a union of all training and test sets.