Deep learning models for semantic segmentation and automatic estimation of severity of foliar symptoms caused by diseases or pests

Juliano de Paula Gonçalves^a, Francisco de Assis de Carvalho Pinto^a, Daniel Marçal de Queiroz^a, Flora Maria de Melo Villar^a, Jayme G.A. Barbedo^b, Emerson M. Del Ponte^{c*}

¹ Departamento de Engenharia Agrícola, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil

² Embrapa Informática Agropecuária, 13083-886, Campinas, SP, Brazil

³ Departamento de Fitopatologia, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil

*Corresponding Author: Emerson M. Del Ponte, e-mail: delponte@ufv.br

ABSTRACT

Measures of percent severity of visible symptoms or injuries caused by diseases or insect pests on plant organs are essential in plant health research. Current color thresholding digital imaging-methods are generally more accurate and reliable than visual estimates. However, these methods perform poorly when scene illumination and background are not uniform, conditions that can be overcome by convolutional neural networks (CNN) for semantic segmentation. In this study, we trained five CNN models for pixel level predictions in images of individual leaves exhibiting necrotic lesions and/or yellowing caused by the insect pest coffee leaf miner (CLM), and two fungal diseases: soybean rust (SBR) and wheat tan spot (WTS). Training was performed in 80% of images annotated for three classes: leaf background (B), healthy leaf (H) and injured leaf (I). Precision, recall, and Intersection over Union (IoU) metrics in the test image set were highest for B, followed by H and I classes, irrespective of the model. When the pixel-level predictions were used to estimate percent severity, Feature Pyramid Network (FPN), Unet and DeepLabv3+ (Xception) performed the best: concordance coefficients were greater than 0.95, 0.96 and 0.98 for CLM, SBR and WTS datasets, respectively. The other three models tended to misclassify healthy pixels as injured, leading to overestimation of percent severity. The accuracy of the predictions by CNN models were comparable with those obtained using a standard commercial software which requires manual adjustments that slows the process.

Keywords: Phytopathometry; Deep learning; Neural network; Image processing; Image segmentation

1. Introduction

Agricultural, forestry and ornamental productions are constantly threatened by pests or pathogens that attack the plants and produce externally visible symptoms (Oerke 2006; Savary et al. 2017). The ability to accurately detect and/or quantify symptomatic area is key for both research and management (Bock et al. 2020). Severity, or the percent symptomatic area, is an essential variable for evaluating treatment effects as well as predicting crop losses (Madden et al. 2007). However, severity is inherently more difficult to obtain than incidence that is based on the presence/absence at the sample unit. Depending on the method, inaccuracy is not uncommon at magnitudes that can lead to poor management decisions (Bock et al. 2016; Bock et al. 2020; Chiang et al. 2016). Therefore, research to develop new methods or improve the existing ones to make them more accurate, reliable and time- and cost-effective are essential (Bock et al. 2016; Franceschi et al. 2020).

Severity is most commonly estimated visually with scores assigned following nominal, ordinal or ratio (percent) scales (Bock et al. 2016; Madden et al. 2017). Given the heterogeneity of the symptoms across pathosystems and subjectivity of the visual estimates, discrepancies among human raters are common and can be minimized via pre-training or use of visual aids during the assessment (Nita et al. 2003; Madden et al. 2007; Bardsley and Ngugi 2013). Standard area diagrams (SADs), a series of pictorial representations of diseased images, each representing a reference severity, generally improve the accuracy of the estimates (Del Ponte et al. 2017). The integration of SADs into smartphone or tablet applications (Pethybridge and Nelson 2015) has opened possibilities of interactive use and improved accuracy (Del Ponte et al. 2019). Although enhancements can be made to disease

scales and SADs, visual assessments are dependent on person's ability and experience - skills that cannot be "ported" to another person.

The use of image analysis as a potential substitute for visual estimations has been envisioned several decades ago (Lindow and Webb 1983), but the first commercial software applications were launched during the mid 2000s (Lamari 2002; Vale et al. 2003). These tools have been developed based on color thresholding images after their acquisition and preparation that demand multiple platform and software. Recently, a mobile app called *Leaf Doctor* allowed to greatly speed up assessments given that image acquisition (smartphone or tablet camera) and manual color-thresholding are conducted in the app using the touch screen technology, but still depend on the rater's ability to segment the image (Pethybridge and Nelson 2015). Semi-automatic or automatic thresholding methods have been proposed and validated for batch-processing images of plant organs affected by disease symptoms (Bock et al. 2009; Barbedo 2016a; Lamari 2002), but their performance is impaired if the images are acquired under less uniform conditions. Complex backgrounds such as those in images obtained in the field, either attached or detached from the plant, have also been an issue. In-field or pre-image processing solutions (black cloth, light box) have been proposed to minimize these issues but at the cost of speed and scalability (Barbedo 2016b; Karisto et al. 2018; Stewart et al. 2016).

Deep learning algorithms, such as convolutional neural networks (CNNs), have emerged in recent years as promising method for image pattern recognition including plant disease (Ferentinos 2018; Barbedo 2019; Ramcharan 2019). However, the great majority of the applications of CNNs target disease identification or classification of pests and diseases (Barbedo 2018; Bock et al. 2020). Thus far, some studies targeting severity estimation have

used CNN models to predict categories or severity intervals, not the percent values (Esgario et al. 2020; Liang et al. 2019; Wang et al. 2017). The aggregation of percent severity into a few (2 to 4) categories may not be useful in field experiments to discriminate among treatments such as fungicides aimed at reducing disease severity. For such situations, percent severity is more informative and is usually provided by visual assessments (Bock et al 2016). Two studies used CNN models to estimate percent severity of foliar symptoms using images obtained under controlled environment (Lin et al. 2019) or field experiments (Tusubira et al. 2020), both making use of semantic segmentation approaches. Also known as pixel-level classification, semantic segmentation models have been tested in a wide range of classification tasks using images of medical (Borne et al. 2020; Y. Zhang et al. 2020) or remote sensing (Guo et al. 2019; Krestenitis et al. 2019) applications, with superior performance compared with the traditional deep learning methods. In agriculture, accuracy and speed of classification were improved using semantic models in land mapping (Du et al. 2019; Yang et al. 2019) and plant segmentation (Xu et al. 2020) studies.

The objective of this study was to evaluate the performance of CNN architectures to segment images of plant leaves into three classes (background, healthy leaf and injured leaf) for further calculation of severity and comparison with measures provided by a commercial software used to manually threshold the image based on color transformations.

98

99 2. Materials and Methods

100 2.1. Image dataset

101 Three image datasets, each representing a distinct external symptom or injury caused by a
102 disease or an insect pest on leaves of three major agricultural crops, were used in this work
103 and obtained at different conditions (Table 1).

104 Coffee leaves damaged by an insect pest, the coffee leaf miner (*Leucoptera coffeella*), were
105 obtained during the conduction of this study at coffee plantations in Viçosa, MG, Brazil.
106 Images of detached coffee leaves exhibiting different levels of severity were taken in the field
107 using a Motorola One XT1941-3 smartphone (Motorola Inc., Chicago, IL, USA) at the
108 resolution of 13 MP. The smartphone was positioned horizontally at a 20 cm distance from
109 each leaf and it was exposed to a wide range condition of illumination and background (Fig
110 1A).

111 Images of the soybean leaflets were obtained from a public repository and had been
112 used in previous study on the development of a standard area diagram for aiding visual
113 estimates of severity (Franceschi et al. 2020). Soybean leaflets displaying a range of severity
114 were detached from plants in the field or from potted plants grown in the greenhouse. In the
115 laboratory, the images were obtained using a Hewlett Packard scanner (Model 2130) at a
116 resolution of 300 dpi. A blue sheet paper was used as image background.

117 Wheat leaves exhibiting the typical tan spot symptoms were obtained from plants of
118 several wheat cultivars grown in the greenhouse and inoculated for the evaluation of wheat
119 genotypes for tan spot resistance. The leaves were placed on top of a black background and
120 leaf sections were photographed using a Nikon D-3000 camera at a resolution of 5.6 MP
121 (Table 1).

2.2. Image annotation

The image annotation was performed manually aided by a custom-made algorithm developed for this work in Python programming language (Gonçalves *unpublished*). Automatic procedures in available plant disease-specific software for severity estimation (Lamari 2002; Vale et al. 2004) could not be used due to difficulties when dealing with complex backgrounds or non-uniform illumination.

Table 1. Information for the image datasets used for training and testing of semantic segmentation models for pixel-level predictions to estimate the severity of external visual symptoms caused by plant diseases and pests

Host species name	Disease or pest name	N. of images	Brazilian state	Geographic coordinates	URL source ¹
<i>Coffea arabica</i>	Coffee leaf miner	406	Minas Gerais	20°46'11"S 42°52'13"W	https://osf.io/ygq82/
<i>Glycine max</i>	Soybean rust	208	Paraná	23°11'30"S 51°10'57"W	https://osf.io/4hbr6/
<i>Triticum aestivum</i>	Wheat tan spot	152	Rio Grande do Sul	28°14'35"S 52°23'32"W	https://osf.io/52cjm

¹ The original and processed images used in this study were made publicly available at Open Science Framework (OSF)

The tool allowed the operator to select, on each image, the most appropriate threshold for discriminating among three classes: background (B), healthy leaf (H) and injured leaf (I) areas. As a result, a three-class mask was generated for each image. The severity was then given by equation (1).

$$Severity = \frac{I}{I+H} \times 100. \quad (1)$$

Figure 1. A. Example images of foliar symptoms caused by coffee leaf miner (A, B) soybean rust (C, D) and wheat tan spot (E,F). The images were individually annotated for three classes of pixels in each image: background (blue), leaf (green) and necrotic symptom (red). The severity shown in the mask image was given by the percent leaf area affected calculated as $\text{red area} / (\text{red area} + \text{green area}) * 100$. B. Distribution of the reference severity on the sample of annotated leaves of coffee (G, $n = 406$), soybean (H, $n = 208$), and wheat (I, $n = 152$).

2.2. Experiment setup

Semantic segmentation algorithms were chosen after searching the literature and based on relevance, availability and reproducibility criteria. Five semantic models were chosen, from oldest to newest publication date: Unet (Ronneberger et al. 2015), SegNet (Badrinarayanan et al. 2017), PSPNet (Zhao et al. 2017), FPN (Lin et al. 2017), and two versions of DeepLabv3+

(Chen et al. 2018). An 80-20 random training-testing split was applied in the original image dataset composed of 612 images. Models were trained with 490 images (all datasets combined) using the free version of Google Colaboratory environment (<http://colab.research.google.com>). The experiments were performed in a computer equipped with an Intel Core i3-4150 CPU (3.50 GHz) and 4 GB RAM.

The algorithms cross-entropy loss function, combinatorial optimization and simulation of rare events were used for estimating the deviation of the predictions from the reference values at the pixel level during training (de Boer et al. 2005). The original loss function was the sum of the cross-entropy between predicted and reference values. Preliminary runs showed that the segmentation models tended to learn healthy and background classes more accurately than the injured class because of the lower number of pixels available for the latter. This unbalance tends to lead to false negatives when the positive cases are few. Hence, although the precision of the predictions was high, recall was low for that class. To minimize this issue, a weighting factor γ was introduced to increase the sensitivity of the model to the injured pixels:

$$L(g_{ij}, \gamma_j, p_{ij}) = - \sum_{i=0}^n \sum_{j=0}^m \gamma_j g_{ij} \log p_{ij} , \quad (2)$$

where g_{ij} denotes the ground truth in the i -th instance (pixel) pertaining to class j , γ_j is the weighting factor pertaining to class j , and p_{ij} represents the probability associated to the i -th

instance of class j , obtained with the softmax function. The defined values for γ_j were 3 for the class I and 1 for the classes B and H . The model weights were updated using Stochastic Gradient Descent (SGD).

The training process of Unet, PSPNet e FPN models was initialized using the Inception-V3 (Szegedy et al. 2016) backbone pre-trained with the ImageNet dataset (Russakovsky et al. 2015). DeepLabv3+ was initialized with the Xception (Chollet 2017) e Mobilenetv2 (Sandler et al. 2018) backbones, pre-trained on the Pascal VOC 2012 dataset (Everingham et al. 2015). All images were resampled prior to training to meet the standard dimensions used in the models (768×480 pixels).

A five-fold cross-validation strategy was adopted to avoid biased results. After a few preliminary runs, the following hyperparameters were adopted: learning rate, 0.045; momentum, 0.9; batch size, 2. All semantic segmentation models converged in less than 10 epochs. To avoid overfitting and increase model's generalization capabilities, the training set was augmented using rotation (90° , 180° , 270°), vertical and horizontal mirroring, and random brightness adjustments, as recommended by Liu et al. (2018). In addition, the Keras (<https://keras.io/>) function *ImageDataGenerator* was used to introduce random rotations (between 0 and 38°), shifts (between 0 and 20% of the image's height and width) and zoom (0 to 20%).

2.3. Pixel level classification performance

Three metrics were used to evaluate the performance of the image segmentation models (Sokolova and Lapalme 2009) all based on the number of true positives (TP: Intersection over

Union, IoU, of the prediction target mask pair exceeds the predefined threshold), false positives (FP: predicted object mask had no associated ground truth object mask) and false negative (FN: ground truth object mask had no associated predicted object mask). Those values were calculated for each class and then averaged.

Intersection over Union (IoU), also known as Jaccard index, was used to quantify the percent overlap between the target mask and the model prediction output. In essence, IoU measures the number of pixels common between the target and prediction masks divided by the total number of pixels present at both masks and are given by $TP/(TP+FP+FN)$. Precision indicates how many predicted objects actually match ground truth annotation and is given by $TP/(TP+FP)$. Recall indicates how many objected annotated in the ground truth were captured as positive predictions and is given by $TP/(TP+FN)$.

2.4. Model-estimated and software-measured severity

We were interested in evaluating how accurate, or close to the reference values (Bock et al. 2016), were the estimates of severity by the semantic models. The reference severity values for the test images were annotated manually using the same criteria of the training set. In addition, we used the commercial software APS (*American Phytopathological Society*) Assess v2.0 (Lamari 2002) to measure severity on each of the test images. APS Assess is one of the most used software in the field of plant pathology for performing several kinds of measures on leaf and crop images, mainly for measuring percent area affected after manually adjusting color thresholding (Del Ponte et al. 2017). Because APS Assess is optimized to segment only lesions and leaves, the background should be of a single color, usually blue. Hence, the background was manually removed from each coffee leaf image before measuring severity,

while the black and blue backgrounds were preserved in the wheat and soybean image sets, respectively.

2.4.1. Accuracy and precision metrics

The quality of the predictions of severity by the semantic models were evaluated based on the Lin's Concordance Correlation Coefficient (CCC) analysis (Lin 1989). This is the most recommended method for evaluating the overall agreement between visual estimates and software measured severity in phytopathometric studies (Bock et al. 2016; Madden et al. 2007). Lin's CCC allows to evaluate the degree to which the estimates fell on the line of concordance (45° , where slope = 1, intercept = 0). The overall agreement (ρ_c) is a product of precision (Pearson's r) and overall bias (C_b), calculated from systematic bias (v) and constant bias (μ). Deviation of ρ_c , r , C_b and v from 1 and μ from zero indicates loss of precision and loss of agreement. Analyses of the LCC statistics were conducted in the R statistical environment using the *epi.ccc* function of the *irr* package (Gamer et al. 2012). In addition, the root Mean Square Error (RMSE) was calculated.

2.4.2. Running time

Running times were compared among the CNN model and included not only model inference, but also the time spent reading and resampling the images and generating the segmentation mask. These were measured using both the Google Colab GPU and the local CPU. CPU times were obtained by averaging the time of 10 independent runs. The APS Assess running times were obtained by summing up the time required for the user to load

the images into the software and select the segmentation parameters. This process was conducted three times and the average time was calculated.

4. Results

4.1 Pixel-level classification performance

The values of precision, recall and IoU of the predictions were highest for the background (B) class, followed by healthy leaf (H) and injured (I) classes, irrespective of the model. For B, all three metrics were consistently close to 98%. Precision and recall were much closer to each other after the introduction of the γ parameter. In general, they differed by four percent points and precision was consistently higher than recall for all but the PSNet model (Figure 2).

The global average of the means of recall and precision, across models, was 98.9% for B, 95.6% for H and 85.2% for I class. When averaging these metrics across the three classes, precision values ranged from 90.4% to 95.6% and recall values ranged from 89.4% to 94.7%. The highest and the lowest average values of the metrics, for both metrics, were obtained for FPN and Deeplabv3+(Mobilenetv2), respectively.

The IoU values for the B class were similarly high (>98%) across the models, with the exception of DeeplabV3+M (94.4%). Differences in IoU among the models became more apparent for the H and I classes, which ranged from 87.5% to 94.1% and 64.7% to 79.4%, respectively (Figure 2). Again, the lowest and highest IoU values were obtained using Deeplabv3+(Mobilenetv2) and FPN, respectively.

Figure 2. Metrics for evaluation of the performance of convolutional network models trained to classify pixels of an image into three predefined classes on the test set of 154 leaf images exhibiting symptoms caused by a fungal disease or an insect pest: background, healthy leaf and injured leaf.

Differences in model performance could be appraised visually when randomly selecting a few images for comparison. Figure 3 shows one soybean and three coffee leaf images, the latter with a complex background. The observation of the masks generated by Unet, FPN and Deeplabv3+ (Xception backbone), as suggested by the metrics, show that these three models produced masks more similar to the reference mask than PSPNet, SegNet and Deeplabv3+ (Mobilenetv2 backbone). For the latter models, a portion of the leaf that was misclassified as background instead of healthy leaf can be noticed.

Figure 3. Example masks for background (blue), healthy leaf tissue (green) and injured leaf tissue (red) predicted by semantic models on test images. (a) Original image; (b) Reference; (c) Unet; (d) SegNet; (e) PSPNet; (f) FPN; (g) Deeplabv3+ com o backbone Xception; e (h) Deeplabv3+ com o backbone Mobilenetv2.

4.2. Severity estimates from pixel-level CNN predictions

The use of the pixel-level predictions to calculate percent severity, separately for each of the three sets of test images, allowed to better understand source of errors of classification specific to crop-injury datasets. As expected, given the pixel-level metrics, all models were capable of providing good quality estimates of severity for most of the model and crop image set combinations. Differences in the performance of the semantic models were noticed both among crop datasets and models within a dataset, which are explained next separately for each dataset.

289

290 4.2.1. Coffee leaf miner

291 The overall agreement coefficients (ρ_c) were generally higher than 0.94 across the models;
 292 ranged from 0.82 (PSPNet) to 0.98 (Unet and FPN). In all models, precision (r) was lower
 293 than generalized bias (C_b), which led to a reduction of its product, ρ_c . These was the case of
 294 predictions by the Segnet and DeepLabv3+ (Mobilenet V2) model where the presence of 6 to
 295 10 relatively larger positive or negative errors caused a decrease in precision and accuracy
 296 (Figure 4). There was a tendency to overestimate severity in most leave, especially for the
 297 PSPNet model that performed the poorest. The overestimation of severity was suggested by
 298 scale-shift (v) values greater than 1.0: 1.27 for PSPNet and 1.08 for DeepLabv3+ (Mobilenet
 299 V2). These models tended to misclassify healthy tissue as symptomatic and showed the
 300 largest RMSE values. Finally, the location-shift values above zero (0.2 and 0.35) for the
 301 predictions of these models are suggestive of the presence of biases that affected the
 302 performance of these models (Figure 4).

303 4.2.2. Soybean rust

304 The overall agreement values were generally high and relatively closer to each other, ranging
 305 from 0.93 to 0.97. Again, Unet and FPN models showed predictions in strongest agreement
 306 to the reference values. Similarly to the coffee images, the precision values (r) were generally
 307 lower than generalized bias (C_b) for all models (Figure 5). The location-shift values (μ) were
 308 generally closer to zero, suggesting no constant bias for all but PSPNet and DeeplabV3+
 309 (Mobilenet V2).

Figure 4. Scatter plot of the best fitted first-order regression line (dashed line) for the relationship between coffee leaf miner severity calculated from pixel-level predictions by semantic segmentation models and a reference severity determined on the same leaves by manually annotating each pixel of the image using a custom-made algorithm. The solid-line represents the perfect concordance line. The statistics represent the Lin's concordance correlation coefficient (ρ_c) and respective 95% confidence intervals, which are given by the product of precision (r) and bias coefficient (C_b). The latter is calculated from measures of location-shift (μ) and scale-shift (v). RMSE is the root-mean square error between the two series.

Together with SegNet, those models showed the presence of systematic biases because of the larger values of scale-shifts ($v > 1$), which indicates that severity is increasingly overestimated as the severity increased in the sample. In fact, these were noticed for most models for which the set of 10 images with severity greater than 50% had their severity predicted consistently above the reference value, with very few exceptions. These models showed the lowest precision ($r < 0.91$) and highest RMSE (> 8) values (Figure 5).

Figure 5. Scatter plot and the best fitted first-order regression line (dashed line) for the relationship between soybean rust severity calculated from pixel-level predictions by semantic segmentation models and a reference severity determined on the same leaves by manually annotating each pixel of the image using a custom-made algorithm. The solid-line represents the perfect concordance line. The statistics represent the Lin's concordance correlation coefficient (ρ_c) and respective 95% confidence intervals, which are given by the product of precision (r) and bias coefficient (C_b). The latter is calculated from measures of location-shift (μ) and scale-shift (u). RMSE is the root-mean square error between the two series.

4.2.2. Wheat tan spot

The overall agreement values were, on average, relatively stronger than those calculated for the coffee and soybean test images. For only one model (DeepLabV3+ Mobilenet V2), the concordance coefficient was < 0.98 . Contrary to the other two crop images, the model with the poorest performance tended to underestimate severity for the large majority of the images, which was evidenced by the negative location-shift and scale-shift values greater than 1.0, suggesting both constant and systematic biases. For the other models, there was a

slight tendency to more frequently underestimate the disease, but with generally small errors and good precision.

Figure 6. Scatter plot the best fitted first-order regression line (dashed line) for the relationship between wheat tan spot severity calculated from pixel-level predictions by semantic segmentation models and a reference severity determined on the same leaves by manually annotating each pixel of the image using a custom-made algorithm. The solid-line represents the perfect concordance line. The statistics represent the Lin's concordance correlation coefficient (ρ_c) and respective 95% confidence intervals, which are given by the product of precision (r) and bias coefficient (C_b). The latter is calculated from measures of location-shift (μ) and scale-shift (v). RMSE is the root-mean square error between the two series.

4.3. Severity measures using APS Assess

The quality of the severity measures obtained using the manually adjusted color thresholding method in APS Assess was considered very good; the overall agreement was greater or equal to 0.94 for the three sets of crop images with small differences (from 1 to 3

percentage points) among them: wheat tan spot > soybean rust > coffee leaf miner (Figure 7A-C). Similar to the CNN model-predicted severity, precision (r) was generally lower than generalized bias coefficient (C_b). There were only four coffee leaves that had severity overestimated with errors larger than four percent points. The location-shift value larger than zero suggested a constant bias due to more frequent overestimation across the entire range of severity (maximum was 16%) (Figure 7A). Similar results were found for soybean rust for which relatively lower agreement was due to overestimation due to constant bias (Figure 7B). Finally, similar to some of the model-predicted severity, wheat tan spot was more consistently underestimated with low precision. The presence of bias, however, did not affect the overall agreement, which was considered very strong ($\rho_c = 0.97$) (Figure 7C).

4.4. Confidence interval-based inferences

The appraisal of 95% confidence intervals for the concordance coefficients suggest that all six models produced predicted severity values that were generally comparable with measures obtained using APS *Assess*. However, in a few cases, such as the coffee leaf miner, the models were superior to *Assess*. Three models, namely FPN, Unet and DeepLabv3+ (Exception) were superior to the other CNN models as well.

These other three models produced predictions of severity that were more similar to APS *Assess*, with the exception of PSPNet that poorly predicted the injured pixels in coffee leaf miner affected images, as demonstrated in the previous section. The relatively higher mean values of overall agreement and overlapping of confidence intervals suggest that the other models can be used to predict severity for soybean rust and wheat tan spot.

Nevertheless, most accurate estimates, or closer to the reference values, were achieved when using the same three models that best-predicted coffee leaf miner severity: FPN, Unet and DeepLabv3+ (Exception) (Figure 7D).

4.3. Inference Time Evaluation

The mean time to process the images and predict severity on each image of the testing dataset is shown in Table 2. Since the models were trained in similar conditions such as same image size, optimizer, learning rate and batch size, the number of model parameters explained the variation in the average time. Those with lower number of parameters such as PSPNet, SegNet and o Deeplabv3+ (Mobilenetv2) used lower time per image to estimate severity. The CNN model that performed poorest was Deeplabv3+ (backbone Xception), but it was 20x faster than an operator using APS Assess (Table 2).

Table 2. Average time for estimating the percent severity, or leaf area affected by fungal diseases in soybean (30 leaves) and wheat leaves (41 leaves) and an insect damage on coffee leaves (81 leaves) using semantic segmentation models and a commercial software that uses a color-thresholding method. The model-based severity estimates are based on pixel-based predictions of three-classes (background, healthy leaf and damaged class).

Method	N. of parameters (millions)	Inference times	
		GPU (s)	CPU (s)
Unet	29.9	1.94	4.79
SegNet	7.8	1.75	3.37
PSPNet	2.8	1.87	3.14
FPN	25.0	1.94	4.94
Deeplabv3+ (Xception)	41.3	1.94	5.14
Deeplabv3+ (Mobilenetv2)	2.1	1.75	3.03
APS Assess	-	-	105.09

Figure 7. Scatter plot for the relationship between severity determined using Assess software and a reference severity determined manually by annotating each pixel of the image aided by a custom-made algorithm for three problems: coffee leaf miner damage (A), soybean rust (B) and wheat tan spot (C). The statistics represent the Lin's concordance correlation coefficient (ρ_c) and respective 95% confidence intervals, which are given by the product of precision (r) and bias coefficient (C_b). The latter is calculated from measures of location-shift (μ) and scale-shift (ν). RMSE is the root-mean square error between the two series. Note the different scale for coffee leaf

415

416 **5. Discussion**

417 We successfully trained and compared the performance of CNN models for semantic
418 segmentation of images of disease leaves, which were capable of accurately estimating
419 percent severity of various kinds of symptoms and leaf characteristics. The performance of
420 the CNN models was variable among the various architectures, but three performed the best:
421 FPN, Unet and Deeplabv3+ with Xception backbone.

422 The FPN architecture employs a top-down approach with a horizontally connected
423 structure that combines strong and weak semantic resources for segmentation of both large
424 and small targets. The Unet employs a fully convolutional architecture with two stages,
425 contraction and expansion. In the contraction section, the output of the feature map becomes
426 smaller as the network deepens, leading to the loss of information linked to small targets.
427 Thus, the slight superiority of FPN over Unet is probably due to variations in the size of the
428 foliar lesions. Deeplabv3+ with Xception backbone showed the best performance over the
429 PASCAL VOC 2012 dataset, but the accuracy of FPN was higher. In the former, atrous
430 convolutional layers are particularly suitable for capturing contextual information, but much
431 of the edge information is lost, reducing its ability to detect small objects (Guo et al. 2019; S.
432 Zhang et al. 2020). In addition, Deeplabv3+ with Xception backbone contains the largest
433 number of parameters in this study, which require higher number of samples required for
434 proper training (Lobo Torres et al. 2020). It is thus possible that the number of training
435 samples was not sufficient for this model to reach its full potential, which deserves further
436 investigation.

The models that performed the poorest, PSPNet, SegNet and Deeplabv3+ with Mobilenetv2 *backbone*, employed a relatively small number of parameters. Our dataset can be considered challenging for the task due to variations in texture, shape, color and scale, which requires larger numbers of parameters to be properly captured. In addition, the SegNet architecture does not utilize the VGG16's pre-trained weights for initialization which are particularly important for small datasets (Mohanty et al. 2016; Wang et al. 2017).

The CNN models showed potential to overcome the limitations of traditional methods that use color transformation methods, including mathematical morphology operations, thresholding and filtering for predicting severity of plant diseases (Barbedo 2016b; Bock et al. 2020). These methods require pre-processing steps for removal and standardization of background (non-target) elements in the image or enhancing contrast. Barbedo (2016b) discusses solutions that help to minimize these issues, but their use tend to slow down the process compared with a well-trained human rater.

The automatic thresholding, by using fixed values of the HIS color space, available in APS Assess has been validated using different dataset of citrus canker leaves (Bock et al. 2009). The authors found that the accuracy was inconsistent across the image sets likely due to variations in a) characteristics of leaves in different datasets; b) photographic conditions such as lighting, surface reflection, saturation, hue or c) leaf characteristics such as symptom or healthy leaf area color that varies according to leaf age, symptom age and progress (Bock et al. 2009). It is no surprise that, although available for longer time, the use of color-thresholding methods has found use mainly in small-scale epidemiology-oriented research and phenotyping under controlled conditions (Schwanck and Del Ponte 2014; Stewart et al. 2016) and design of SADs (Del Ponte et al. 2017). For field research, visual estimates are largely dominant in spite of the known problems related with subjectivity and inaccuracy of

the estimates, which can be minimized via training or visual aids (Bock et al. 2016). An automated color-thresholding method has been used successfully adjusted for batch processing large number of images for specific contexts, but standard image acquisition under uniform conditions was required such as that provided by a flatbed scanner (Karisto et al. 2018).

Among the limitations that slow down the development of more accurate automated systems for severity estimation, good quality and properly annotated image datasets for training the models are not available. Conversely, a vast amount of images captured under a wide variety of conditions have been made available in public databases in the recent years. In those, images are annotated disease or injury name or type and such datasets have been of great value to support the development of disease/pest classification problems (Barbedo et al. 2018; Hughes and Salathe 2016; Krohling 2019; Tian 2020). The manual annotation by grouping pixels according to different classes that include background is more time consuming.

The availability of good quality image datasets of plant organs displaying a wide range of severity and their respective masks may be greatly useful for further improvement of the models combining them with new datasets. The images used (and made available for further use) in this study cover a wide variety of challenging conditions that were explicitly sought during imaging to make the dataset as comprehensive as possible. In particular, three features that are known to affect the performance of segmentation algorithms: complex background, specular reflection and the simultaneous presence of shadow and direct lighting, according to Barbedo (2016b) were present in most of the images used in our study.

The severity estimates by CNN models were further compared with APS Assess, which required manually background removal, for the case of coffee leaf images. Previous

work showed that manual adjustments tend to produce more accurate measures than semi-automatic and automatic color transformation-based algorithms, given the already discussed limitations of the latter approach (Bock et al. 2009; Barbedo 2016b). The CNN models outperformed APS Assess without the need to manually remove the background depending on the dataset. This can be partially explained by the fact that CNN models are capable of extracting and learning context information by successive convolution and, as a result, can segment images in a variety of conditions. Moreover, the processing time may be a problem in situations requiring a large number of observations. When using APS Assess or Leaf Doctor for manually thresholding the image the user takes from 30 to 50 seconds to recolor and/or segment appropriately (Bock et al. 2009; Barbedo 2016a; Pethybridge and Nelson 2015). In fact, CNN models have shown to be faster than manual approaches but its speed is usually affected by the number of parameters as we found in this study; the greater the number the slower the inference time (Lobo Torres et al. 2020; Wang et al. 2019).

The CNN-based algorithms developed in this study do not require manual adjustments or previous user's experience. They learned features of symptomatic tissues of various types as well as healthy (green) tissue under a wide range of conditions, which expands its use beyond the controlled environment. Even for controlled conditions, depending on experimental-specific variations, color-thresholding methods in commercial or custom-made software may require adjustments. In the case of CNN, new images can be used to augment the database and further train the models. For example, the wheat leaves were obtained from plants grown in the greenhouse during experiments for genotype selection. The use of an automated method that provides accurate severity measures could potentially reduce costs in training of new personnel for visual assessment, with the advantage of establishing even larger datasets for further use. In fact, the potential use for

field images is the most innovative approach that could potentially substitute visual assessments in field research. For example, uniform field trials such as those for multi-site and multi-year evaluation of treatments (fungicides, for example) (Dalla Lana et al. 2018; Machado et al. 2017), could greatly benefit from measuring severity with the proposed framework. A more accurate and reliable method requires a lower sample size as shown by a simulation study on soybean rust severity estimated visually or using image analysis (Franceschi et al., 2020). In the field, while automation of sampling is not possible to capture good quality images, raters could focus efforts good sampling and use image or video acquisition platforms. In general, instantaneous measures of severity in the field are not critical for evaluating treatment effects, and so the images can be stored for further analysis using a CNN model.

The field images obtained in this study, although more complex than those obtained under controlled conditions, were captured after the leaves were detached from the host plant, which can be undesired for some situations. Future work should consider the use of images of non-detached leaves as well as those exhibiting multiple types (causes) of symptoms, obtained using portable imaging or computer vision systems. Their implementation could be combined with object detection algorithms employing bounding boxes like YOLO (You Only Look Once) (Redmon et al. 2016).

527

528 Declaration of Competing Interest

529 The authors declare that they have no known competing financial interests or personal
530 relationships that could have appeared to influence the work reported in this paper.

531

532

533 Acknowledgments

534 The first author thanks the National Council for Scientific and Technological Development
535 (CNPq)/Brazil for a scholarship. This study was financed in part by the Coordenação de
536 Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001.

537

538 Data availability

539 The data and scripts for performing the analysis and produce the graphs are publicly
540 available as research compendium at <https://osf.io/hx2um/>. The data are stored as text files
541 and the R codes were prepared as RMarkdown documents. Webpages were produced to
542 facilitate navigating through the commented scripts and outcomes.

543

544 Authors' contribution

545 JPG conceptualized the work, conducted experiments, analyzed the data and wrote the
546 manuscript. FACP, DMQ and FMMV discussed the experimental design and revised the
547 manuscript. EMD discussed the experiments, performed analysis and wrote the manuscript.
548 JGAB discussed the approaches and revised the manuscript.

549

550 **References**

- 551 Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-
552 decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39:
553 2481–2495.
- 554 Barbedo, J.G.A., 2018. Factors influencing the use of deep learning for plant disease
555 recognition. *Biosyst. Eng.* 172:84–91.
- 556 Barbedo, J.G.A., 2016a. A novel algorithm for semi-automatic segmentation of plant leaf
557 disease symptoms using digital image processing. *Trop. Plant Pathol.* 41:210–224.
- 558 Barbedo, J.G.A., 2016b. A review on the main challenges in automatic plant disease
559 identification based on visible range images. *Biosyst. Eng.* 144:52–60.
- 560 Barbedo, J.G.A., Vieira Koenigkan, L., Almeida Halfeld-Vieira, B., Veras Costa, R., Lima
561 Nechet, K., Vieira Godoy, C., Lobo Junior, M., Rodrigues Alves Patricio, F., Talamini, V.,
562 Gonzaga Chitarra, L., Alves Santos Oliveira, S., Nakasone Ishida, A.K., Cunha Fernandes,
563 J.M., Teixeira Santos, T., Rossi Cavalcanti, F., Terao, D., Angelotti, F., 2018. Annotated plant
564 pathology databases for image-based detection and recognition of diseases. *IEEE Lat. Am.*
565 *Trans.* 16:1749–1757.
- 566 Barbedo, J. G. A., 2019. Plant disease identification from individual lesions and spots using deep
567 learning. *Biosyst. Eng.* 180:96-107.
- 568 Bardsley, S. J., & Ngugi, H. K., 2013. Reliability and accuracy of visual methods to quantify severity
569 of foliar bacterial spot symptoms on peach and nectarine. *Plant Pathol.* 62:460-474.
- 570 Bock, C. H., Cook, A. Z., Parker, P. E., and Gottwald, T. R. ,2009. Automated image analysis
571 of the severity of foliar citrus canker symptoms. *Plant Dis.* 93:660-665.
- 572 Bock, C. H., Chiang, K., & Del Ponte, E. M., 2016. Accuracy of plant specimen disease
573 severity estimates: concepts, history, methods, ramifications and challenges for the future.
574 *CAB Rev. Perspect. Agric. Vet. Sci. Nutr. Nat. Resour.* 11:1-21.
- 575 Bock, C.H., Barbedo, J.G.A., Del Ponte, E.M., Bohnenkamp, D., Mahlein, A.-K., 2020. From
576 visual estimates to fully automated sensor-based measurements of plant disease severity:
577 status and challenges for improving accuracy. *Phytopathol. Res.* 2:9.

- 578 Borne, L., Rivière, D., Mancip, M., Mangin, J.F., 2020. Automatic labeling of cortical sulci
579 using patch- or CNN-based segmentation techniques combined with bottom-up geometric
580 constraints. *Med. Image Anal.* 62:101651.
- 581 Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with
582 atrous separable convolution for semantic image segmentation. Presented at the Proceedings
583 of the European Conference on Computer Vision (ECCV), pp. 801–818.
- 584 Chiang, K.S., Bock, C.H., Lee, I.H., El Jarroudi, M. and Delfosse, P., 2016. Plant disease
585 severity assessment—how rater bias, assessment method, and experimental design affect
586 hypothesis testing and resource use efficiency. *Phytopathology*, 106:1451–1464.
- 587 Chollet, F., 2017. Xception: Deep Learning with Depthwise Separable Convolutions, in: 2017
588 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017
589 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI,
590 pp. 1800–1807. <https://doi.org/10.1109/CVPR.2017.195>
- 591 Dalla Lana, F., Paul, P.A., Godoy, C.V., Utiamada, C.M., da Silva, L.H.C.P., Siqueri, F.V.,
592 Forcelini, C.A., Jaccoud-Filho, D. de S., Miguel-Wruck, D.S., Borges, E.P., Juliatti, F.C.,
593 Campos, H.D., Nunes, J., Carneiro, L.C., Canteri, M.G., Ito, M.F., Meyer, M.C., Martins, M.C.,
594 Balardin, R.S., Furlan, S.H., Carlin, V.J., Del Ponte, E.M., 2018. Meta-Analytic modeling of the
595 decline in performance of fungicides for managing soybean rust after a decade of use in
596 Brazil. *Plant Dis.* 102:807–817.
- 597 de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-
598 entropy method. *Ann. Oper. Res.* 134, 19–67.
- 599 Del Ponte, E.M., Nelson, S.C., Pethybridge, S.J., 2019. Evaluation of app-embedded disease
600 scales for aiding visual severity estimation of cercospora leaf spot of table beet. *Plant Dis.*
601 103:1347–1356.
- 602 Del Ponte, E.M., Pethybridge, S.J., Bock, C.H., Michereff, S.J., Machado, F.J., Spolti, P., 2017.
603 Standard area diagrams for aiding severity estimation: Scientometrics, pathosystems, and
604 methodological trends in the last 25 Years. *Phytopathology* 107:1161–1174.
- 605 Du, Z., Yang, J., Ou, C., Zhang, T., 2019. Smallholder crop area mapped with a semantic
606 segmentation deep learning method. *Remote Sens.* 11: 888.
- 607 Esgario, J.G.M., Krohling, R.A., Ventura, J.A., 2020. Deep learning for classification and
608 severity estimation of coffee leaf biotic stress. *Comput. Electron. Agric.* 169:105162.

- 609 Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2015.
610 The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* 111:98–136.
- 611 Ferentinos, K.P., 2018. Deep learning models for plant disease detection and diagnosis.
612 *Comput. Electron. Agric.* 145:311–318.
- 613 Franceschi, V.T., Alves, K.S., Mazaro, S.M., Godoy, C.V., Duarte, H.S.S., Del Ponte, E.M.,
614 2020. A new standard area diagram set for assessment of severity of soybean rust improves
615 accuracy of estimates and optimizes resource use. *Plant Pathol.* 69:495–505.
- 616 Gamer, M., Lemon J., Fellows, I., Singh, Puspendra, 2019. irr: Various coefficients of interrater
617 reliability and agreement. R package version 0.84.1. <https://CRAN.R-project.org/package=irr>
- 618 Guo, S., Jin, Q., Wang, H., Wang, X., Wang, Y., Xiang, S., 2019. Learnable gated convolutional
619 neural network for semantic segmentation in remote-sensing images. *Remote Sens.* 11:1922.
- 620 Hughes, D.P., Salathe, M., 2016. An open access repository of images on plant health to
621 enable the development of mobile disease diagnostics. ArXiv151108060 Cs.
622 <https://arxiv.org/abs/1511.08060>
- 623 Karisto, P., Hund, A., Yu, K., Anderegg, J., Walter, A., Mascher, F., McDonald, B.A.,
624 Mikaberidze, A., 2018. Ranking quantitative resistance to *Septoria tritici* blotch in elite wheat
625 cultivars using automated image analysis. *Phytopathology* 108:568–581.
- 626 Krestenitis, M., Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S., Kompatsiaris, I.,
627 2019. Oil spill identification from satellite images using deep neural networks. *Remote Sens.*
628 11:1762.
- 629 Krohling, R.A., 2019. BRACOL - A Brazilian Arabica coffee leaf images dataset to
630 identification and quantification of coffee diseases and pests. Mendeley Data, v1.
631 <http://dx.doi.org/10.17632/yy2k5y8mxg.1>
- 632 Lamari, L., 2002. ASSESS: Image analysis software for plant disease quantification. APS
633 Press.
- 634 Liang, Q., Xiang, S., Hu, Y., Coppola, G., Zhang, D., Sun, W., 2019. PD2SE-Net: Computer-
635 assisted plant disease diagnosis and severity estimation network. *Comput. Electron. Agric.*
636 157:518–529.

- 637 Lin, K., Gong, L., Huang, Y., Liu, C., Pan, J., 2019. Deep learning-based segmentation and
638 quantification of cucumber powdery mildew using convolutional neural network. *Front.*
639 *Plant Sci.* 10.
- 640 Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility.
641 *Biometrics* 45:255.
- 642 Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid
643 Networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern
644 Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern
645 Recognition (CVPR), pp. 936–944.
- 646 Lindow, S.E., Webb, R.R., 1983. Quantification of foliar plant disease symptoms by
647 microcomputer-digitized video image analysis. *Phytopathology* 73:520–524.
- 648 Lobo Torres, D., Queiroz Feitosa, R., Nigri Happ, P., Elena Cué La Rosa, L., Marcato Junior,
649 J., Martins, J., Olã Bressan, P., Gonçalves, W.N., Liesenberg, V., 2020. Applying fully
650 convolutional architectures for semantic segmentation of a single tree species in urban
651 environment on high resolution UAV optical imagery. *Sensors* 20:563.
- 652 Machado, F.J., Santana, F.M., Lau, D., Del Ponte, E.M., 2017. Quantitative review of the
653 effects of triazole and benzimidazole fungicides on *Fusarium* head blight and wheat yield in
654 Brazil. *Plant Dis.* 101:1633–1641.
- 655 Madden, L.V., Hughes, G., van den Bosch, F., 2007. The study of plant disease epidemics.
- 656 Mohanty, S.P., Hughes, D.P., Salathe, M., 2016. Using deep learning for image-based plant
657 disease detection. *Front. Plant Sci.* 7:1419
- 658 Nita, M., Ellis, M.A., Madden, L.V., 2003. Reliability and accuracy of visual estimation of
659 *Phomopsis* leaf blight of strawberry. *Phytopathology* 93:995–1005.
- 660 Oerke, E.-C., 2006. Crop losses to pests. *J. Agric. Sci.* 144:31–43.
- 661 Pethybridge, S.J., Nelson, S.C., 2015. Leaf Doctor: A new portable application for quantifying
662 plant disease severity. *Plant Dis.* 99:1310–1316.
- 663 Ramcharan, A., McCloskey, P., Baranowski, K., Mbilinyi, N., Mrisho, L., Ndalawwa, M., Legg, J.,
664 Hughes, D.P., 2019. A mobile-based deep learning model for cassava disease diagnosis. *Frontiers in*
665 *Plant Science* 10:272.

- 666 Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time
667 object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition
668 (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition
669 (CVPR), IEEE, Las Vegas, NV, USA, pp. 779–788.
- 670 Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical
671 image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical
672 Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in
673 Computer Science. Springer International Publishing, Cham, pp. 234–241.
- 674 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A.,
675 Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual
676 Recognition Challenge. *Int. J. Comput. Vis.* 115:211–252.
- 677 Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: Inverted
678 residuals and linear bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and
679 Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and
680 Pattern Recognition (CVPR), IEEE, Salt Lake City, UT, pp. 4510–4520.
- 681 Savary, S., Bregaglio, S., Willocquet, L., Gustafson, D., Mason D’Croz, D., Sparks, A., Castilla,
682 N., Djurle, A., Allinne, C., Sharma, M., Rossi, V., Amorim, L., Bergamin, A., Yuen, J., Esker,
683 P., McRoberts, N., Avelino, J., Duveiller, E., Koo, J., Garrett, K., 2017. Crop health and its
684 global impacts on the components of food security. *Food Secur.* 9:311–327.
- 685 Schwanck, A.A., Del Ponte, E.M., 2014. Accuracy and reliability of severity estimates using
686 linear or logarithmic disease diagram sets in true colour or black and white: a study case for
687 rice brown spot. *J. Phytopathol.* 162:670–682.
- 688 Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for
689 classification tasks. *Inform. Process. Manag.* 45:427–437.
- 690 Stewart, E.L., Hagerty, C.H., Mikaberidze, A., Mundt, C.C., Zhong, Z., McDonald, B.A., 2016.
691 An improved method for measuring quantitative resistance to the wheat pathogen
692 *Zymoseptoria tritici* using high-throughput automated image analysis. *Phytopathology*
693 106:782–788.
- 694 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception
695 architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern
696 Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern
697 Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 2818–2826.
698 <https://doi.org/10.1109/CVPR.2016.308>

- 699 Tian, K., 2020. Tomato leaf image dataset. <https://doi.org/10.17632/369CKY7N39.1>
- 700 Tusubira, J.F., Akera, B., Nsumba, S., Nakatumba-Nabende, J., Mwebaze, E., 2020. Scoring
701 root necrosis in cassava using semantic segmentation. *ArXiv200503367 Cs Eess.*
702 <https://arxiv.org/abs/2005.03367>
- 703 Vale, F.X.R., Fernandes Filho, E.I., Liberato, J.R., 2003. QUANT. A software for plant disease
704 severity assessment., in: In: 8th International Congress of Plant Pathology. Christchurch New
705 Zealand, p. 105.
- 706 Wang, G., Sun, Y., Wang, J., 2017. Automatic image-based plant disease severity estimation
707 using deep learning. *Comput. Intell. Neurosci.* 2017: 2917536
- 708 Wang, T., Rostamza, M., Song, Z., Wang, L., McNickle, G., Iyer-Pascuzzi, A.S., Qiu, Z., Jin, J.,
709 2019. SegRoot: A high throughput segmentation method for root image analysis. *Comput.*
710 *Electron. Agric.* 162:845–854.
- 711 Xu, L., Li, Y., Xu, J., Guo, L., 2020. Two-level attention and score consistency network for
712 plant segmentation. *Comput. Electron. Agric.* 170, 105281.
- 713 Yang, Q., Liu, M., Zhang, Z., Yang, S., Ning, J., Han, W., 2019. Mapping plastic mulched
714 farmland for high resolution images of unmanned aerial vehicle using deep semantic
715 segmentation. *Remote Sens.* 11:2008.
- 716 Zhang, S., Ma, Z., Zhang, G., Lei, T., Zhang, R., Cui, Y., 2020. Semantic image segmentation
717 with deep convolutional neural networks and quick shift. *Symmetry* 12:427.
- 718 Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network, in: 2017 IEEE
719 Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE
720 Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239.
721 <https://doi.org/10.1109/CVPR.2017.660>

722

723