DeepWheat: Estimating Phenotypic Traits from Crop Images with Deep Learning

Shubhra Aich[1], Anique Josuttes[2], Ilya Ovsyannikov[1], Keegan Strueby[2], Imran Ahmed[1], Hema Sudhakar Duddu[2], Curtis Pozniak[2,3], Steve Shirtliffe[2], and Ian Stavness*[1]

[1]Dept. Computer Science, [2]Dept. Plant Sciences, [3]Crop Dev. Centre, Univ. Saskatchewan, Canada

Abstract

In this paper, we investigate estimating emergence and biomass traits from color images and elevation maps of wheat field plots. We employ a state-of-the-art deconvolutional network for segmentation and convolutional architectures, with residual and Inception-like layers, to estimate traits via high dimensional nonlinear regression. Evaluation was performed on two different species of wheat, grown in field plots for an experimental plant breeding study. Our framework achieves satisfactory performance with mean and standard deviation of absolute difference of 1.05 *and* 1.40 *counts for emergence and* 1.45 *and* 2.05 *for biomass estimation. Our results for counting wheat plants from field images are better than the accuracy reported for the similar, but arguably less difficult, task of counting leaves from indoor images of rosette plants. Our results for biomass estimation, even with a very small dataset, improve upon all previously proposed approaches in the literature.*

1. Introduction

Measuring the phenotypic traits of crops, which are the differences in plant characteristics caused by the interaction of the plant's genetics and the environment, is important in plant breeding research as it allows the breeders to select crop varieties with desirable physical characteristics, such as high yield, resistance to stress, and ability to be easily harvested. Traditionally, phenotypic measurements are made manually in the field, which is both labor intensive and potentially inaccurate due to substantial sub-sampling involved. To overcome these drawbacks, image-based automated phenotypic traits estimation is emerging as an important area of applied computer vision research with the goal of capturing more accurate information at a large scale for better crop production.

*Corresponding Author: ian.stavness@usask.ca

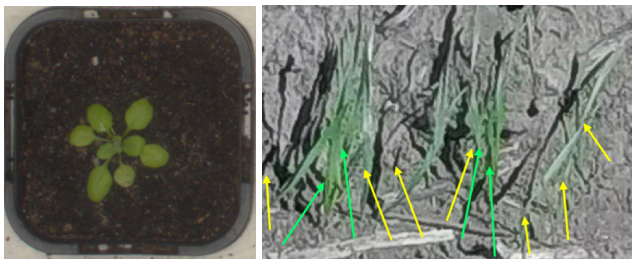

Figure 1: Eleven leaves in an image from the standard leaf counting dataset [9] (left) and eleven wheat plants in an outdoor image used for emergence counting in this paper. Counting plants from the right image is more challenging to due to variable number of leaves per plant and occlusion.

In many crops, including wheat, emergence (the density of plants within the field) and biomass (the total mass of each plant) are important phenotypes. Emergence is important because a vigorous and uniform crop stand is needed to compete for moisture, nutrients, and sunlight. Plants that emerge late will have a lower yield than the early emerging ones due to the increase in competition for sunlight and essential nutrients [21]. Determining biomass in different crop varieties is important because it is correlated with yield [32] and photosynthetic activity, and is an indicator of overall plant health [12]. These phenotypes are labour intensive and destructive to measure manually: emergence typically requires physically touching plants in the field to determine which leaves belong to which plant, and biomass measurements are made by cutting out plants from the field and measuring their mass. Furthermore, these phenotypes are traditionally measured on only a small sub-sample of the experimental plot area, which can result in sampling error. The combination of high importance and high measurement difficulty makes these phenotypes good candidates for image-based phenotyping in any crop breeding programs.

Counting plants is related to the well-studied problem of counting leaves from plant images [14, 7], but much more

978-1-5386-4886-5/18/$31.00 ©2018 IEEE
DOI 10.1109/WACV.2018.00042

challenging. Wheat seeds are planted in close proximity, therefore, the plants grown from these seeds are highly occluded by each other in the image. To illustrate the level of difficulty, Figure 1 shows a sample image from the standard leaf counting dataset [9] and another image from the dataset we are using for wheat emergence counting. Both images have the same label: 11 leaves in the left image, and 11 wheat plants in the right image. In the left image, the number of leaves is unambiguous despite a few small leaves in the center, which is not the case for plant count in the right image. According to the plant science experts who generated the ground truth counts and who have experience counting plants in the field, while counting from the images, they looked at the stems as close to the ground as possible. When a stem seemed unreasonably thick, they presumed that there were more plants behind the visible ones. Plant bases indicated by the yellow arrows in the figure are easy to count. However, in regions denoted by the green arrows, it may look like there is one plant, based on the thickness of the plants, amount of leaves, and age of plants, the count of plants was estimated by the raters as more than one. Hence, both intuition and experience play a role in accurate emergence counting, making it a difficult image analysis task.

In this paper, we propose completely data-driven frameworks for emergence counting and biomass estimation. We develop generalized architectures for phenotypic traits estimation blending the concepts of learning sparse structure via dense, multiscale representations [33] and residual or shortcut connections [15]. We train our models from scratch to keep our phenotypic estimation tasks independent of the other large-scale machine learning tasks pursued with very large models. For this reason, to efficiently train the data-hungry deep models with a few training samples, we also propose a novel data augmentation strategy based on randomized minimal region swapping of the superpixels in an image, which can be used to augment low to medium resolution images. Also, we examine the quality of learning of the emergence counting architecture qualitatively by visualizing salient regions using the class activation mapping (CAM) [35] approach. We find that the learned network features focus on image regions that are responsible for counting, notably the base of each leaf-cluster, and the dense regions of leaves, according to the plant breeding experts who provided the ground truth counts.

To the best of our knowledge, this is the first work on image-based phenotypic trait estimation of crops with deep learning. The name *DeepWheat* refers to our overall system because of the first use of deep learning in this domain and since we have used the image dataset of two species of wheat for the evaluation. Although we evaluate our approach on wheat, our design allows the frameworks to be generalized to other types of crops with minimal additional manual intervention.

2. Related Work

Despite the significance of emergence and biomass in crop breeding, little computer vision research has been done on the automated estimation of these traits from images. Leaf counting has been studied in more detail due to a standardized dataset of rosette plants and previous computer vision competitions [5]. Recent approaches to leaf counting have employed convolutional neural networks to count by regression [7]. We adopt a similar approach in this study to evaluate if it extends to much more difficult phenotyping tasks such as plant and biomass counting from field images.

A few studies have looked at plant density estimation in maize [31, 30, 29] and wheat [23, 16] from RGB images. All of these previous methods employ a traditional image processing pipeline that requires hand-tuned parameters tailored to the specific crop of interest. In the wheat studies, the plant counting algorithm depends on the accurate segmentation of leaves, followed by extracting regional properties of the leaves as features, and then training a simple artificial neural network (ANN) [23] or a support vector machine (SVM) [16]. In both papers, the initial segmentation of the plant foreground from the soil background is accomplished with simple naive approaches: Otsu thresholding on the "b" channel of Lab image or a predefined RGB transformation channel $(2G - 2B - 2.4R)$. However, simple threshold-based segmentations are not robust to variable illumination in different field environments. Indeed, these segmentation approaches are found to give very poor results for the images used in our study and are therefore not useful benchmarks for comparison.

A number of previous studies have attempted to estimate biomass, but most have done so from field-based measurements and are therefore not applicable to image datasets. A few studies have used aerial images as a basis for biomass estimation. In [27], naive linear regression models are fitted on plant height and plant coverage in aerial images. In [26], different linear and nonlinear combinations of height measured with an ultrasonic sensor, leaf area index measured with a plant canopy sensor, and vegetation indices from canopy reflectance obtained using a portable spectrometer are used as the predictors and biomass is used as the response of the multiple linear regression model. The product of leaf area index and dry matter content per leaf area is regarded as the estimation of above-ground biomass (AGB) in [25]. The authors also provide a comparison against the models developed using exponential regression, partial least square regression and simple artificial neural networks. In [20], AGB was estimated from height information obtained from the Digital Terrain Model (DTM) derived from LiDAR data. For each plot, simple statistical measures of height, such as mean, quadratic mean, standard deviation, skewness, kurtosis, and percentile of height along with height bins at fixed intervals, are used as the predictors for regres-

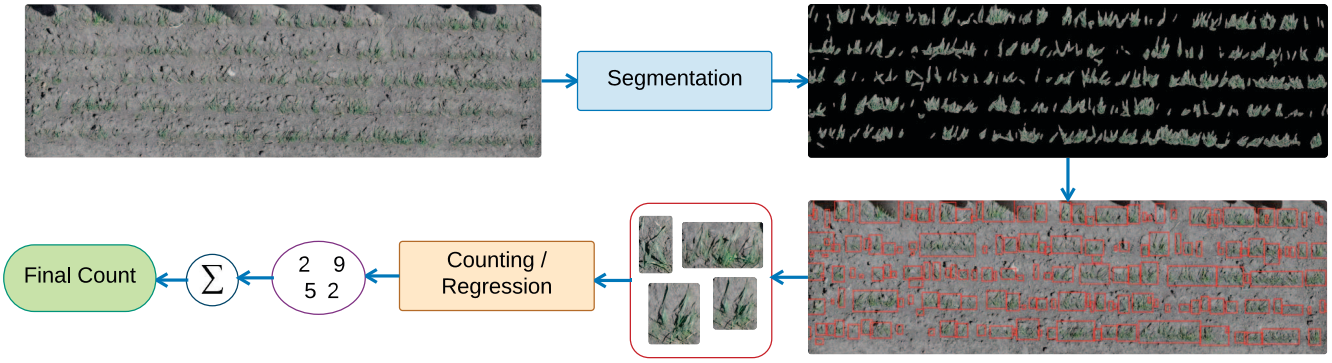

Figure 2: Workflow for emergence counting: 1) loosely segment the plant regions from RGB plot images with the segmentation module, 2) extract small patches containing plants via connected component analysis, 3) use counting module for individual counts on each patch, 4) sum all the patches to get the overall emergence count for a single plot.

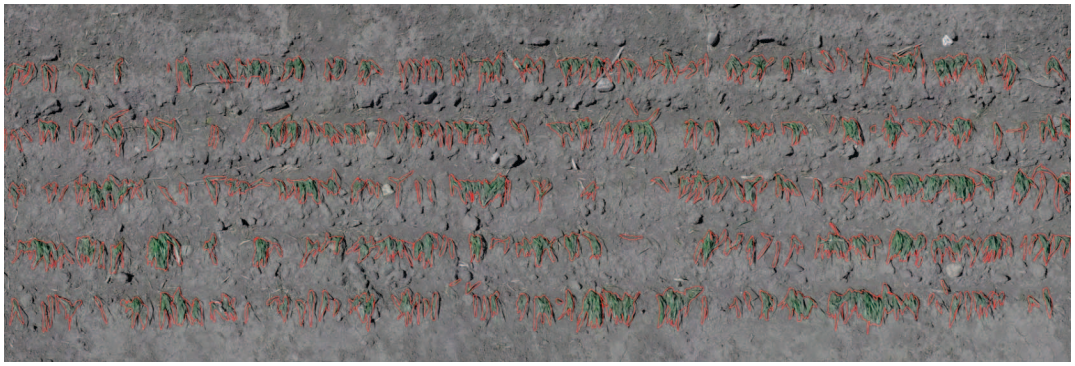

Figure 3: Manual ground-truth generated for relaxed segmentation of plants showing manually drawn contours around plant regions (red). Later, contours are filled with simple morphological hole-filling to create the binary segmentation mask.

sion modeling. A similar approach is taken in [19] with additional vegetation indices extracted from hyperspectral data. In terms of the list of predictor variables, the approach in [10] can be considered an extended version of the other two [19, 20] with height information plus the vegetation indices based on both hyperspectral and unmanned aerial vehicle (UAV) images.

3. Our Approach

In this section, we describe the design of both emergence count and biomass estimation frameworks in detail. Although both traits are estimated by convolutional networks performing regression, the architectures and overall workflows are different.

3.1. Emergence Counting

Figure 2 depicts the overall computational procedure for counting crop emergence. First, we loosely segment the plant regions from the RGB plot images through the segmentation module described later. Next, we extract all the segmented patches from the whole image, as indicated by the red rectangles in Figure 2 and input each patch image to the counting module to get the individual emergence counts for each patch. Finally, we sum up all the predicted counts for a single plot image to get the overall prediction for emergence count for that particular plot. In this framework, both the segmentation and the counting modules comprise deep architectures which we describe below.

3.1.1 Segmentation

Our motivation for segmenting plot images into smaller patches is twofold. First, due to the very high resolution of plot images ($\sim 2500 \times 7500$), it is not computationally feasible to do the emergence counting task on the whole image at once. Instead, either sequential or parallel counting over disjoint plant regions is required. Second, data-driven approaches, like deep learning, require many training samples, whereas we have only a few high-resolution plot images available for that purpose. Therefore, we generate non-overlapping patches of segmented plant regions to provide us with more than a hundred subsamples from each plot image for further training of the counting model.

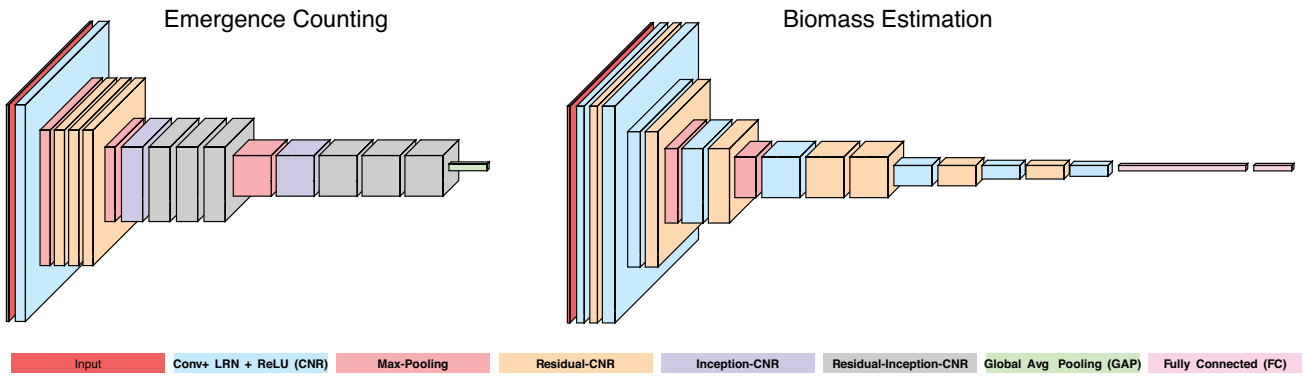

Figure 4: Emergence and biomass estimation architectures. We use 7×7 receptive fields in the initial CNR block with unit stride. The number of filters after each max-pooling operation is doubled, except the first one for emergence counting. *residual*-CNR is a simplified version of the residual block described in [15], where we keep the number of receptive fields constant inside the block. We use a simplified "Inception" module [34], where the number of input and output receptive fields are the same. Inside our Inception block, we employ half of the size of filters for 3×3 convolution, a quarter of the input size for the equivalent 5×5 convolution, and half of the rest for pooling and unit convolution each. For the emergence network, to visualize the representations learned by our model, we use global average pooling (GAP) [22].

From the design perspective, we relax the output of the segmentation module from exact segmentation to a soft or relaxed segmentation for several reasons. First, generating the exact ground-truth manually for images like the ones shown in Figure 3 is a more tedious and time-consuming process than defining loose or relaxed contours around plants. Moreover, for deep networks, learning to count from the subsamples with exact vs. loose segmentations is similar since the background is uniform and so, it is unlikely that the model would pick up distinctive features from the background region. This claim is also validated by CAM [35] visualizations of the network in the *Experiments* section that show saliency in foreground regions. In addition, the wheat leaves are thin and partly occluded; therefore, going for precise segmentation could result in missing very thin or hard-to-detect regions of the plants which could deteriorate the counting performance since the model responsible for counting would assume the segregated leaves as different instances rather than a single one.

To perform soft segmentation with deep learning, we use the SegNet architecture [8, 7] rather than deconvolutional networks containing fully connected (FC) layers [24] with a many more training parameters. This is because the problem we are dealing with is easier than the exact segmentation and much simpler than general multi-class semantic segmentation both in terms of the cardinality of the output categories and the nature of the domain since the diversity of the pixel intensities in a single plot image is highly restrained compared to that of natural images. Furthermore, our concern is not to get an overall-high precision segmentation mask, rather we are concerned with not missing plant regions in the image for the counting model afterward.

3.1.2 Counting by Regression

In this paper, we focus on different species of the crop wheat, which except the very late season, resembles mostly to grass crops. The leaves of such plants are the most deformable among all kinds of plants and crops, and so, a set of wheat plants in an image might appear in a combinatorially large number of variations. Thus, to successfully count the number of plants in the image, the deep model must be able to deal with such combinatorial number of deformations and resulting occlusions as much as possible.

As argued in the NIN paper [22], a simple stack of convolutional layers with an over-complete set of filters followed by nonlinearity and pooling serve well when the underlying concepts to be learned via abstract representation are linearly separable. However, for highly nonlinear latent concepts, replacing plain convolutional blocks with small networks inside the basic architecture is already proved to be useful in several large-scale image classification tasks [33, 34]. Hence, we take inspiration from these works, where the representation in each layer is approximated from the dense multi-scale feature responses learned in the previous layer. Also, we incorporate the concept of residual learning [15] in our architecture, which we experimentally found to be useful for faster training in case of stacked-convolutional architecture for our task.

Therefore, in the design of our network as depicted in Figure 4, four different convolutional blocks are used. Our initial convolutional block (CNR) is a simple convolution operation followed by local response normalization and rectified nonlinearity [18]. Next, we use a simplified residual version of the original residual block described in [15],

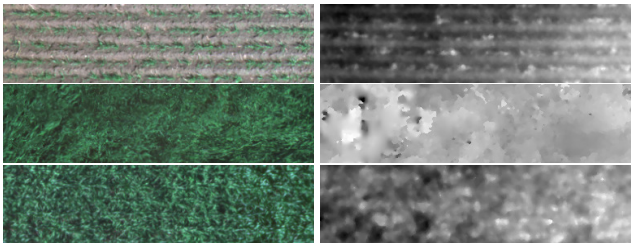

Figure 5: Sample RGB plot images (left) with corresponding DEMs (right) showing wheat plants from emergence as individual plants (top) to full crop canopy (middle) and during the reproductive stage (bottom). DEM values (height) converted to grayscale for visualization.

in the sense that the number of feature maps is constant throughout the block from input to output. Also, for deeper layers, where the number of receptive fields is comparatively higher, we incorporate the "Inception" version of CNR followed by the *residual*-Inception version. All these modules are crafted to have the same input-output capacity. Finally, for the ease of visualization of the salient regions detected by our model, we simply use the global average pooling (GAP) [22] layer. We experimented with different setups of fully connected layers instead of GAP and got slightly improved performance. However, we prefer visualization over those minor improvements to encourage further research based on visualization. Lastly, we have not used any pre-trained model because unlike classification problems, the capacity of the final layer does not scale up with the complexity of the counting task. In addition, opening up the full network for finetuning might result in significant overfitting due to comparatively smaller datasets.

3.2. Biomass Estimation

For biomass estimation, we have both 5 channel orthomosaics (Blue, Green, Red, Near-infrared, red-Edge) and digital elevation maps (DEM). Sample RGB images are shown in Figure 5. The pixel values of the DEM files indicate the elevation of plants from the ground. Note that, the RGB images of the plots available for emergence counts in the previous section and biomass estimation here are from different sources. The plot images for biomass estimation are lower resolution ($\sim 120 \times 480$) than those used for emergence counting (see Section 4.1).

above-ground biomass refers to the weight of all plant material above the ground. We expect that there is a relationship between biomass and height or elevation values of the DEM images, but this relationship is difficult to observe from simple biomass versus elevation graphs. However, representing values from each plot as a different dimension in $\mathbb{R}^n$ space, we have found small angles ($[30^\circ - 32^\circ]$ in our dataset) between the normalized elevation vector and the biomass vector. This suggests a non-linear relationship between these two quantities and we take this as motivation for further computational analysis.

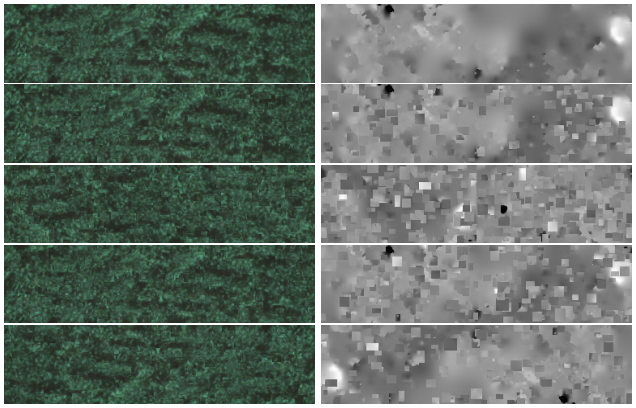

Figure 6: Sample RGB plot images (left) with corresponding DEMs (right) showing the original image (top row) and images generated by our *RMRS* data augmentation procedure (other rows). DEM values (height) converted to grayscale for visualization.

Now, to apply any data-hungry models like deep learning to estimate biomass from these images, one of the main obstacles is the extremely low number of available samples (~ 100) for training and testing. One of the obvious ways to overcome this drawback is to figure out a suitable data-augmentation strategy. In this paper, we have devised a novel, simple and effective randomized data augmentation scheme that can be utilized to generate a sufficiently large number of augmented samples from each image. The idea is based on swapping similar superpixels in the image randomly. We call this approach the *randomized minimal region swapping (RMRS)* algorithm. The steps of the *RMRS* algorithm are as follows:

1. Get the list of K superpixels from RGB to gray-converted image and sort by their mean values.

2. Generate a randomized list of length N of the number of random swaps needed to generate the pool of N augmented samples from a single image. The random integer values are in the range $[low, \lfloor K/2 \rfloor]$, where low is the predefined threshold for the minimum number of swaps needed to create an augmented sample.

3. For each number r in the list generated in step 2, generate a randomized list of length r of either even or odd superpixel indices in the range $[1, \lfloor K/2 \rfloor]$ and swap minimal rectangular regions between those even(odd) superpixels and their consecutive odd(even) counterparts in the sorted list. Even-odd consideration is necessary to avoid unaugmentation by repeated swaps.

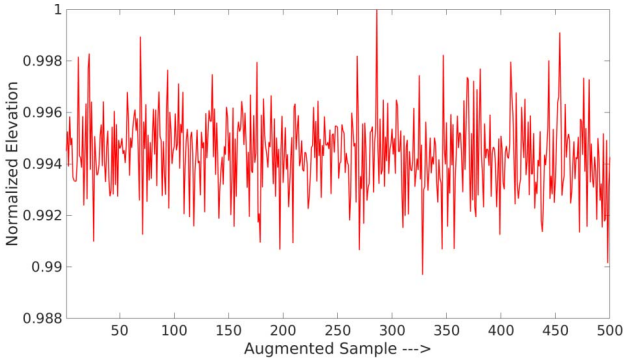

Figure 7: Normalized summation of the elevation for the samples augmented from a single image. The first point represents the elevation of the original sample and the rest (499) are the augmented ones. The range of normalized elevation is in the range $[\sim 0.99, 1.0]$ indicating that the total elevation for all the samples are similar to the original.

In our implementation, we use SLIC [6] as the superpixel algorithm. Figure 6 shows sample augmentation results for a single image along with the original one. As can be seen, it is impossible to identify the augmented samples as the artificial ones by looking only at RGB images, even though the corresponding DEMs appear to be highly discretized. Hence, as part of a further exploratory analysis, we plot the normalized summation of all pixel values or elevations of each DEM file for all the augmented samples along with the original one. Figure 7 shows this normalized elevation plot for a single image and its augmented samples. As you can see, the normalized elevation varies in the range $[\sim 0.99, 1.0]$, which means that although the augmented DEM files look different and discretized, the contents of the DEM pixels remain nearly constant after being augmented by the *RMRS* algorithm.

In addition to increasing the number of training samples, augmenting data this way has another advantage as a byproduct. We hypothesize that the spatial relationships among the pixels in DEM images have little to do with the prediction of biomass since plants can be found in almost any region in the plot images. Therefore, the counting model should learn to map the pixel values from DEM images into the real-valued space of biomass in an almost spatially invariant manner. For data augmentation by *RMRS* algorithm, new samples are just different permutations of the original one. From the practical standpoint, the interpretation might be that to generate an augmented sample, we swap the plants with similar color information within the plot. Thus, by learning to predict from this augmented dataset, the model may intrinsically learn a spatially invariant mapping from color and elevation to biomass.

Finally, we use a similar network architecture for biomass estimation (Figure 4). The only difference between this model and the emergence count one is that the parameters and the placement of the computational blocks or layers are slightly modified to fit the model into this problem.

4. Experiments

This section contains the experimental details of our work. First, we describe the datasets used for both tasks. Next, training procedure and implementational details of the networks are provided. Finally, the evaluation metrics are described and the evaluation results are reported in comparison to previous work along with the qualitative visualization of the salient regions.

4.1. Datasets

The dataset used for emergence count consists of 274 wheat *(Triticum durum)* plots of $1.5m \times 3.7m$ area. High-resolution aerial images ($\sim 2500 \times 7500$ pixels per plot) were captured for each plot by walking through the field with a GoPro Hero 5 camera [3] mounted on a monopod with a gimbal for stabilization. Covering plots with this device has the advantage of getting very high-resolution images appropriate for detailed computational analysis compared to other remote sensing technologies.

For biomass estimation, we used aerial drone images for 48 wheat (Triticum aestivum) plots for two dates: June 27 and July 20, 2016. The UAV images have been captured using a MicaSense RedEdge camera [4] on a DraganFly Commander drone [2]. The RedEdge camera includes five different sensors, one for each band: $Blue(\sim 465 - 485nm)$, $Green(\sim 550 - 570nm)$, $Red(\sim 658 - 678nm)$, $NIR(\sim 820 - 860nm)$, and $RedEdge(\sim 707 - 727nm)$. The output from these sensors was post-processed using the Agisoft Photoscan [1] to generate an orthomosaic image and digital elevation map. For each of these dates, manual ground truth measurement of biomass have also been conducted. For manual counting, plants were cut randomly from the plots at ground-level using sickles, dried, and then weights of those plants were noted. The dataset is randomly split into two equal subsets for training and testing.

4.2. Training and Implementation

We used Torch [11] as the deep learning framework. To train the segmentation network, we generated $0.25M$ subsamples of size 224×224 from 10 high-resolution plot images. The network was trained for 30 epochs over this augmented dataset. SGD-momentum was used as the optimizer with a fixed learning rate, momentum, and weight decay of $0.01, 0.9$, and 0.0001 respectively, over the training period.

Both the emergence count and biomass estimation networks were trained with similar parameter settings. Adam optimizer [17] was used with learning rate and weight decay both set to 0.0001. Absolute value and Smooth L1 measures [13] are used as the error criteria (loss functions) for

training emergence and biomass models, respectively. For emergence network training, we slowed down the training rate later based on our observation of the training statistics. Training for the emergence network was conducted for 100 epochs, whereas the biomass estimation network was trained with different combinations of input channels for 50 epochs with the same initial parameter settings. We will provide the link for pre-trained models and codes in the final version of this paper.

Note that the emergence count network was trained on 7855 patches extracted from 37 images and their slightly augmented versions. On the other hand, the biomass network was trained with about $0.15M$ augmented training samples generated by the *RMRS* algorithm from 48 plot samples. Codes, pre-trained models, and datasets are publicly available here. [1]

4.3. Evaluation

Here, we provide three evaluations of our approach. First, we assess the performance of our segmentation network for generating relaxed binary segmentations. Next, both emergence count and biomass estimation networks are evaluated based on the metrics listed in Equation 1 below. Among these metrics, we take *MAD* and *SDAD* from the leaf counting benchmark [7]. The other is simply a variant of these measures. In addition, we provide CAM visualization for the emergence counting model.

Precision $= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	85.59
Recall $= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$	83.76
Accuracy $= \frac{\text{True Positive} + \text{True Negative}}{All}$	93.76

Table 1: Binary segmentation results

Emergence evaluation: Precision, recall, and accuracy are measured to evaluate the segmentation network (Table 1). Results for precision ($\sim 86\%$) and recall ($\sim 84\%$) are a somewhat low because the ground truth segmentations are not precise, but loosely defined contours covering all the plant regions in the images. To justify our outputs, we have visually checked almost all the test segmentation results and find almost no plant regions undetected by the network.

$$\begin{cases} a_i, t_i = \text{actual and target counts for } i^{th} sample \\ N = \text{Number of samples} \\ \textit{\%Difference(\%D)} = \frac{\sum_i |a_i - t_i| I_{[a_i - t_i \neq 0]}}{\sum_i t_i} \\ \textit{Mean Absolute Difference (MAD)} = \frac{\sum_i |a_i - t_i|}{N} \\ \textit{Std Absolute Difference (SDAD)} = \sqrt{\frac{\sum_i (|a_i - t_i| - MAD)^2}{N-1}} \end{cases} \quad (1)$$

[1] https://github.com/p2irc/deepwheat_WACV-2018

Problem	MAD	SDAD	%D
Prev. Leaf Counting [7]	1.62	2.30	-
Plain Architecture	1.13	1.42	27.04
Inception Architecture	1.08	1.38	25.78
Our Emergence Counting	**1.05**	**1.40**	**25.08**

Table 2: Evaluation metrics for the emergence count model

Table 2 lists the evaluation metrics for our emergence counting network. As stated in the introduction, we did not find appropriate literature to benchmark our approach. The closest approach is the one used for Arabidopsis and Tobacco leaf counting problem [7]. We achieve *%D* of 25% and *MAD* and *SDAD* of 1.05 and 1.40 which is more accurate than previously reported results for one of the best leaf counting system currently available. These results are notable because counting wheat plants with thin, overlapping leaves from outdoor images is substantially more difficult than counting leaves from indoor images of rosette plants (as discussed in the Introduction and illustrated in Figure 1). We have also included the results for the corresponding plain and Inception-only version to justify the additional complexity of our final model. The plain network was trained for twice the number of epochs than others.

The salient regions detected by our counting model for sample RGB images are shown as heatmaps, generated by CAM [35], in Figure 8. Although in the original paper, CAM is used to visualize class-specific mapping of the salient regions, for our counting task, it can also be used for visualizing the regions responsible for making the counts. As already discussed, the bases of leaf-clusters are the most significant parts for successful counts followed by dense regions of overlapping leaves. The sample heatmaps also follow this counting strategy. In the heatmaps, the bases of the plants are marked with red (highest saliency) followed by the leaves with yellow, which clearly indicates that our model is capable of identifying the correct regions in the images responsible for counting. Nonetheless, our percentage deviation is a bit high because of the inherent difficulty of counting the plants due to severe occlusion and large leaf deformations. To enable CAM visualization, we cut out additional fully connected layers, which had provided a slight performance boost, but the resulting visualization provides more valuable insight into the learning process for plant counting.

Biomass evaluation: Table 3 contains the same metrics as in Equation 1 for biomass models trained with different input channel combinations. Here, H, R, G, B, N, and E stand for DEM, Red, $Green$, $Blue$, NIR, and $RedEdge$ channels, respectively.

As can be seen, the model trained with only $H(DEM)$ as input gives *%D* of $\sim 26\%$, which is $\sim 4\%$ and $\sim 2\%$ lower than the model trained with $RGBH$ and all the chan-

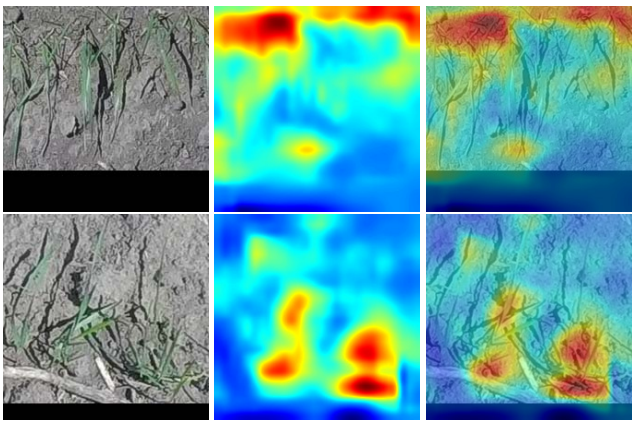

Figure 8: Sample RGB images (left), their CAM [35] visualizations (middle), and superimposed images (right). Note that, RGB images are padded by black to maintain a constant size of 224×224. Red and blue indicate the most and the least significant regions responsible for emergence counting. As you can see, the plant bases are detected as the most salient regions (red) which the experts also use for counting followed by the leaves (yellow).

nels. At this point, it is unclear whether the deep learning model takes care of any of the RGB texture in the biomass image. Intuitively, although color information or greenness of the RGB image might be important, the texture information is not that significant for biomass estimation. However, there is a high variance in the color information under different weather conditions. For instance, if the weather is overcast, crops will appear dark-green, for sunny weather, it will be yellowish-green, and so on. Another critical issue is that after augmenting data using *RMRS* algorithm, albeit the very local texture property and the total energy of the images are more or less preserved, semi-local texture property is destroyed. We are not sure whether this lack of semi-local texture causes the network trained with RGBE input to perform poorer than the one with only DEM input. This issue can only be explored further if sufficient raw training samples are available in future.

On the other hand, the fact that the model works better when two extra non-visible wavelengths, such as, NIR and $RedEdge$, are provided along with RGB, is consistent with the plant science literature [28] where vegetation indices extracted from hyperspectral and visible wavelength data are used as strong indicators of photosynthetic measurements of plants. However, the utility of hyperspectral data for biomass estimation is still an open question.

In Table 3, we provide a comparison against the recent literature. We implemented the methods described in [20, 19, 10] on our data for comparison. These three papers reported the effect of different feature combinations from the set of simple statistical features based on height and different vegetation indices as the predictor variables for their regression models. In this table, we use the combination of features that performed best on our dataset. H_1, H_2, and H_3 indicates slightly different variations statistical height features and OH_3 stands for the combination of H_3 and *Optimized Soil-Adjusted Vegetation Index (OSAVI)*. Also, *MARS (Multivariate Adaptive Regression Splines)*, *PLS (Partial Least Squares)*, and *MLR (Multivariate Linear Regression)* are different linear and nonlinear regression algorithms. As can be seen, even with such tiny amount of original training data, the best performance of our deep model (trained with $DEM(H)$) is $\sim 4\%$ better than the recent nonlinear regression model for biomass.

Method	MAD	SDAD	%D
H_1+MARS[20]	1.66	2.03	29.61
H_2+PLS[19]	3.86	2.72	68.92
H_2+MARS[20, 19]	1.74	2.07	30.96
OH_3+MLR[10]	1.67	1.63	29.67
Ours ($RGBH$)	1.67	2.05	29.75
Ours ($RGBNEH$)	1.53	**1.62**	27.38
Ours (H)	**1.45**	2.05	**25.88**

Table 3: Comparison of biomass estimation metrics to other methods and with different input channels ($H \equiv DEM$, Red, Green, Blue, Near-infrared, and red$E$dge)

Acknowledgment

This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund and the Natural Sciences and Engineering Research Council (NSERC) of Canada. We also thank Seungbum Ryu and the USask field crew for providing biomass data.

5. Conclusion and Future Work

In this paper, we have developed three different deep learning models for segmenting plant regions, counting plants, and estimating biomass from aerial field images. Our results show better biomass estimation accuracy than previous methods and better accuracy for outdoor emergence counting as compared to previous studies of indoor leaf counting. Although we have only evaluated our model on particular species of wheat, we expect that our design methodology allows for generalization of these models to other types of crops with minimal changes. As future work, we plan to evaluate our networks with other crops that have different plant morphologies, such as pulses and oilseeds. We also plan to further investigate if estimation accuracy for these phenotypic traits can be improved with larger datasets in subsequent growing seasons, as well as the use of digital elevation maps together with non-visible wavelengths of light as input for biomass estimation.

References

[1] Agisoft. http://www.agisoft.com/. 6

[2] DraganFly. http://www.draganfly.com/. 6

[3] GoPro. https://gopro.com/. 6

[4] Micasense-Rededge Camera. https://www.micasense.com/rededge/. 6

[5] Leaf Counting Challenge. https://www.plant-phenotyping.org/CVPPP2017-challenge, 2017. 2

[6] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012. 6

[7] S. Aich and I. Stavness. Leaf counting with deep convolutional and deconvolutional networks. *CoRR*, abs/1708.07570, 2017. 1, 2, 4, 7

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. 4

[9] J. Bell and H. M. Dee. Aberystwyth leaf evaluation dataset, Nov. 2016. 1, 2

[10] J. Bendig, K. Yu, H. Aasen, A. Bolten, S. Bennertz, J. Broscheit, M. L. Gnyp, and G. Bareth. Combining uav-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, 39:79 – 87, 2015. 3, 8

[11] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 6

[12] J. Dai, B. Bean, B. Brown, W. Bruening, J. Edwards, M. Flowers, R. Karow, C. Lee, G. Morgan, M. Ottman, J. Ransom, and J. Wiersma. Harvest index and straw yield of five classes of wheat. *Biomass and Bioenergy*, 85:223 – 227, 2016. 1

[13] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Dec 2015. 6

[14] M. V. Giuffrida, M. Minervini, and S. Tsaftaris. Learning to count leaves in rosette plants. In S. A. Tsaftaris, H. Scharr, and T. Pridmore, editors, *Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP)*, pages 1.1–1.13. BMVA Press, September 2015. 1

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 2, 4

[16] X. Jin, S. Liu, F. Baret, M. Hemerl, and A. Comar. Estimates of plant density of wheat crops at emergence from very low altitude uav imagery. *Remote Sensing of Environment*, 198:105 – 114, 2017. 2

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 4

[19] G. V. Laurin, Q. Chen, J. A. Lindsell, D. A. Coomes, F. D. Frate, L. Guerriero, F. Pirotti, and R. Valentini. Above ground biomass estimation in an african tropical forest with lidar and hyperspectral data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 89:49 – 58, 2014. 3, 8

[20] G. V. Laurin, N. Puletti, Q. Chen, P. Corona, D. Papale, and R. Valentini. Above ground biomass and tree species richness estimation with airborne lidar in tropical ghana forests. *International Journal of Applied Earth Observation and Geoinformation*, 52:371 – 379, 2016. 2, 3, 8

[21] K. Lawles, W. Raun, K. Desta, and K. Freeman. Effect of delayed emergence on corn grain yields. *Journal of Plant Nutrition*, 35(3):480–496, 2012. 1

[22] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 4, 5

[23] S. Liu, F. Baret, B. Andrieu, P. Burger, and M. Hemmerl. Estimation of wheat plant density at early stages using high resolution imagery. *Frontiers in Plant Science*, 8:739, 2017. 2

[24] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, Dec 2015. 4

[25] X. Quan, B. He, M. Yebra, C. Yin, Z. Liao, X. Zhang, and X. Li. A radiative transfer model-based method for the estimation of grassland aboveground biomass. *International Journal of Applied Earth Observation and Geoinformation*, 54:159 – 168, 2017. 2

[26] B. Reddersen, T. Fricke, and M. Wachendorf. A multisensor approach for predicting biomass of extensively managed grassland. *Computers and Electronics in Agriculture*, 109:247 – 260, 2014. 2

[27] M. Schirrmann, A. Hamdorf, A. Garz, A. Ustyuzhanin, and K.-H. Dammer. Estimating wheat biomass by combining image clustering with crop height. *Computers and Electronics in Agriculture*, 121:374 – 384, 2016. 2

[28] P. J. SELLERS. Canopy reflectance, photosynthesis and transpiration. *International Journal of Remote Sensing*, 6(8):1335–1372, 1985. 8

[29] D. Shrestha and B. Steward. Shape and size analysis of corn plant canopies for plant population and spacing sensing. *Applied Eng. in Agric*, 21(2):295–303, 2005. 2

[30] D. Shrestha, B. Steward, and S. Birrell. Video processing for early stage maize plant detection. *Biosystems engineering*, 89(2):119–129, 2004. 2

[31] D. S. Shrestha and B. L. Steward. Automatic corn plant population measurement using machine vision, paper number 011067. In *2001 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers, 2001. 2

[32] J. M. Soriano, M. Malosetti, M. Rosell, M. E. Sorrells, and C. Royo. Dissecting the old mediterranean durum wheat genetic architecture for phenology, biomass and yield formation by association mapping and qtl meta-analysis. *PLOS ONE*, 12(5):1–19, 05 2017. 1

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015. 2, 4

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, June 2016. 4

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, June 2016. 2, 4, 7, 8