Supervised and weakly supervised deep learning for instance segmentation and counting of plant parts

by

SHRINIDHI SHRIDHAR ADKE

(Under the Direction of Changying Li)

Abstract

The advent of deep learning in computer vision has enabled the detection and segmentation of plant phenotypic traits with increased accuracy and efficiency. The challenges involved in segmenting the individual instances of plant traits using traditional image processing can be addressed with the help of supervised instance segmentation techniques such as Mask R-CNN. This thesis presents two instance segmentation applications to extract traits from 2-D images of corn cob and cotton plant. The former develops a deep learning-based image processing pipeline that aims to estimate the consumption of corn by identifying corn and its bare ear, which will aid in testing the wild animals' preference for genetically modified corn. The estimation results of these models were included and compared with manually labeled test data with $R^2 = 0.99$, which showed that the use of Mask R-CNN model provides highly accurate results, thus, allowing it to be used further on all collected data. The later study uses Mask R-CNN for extracting three of the most crucial cotton plant phenotypic traits: main stalk height, node, and boll count. The instance segmentation of the main stalk and nodes was proved to be a challenging task for supervised methods due to the lack of sufficient annotated data and feature similarity in the cotton plant architecture. In order to find a solution to the challenge of data scarcity, two weakly supervised approaches: Weakly Supervised Counting (WS-COUNT) and CountSeg, were demonstrated to carry out the cotton boll counting task. The results showed that weakly supervised counting approaches based on peak response maps such as CountSeg (RMSE = 1.284 ± 0.08) yield comparable results as of Mask R-CNN (RMSE = 1.175 ± 0.20). In terms of data annotation, the weakly supervised approaches were found to be at least 10 times efficient compared to the supervised approach for the boll counting task. In the future, the weakly supervised approach allows us to improve the current supervised frameworks in the absence of quality mask annotations by leveraging the density maps obtained from weak supervision and can be extended to various cotton organs or other crops.

Index words: GMO Corn, Cotton phenotyping, Boll counting, instance segmentation, image processing, Supervised learning, Mask R-CNN, Weakly supervised learning, class activation map, multiple instance learning;

Supervised and weakly supervised deep learning for instance segmentation and counting of plant parts

by

SHRINIDHI SHRIDHAR ADKE

B.Engg. (Electrical), University of Pune, India, 2015

A Thesis Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Science

Athens, Georgia

2021

©2021

SHRINIDHI SHRIDHAR ADKE

All Rights Reserved

Supervised and weakly supervised deep learning for instance segmentation and counting of plant parts

by

SHRINIDHI SHRIDHAR ADKE

Major Professor:	Changying Li
Committee:	Khaled Rasheed
	Frederick Maier

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2021

Acknowledgments

I would like to thank all my committee members for their time and the numerous suggestions they provided throughout my research. In particular, I would like to thank Dr. Li for his guidance, constant encouragement and motivation to keep me going. I would like to thank all the current and past lab members at *Bio-sensing and Instrumentation Lab* for their contribution to the data and ground work for this study. I would like to thank everyone at the *Institute for Artificial Intelligence* who presented me with an opportunity to be a part of this project. In the end, I'm grateful for all the support and strength that my family and friends have provided to me during the course of this degree.

Contents

Acknowledgments **iv**

List of Figures **vi**

List of Tables **viii**

1 Introduction **1**

2 Instance Segmentation to Estimate Consumption of Corn Ears by Wild Animals for GMO Preference Tests **9**

2.1 Abstract 10

2.2 Introduction 10

2.3 Materials and methods 13

2.4 Results and Discussion 23

2.5 Conclusions 33

3 Supervised and weakly supervised deep learning for segmentation, detection and counting of cotton organs using proximal imagery. **35**

3.1 Abstract 36

3.2 Introduction 37

3.3 Materials and methods 42

3.4 Results and Discussion 59

3.5 Conclusions 73

4 Conclusion **75**

Appendices **77**

A Supplementary Materials for Corn Image Instance Segmentation **77**

A.1 Supplimentary Data 77

A.2 Supplementary Tables and Figures 77

Bibliography **80**

List of Figures

2.1 An overview of stages of consumption estimation from raw images. 14

2.2 Illustration of two data labelling approaches. 17

2.3 Segmentation using two approaches. 23

2.4 PR curves. 26

2.5 Segmentation results. 27

2.6 Scatter plots of predicted vs. ground truth values of consumption ratio 28

2.7 Failed segmentation examples. 29

2.8 Confusion matrix for left-right corn ear consumption comparison. 30

2.9 Effect of background and resolution change. 31

3.1 Annotation Types 46

3.2 Schematic representation of WS-COUNT architecture 49

3.3 CountSeg for boll counting 52

3.4 Overview of boll counting 54

3.5 Main stalk and node instance segmentation 56

3.6 Error histograms from median predictions given by each method 60

3.7 Bubble plots for boll counts 61

3.8 Comparison of CountSeg and Mask R-CNN 64

3.9 Annotation times comparison by label type 66

3.10 Segmentation of main stalk and nodes 67

3.11 **Outlier cases in segmentation and detection** 68

3.12 **Scatter plot for main stalk height calculation** 70

3.13 **Scatter plot for node counting** 71

List of Tables

2.1 Image categorization after pre-screening. 15

2.2 Data partitioning and number of masks for approach 1. 17

2.3 Labelling approach comparison 23

2.4 Effect of training sample size on segmentation performance 25

3.1 Summary of raw images with entire plant 43

3.2 Summary of raw images selected for boll counting task 44

3.3 Summary of boll count per tile in different counting ranges 45

3.4 Testing set boll counting mean RMSE 61

3.5 Hold-out Test set average boll count 63

A.1 Statistical Analysis using t-test on overall segmentation mAP with various sample size (t-value, P-value) 78

A.2 Statistical Analysis using t-test on segmentation IoU for Whole corn with various sample size (t-value, P-value) 78

A.3 Statistical Analysis using t-test on segmentation IoU for Bare cob with various sample size (t-value, P-value) 78

A.4 Effect of training sample size on segmentation performance of Mask R-CNN and SoloV2 model 79

Chapter 1

Introduction

Machine learning and deep learning along with computer vision is playing a vital role in modern agriculture in several areas (Jiang, Li, et al., 2020; Uddin & Bansal, 2021). Some of the widely studied applications include plant disease classification (Saleem et al., 2019; Sibiya & Sumbwanyambe, 2019; Sladojevic et al., 2016), yield prediction (Y. Chen et al., 2019; Maimaitijiang et al., 2020; Van Klompenburg et al., 2020), organ detection and counting (Atanbori et al., 2019; Koirala et al., 2019; Xia et al., 2019). In addition, recent developments in the instance segmentation techniques such as Mask R-CNN (He et al., 2017), SoloV2 (X. Wang et al., 2020), Yolact (Bolya et al., 2019a) helped researchers to study beyond above applications (Hamuda et al., 2016; Ni et al., 2021; Sukmana & Rahmanti, 2017; Ward et al., 2018). Inspired from the success of these methods, this thesis focuses on two phenotypic applications of image instance segmentation using deep learning. The first application deals with segmentation of corn cob images to estimate the consumption, while the second explores the cotton plant organ segmentation to extract main stalk height, node count and boll count. Both the applications develop a deep learning based pipeline to segment the 2-D images that helps researchers to further study the various relationships between these phenotyoic traits and their extraction methods.

Corn is one of the world's most important crops and is produced both traditionally and with genetically modified organisms (GMO) (FAOSTAT, 2018). To obtain certain agriculturally-desirable traits, such as resistance to pests, herbicide tolerance, drought tolerance, specific corn varieties have been genetically engineered. Despite research on its safety and equivalence to traditional varieties, questions continue to be raised by members of the public regarding its safety and edibility. Since its introduction, there have been mixed views on GMO foods and crops and GMO corn is no exception for this. Over the past few decades, studies in favor and against GMO corn were done which is collectively reviewed by (Chassy & Tribe, 2010) and some of them clarified that *"animals are not biased to organic corn."*

During this time, the hypothesis was formulated that wild animals, specifically squirrels and deer, can sense differences between GMO and non-GMO corn and prefer non-GMO corn when given a choice between the two. To test this hypothesis, a nation-wide community science project called "The GMO Corn Experiment" was started in 2015 that gathered the image data of GMO and non-CMO corn set outside by volunteers in their yards (Haro von Mogel & Bodnar, 2015). The next challenge of this experiment was to estimate the precise consumption of all the GMO and non-GMO corn samples based on the complex image data obtained to address the hypotheses being tested in the study. The purpose was to measure which of the two types of corn was consumed more in all the collected samples. On a larger scale, this study can serve as a model for studying animal preferences of GMO versus non-GMO food in a publicly-accessible manner. In many cases visual observations were unreliable. It is also time consuming and tedious to observe accurately the consumption of each corn ear used in the experiment because there are hundreds of images and identifying them individually takes significant amount of time for a human observer. Furthermore, one observer may not be as precise as another, and thus more observers are needed to get an accurate estimation of consumption. Any error in this process can result in supporting a false hypothesis and will impact corn related studies and the corn industry at large scale.

Image instance segmentation was best suited to estimate the consumption of corn because it is crucial to know the exact area of the individual ear of corn and the consumed part of the corn. Therefore, the first study is intended to focus mainly on the development of the image analysis methods that will allow us to better address the hypotheses about wild animal preferences and avoidance of the two maize genotypes, which will be tested and submitted in a subsequent works along with other lines of evidence gathered in the project. Thus, the in the chapter chapter 2, we propose an automated algorithm that uses the masks produced by the Mask R-CNN method to estimate the area of the consumed part of the corn as well as the entire corn, which results in a percentage consumption/eaten.

On the other hand, Cotton (*Gossypium hirstum L.*) is one of the most grown cash crop around the world (FAOSTAT, 2019). Efforts involving the study, improvement of existing or to develop new methodologies that aids the production essentially requires measurement of plant traits (Fahlgren et al., 2015). It is an arduous task to obtain such measurements manually and may result in ruining the agricultural field. To carry-out these tasks efficiently, researchers are developing artificial intelligence based systems with the help of robotics (Oberti & Shapiro, 2016), machine learning (Liakos et al., 2018) and deep learning (N. Zhu et al., 2018). In the case of cotton, the field and the plant architecture makes it even more difficult. Furthermore, based on the cotton plant growth cycle the important morphological traits differ (Ritchie et al., 2007). For example, in the germination stage monitoring seedlings and identifying weed is an important task, while, in blooming stage the flowering pattern/stages are essential for further growth. Identifying and monitoring these traits can be efficiently done with the help of in-field, non-destructive high throughput phenotyping (HTP) techniques (Normanly, 2012).

In a fully grown stage, one of the most important phenotypic trait of a cotton plant is the total boll count. For growers, boll count is the primary indicator of potential yield from the field and is hard to accurately predict manually. In addition, boll count can provide a better understanding about the physical and genetic conditions of the crop that can lead growers to assess growth conditions and take crucial decisions (Pabuayon et al., 2021). Apart from boll count, the cotton plant skeleton can be studied for obtaining structural information (Ritchie et al., 2007). Main constituents of the cotton plant skeleton are its main stalk, primary branches that starts from main stalk at the point called node, and the secondary branches(Twine & Redfern, 2021). Identification of main stalk and nodes can provide vital architectural information- such as the average plant size, inter-node distance, branch angles, etc.- that can be used for plant growth analysis as well as managing in-field operations (McCarthy et al., 2009). Therefore, obtaining such traits using HTP techniques that requires less time and labor are needed as well as widely studied (Pabuayon et al., 2019).

In the recent years, a vast amount of research has been carried out to extract various phenotypic traits of cotton plant with the help of deep learning and computer vision. (Jiang et al., 2020) proposed a method to detect and count emerging cotton blooms that can help characterizing flowering patterns efficiently. Apart from these, with the help of deep learning based methods different plant traits can be tracked and measured over time (Jiang et al., 2019; Petti & Li, 2021). (S. Sun et al., 2021; S. Sun et al., 2018) shows that along with the 2-D image data, in-field HTP can make the use of 3-D measurements to perform growth analysis. As the application changes, the underlying deep learning model gets complex, thus, managing the training along with inference time becomes challenging and can be addressed via speed-accuracy tradeoff (J. Huang et al., 2017). Furthermore, most of these methods perform extensive data acquisition and pre-processing to achieve the desired task which heavily rely on the highly accurate data annotations that are complex and time consuming. In plant phenotyping, though there are multiple ways to collect high-quality raw data in abundance, the precise annotation of this data has been a challenge due to factors such as domain knowledge, multi-modal input data, and application specific annotations.

To tackle the problem of scarce data annotations, several imaging methods rely on three major approaches based on low-quality labels or certain abstractions of the problem: using traditional image processing, unsupervised learning, and weakly supervised learning. Traditionally, these tasks are carried out with the help of various features of the object in the context such as color and shape information (J. Liu et al., 2011; Wei et al., 2008). To detect the cotton boll and predict the yield using that count, S. Sun et al., 2019 implemented a boll recognition and counting pipeline based on traditional image-processing algorithms; achieving 84.6% accuracy in boll counting. On the other hand, in the absence of data annotations, unsupervised methods aim to group the features using several clustering methods such as k-means (Ashapure et al., 2019), DB-SCAN (Li et al., 2016), and neighboring component analysis (Li et al., 2020). Most these methods require a little or no training at all, and thus, depends completely upon the input image features. In terms of in-field data collection, the imaging platforms such as robots, UAVs generates a higher amount of occlusions, noisy backgrounds, uncontrollable lighting conditions as compared to lab-based imaging. Additionally, to collect the cotton plant features with specific instruments such as

LIDARs, hyperspectral cameras requires extra cost in terms of hardware, working skill, data processing. In most of the cases unsupervised methods are tailor-made for a specific representation of data, but in case of cotton plant, identifying specific features and developing the algorithm that fits all the features is difficult. Moreover, the results obtained with such techniques are not guaranteed to meet supervised learning standards and performance can be enhanced significantly with the minor efforts in data annotations.

In the recent years, another promising way of addressing the availability of data annotations is to provide a weak supervision either with a partially annotated data or with pseudo-labels obtained from lower level labels. When the lower-level labels (image-level class annotations) are available, there are a plethora of approaches to localize and detect object instances (Zhang et al., 2021) that leverages two prominent paradigms: multiple-instance learning (MIL) (Andrews et al., 2003) and class feature activation maps (CAMs) (B. Zhou et al., 2016). MIL is based on learning object instances from positive (one or more instances present) or negative (no instances at all) bins that contains of samples from the given dataset. For example, an image is considered to be in a positive bin if at least one object instance is present in it, otherwise it is in the negative bin. While MIL based methods are widely used for weakly supervised object detection (C. Lin et al., 2020; H. Wang et al., 2021), using intermediate classifier activation layer's feature maps i.e. CAMs along with pseudo-label generation is gaining popularity due to its versatility for object detection as well as segmenting individual instances (Durand et al., 2017; J. Wang et al., 2018; Y. Zhou et al., 2018). CAMs allow us to understand the relation between a deep learning model and its output for the given image by inspecting the specific pixels that contributed more in the output. Most of the CAM based approaches require external proposal generator and a heuristic-based scoring mechanism to predict the final output that lack a unified framework. These issues can be tackled in several ways such as the introducing self-supervised deep neural nets that generates proposals (Shen et al., 2020), providing attention to generate pseudo-labels (L. Chen et al., 2021; Ou et al., 2021), using pseudo-labels to train stronger models such as Mask R-CNN (Laradji et al., 2019). All of these improvements are specific to the application and differ according to the domain of the data, however, they perform sufficiently to address the posed problem.

Agricultural researchers have started to make use of weakly supervised models to reduce the annotation efforts in a variety of applications that includes disease classification, yield estimation and counting, etc. Bollis et al., 2020 designed a CNN-based algorithm to automatically select the regions of interest (ROI) from citrus crop images that were damaged by pests and diseases. The algorithm uses MIL paradigm to classify those crops with the help of Saliency maps that significantly reduced the annotation costs. Ghosal et al., 2019 performed head detection and counting to understand the relation between phenotypic and genotypic traits of the sorghum crop. They demonstrated that it was possible to alleviate the annotation costs with the help of partially annotated dataset in a weakly supervised learning settings without compromising the final model performance. TasselNet (Lu et al., 2017) and its successors (Xiong et al., 2019) used a regressor to count maize tassels with the help of local density maps that achieved robust in-field counting with high accuracy. The idea of working in smaller sub-windows of an image so as to obtain fine density maps was proved to be efficient and integrated in our study. Another weakly supervised counting network PSSNet (Tong et al., 2021) used point-supervision to segment and count the trees in aerial images that converts feature maps to masks. This network outperformed state-of-the-art methods in most of the challenging conditions and greatly reduced human labor by generating masks automatically. As discussed previously, these approaches makes use of either CAM or MIL paradigms but (Yu et al., 2020) proposed a Minirhizotron image segmentation approach combining both MIL and CAM that outperformed standard weakly supervised semantic segmentation frameworks.

The effectiveness of above discussed MIL and CAM paradigms is totally based on domain of the application, and thus at this point, it will not be fair to compare them just on the basis of separate applications. Therefore, to test the suitability of either of the framework, our study uses methods based on both of them. Bellocchio et al., 2019 proposed a novel fruit count method for yield estimation based on spatial consistency loss which falls under MIL due to the nature of weak learning provided to the model. This weakly supervised counting (WS-COUNT) architecture performed exceptionally well in high density fruit counting and was able to achieve a performance similar to its fully supervised counterparts. Thus, for the boll counting application we chose WS-COUNT as one of the frameworks to study in detail. At the time of submission, there are a few studies involving some paradigm of weakly supervised learning applied to cotton plant phenotyping, for example, Z. Huang et al., 2020 proposed an algorithm based on density classification for in-field cotton boll counting with the help of high-dimensional feature maps. Though this method achieves better performance than most of the counting methods, this method is computationally expensive on both training and inference levels. Furthermore, the method uses the high-resolution images of cotton field taken from above the ground, and thus, counts the bolls that are visible in the top view making the method erroneous in the presence of dense boll distribution and occlusion. However, the success in counting bolls in the in-field images motivates us to test the more advanced weakly supervised paradigms for the cotton plant phenotyping.

Overall, cotton phenotyping from 2-D images faces several challenges with respect to data annotations and preparing the multi-purpose labelled dataset for various trait extraction. The availability of cotton plant datasets with partial or low-level annotations will enable the use of weakly supervised learning in alleviating these challenges. Furthermore, with a careful comparison between fully supervised methods, the potential of these weakly supervised methods to extract complex cotton plant phenotypic traits should be tested. Therefore, chapter chapter 3 aims to study the various traits of cotton plant extracted using deep learning with the limited availability of labelled data for training. Additionally, the application of weakly supervised learning based on partial labels, CAMs and MIL (Bellocchio et al., 2019; Cholakkal et al., 2019) is explored for counting the cotton bolls from the front view of cotton plant images taken in various environmental conditions at different resolutions.

Organization of Thesis

The rest of the thesis is organised as follows.

- In chapter chapter 2, *"Corn Image Instance Segmentation"* the study of image instance segmentation of corn images collected from GMO Corn experiment is presented. Based on image data annotation, we discussed two approaches for segmentation: identifying whole corn ears and bare ear parts with and without corn kernels. The Mask R-CNN model was trained for both approaches and segmentation results were compared. The ablation experiments were performed with the latter approach to obtain the best model with the available data. The estimation results of these models were included and compared with manually labelled test data with $R^2 = 0.99$ which showed that use of the Mask R-CNN model to estimate corn consumption provides highly accurate results, thus, allowing it to be used further on all collected data and find the answer to the raised hypothesis.

- In chapter chapter 3, *"Cotton Plant Phenotypic Trait Extraction"* the study of extracting cotton plant phenotypic traits with the help of supervised and weakly supervised methods is presented. This study demonstrates the use of Mask R-CNN as a supervised instance segmentation method to extract cotton plant phenotypic traits such as main stalk height, node and boll count. The study further focuses on the use of two weakly supervised methods: WS-COUNT and CountSeg to extract the cotton boll count under the presence of weak labels. Furthermore, the comparison between supervised and weakly supervised methods was done on the basis of counting performance as well as the efforts required in the data annotation. This allows us to improve the current supervised frameworks in the absence of high level instance annotations by leveraging the density maps obtained from weak supervision and can be extended to various cotton plant organs.

- In chapter chapter 4, we conclude the thesis while exploring the potential of the methods above and outlining the directions for the future work.

Chapter 2

Instance Segmentation to Estimate Consumption of Corn Ears by Wild Animals for GMO Preference Tests[1]

[1] Adke, Shrinidhi, et al. "Instance Segmentation to Estimate Consumption of Corn Ears by Wild Animals for GMO Preference Tests." Frontiers in artificial intelligence 3 (2020): 119.

2.1 Abstract

The Genetically Modified (GMO) Corn Experiment was performed to test the hypothesis that wild animals prefer Non-GMO corn and avoid eating GMO corn, which resulted in the collection of complex image data of consumed corn ears. This study develops a deep learning-based image processing pipeline that aims to estimate the consumption of corn by identifying corn and its bare ear from these images, which will aid in testing the hypothesis in the GMO Corn Experiment. This approach uses Mask Regional Convolutional Neural Network (Mask R-CNN) for the instance segmentation task. Based on image data annotation, two approaches for segmentation were discussed: identifying whole corn ears and bare ear parts with and without corn kernels. The Mask R-CNN model was trained for both approaches and segmentation results were compared. Out of the two, the latter approach, i.e. without the kernel, was chosen to estimate the corn consumption because of its superior segmentation performance and estimation accuracy. The ablation experiments were performed with the latter approach to obtain the best model with the available data. The estimation results of these models were included and compared with manually labelled test data with $R^2 = 0.99$ which showed that use of the Mask R-CNN model to estimate corn consumption provides highly accurate results, thus, allowing it to be used further on all collected data and help test the hypothesis that is the focus of the GMO Corn Experiment. These approaches may also be applied to new hypotheses relevant to crop yield, biotic and abiotic stresses, and phenotypes that are difficult to examine through traditional methods.

Keywords: deep learning, Mask R-CNN, instance segmentation, GMO, image processing

2.2 Introduction

Corn is one of the world's most important crops and is produced both traditionally and with genetically modified organisms (GMO) (FAOSTAT, 2018). To obtain certain agriculturally-desirable traits, such as resistance to pests, herbicide tolerance, and drought tolerance, specific corn varieties have been genetically engineered. Despite research on its safety and equivalence to traditional varieties, questions continue to be raised by members of the public regarding its safety and edibility. Since its introduction, there have been mixed views on GMO foods and crops and GMO corn is no exception for this. An early study summarizes the environmental benefits and risks of GMO corn (Gewin, 2003). In 2008, a United States (US) grower observed that mice preferred non-GMO corn over GMO corn (Roseboro, 2008). To test this, another grower repeated the experiment and published his results online stating that "The squirrel could have switched to GMO, but it did not. It knew it was different" (Roseboro, 2013). Another study done by an Italian group claimed GMO corn to be toxic (Séralini et al., 2012). This led to several studies and debates amongst research community and the corn industry (Butler, 2012). The initial study was retracted because of a lack of detailed analysis and insufficient evidences, but this did not stop further studies and debates on GMO corn. Further studies in favor and against GMO corn were done which is collectively reviewed by (Chassy & Tribe, 2010) and some of them clarified that *"animals are not biased to organic corn."*

During this time, the hypothesis was formulated that wild animals, specifically squirrels and deer, can sense differences between GMO and non-GMO corn and avoid GMO corn or prefer non-GMO corn when given a choice between the two. To test this hypothesis, a nation-wide community science project called "The GMO Corn Experiment" was started in 2015 that gathered the image data of GMO and non-CMO corn set outside by volunteers in their yards (Haro von Mogel & Bodnar, 2015). The next challenge of this experiment was to estimate the precise consumption of all the GMO and non-GMO corn samples based on the complex image data obtained to address the hypotheses being tested in the study. In the experiment, there were two choices of corn for animals kept side-by-side with bar-coded labels to keep volunteer community scientists blinded to the identity of the ears. On a larger scale, this study can serve as a model for studying animal preferences of GMO versus non-GMO food in a publicly-accessible manner. The experiment recruited volunteers ranging from families with children, to school classrooms, and adults, which necessitated striking a balance between simplicity and thoroughness in the data collection strategy. Volunteers were asked to take images of one side of the ears of corn before and after 24 hours, but had the option to add additional data and observations. This helped provide consistent samples of each ear of corn, however, the image data still presented challenges to analyze. Visual observations of the ears could be unreliable, as well as time-consuming and tedious to accurately estimate the consumption of each ear used in the experiment due to the large number of images and the time required for each image. Multiple observers would be needed to overcome individual biases, and any error could result in supporting a false hypothesis and affect future research. Finally, the wide range of image orientations, dimensions, quality, and lighting conditions would make traditional computer analyses difficult to perform, so a more robust method of analyzing community science-generated images was needed.

Computer vision can play an important role in estimating the consumption rate automatically. One of the challenges is to identify corn ears from the images and distinguish between the consumed and the non-consumed parts of the corn. Thus, a detection algorithm is needed that will distinguish between different parts of the corn, and based on its detection, compute the consumption percentage. Object detection and instance segmentation are two common categories in computer vision used to detect, classify, and segment images based on predefined/labelled classes. In object detection, the object's location in the given image is identified, whereas instance segmentation detects and delineates each distinct object of interest with the corresponding pixels in the image. There have been many studies on segmentation of plants/crops to detect different diseases and various image processing techniques have been reviewed by (Hamuda et al., 2016). For instance, one study used color transformation from the RGB to CIELAB color space to segment blight in corn leaves (Sukmana & Rahmanti, 2017) while another study used color features and K-means clustering to segment and identify crop diseases (Kumar & Jayasankar, 2019). Compared to traditional image processing, deep learning has shown promising results in fruit detection, plant phenotyping, and yield estimation tasks (Koirala et al., 2019) (Jiang, Li, et al., 2020). For example, one study proposed a method to augment training images using synthetic data to train a deep learning model Mask R-CNN to segment individual leaves of a plant (Ward et al., 2018).

Image instance segmentation was best suited to estimate the consumption of corn because it is crucial to know the exact area of the individual ear of corn and the consumed part of the corn. The problem is with the subset of instance segmentation, in which there may be more than one label for single pixel, a phenomenon known as multi-label segmentation. This poses challenges as most current algorithms cannot handle this type of task. Recent advances in Convolutional Neural Networks led to a variety of frameworks that can be used to perform instance segmentation on different levels (Hafiz & Bhat, 2020). One of the most successful and popular approach was the Mask R-CNN framework which efficiently detects the object while simultaneously generating a high quality segmentation mask for each instance (He et al., 2017). It achieved an average precision (AP) of 37.1% with a speed of 5 fps on benchmark datasets. Moreover, Mask R-CNN has proven efficient in segmenting leaves (Ward et al., 2018) and nuclei (Johnson, 2018) which relied upon a limited amount of data for training. In this paper, an automated algorithm is proposed that uses the masks produced by the Mask R-CNN method to estimate the area of the consumed part of the corn as well as the entire corn, which results in a percentage consumption/eaten. Specific objectives of this study were to:

1. Compare the data labelling approaches that were used for model training.
2. Perform ablation experiments for model training parameters and identify an optimal training data size.
3. Evaluate the model performance in segmentation using manually labelled ground truth.

2.3 Materials and methods

The proposed workflow is presented in Fig. 2.1 to estimate corn consumption from raw images collected. This involves three primary tasks : 1) Raw data labelling, 2) Training Mask R-CNN to segment corn instances, 3) Consumption estimation from segmentation results.

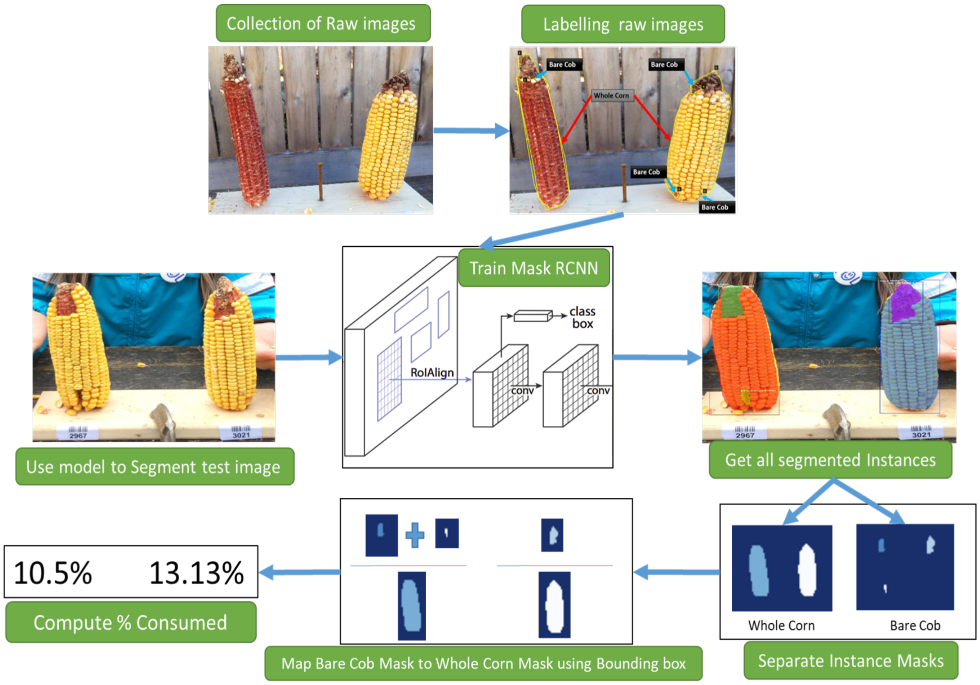

Figure 2.1: **An overview of stages of consumption estimation from raw images.** Collected raw images are labelled manually and used to train the Mask R-CNN model. The instances segmented by the model and their masks are then processed to estimate the consumption.

2.3.1 Image Collection and Annotation

Data Source and Pre-screening

Experiment kits were distributed to volunteer community (non-academic) scientists in the United States containing two experiments, each consisting of a pair of size-matched GMO and non-GMO corn, a feeding stand, and instructions for conducting the experiments. They placed experimental setups in their backyard or some kind of open space, thereby offering one GMO and one non-GMO corn to wild animals in the same environmental conditions, taking observations at 24 hour intervals. Each ear of corn was labelled with an unique bar code without cultivar information to avoid potential human bias. A total of 630 images were provided at the start of the project, which needed to be analyzed in a reliable and repeatable manner. Based on visual observations, pairs of before/after images from each experiment were pre-screened, and if both of the ears of corn in the experiment were consumed between 0-5%, 5-95% or 95-100%, that image was placed in its respective category. In the case of mixed consumption, they were labeled as 5-95%. This approach allowed for the consideration of the potential variations in consumption for training the segmentation model. Table 2.1 shows the category-wise statistics of the images considered for further annotation stage.

Table 2.1: **Image categorization after pre-screening.**

Category	Total Images	Selected for Annotation	Selected for Training
0 - 5%	210	143	95
5 - 95%	370	257	175
95 -100%	50	50	30
Total	630	450	300

Multiple images were gathered from a single kit over the course of 24 hour intervals (e.g., the same corn ear would be consumed more on the second day than on the first day). For training purposes, images of the same ears of corn at different phases of the experiment were considered (initial, intermediate, final consumed image). This allowed more data to be collected with fewer experiment kits. After collecting the raw images in various conditions, a total of 450 images were selected from these categories for manual annotation and labelling. In this dataset, ambiguous images are those in which it could not be identified whether the corn is present or not, and if present,consumption could not be estimated. Based on these considerations, images for training was was selected based on following criteria:

- Corn ears present in the image should be identified easily by the human eye.
- The image should be high-resolution and have legible brightness and contrast.
- The skewing of the image (i.e., rotating the image at various angles, changing the brightness/contrast, applying blur) should not result in ambiguous image.
- The image can contain other objects than corn such as a chair, table, person, toy, etc.

Labelling Approaches

The sorted image data were then labelled using VGG's Image Annotator(VIA) tool (Dutta et al., 2016). For this study the whole corn in the image as well as the consumed parts of the corn were needed to be identified. To achieve this, the masks were labelled in many different ways. However, predicting exact and adequate segments posed the biggest challenge to the estimates. The output of the model is based on the segments and class masks given for training. To estimate consumption, it is possible to compare the eaten part, i.e., the bare part of corn ear, with the total visible part of the corn ear. Also, an individual corn kernel or cluster of kernels might have been considered as separate classes. Two of many possible approaches that consider two distinct classes were attempted.

First, the whole corn and bare corn ear were considered as two distinct classes (Approach 1), while in the second approach, the clusters of intact corn kernels along with bare corn ears were considered as two separate classes to segment (Approach 2). The two approaches differ at the image labelling level. For Approach 1, masks were drawn for whole corn ears and bare ear parts were used for consumption estimation, while in Approach 2, masks were drawn for corn kernels and bare ear parts, the sum of which equals a whole visible corn ear (which is not considered as a separate class for segmentation but computed later by adding these two classes). Fig. 2.2 illustrates sample images from the training dataset for both approaches.

After performing the experiments with these approaches, which are discussed in later sections of this paper, manual labeling of all 450 images was continued for Approach 1. Table 2.2 provides the details of data partitions and the corresponding number of labelled masks for Approach 1. It should be noted that a normal image in the dataset has two corn ears. The table shows that there are certain images present in the dataset, which contains only one corn ear. These images may not be useful for comparing two corn ears, but they were selected to improve image segmentation accuracy and to avoid overfitting the model to specific features in images having two corn ears.

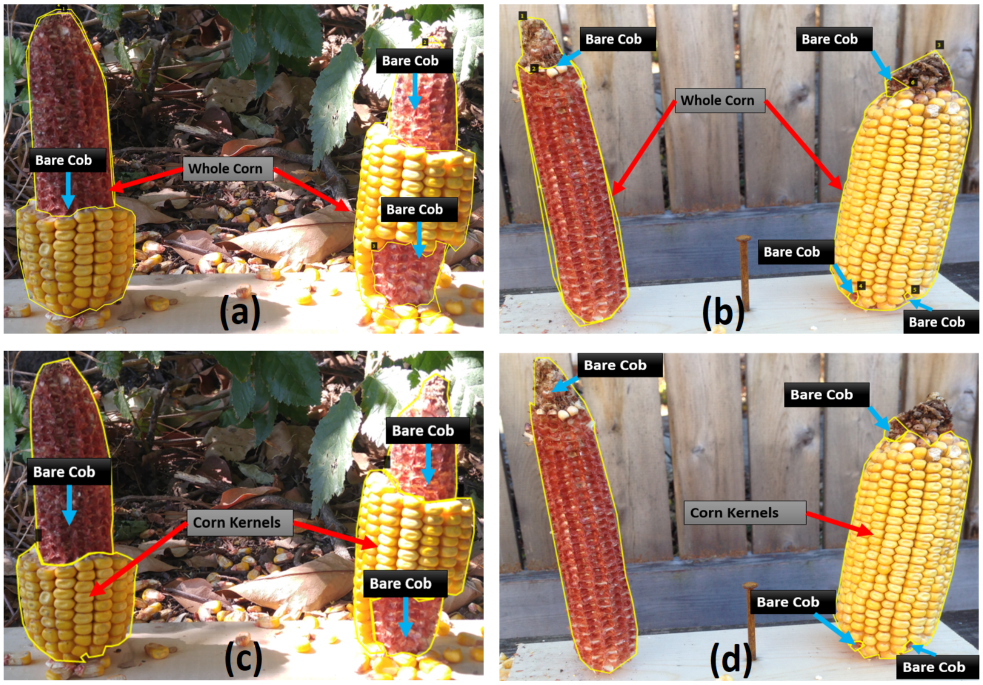

Figure 2.2: **Illustration of two data labelling approaches.** Representative images with different levels of consumption and corresponding masks. In Approach 1 (**a-b**), the visible part of a whole corn ear and distinguishable bare part segments are labelled as two mask classes. In Approach 2 (**c-d**), intact corn kernels and bare part segments are labelled as two mask classes.

Table 2.2: **Data partitioning and number of masks for approach 1.**

Labelled dataset	**# of images**	**# of corn masks**	**# of bare ear masks**
Training	300	593	846
Validation	125	248	408
Testing	25	50	79

2.3.2 Corn Segmentation

Mask R-CNN

As the Mask R-CNN architecture has proven successful in a wide range of applications requiring instance segmentation, it was chosen for the present study. It consists of two stages: the first stage scans the image and generates areas with a high probability of containing an object of interest, often referred to as *proposals*, and the second stage is responsible for the classification of these proposals to generate bounding boxes and *masks* for each detected object. To build the basic network, we used Matterport's (Abdulla, 2017) implementation of Mask R-CNN and performed ablation experiments with modified configurations.

Based on experiments with this implementation for different segmentation tasks such as deep leaf segmentation (Ward et al., 2018) and nuclei segmentation (Johnson, 2018; Naylor et al., 2018), widely used common parameters were chosen for the training. Also, ResNet-50 was (He et al., 2016) as the backbone network to detect features. To improve standard feature extraction, a feature pyramid network (FPN) (T.-Y. Lin et al., 2017) was added to the network. Preliminary trials with small samples were conducted to tune the hyper-parameters such as learning rate, non-max suppression threshold, and training ROIs per image. The remaining parameters were left unchanged from Matterport's original implementation. The training followed a predetermined backpropagation schedule with a stochastic gradient descent(SGD) optimizer, L2 Regularization with a weight decay of 0.0001, a learning rate of 0.001, and cross entropy loss functions for various losses in the network.

During the preliminary trials, the model tended to overfit because of a small sample size of 50, as the model learned the features only specific to those training images, such as the number of corn ears present and the position and alignment of corn ears in the picture. Therefore, to prevent model overfitting, image augmentation (Jung, 2018) methods were used to enhance the dataset diversity. These augmentations include random image flips (left/right/up/down) and rotations (90°, 180°, 270°) along with Gaussian blur, color multipliers. The gaussian blur and color multiplier takes care of variations in the dataset such as image focus, distance(pan/zoom) of object from camera, color variations in objects as well as background. Also, to avoid having the images with similar test-kit positions i.e. two vertical ears in the center of the image, we considered the images with skewed ear positions, various image angles and kits having only one corn ear in the training dataset.

The dataset has images of different resolutions ranging from 640×480 to 4080×3072 with an average of 1280×760. For training, the input image size was limited to 1024×1024 with the help of Matterport's utility methods that uses the standard bilinear interpolation to resize the image. A batch size of 2 was used because of the GPU memory limit (NVIDIA GeForce GTX1080Ti) and generally, steps per epoch are decided by batch size along with number of training samples. In this work, training configurations were optimized by observing the differences between training and learning methods and to observe which one can perform better in terms of estimating the corn consumption.

Ablation Experiments

To begin the consumption estimation, a better segmentation model was trained by performing various ablation experiments, which led to the effective end model used for testing. Labelling approach and training sample size were the two primary factors considered in the ablation experiments.

1) Labelling Approach comparison.

The overall segmentation problem was simplified at the data labelling level. All the raw images were labelled according to one of the two approaches explained in the previous section. Instead of labelling the *entire* dataset twice to arrive at a better labelling approach for further experiments, this test was performed at the beginning with a smaller dataset. For initial comparison purposes, out of the 450 raw images, 70 were selected and manually labelled using both approaches. For this experiment, 50 training images and 20 validation images were used. Identical network configurations were selected for both labelling methods. This experiment indicated which labelling approach to follow for the remaining raw images.

Labelling each image for the two approaches took considerable time. In a standard image with two corn ears, both class instances could be labelled in one image within 3 minutes on average using Approach 1, while Approach 2 took 4.5 minutes on average. This was because, in a single image total instances of corn kernels can be more than total whole corn instances. There will be at most two whole corn instances but can be zero or multiple corn kernels present in one image, labelling multiple corn kernel instances contributed to more time in Approach 2. In the end, the Mask R-CNN model trained using images labelled by approach 1 were referred to as *Model 1* while the one by approach 2 as *Model 2*.

2) Training sample size effect.

One of the challenges in training effective deep learning models is the limited amount of training data. The sparsity of labelled images in the agricultural domain is a common problem for segmentation model failures resulting from overfitting to a small sample size. The number of training images required is not fixed, but they are domain and application specific. To ascertain the minimum number of training images for a good segmentation performance, this experiment was performed.

To answer this question, multiple models were trained with a different number of training images with an approach selected from the above comparison. Starting with 50 images, models were trained on increments of 50 images, up to 300 images, while keeping an uniform size in the validation dataset. In preliminary tests, it was observed that, when the selected training samples contained only the images from a certain category (e.g., “95-100% consumed"), the resultant model performed poorly on the remaining categories, which resulted in inaccurate image segmentation. To address this data imbalance, each training procedure was performed five times by selecting the training images randomly from each category. Thus, a total of 30 different segmentation models were trained with 6 different sample sizes.

Apart from the major experiments mentioned above, the transfer learning phenomenon applied to this use case was also examined. During preliminary testing, one of the Mask R-CNN model was trained by initializing random weights at the beginning. Then, this model was compared to a model trained on pre-trained weights on the MS-COCO dataset (T.-Y. Lin et al., 2014). For better segmentation of the background from the corn ears in the image, further models were trained using COCO initial weights.

Evaluation Metrics

To evaluate the performance of the above ablation experiments, a comparison was made of the training procedures as well as the results on labelled test dataset. Below are the metrics considered for evaluating these experiments.

1) Jaccard Index.

It is important to predict the mask accurately as the area to be calculated is based on the mask. This can be verified by standard metrics of Jaccard Index, a.k.a. Intersection over Union (IoU), which is the ratio between the overlap of a predicted mask and the actual ground truth mask and the area of union between the predicted mask and the ground truth mask. Then the weighted mean IoU was computed for all the predicted instances of each class in an image and average all the images in the dataset. In both models, class *"bare ear"* is present, while *Model 1* has *"whole corn"* and *Model 2* has *"corn kernel"* as the second class, respectively. Mean IoU will be a key metric used to evaluate the segmentation performance of these models.

2) Mean Average Precision.

For instance segmentation, it is important to compute the precision and recall achieved by the model in addition to IoU. Precision is defined as a ratio of true positives over both true and false positives. Recall is defined as the ratio of true positives over both true positives and false negatives. Precision and recall are computed over a range of different IoU thresholds (typically 0.5 to 0.95 in steps of 0.05). The average precision (AP) is the averaged precision for all classes for one input image. The mean of APs is the mean average precision (mAP) over all images.

3) PR Curve.

Along with precision values, the recall of these models can be visualized better in terms of a precision-recall (PR) Curve. The area under the PR curve for a certain IoU threshold is nothing but the mAP for that model. We can plot the PR curve for different IoU threshold values that are especially close to the model's mean IoU to ascertain how well it is performing on all of the ground truth instances.

The main criteria for selection of a better labelling approach involves a high mean IoU value, a near ideal PR curve for different IoU thresholds, and thus a high mAP value. Additionally, it is important to consider the complexity in computing the consumption ratio with segmentation masks of two different labelling approaches. The ideal approach should be less complex in terms of detection of instances. It should present a significant overlap of predictions and ground truth masks.

2.3.3 Corn Consumption Estimation and Comparison

Calculation of Corn Consumption Ratio

Once a network was trained to segment the required classes, distinct corn ears in the image would need to be identified to estimate the consumption. For this, a straightforward method was used for preparing a list of all distinct class instances and compute the sum of individual pixels. The detection results of the model provides all instances of a whole corn, the consumed part of the corn, and corn kernels (in Approach 2). Each instance has its mask pixels and the bounding box coordinates. This output was further processed using custom Python scripts to arrive at final consumption estimations.

After segmenting the image into the whole corn and its corresponding parts — bare ear part and corn kernels (as illustrated in Fig. 2.1), the next step is to group the segmentation results to map all parts to respective corn ears. This is done by using the bounding box created by Mask R-CNN's bounding box detection layer. For example in Approach 1, the individual whole corn instances are first separated and then grouped among all other segmented instances of bare ear parts in their bounding box. This provides mapping of all whole corn instances and their bare ear parts.

Finally, the consumption was determined by calculating the ratio of *total pixels of all individual consumed or bare ear parts of the corn* over *total pixels of that corn*. (In Approach 2, the total pixels of corn is the sum of the pixels of all bare ear parts and the pixels of all kernel parts).

Evaluation Methods

During the pre-screening and categorizing stage, the images were manually annotated by authors. To verify the consumption value obtained from manual annotations, five human observers rated the test images. The averages of all observed consumptions were verified with the values obtained from manual annotations. In this comparison, the manual annotations were considered as ground truth for consumption estimation.

The consumption was calculated individually for each segmented corn ear. It should be noted that a variation of $\pm 4\%$ in consumption estimation was observed because of configurations of the test environment. This can be visualized by a scatter plot with a linearly fitted line. We observed failure in segmentation for a few of the models trained on less samples, which led to inaccurate consumption estimations. These cases were treated as outliers while evaluating consumption estimations. The outliers were not considered in fitting the line. The consumption of left and right corn ears was then compared, and the results can be viewed in the confusion matrix compared from human observations.

2.4 Results and Discussion

2.4.1 Labelling approach comparison

The metric values for the two annotation approaches were compared and the aim was to select one labelling approach that will be used for further experimentation and labelling the required training data. It can be seen that *Model 1* gave better results on the overall test dataset (Table 2.3). Figure 2.3 shows example outputs using the two approaches. Approach 2 missed a significant portion of both classes that can be seen by observing the instance masks. In addition, Approach 2 segmented the inner bare part as both classes due to its relatively small size and surrounding kernels, which was a false positive for corn kernel class and can result in an inaccurate estimation of consumption.

Table 2.3: **Labelling approach comparison**

Metric	Model 1	Model 2
mean IoU for bare ear	0.61	0.51
mean IoU for corn kernel	-	0.77
mean IoU for whole corn	0.87	0.64
mAP	0.57	0.48

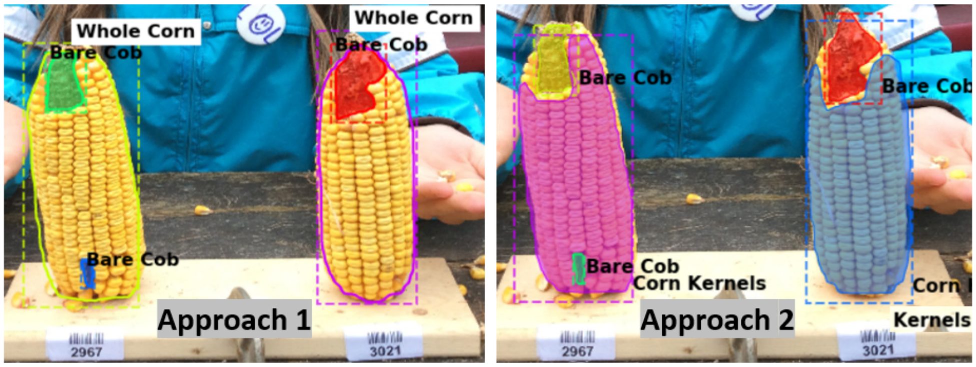

Figure 2.3: **Segmentation using two approaches.** For better visualization masks are not shown for the whole corn class in Approach 1 (left), while both class masks are shown in Approach 2 (right).

Overall, labelling Approach 1 led to a better performance than Approach 2 which can be seen from above results. Compared with Approach 2, Approach 1 increased the segmentation accuracy by approximately 10% and 23% for bare ear and whole corn, respectively, which were substantial improvements for a two-class segmentation problem. This occurred primarily because segmentation of bare ear and whole corn was simpler than that of bare ear and corn kernels. First, a whole corn ear had a relatively predictable conical shape regardless of how much of it is consumed, whereas corn kernel parts could be in any shapes and locations based on the consumption of the ear. Given the same number of training images for a whole corn ear, a Mask R-CNN model could easily learn adequate feature representations, resulting in better segmentation accuracy. Second, Mask R-CNN could not achieve a perfect segmentation of objects with complex boundaries, such as bare ear and corn kernel parts both of which had variable instance mask boundaries that were not as obvious as the conical whole corn. In particular, the boundaries of the two parts were dramatically variable because of natural uncertainties in the experiments such as ear placement height and ear size. Additionally, since it was not possible to predict how the corn ears were consumed by animals and thus the remaining bare parts and kernel parts could vary. Approach 2 included both classes (bare ear and corn kernels) and increased the difficulty of training a model for such a segmentation task. In addition, the two annotation methods had different labeling cost and model training time.

Proceeding with Approach 1, it was determined that the consumption estimation was faster for all the different image types because of the easy computation involved in detecting the whole corn area. Approach 2 required extra computations to identify the appropriate kernels and bare parts for a single corn based on its bounding box. This was not the case with Approach 1, since it gives the bounding box of a whole corn that already has the bare part. Thus for initial experimentation purposes, Approach 1 performed considerably well.

2.4.2 Effect of sample size

Various models with different training sample sizes were compared in terms of key performance metrics (Table 2.4). As more images were added to training, the segmentation improved generally. For bare ear class, the model with a sample size of 150 slightly under performed compared to model with 100 samples and thus lowering the total mAP by 0.014. With two-sided t-tests (Table S1-S3), the performance metrics were statistically analyzed to find the significant improvements corresponding to various sample sizes. Sample increment in most of the early sample sizes found to be insignificant with respect to IoU. However, we achieved significant improvement in mAP with all 300 images. Therefore, the use of all 300 training images is beneficial to the best performance in the present study. In the future, annotating 50 to 150 images would initiate a good baseline model at an affordable cost. With active learning methods, additional instances important to model performance improvements could be identified and labeled with minimized human efforts.

Table 2.4: **Effect of training sample size on segmentation performance (mean±standard deviation)**

Sample size	**50**	**100**	**150**	**200**	**250**	**300**
Bare ear IoU	0.606±0.04	0.636±0.02	0.626±0.03	0.661±0.02	0.656±0.04	0.670±0.02
Whole corn IoU	0.868±0.01	0.876±0.01	0.885±0.01	0.886±0.01	0.888±0.01	0.893±0.01
mAP	0.574±0.01	0.592±0.01	0.588±0.02	0.608±0.01	0.609±0.01	0.610±0.01

Drawing the PR curves using these models can provide a more informative comparison of the segmentation performance of the models. Figure 2.4 shows the PR curve at higher IoU threshold values than the one used to obtain the mean IoU for both the classes shown in the above table. This was specifically done to observe the model performance in case a high IoU thresholds are considered at the time of segmentation. A model with 100 images can be considered as poor compared to a model with 300 training images, but models with 250 and 300 images performs nearly identical to each other in the segmentation task.

From these observations, intuitively it is possible to answer the question of the number of images required for training. For segmentation of the dataset, all of the considered metrics showed that there was little difference in performance from 250 images to 300 images. There is no standard (fixed) ratio for selecting this number in most of the domains and thus we find the sufficient training size by performing these experiments. Hence, it can be said that, about 200 to 300 images are required for adequate segmentation in this case. Figure 2.5 shows some of the segmentation outputs from the test dataset. It can be seen that this model identified the bare parts very accurately, which is the key for calculating the consumption of corn ears.

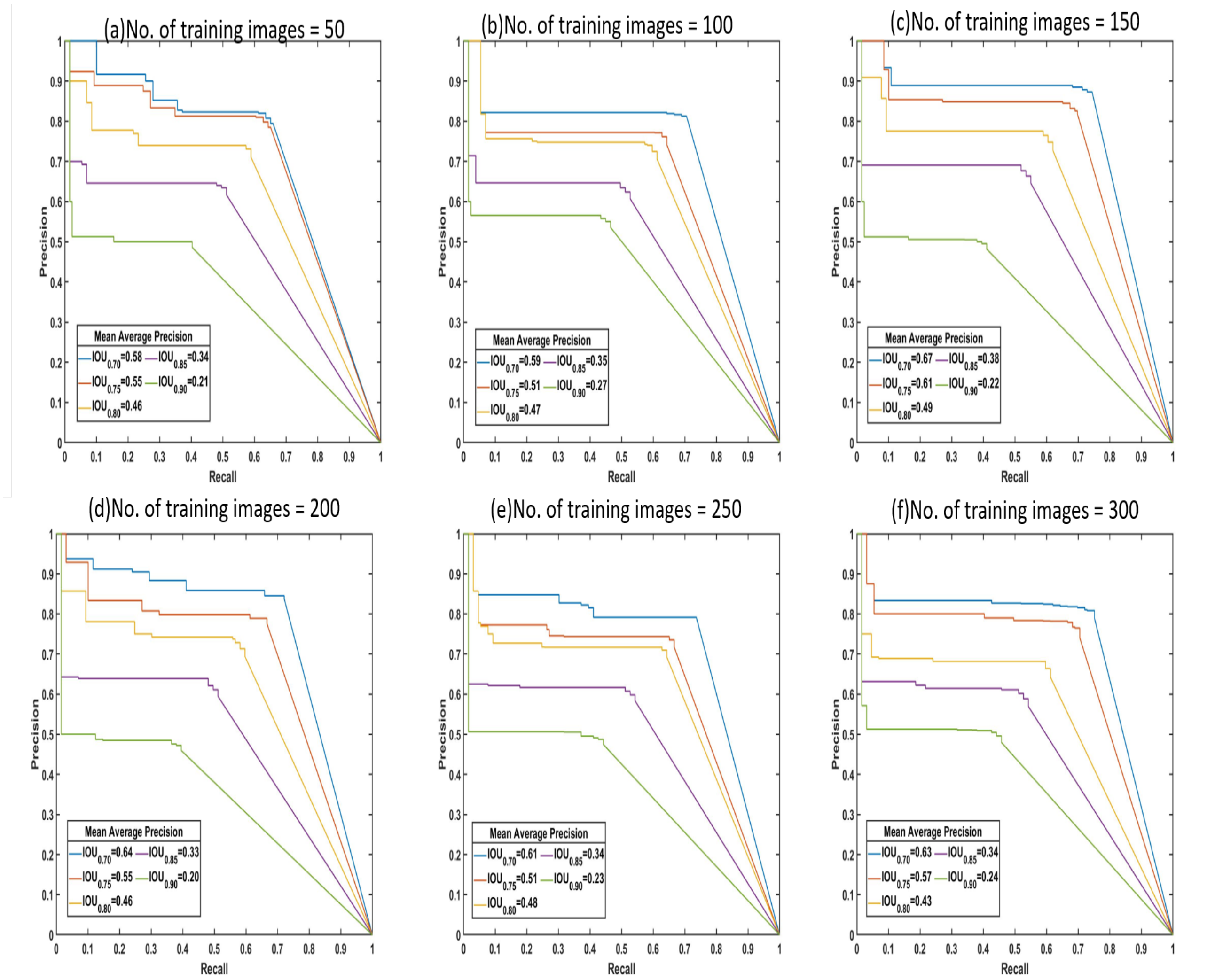

Figure 2.4: **PR curves.** These are standard PR curves with IoU thresholds and corresponding AP values. As we increase the threshold, the area under the curve decreases, and after a certain number of training images, there is less variance.

2.4.3 Corn consumption estimation

The consumption values of 50 individual corn ears were obtained from the segmentation results of 25 test images and comparison was done with the average of human labelled ground truth values. The test

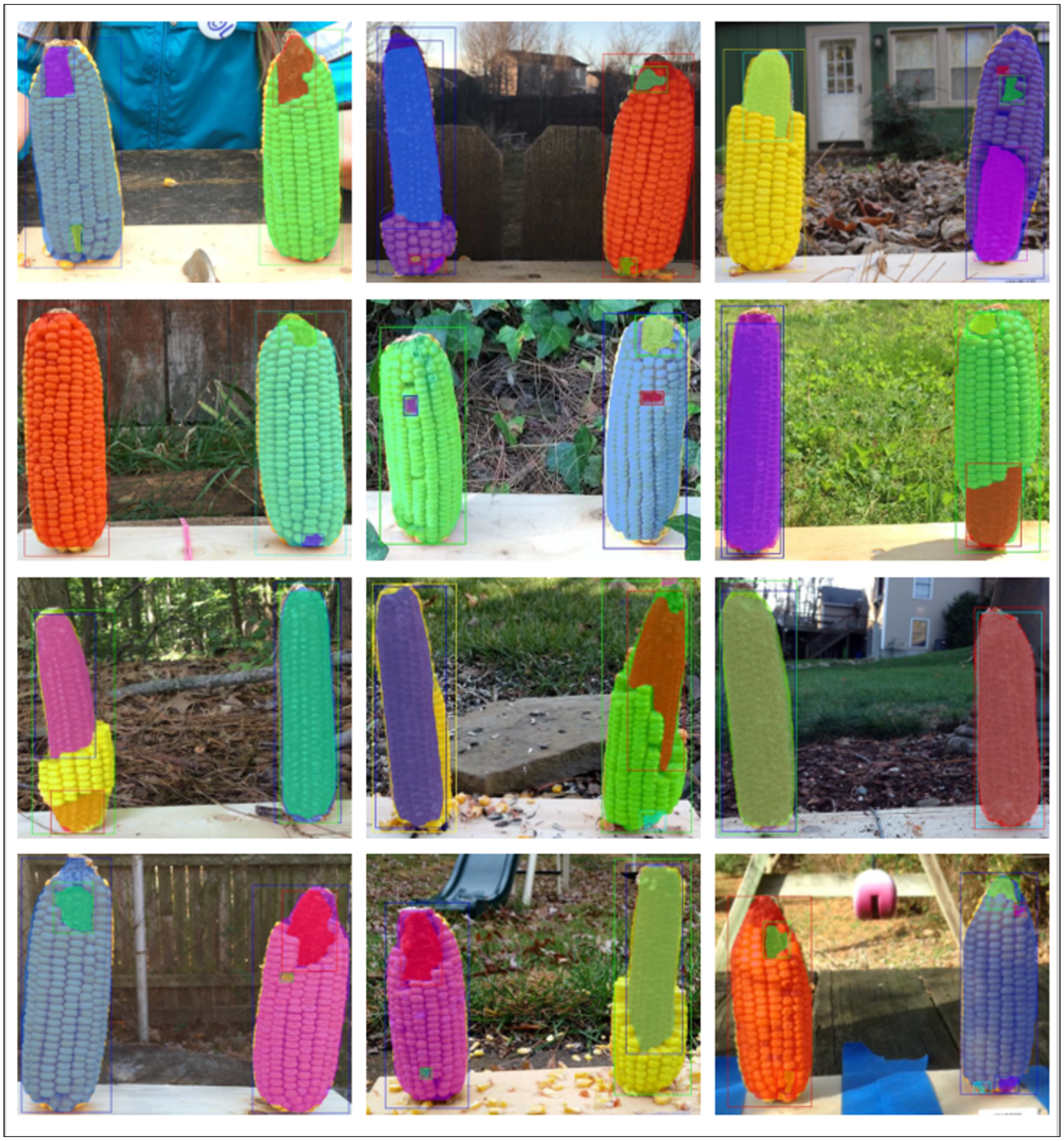

Figure 2.5: **Segmentation results.** Representative examples from test data with predicted masks.

images were annotated with Approach 1 and the consumption was calculated using these manually labelled masks. The results of this comparison showed that all of these models performed well on the test dataset but the models trained with lower number of images had a few outliers (Figure 2.6). The best R^2 value of 0.9929 was achieved with models trained on 300 samples. There was not much difference in consumption calculations with these models, but it can be seen that there were more instances of failure in segmentation for the models trained with fewer images.

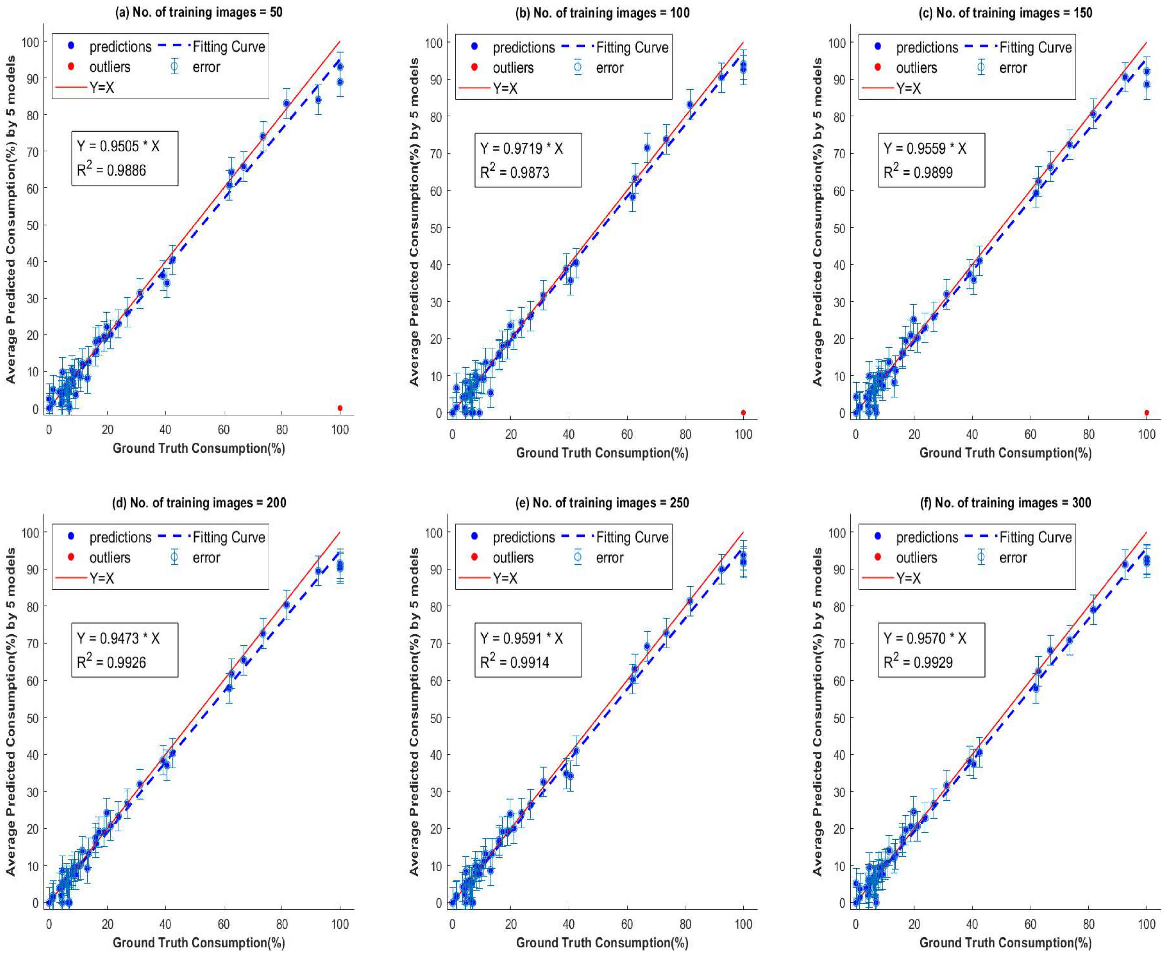

Figure 2.6: **Scatter plots of predicted vs. ground truth values of consumption ratio using models with different number of training images**. The outliers shown in red are estimations when the model has failed to predict the class instance mask and they were not considered for computing the R^2 values.

Outlier analysis

When the training set was smaller in size, there is a missed segmentation of bare ear with *Approach 1*. Figure 2.7 shows the 2 images in which the right corn ear was not identified as *Whole corn* and *Bare ear* (which are the two classes we have considered). As per Approach 1, these corn ears belong to both the categories but the models with fewer training samples could identify that corn ear in only one of the classes, i.e., whole corn, because of which the predicted consumption is zero. As the segmentation itself failed, this is an outlier in consumption estimation and thus omitted from the scatter plots discussed above.

It can be seen that the right side corn ear in both images has a similar shaped structure in the background — a tree stem attached directly to the corn ear. This could be one of the reasons it is ambiguous at the pixel level. Also, these test images belong to 95-100% consumed category, for which we did not have abundant training data. In the total of 300 training images, we only had 10% of such images, and may be fewer than this when we randomly select a smaller sample size (50-150). This imbalanced training data can be fixed by adding more of such training samples, as shown in the model trained with 300 images, where we did not observe outliers.

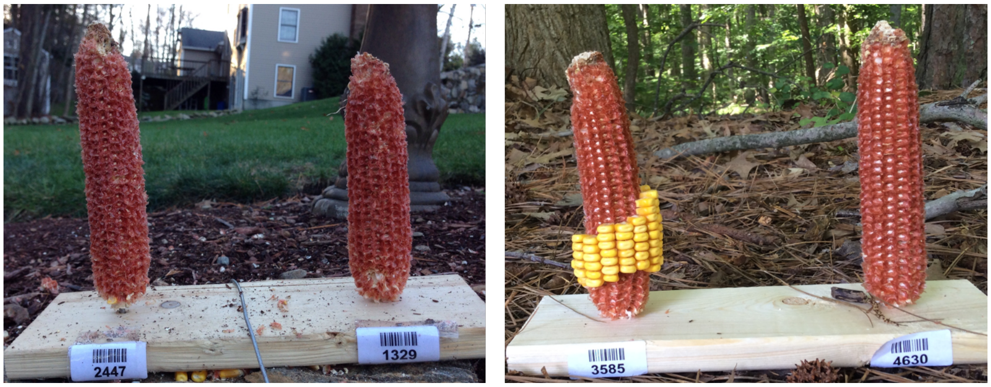

Figure 2.7: **Failed segmentation examples.** The right ears in both images were completely consumed, but the models failed to identify it as both whole corn and bare ear part.

Left-right comparison

The GMO and Non-GMO ears were randomly placed by participants on each feeding stand, and the identity of both corn ears in each image was blinded. To perform an unbiased comparison between ears, the model's output was compared for the consumption of left and right corn ears within each image. The 25 test images were then classified in three categories— right, left, and equal— representing on which side the corn consumed more. The results can be seen in the confusion matrix — among the 25 test images, 22 classified images matched with the ground truth (Figure 2.8). It should be noted that the remaining three images were misclassified by marginal differences between ground truth and predicted consumption estimation. For example, the image with equal consumption had both the corn ears 100% eaten, while the predicted consumption was 99% and 98% for the right and left ear respectively.

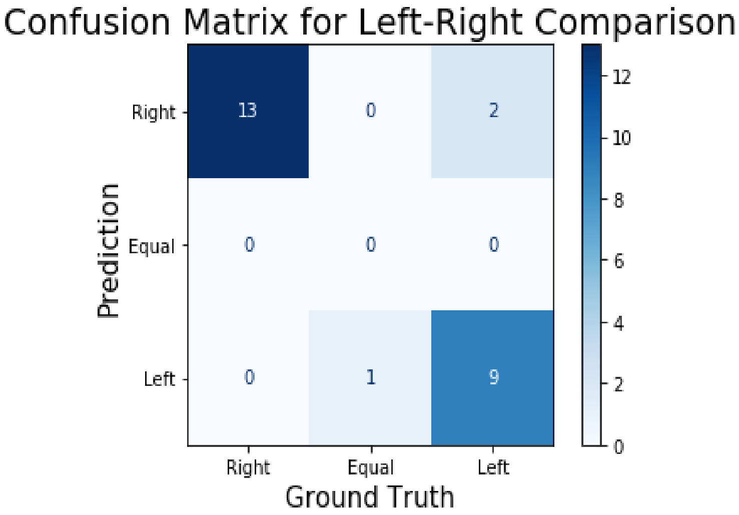

Figure 2.8: **Confusion matrix for left-right corn ear consumption comparison.** In ground truth, there were 13 images in which the right ear is eaten more, one image in which both are equally consumed (as shown in Figure 2.7), and 11 images in which left ear is consumed more.

2.4.4 Segmentation of ambiguous data

As stated in section 2.3.1, the images that fits certain criteria were used to prepare the models discussed so far. Furthermore, the images in this actual case can have various abnormalities, such as varying light exposure, ambiguous corn ears, varying backgrounds, among others. To verify the performance of this model, a manually generated set of 20 images having corn ears in varying conditions was used and performed consumption estimations of these images. When the consumption of the same set of corn was computed from images taken from different perspectives, it was observed that there was a greater difference between the predicted consumption value and the ground truth. The R^2 value achieved was 0.88 for this dataset with more than 10 outliers.

Representative segmentation results illustrate that the effect of change in perspective as well as effect of brightness on segmentation (Figure 2.9). The top 2 images are from the same corn ear set, but in the

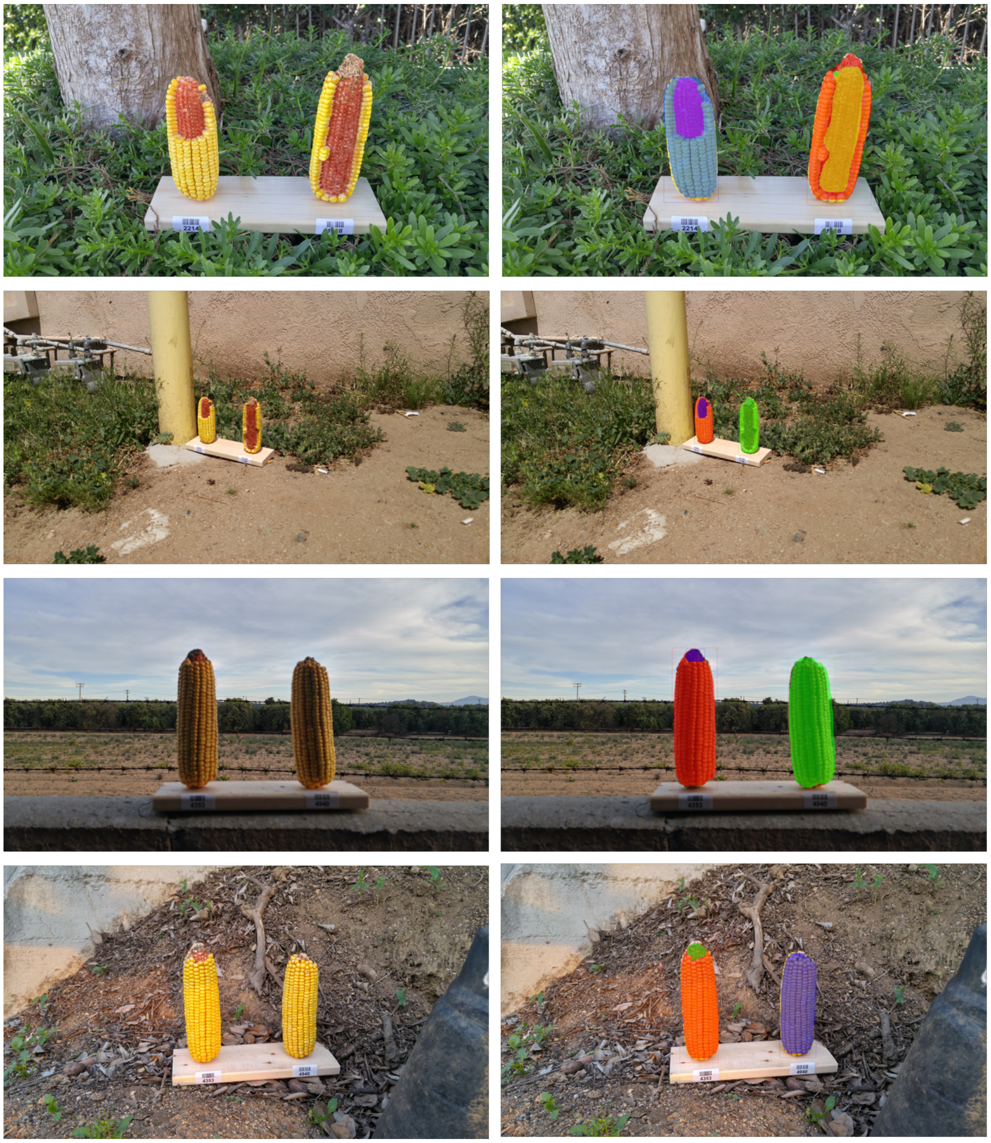

Figure 2.9: **Effect of background and resolution change.** Left column shows the original image while right shows the corresponding masks. Top two rows are the same set of ears and bottom two rows are the same, but with different background and resolutions.

2nd row, we can see that the model failed to segment the right side corn due to its low resolution of the actual corn ear (the corn ears only take a very small portion of the whole image). On the other hand, the bottom 2 rows shows that if an image is taken from an appropriate distance, the model performs well in different conditions of light. Though the training data do not contain such images, the model can still correctly segment these instances and the performance of the models can be improved further by adding more of such images in training.

2.4.5 Discussion

Since the inception of this project in 2018, a variety of new instance segmentation frameworks have been proposed as surveyed by (Hafiz & Bhat, 2020). For example, Mask Scoring RCNN that also makes segmentation based on detection of region proposals extends further with the inclusion of mask overlap scores and has slightly surpassed Mask R-CNN to achieve 39.6% AP on the same datasets (Z. Huang et al., 2019). Another region proposal-based Path Aggregation Network–PANet achieved top performance in segmentation that enhances feature hierarchy by path augmentation (S. Liu et al., 2018). Although the above mentioned frameworks perform better in terms of accuracy, the speed of detection remains an issue when the real time segmentation is to be performed. YOLACT (Bolya et al., 2019a) and YOLACT++ (Bolya et al., 2019b) addresses the segmentation speed at the cost of a reduction in AP by prototyping the masks and producing the instance masks with previously predicted mask coefficients. Most recent methods SOLO (X. Wang et al., 2019) and SOLOv2 (X. Wang et al., 2020) that addresses both speed and AP provides a simple, fast yet strong segmentation framework. This framework follows a rather unconventional approach to assign each pixel a “instance-category" to modify segmentation into a classification-solvable problem. To explore other segmentation approaches that are more recent and are lighter than Mask R-CNN, we performed experiments on available datasets using SoloV2 (Table S4). It was observed that SoloV2 segments the Whole Corn class with good precision but does not perform well for the Bare Ear class. The result suggests that Mask R-CNN still remains to be a robust instance segmentation method and it performed satisfactorily with a small amount of data. In future studies, however, other newer instance segmentation models could be further explored to improve the segmentation performance.

In this study, the proposed instance image segmentation approach measured how much corn was removed from an ear by comparing bare ear to kernels. The same method could be used to measure yield losses where ears of corn are damaged by hail or partially eaten by pests and wild animals. If further developed by providing appropriate training data, the trained deep learning model could learn to identify particular types of damage. In participatory breeding, some field trials are conducted at remote locations and there is the challenge of measuring phenotypes that would otherwise be easy to measure with people and equipment available at one's home institution (Ceccarelli, 2015). The method developed in this study would allow farmers and others who are monitoring remote locations to be able to collect images that could be turned into useful data after the images are analyzed by similar deep learning models presented in this study with further hyper-parameter tuning. A tool could be developed from this study that would allow farmers to take photos of damaged ears and quantify how much yield loss was caused by pests and disease during ear development (Steinke et al., 2017). It should be noted that plant breeding is already incorporating machine learning approaches to analyze and predict phenotypes (Jiang, Li, et al., 2020; Singh et al., 2016), but the proposed approach is unique because it can utilize field-collected image data with varying angles, orientations, and lighting with non-standardized resolutions and uncontrolled background. Therefore, this approach can be useful for decentralized and participatory crop research and breeding.

2.5 Conclusions

In this work, a deep learning based framework to quantify the consumption of corn with a relatively small number of images collected by community scientists was presented and evaluated. The Mask R-CNN model was demonstrated to produce high quality results of pixel-wise segmentation for the challenging task of multi-label segmentation of consumed corn. The two approaches for labelling the ground truth were presented, and it was found that segmenting only the whole corn and its consumed part is sufficient for estimating consumption. The best results were obtained when the training data were sufficient and labelled with high accuracy. The effects of varying light conditions and background were examined and it was found that the Mask R-CNN model, which was not trained with such images, was able to identify certain segmentation instances accurately, and can be improved upon by including such images for further training. The framework developed in this study can be used to predict more samples collected in the

GMO Corn Experiment and will produce reliable results more efficiently than manual labeling. Future work will be directed at improving the variation in accuracy as well as testing the visually challenging images, and toward applying the methods developed here along with additional lines of evidence to test the hypotheses that are the focus of the GMO Corn Experiment.

Chapter 3

Supervised and weakly supervised deep learning for segmentation, detection and counting of cotton organs using proximal imagery.[1]

[1] Adke, Shrinidhi, and Li, Changying. To be submitted to MDPI Sensors Journal.

3.1 Abstract

Cotton plant phenotyping has many advantages in terms of cotton yield estimation and harvesting. Study of plant organ architecture such as node count, main stalk height along with boll count provides important phenotypic traits that helps researchers as well as growers in planning. With the advances in image processing and modern computer vision, numerous supervised learning approaches have been implemented to perform these tasks from field images. The supervised segmentation requires large amount of precisely labelled image data. However, it is not always possible to gather and label required amount of images. In such cases, weakly supervised algorithms can be employed and used to obtain quality results. In this study, it is shown that with a small labelled data of 175 images, supervised approaches such as Mask R-CNN are effective for detection and segmentation of cotton boll, plant main stalk, nodes to count and measure the plant traits. Although these models perform an acceptable instance segmentation, the scarcity of annotations along with complex features of main stalk and nodes affect the performance of node counting (RMSE = 3.346 nodes) and main stalk height measurement (RMSE = 284 pixels). To find possible solutions to this, the comparison of supervised and weakly supervised approaches is done in order to predict cotton boll count on the basis of image processing performance, and labeling costs. The results suggest that, weakly supervised counting approaches based on peak response maps such as CountSeg (RMSE = 1.284 ± 0.08) yields comparable results as of Mask R-CNN (RMSE = 1.175 ± 0.20). In terms of data annotation, the weakly supervised approaches were found to be at least 10 times cost efficient compared to supervised approach for boll counting task. In future, this allows us to improve the current supervised frameworks in the absence of quality mask annotations by leveraging the density maps obtained from weak supervision and can be extended to various cotton plant organs such as primary and secondary branches.

Keywords: Cotton phenotyping, Boll counting, Supervised learning, Mask R-CNN, Weakly supervised learning, class activation map, multiple instance learning;

3.2 Introduction

Cotton (*Gossypium hirstum L.*) is one of the most grown cash crops around the world (FAOSTAT, 2019). Efforts involving the study, improvement of existing or to develop new methodologies that aids the production essentially requires measurement of plant traits (Fahlgren et al., 2015). It is an arduous task to obtain such measurements manually and may result in ruining the agricultural field. To carry-out these tasks efficiently, researchers are developing artificial intelligence-based systems with the help of robotics (Oberti & Shapiro, 2016), machine learning (Liakos et al., 2018) and deep learning (N. Zhu et al., 2018). In the case of cotton, the field and the plant architecture makes it even more difficult. Furthermore, based on the cotton plant growth cycle the important morphological traits differ (Ritchie et al., 2007). For example, in the germination stage monitoring seedlings and identifying weed is an important task, while, in blooming stage the flowering pattern/stages are essential for further growth. In a fully-grown stage, one of the most important phenotypic traits of a cotton plant is the total boll count. For growers, boll count is the primary indicator of potential yield from the field and is hard to accurately predict manually. In addition, boll count can provide a better understanding about the physical and genetic conditions of the crop that can lead growers to assess growth conditions and take crucial decisions (Pabuayon et al., 2021). Apart from boll count, the cotton plant skeleton can be studied for obtaining structural information (Ritchie et al., 2007). Main constituents of the cotton plant skeleton are its main stalk, primary branches that starts from main stalk at the node, and the secondary branches(Twine & Redfern, 2021). Identification of main stalk and nodes can provide vital architectural information- such as the average plant size, inter-node distance, branch angles, etc.- that can be used for plant growth analysis as well as managing in-field operations (McCarthy et al., 2009). Therefore, identifying and monitoring these traits can be efficiently done with the help of in-field, non-destructive high throughput phenotyping (HTP) techniques that requires less time and labor (Normanly, 2012; Pabuayon et al., 2019).

Machine learning and deep learning along with computer vision is playing a vital role in modern agriculture in several areas (Jiang, Li, et al., 2020; Uddin & Bansal, 2021). Some of the widely studied applications include plant disease classification (Saleem et al., 2019; Sladojevic et al., 2016), yield prediction (Van Klompenburg et al., 2020), organ detection and counting (Koirala et al., 2019). In addition, recent developments in the instance segmentation techniques such as Mask R-CNN (He et al., 2017), SoloV2 (X. Wang et al., 2020) helped researchers to study beyond above applications (Hamuda et al., 2016; Ni et al., 2021). Jiang et al., 2020 proposed a method to detect and count emerging cotton blooms that can help characterizing flowering patterns efficiently. Apart from these, with the help of deep learning based methods different plant traits can be tracked and measured over time (Jiang et al., 2019; Petti & Li, 2021). To detect the cotton boll and predict the yield, S. Sun et al., 2019 implemented a boll recognition and counting pipeline based on traditional image-processing algorithms; achieving 84.6% accuracy in boll counting. (S. Sun et al., 2021; S. Sun et al., 2018) further shows that along with the 2-D image data, in-field HTP can make the use of 3-D measurements to perform growth analysis. As the application changes, the underlying deep learning model gets complex, thus, managing the training process along with inference time becomes challenging and sometimes can be addressed via speed-accuracy trade-off (J. Huang et al., 2017). Furthermore, most of these methods requires extensive data acquisition experiments and pre-processing to achieve the desired task that heavily rely on the highly accurate, complex data annotations consuming considerable amount of time. In plant phenotyping, though there are multiple ways to collect high-quality raw data in abundance, the precise annotation of this data has been a challenge due to factors such as domain knowledge, multi-modal input data, and application specific annotations. The current state-of-the-art methods perform best when trained with higher amount of labeled data as it helps learning diverse features in the data and reduce overfitting. In the case of a single cotton plant, even a single 2-D image contain 60 to 70 instance masks on an average. These masks represent main stalk, nodes, and bolls, with irregular shaped polygons that requires intricate labelling consuming at least 40 seconds for a simpler mask with a skilled annotator. Furthermore, there are not many public datasets such as ImageNet, MS-COCO are available for agricultural domain. This creates a burden of annotation on HTP researchers that needs to be addressed in order to produce high quality deep learning models.

One promising way of addressing the availability of data annotations is to provide a weak supervision either with a partially annotated data or with pseudo-labels obtained from lower-level (image-level class annotations) labels. With the vast availability of image datasets, it is not always possible to have each and every image annotated with high quality labels. In such cases the images are missing some or all of the ground truth annotations and contributes to the false assumption of having all the correct instance annotations. These datasets are referred to as partially annotated and can certainly mislead the overall learning. A deep convolutional network can be scaled-up for multi-label classification with missing annotations when trained with a loss that exploits proportion of known labels (Durand et al., 2019). With the introduction of complex loss functions and processing the intermediate outputs such as activation layer output , partially-labelled datasets can be used to perform more advanced tasks such as semantic segmentation (Kalluri et al., 2019), object instance counting and segmentation (Cholakkal et al., 2019) that can achieve comparable results to that of supervised methods. However, for object detection and/or segmentation tasks, such methods still require considerable annotation efforts since the localization of all the object instances are essential. Additionally, these methods are more susceptible to noisy, occluded, and imbalanced data. Thus, weakly supervised approaches based on partially annotated data are used in applications with lower variance in the object features and there is a vast scope of improvement.

On the other hand, when the lower-level labels are available, there are a plethora of approaches to localize and detect object instances (Zhang et al., 2021) that leverages two prominent paradigms: multiple-instance learning (MIL) (Andrews et al., 2003) and class feature activation maps (CAMs) (B. Zhou et al., 2016). MIL is based on learning object instances from positive (one or more instances present) or negative (no instances at all) bins that contains of samples from the given dataset. For example, an image is considered to be in a positive bin if at least one object instance is present in it, otherwise it is in the negative bin. While MIL based methods are widely used for weakly supervised object detection (C. Lin et al., 2020; H. Wang et al., 2021), using intermediate classifier activation layer's feature maps i.e. CAMs along with pseudo-label generation is gaining popularity due to its versatility for object detection as well as segmenting individual instances (Durand et al., 2017; J. Wang et al., 2018; Y. Zhou et al., 2018). CAMs allow us to understand the relation between a deep learning model and it's output for the given image by inspecting the specific pixels that contributed more in the output. Most of the CAM based approaches require external proposal generator and a heuristic-based scoring mechanism to predict the final output (either bounding boxes or instance masks) that lack a unified framework . These issues can be tackled in several ways such as the introducing self-supervised deep neural nets that generates proposals to fill in the missing parts on its own (Shen et al., 2020), providing attention to generate pseudo-labels (L. Chen et al., 2021; Ou et al., 2021), using pseudo-labels to train stronger models such as Mask R-CNN (Laradji et al., 2019). All of these improvements are specific to the application and differ according to the domain of the data, however, they perform sufficiently to address the posed problem.

Due to the success of weakly supervised approaches to detect and segment object instances with high precision, they are gaining researcher's interest from various fields in which the labelled data is hard to produce. In most of the cases, such applications fall under medical and cellular biology (Chamanzar & Nie, 2020; Qu et al., 2019). Apart from that, Laradji et al., 2021 showed that using affinity and localization based counting loss for fully convolutional networks (A-LCFCN) (Laradji et al., 2018) gives closer results to fully supervised segmentation methods with fixed annotation cost. Agricultural researchers have started to make use of weakly supervised models to reduce the annotation efforts in a variety of applications that includes disease classification, yield estimation and counting, etc. Bollis et al., 2020 designed a CNN-based algorithm to automatically select the regions of interest (ROI) from citrus crop images that were damaged by pests and diseases. The algorithm uses MIL paradigm to classify those crops with the help of Saliency maps that significantly reduced the annotation costs. Ghosal et al., 2019 performed head detection and counting to understand the relation between phenotypic and genotypic traits of the sorghum crop. They demonstrated that it was possible to alleviate the annotation costs with the help of partially annotated dataset in a weakly supervised learning settings without compromising the final model performance. TasselNet (Lu et al., 2017) and its successors (Xiong et al., 2019) used a regressor to count maize tassels with the help of local density maps that achieved robust in-field counting with high accuracy. The idea of working in smaller sub-windows of an image so as to obtain fine density maps was proved to be efficient and integrated in our study . Another weakly supervised counting network PSSNet (Tong et al., 2021) used point-supervision to segment and count the trees in aerial images that converts feature maps to masks. This network outperformed state-of-the-art methods in most of the challenging conditions and greatly reduced human labor by generating masks automatically . As discussed previously, these approaches make use of either CAM or MIL paradigms but Yu et al., 2020 proposed a Minirhizotron image segmentation approach combining both MIL and CAM that outperformed standard weakly supervised semantic segmentation frameworks. Bellocchio et al., 2019 proposed a novel fruit count method for yield estimation based on spatial consistency loss which falls under MIL due to the nature of weak learning provided to the model. This weakly supervised counting (WS-COUNT) architecture performed exceptionally well in high density fruit counting and was able to achieve a performance similar to its fully supervised counterparts. Thus, for the boll counting application we chose WS-COUNT as one of the frameworks to study in detail. At the time of submission, there are a few studies involving some paradigm of weakly supervised learning applied to cotton plant phenotyping, for example, Z. Huang et al., 2020 proposed an algorithm based on density classification for in-field cotton boll counting with the help of high-dimensional feature maps. Though this method achieves better performance than most of the counting methods, this method is computationally expensive on both training and inference levels . Furthermore, the method uses the high-resolution images of cotton field taken from above the ground, and thus, counts the bolls that are visible in the top view making the method erroneous in the presence of dense boll distribution and occlusion. However, the success in counting bolls in the in-field images motivates us to test the more advanced weakly supervised paradigms for the cotton plant phenotyping.

Overall, cotton phenotyping from 2-D images faces several challenges with respect to data annotations and preparing the multi-purpose labelled dataset for various trait extraction. The availability of cotton plant datasets with partial or low-level annotations will enable the use of weakly supervised learning in alleviating these challenges. Furthermore, with a careful comparison between fully supervised methods, the potential of these weakly supervised methods to extract complex cotton plant phenotypic traits should be tested. Therefore, the overall goal is to extract the various traits of cotton plant, such as main stalk height, node and boll count, using supervised and weakly supervised deep learning with the limited availability of labelled data for training. Additionally, the application of weakly supervised learning based on partial labels, CAMs and MIL (Bellocchio et al., 2019; Cholakkal et al., 2019) is explored for counting the cotton bolls from the front view of cotton plant images taken in various environmental conditions at different resolutions. The specific objectives of this study were to:

1. Develop weakly supervised methods based on class activation maps and MIL to segment cotton bolls from both indoor and infield 2D images.
2. Develop supervised methods to segment cotton bolls, nodes and main stalk of the plant from 2D images both indoor and in the field.
3. Compare the supervised and weakly supervised methods in their performance on phenotypic trait (boll number, main stalk height, and node number) measurement and efficiency.

3.3 Materials and methods

3.3.1 Data Source and Pre-screening

The dataset used for this study consists raw 2D cotton plant images from both potted (indoor and outdoor) and in-field settings representing variations in plant conditions, traits, and background. A pre-screening of data was performed to select good quality images resulting into 290 images with the resolution ranging from 800px to 4000px across both dimensions. This resolution was not a problem for supervised methods such as Mask R-CNN, as they can learn high resolution images along with training masks even when scaled down/up to a fixed smaller (between 500px to 1000px) resolutions. Thus, for segmentation of entire plant, only 203 raw images were labelled with instance masks of main stalk, node, bolls and branch. The summary of masks is given in table 3.1. This dataset was considered as full plant image (one image contains entire plant plus background) dataset for extracting main stalk, node instances with supervised models and will be used for further experimentation of these traits. Initial experiments with the supervised model for obtaining boll masks missed significant number of instances due to above mentioned scaling operation, and thus, it was necessary to refine this dataset only for boll counting before using it further.

Table 3.1: **Summary of raw images with entire plant** and corresponding instance mask labels used to extract all traits from entire plant image with Mask R-CNN. A total of 203 images were selected and labelled with precise polygonal instance masks for main stalk (one per image), nodes, and cotton bolls.

Image Set	No. of Images	No. of Main Stalk masks	No. of Node masks	No. of Boll masks
Training	175	175	1940	3147
Testing	28	28	408	467
Total	**203**	**203**	**2348**	**3608**

The boll counting task mainly falls under object detection and it was observed that if the single boll instance occupies less pixels compared to an entire image, then the models that learns the pixel representation were under-performing in the counting task. For example, in weakly supervised methods, main aspect of the learning is the pixel density map for a particular class i.e. Class Activation Maps (CAM). Once the CAMs are obtained, the specific object instances or locations can be learnt through the pixels with the help of various techniques such as peak aggregation, multiple instance learning (MIL) regression, etc. The peak aggregation methods rely on these maps to find peaks of the classes and then to aggregate the peak count thus yielding a total object count. Due to this nature of MIL-CAM based method, the object resolution in an image along with the pixel density for a single instance becomes an important factor in the learning process.

For weakly supervised methods, the under-performance was evident in the initial models that were trained on the above-mentioned dataset annotated without pre-processing the images. The counting performance was not even close to 50% of that was achieved by the poor supervised counts. In most of the cases the count was underestimated and at a certain point the model was not able to count more than a certain number, which can be termed as model's subitizing range. The main reason, as explained above, can be predicted by observing the class activation maps and corresponding peak maps obtained via peak aggregation. The peak maps obtained from full size images showed that the crowding of multiple bolls in an area of the image contributes to lower number of peaks in that area. These lower peaks thus resulted in lower boll count when the full-size image was processed through the feature extractor.

To fix this, the entire dataset was re-screened and pre-processing steps were taken to generate a new uniform raw dataset that can be labelled in order to train the models for boll counting. 285 out of 290 images were selected for generating training and test (validation) sets with uniform image size. 5 plant images with variable features were held out for testing the entire processing pipeline. Table 3.2 summarizes this dataset and corresponding boll counts per image. The background and other adjacent cotton plants (if any) are cropped out to focus on the subject plant and reduce the noisy data while training. Furthermore, a fixed input image window of 500 × 500 was chosen to train the models and thus the images were cropped with zero-padding along the corners. This resulted in 4266 image tiles from 285 images while 84 tiles from 5 held out images, all with fixed size of 500 × 500.

Table 3.2: **Summary of raw images selected for boll counting task** and corresponding tiles generated after zero-padding along with sum of boll count per tile. A total of 285 full plant images were selected to obtain image tiles that will be used for training and testing of weakly supervised methods with random shuffling. Remaining 5 full plant images were held out for testing all the methods on a full scale plant image.

Image Set	No. of Images	No. of tiles generated	No. of Bolls
Training + Testing	285	4266	23651
Full Plant (held-out) Testing	5	84	217
Total	**290**	**4350**	**23868**

3.3.2 Annotation Approaches

In general, the instances to be detected are complex in nature such as variable bolls in different size, shape, appearance, and conditions. In the dataset explained above, on an average, a typical cotton plant consists of 15 nodes and 35 bolls. Considering the scale of object instances that are required to be annotated, it is not practical to annotate the entire raw dataset. Instead, the previous studies use a small portion for labelling in the learning process as it is convenient and most of the times sufficient to achieve the objectives of the model.

In supervised learning, the choice of labelling approach and type of annotations differs based on the objective in focus. For example, to count the number of cotton bolls, a simple regression method can be trained with a total count in the image used as fully supervised ground truth. On the other hand, to obtain the pixel-level information such as boll density or main stem height, an entire instance mask is needed for the supervised approach. This demands highly skilled annotators and their inexhaustible efforts both in terms of precision and time. To train the supervised methods, all the ground truth object instances should be accurately provided with pixel masks. This type of labelling is typically known as pixel-wise annotations or mask annotations. For this, each object instance in the image will be labelled along its contour and the pixels enclosed in it represents the instance mask.

On the other hand, weakly supervised methods tend to reduce the amount of supervision required to achieve the same objective. For example, instead of providing total object count as a ground truth, weakly supervised methods for the counting task are trained using a classification or class-level annotations. This reduces the annotation efforts significantly as the annotator is now required to detect the presence of any single instance and give the classification label as class "present or absent". Images are further annotated using point and mask labels only. Initially, the dataset was labelled with point annotations (Cheng et al., 2021) for each boll instance where an instance of cotton boll is marked in the image with a point. Although only the boll presence label is required for learning, point label helps to maintain the ground truth count for the image. The classification labels were obtained from these point labels. Figure 3.1 shows the basic difference between these labels and represents one sample each from field data, potted outdoor and potted indoor data. In this study, all the annotations are performed using VGG Image Annotator online tool (Dutta et al., 2016).

Table 3.3: **Summary of boll count per tile in different counting ranges** for each of the train, test (validation) and held-out test dataset. Train and test (validation) set contains tiles obtained from 285 full-scale raw images while held-out test set represents tiles from 5 held-out images. Due to the dense boll population the tiles containing more than 15 bolls cannot be counted accurately. Hence, these were not considered in training as well as did not consider (DNC) in the total count for the train and test (validation) sets.

Boll count per image	**Training tiles**	**Test tiles**	**Held-out Test tiles**	**Total**
0	919	50	21	990
[1,5]	1852	102	48	2002
[6,10]	710	42	6	758
[11,15]	231	6	3	240
above 15	300 (DNC)	54 (DNC)	6	360
Total	3712	200	84	4350

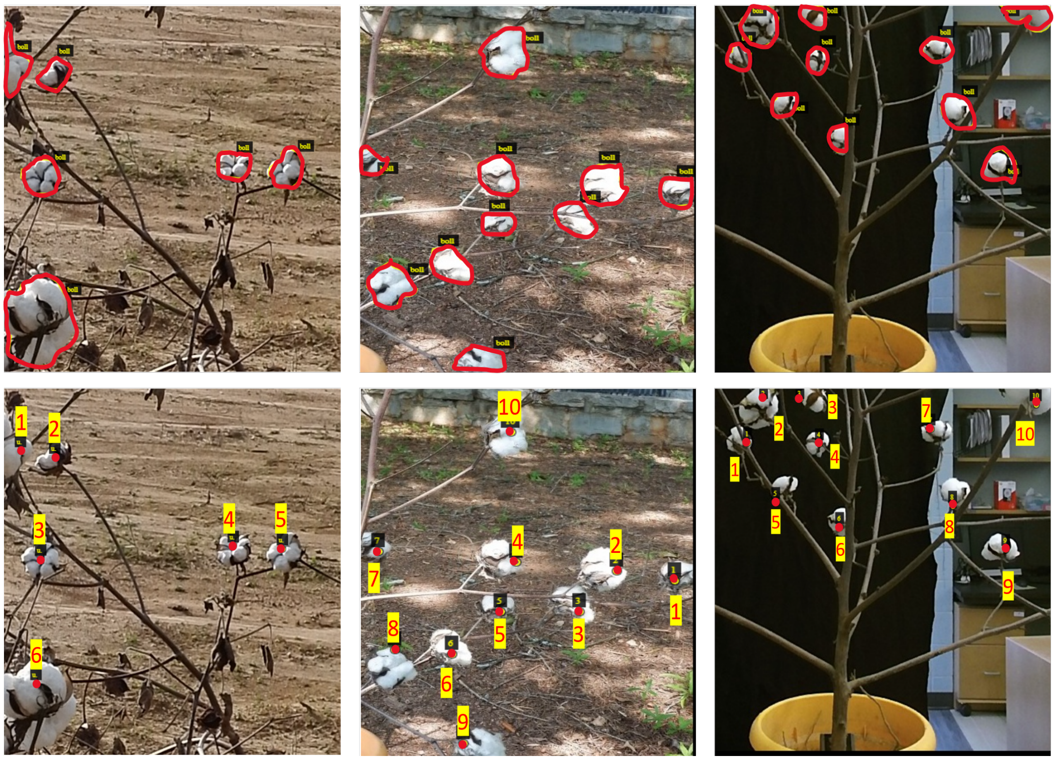

Figure 3.1: **Annotation Types**. The two main types of annotations were used in this study. **Top Row** represents instance masks, while **Bottom Row** represents point labels for the same image. **First column** shows a sample tile from in-field plant image, while **second and third column** shows potted plant in outdoor and indoor conditions respectively.

Table 3.3 gives a summary of the total 4350 image tiles that were labelled with point annotations for the main boll category. The images that do not contain bolls (background or other plant parts) were kept as negative category images i.e class label = 0. Even with pre-processing, in certain images the presence of background bolls from nearby plants contributed to noisy boll-count. As following the literature for weakly supervised counting, a subitizing range was considered as [0,10], and range [11,15] was categorized as challenging examples. The images above 15 were discarded as most of them had large boll count due to bolls in the background or some of them were visually indistinguishable from the 2D image. This process curbed the experiment dataset to 3912 images. Furthermore, as Table 3.3 shows, the images with boll count in the range [1,5] skews the data and this imbalance proved to be a limiting factor for the weakly supervised models. The initial models tend to restrict the count between this range and were unable to yield a count above 5. Thus, to avoid this imbalance, the 3712 training images were manually augmented with 1758 randomly sampled images from range [6,15] by applying 90°, 180°, 270°rotations. This training set with 5470 image tiles was used as the final dataset for weakly supervised counting.

Although the mask labels for bolls were available from the dataset described in Table 3.1, they cannot be directly used to perform a fair comparison between supervised and weakly supervised methods as those images were full-scaled (without pre-processing) and not cropped. Therefore , 350 image tiles from Table 3.2 were annotated with additional instance masks. 300 of them were used for training and 50 were used as a validation set to store the best model.

3.3.3 Fully Supervised Learning Approaches

Mask R-CNN

Mask R-CNN is considered as a strong instance segmentation model and was chosen. It is an extension to the popular object detection framework Faster R-CNN that is used for supervised counting tasks and widely used as a baseline method for comparing the counting results with weakly supervised methods. Mask R-CNN operates on the regions proposed by the region proposal network (RPN) in order to predict the class label, bounding box and the corresponding instance pixel masks from multiple regions of interest (ROIs). The ROI Align layer of Mask R-CNN is one of the key improvements from Faster R-CNN and allows it to achieve better object detection results than Faster R-CNN. This is one of the reasons that Mask R-CNN is used for both segmentation and object detection/counting in this study. The choice of Mask R-CNN was inspired by the flexibility this method offers in terms of output. It can produce both instance masks as well as bounding boxes. Also, it has been proven superior at localization task both in terms of precision and accuracy.

Supervised Count Regression: S-COUNT

The S-COUNT network is an end-to-end fully supervised method that was trained as a regression model for counting bolls from the image tile. The network consists of ResNet-101 as feature extractor without the final fully connected layer. It was replaced with an additional 1 × 1 convolution layer with N filters. This produces **N** filter maps which were treated as output response maps and fed to the fully connected layer to regress the boll count. The choice of N varies as per dataset. In this case, both $N = 6$ and $N = 8$ were experimented with variable random seed value to initialize the weights and data samples. Except the output layer, all layers have ReLU activation with batch normalization layers. Finally, the network is optimized with standard mean squared error (MSE) loss.

3.3.4 Weakly Supervised Learning

MIL-CAM based Weakly Supervised Counting : WS-COUNT

Weakly Supervised Counting (WS-COUNT) (Bellocchio et al., 2019) is one of the promising weakly supervised architecture for counting based on classification label supervision. The counting using weakly supervised methods is usually based on class activation maps (CAMs) and obtaining those is an easy procedure for dense CNNs such as ResNets. This architecture is comprised of two sub networks: a presence-absence-classifier and a regression network to yield count. The presence-absence classifier takes the image as an input to detect the presence of a class in it (Binary classifier) which is trained in a supervised manner using the classification labels. The second part is the counting branch that regresses the object count based on the feature map and trained using the output of above classifier. The schematic layout of this framework is shown in Figure 3.2

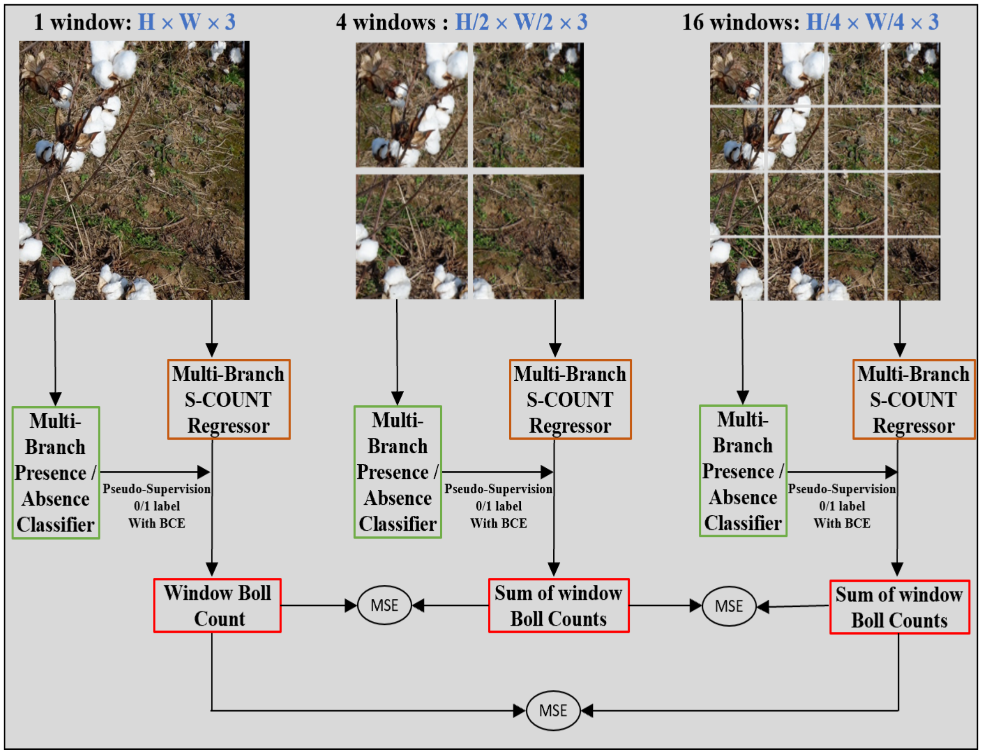
1 window: H × W × 3
4 windows : H/2 × W/2 × 3
16 windows: H/4 × W/4 × 3
Multi-Branch Presence / Absence Classifier
Multi-Branch S-COUNT Regressor
Pseudo-Supervision 0/1 label With BCE
Window Boll Count
Sum of window Boll Counts
Sum of window Boll Counts
MSE
MSE
MSE

Figure 3.2: **Schematic representation of WS-COUNT architecture**. An image is divided into 4 windows which then further divided into 16 windows. A total of 21 images are passed to the main two networks that are responsible for boll counting. **Presence Absence Classifier** (PAC) detects the presence of boll in the patch and thus provides a weak supervision for the regression network, while **counting network** (S-COUNT) estimates a boll count for that patch with the help of additional fully connected layers. The processing of 21 image patches in parallel makes the individual PAC and S-COUNT networks to be multi-branched (MB) and predicts unique output count for each of the 21 patches. The count predictions are kept in accordance to the classifier supervision and the total count loss is optimized through all the image levels.

One of the important parts of WS-COUNT architecture, presence-absence classifier (PAC), is based on multiple instance learning that needs to be trained separately. PAC has a similar architecture to S-COUNT except the output response maps were not fed to fully connected layer. Instead, it is fed to a pooling layer that combines the output features of the response maps which in turn produces a class prediction with the help of sigmoid function. The regression network (S-COUNT) in the WS-COUNT architecture was trained by the weak supervision of these class labels.

During the training of WS-COUNT network, the input image was processed at three different scales (see Figure 3.2), first of which is the entire image tile while the second, third are 4, 16 sub-windows of the input image tile, respectively. This scaling makes this architecture a multi-branch (MB) network with a total of 21 branches being trained simultaneously. Each branch processes a sub-window and produces the count which is then weakly supervised by the PAC prediction of same sub-window. The learning objective for this method is to minimize the classifier consistency loss $\mathcal{L}_{PAC\text{-}C}$ and a spatial consistency loss $\mathcal{L}_{SP\text{-}C}$. Classifier consistency loss maintains the coherence between classifier output and counts of each branch. Ideally, if the classifier predicts the presence of boll then the count network should produce a count greater than zero and vice versa. The main purpose of spatial term is to bring consistency between the total count at three different scales. Eq. 3.1 gives combined loss function that was optimized during the training of WS-COUNT multi-branch network.

$$\mathcal{L}_{WS\text{-}COUNT} = \text{Classifier consistency loss}(\mathcal{L}_{PAC\text{-}C}) + \text{Spatial consistency loss}(\mathcal{L}_{SP\text{-}C}) \tag{3.1}$$

To perform a comparative study, initially, both the networks, S-COUNT and classifier (PAC), are trained separately to yield a count for every image. In terms of object counting, these can be viewed as one method with "knows what to count but not How" and other method that "knows how to count but doesn't know what or where". These two networks are called MB-PAC and MBS-COUNT respectively after the introduction of multi-branch **(MB)** architecture. The combination of both trained together that "knows what and how to count" gives the final WS-COUNT architecture as shown in Figure 3.2. It is important to understand the effect of a weak supervision provided by a naive model to a fully capable model. The performance of MB-PAC was expected to be off by a large margin as it just classifies the image patches and yields the sum of classification scores as the final count, but MBS-COUNT is expected to give a very closer estimate for boll-count as it was trained in an end-to-end supervised manner. Thus, the WS-COUNT tries to achieve the performance gain in MB-PAC and goes as close as possible to the MBS-COUNT results.

CAM based Counting with partial labels : CountSeg

CountSeg (Cholakkal et al., 2019) is another method that relies on class activation maps to count the objects in the image as well as retaining their spatial information which is helpful for segmentation of these object instances. This method builds upon the works of Peak Response Mapping (PRM)(Y. Zhou et al., 2018) architecture by introducing a new supervision method and novel loss function to predict the global counts of objects. This method utilizes the property of subitizing range (Cholakkal et al., 2020) to annotate the object counts in the image and use these image-level counts as supervision.

CountSeg requires annotator to annotate the counts only between subitizing range [0,5] which is different than the initial consideration of [0, 10] and can predict the counts beyond this range with a high accuracy. Figure 3.3 illustrates the general network architecture for boll counting using CountSeg. The input image is passed through the ResNet-50 backbone feature extractor and applies 1x1 convolution to feed the two main branches of this network. Out of available features, half the features are fed to image classification branch while the other half are given to density branch as an input. Image classification branch performs simple convolution operation to predict the presence or absence of the boll in the image while the density branch is responsible for predicting the global boll count as well as its spatial distribution for that image by constructing a density map.

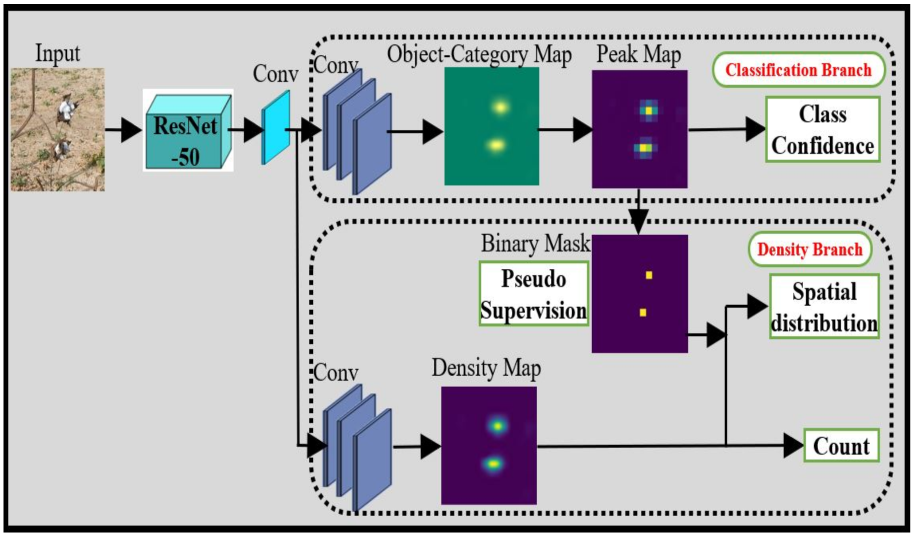

Figure 3.3: **CountSeg for boll counting**. The two branches: classification branch and density branch are jointly trained using image-level lower-counts (ILC) supervision. The pseudo ground truth is generated by classification branch to supervise the output of density map with the help of spatial and global loss functions.

CountSeg architecture is jointly trained in an end-to-end learning schedule with the help of image-level lower-counts (ILC) supervision (Cholakkal et al., 2020). The main objective of the joint training is to minimize the total loss given by Eq. 3.2. The term $\mathcal{L}_{class}$ denotes the loss of classification branch which is trained with the supervised class labels. The class labels were extracted from earlier point labels, where a count greater than zero was considered as presence label and count equal to zero was given an absence label. Due to the unawareness of classifier in delineating object instances, the peaks generated from the classifier alone contains large number of false positives. For this, the lower-level count information is incorporated to generate the pseudo ground truth for training the density map branch. The term $\mathcal{L}_{global}$ helps reducing the error between predicted count and ground truth count while the term $\mathcal{L}_{spatial}$ makes sure that all the individual object instances are localized properly. For detailed implementation on these loss functions readers are referred to (Cholakkal et al., 2019).

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{spatial} + \mathcal{L}_{global} \tag{3.2}$$

The CountSeg network is trained end-to-end with two stages. The first stage involves training classification branch along with density branch considering only $\mathcal{L}_{class}$ and $\mathcal{L}_{global}$. Since the spatial loss requires pseudo ground truth from classification branch- which is still under training- it is considered in the stage two of training. Furthermore, CountSeg architecture can be used to perform instance segmentation base on the peak response maps obtained from the peak backpropagation process. If off-the-shelf object proposals are available, then it can be ranked based on the peaks. This study does not include the segmentation as training the object proposal network required full supervision with enough instance labels. Thus, the scope of CountSeg in this study was limited to counting application.

3.3.5 Boll Counting

Figure 3.4 shows overall workflow for boll counting. The full-scale raw images were pre-processed into image tiles (section 3.3.1) and labelled (section 3.3.2) based on the learning approach. These tiled image dataset (table 3.3) was then used to train and test the individual boll counting method discussed in above section. The total count for an entire plant image was considered to be the sum of counts from all the tiles. Though the question of counting a single boll multiple times may arise, a very few instances of such case were actually observed in the raw dataset. To verify this the 5 held out images were tested in an end-to-end fashion. In the later stage, the performance of fully supervised and weakly supervised methods for boll counting task with respect to the annotation cost and availability of labelled data was demonstrated.

3.3.6 Main Stalk and Node Segmentation

With the help of Mask R-CNN various plant organs can be segmented from the 2D images. These segments then can be post-processed to obtain important plant phenotypic traits. In a cotton plant, identifying main stalk is a visually challenging task in most of the 2D images and once identified can be a helpful trait. Computing a precise main stalk height from image requires calibrated data, but in most of the times the images has varying properties such as background, illumination, field conditions, etc. However, supervised Mask R-CNN approach was able to overcome these challenges and the instance

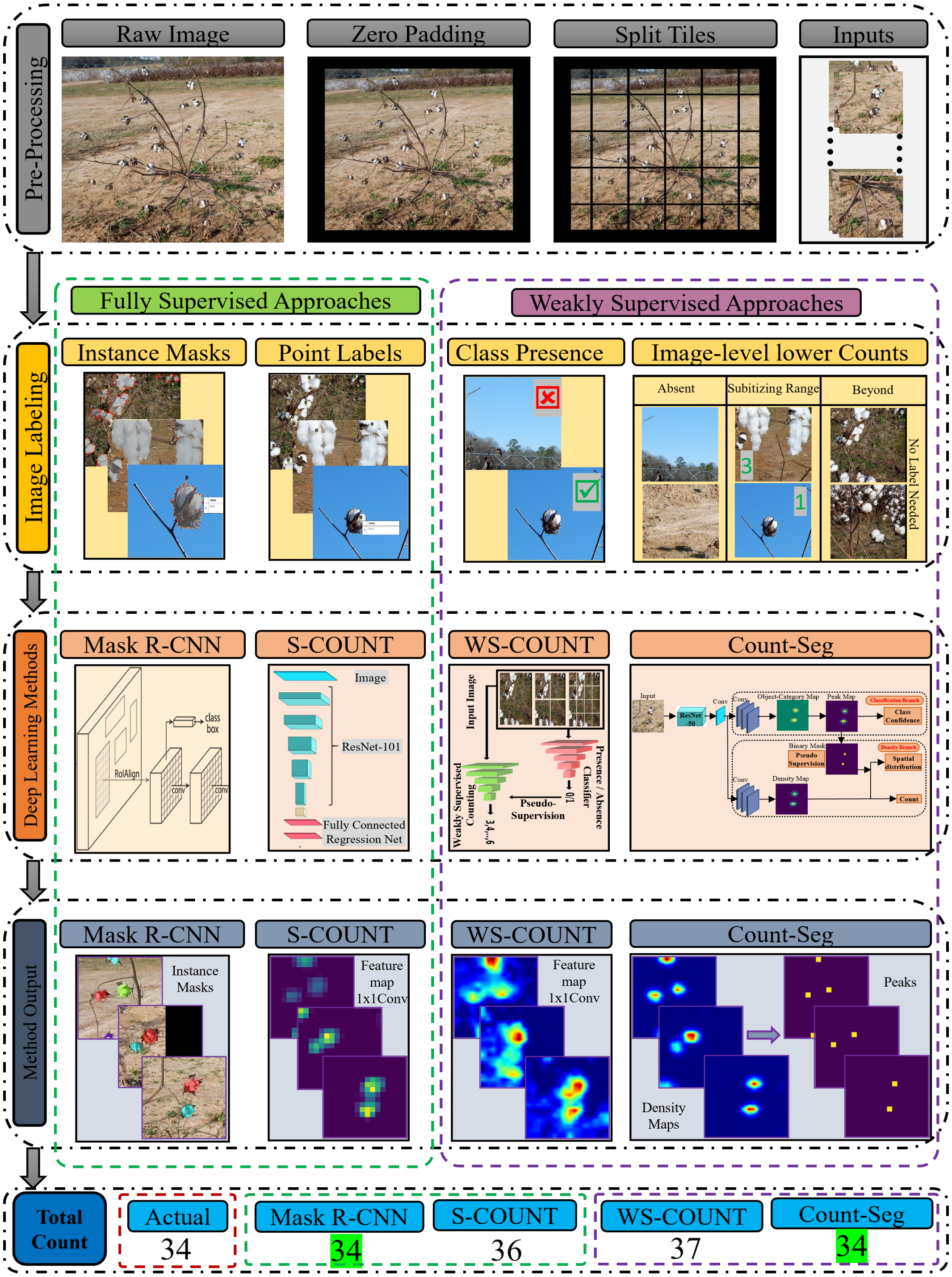

Figure 3.4: **Overview of boll counting**. The image tiles generated from pre-processing steps explained in section 3.3.1 were labelled with point and mask labels. The classification labels and image-level counts were derived from point label counts. Two fully supervised and two weakly counting methods were trained on the image tiles training set explained in Table 3.3. The intermediate and final stage output of each methods can be visualized by instance masks and feature maps that will be used to obtain final boll count.

masks were further processed to obtain the main stalk height in pixels. This pixel height could be converted to actual height with the help of a calibration target. This is not considered in this study as the dataset does not contain any calibrated data.

Figure 3.5 demonstrates overall workflow for obtaining the main stalk height and node count from the input 2D plant image with the help of Mask R-CNN and PlantCV library (Gehan et al., 2017). Initially, the Mask R-CNN model was trained with a smaller sample size to adjust the hyperparameters. It was observed that the segmentation of main stalk and nodes via a single network with these two as output classes could not perform as per expectations. This was due to the overlapping nature of the two classes as nodes are the part of main stalk. This attributed to missing many of the node instances as well as wrong identification of few branches as main stalk. To keep the model simpler and improve the instance identification performance, two separate networks were trained for each class. Once the instance masks are obtained, they can be processed for trait extraction. Node count can be directly obtained by counting identified node instance masks.

In most of the cases main stalk is visually occluded by other branches or cotton bolls and thus the predicted instance masks may not be continuous. In such cases, there can be multiple instances predicted by Mask R-CNN which needs to be handled in the post-processing steps. Therefore the masks obtained were combined together in a single binary mask and dilated to remove any inconsistent main stalk patch. Afterwards, using the built-in skeletonize and segment functions of PlantCV library main stalk height in pixels was computed. The images in this dataset were not calibrated to the global scale and thus the physical measurements were not obtained. Nevertheless, in future, if the cameras are calibrated and images are taken with fixed aspect ratio, one can translate the pixel height to physical quantity.

3.3.7 Evaluation Metrics

To evaluate the performance of main stalk height measurements along with node and boll counting, a comparison between errors of above models' predicted counts can be done. The main stalk height measurement and node counting was evaluated on testing set given in Table 3.1. The boll counting task

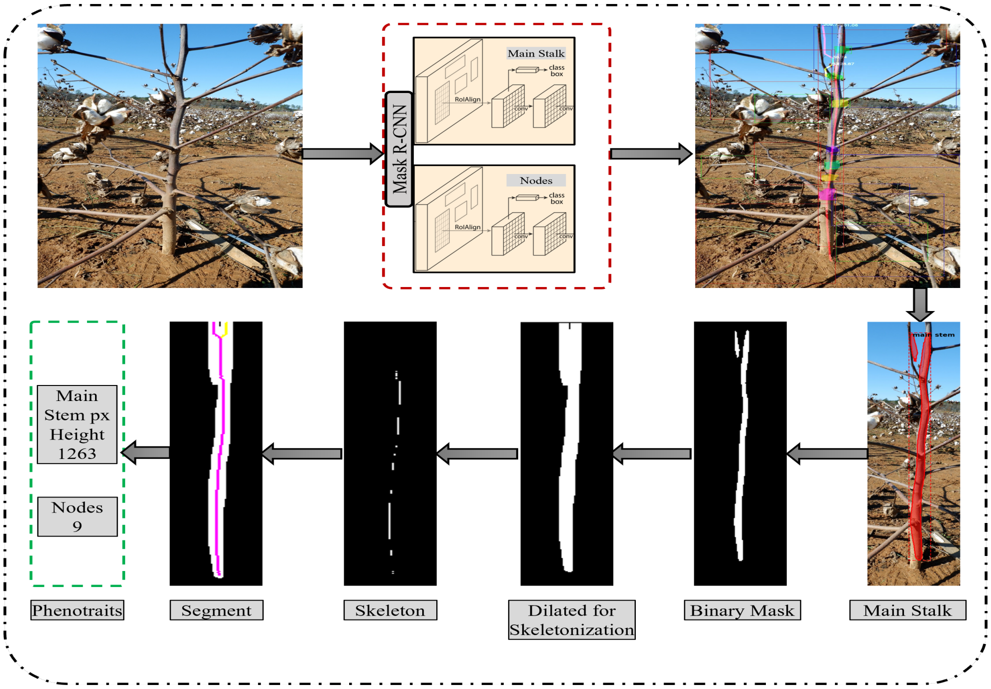

Figure 3.5: **Main stalk and node instance segmentation**. The figure illustrates trait extraction from plant images using instance segmentation. Two individual Mask R-CNN networks are employed to obtain instance masks of nodes and main stalk. The main stalk instance was further processed to obtain the height in pixels by performing morphological transformations.

was evaluated on test (validation) and held-out test sets explained in Table 3.3 with 200 and 84 image tiles in each set respectively. Five methods for boll counting were employed, namely S-COUNT, MB-PAC, WS-COUNT, CountSeg and Mask R-CNN. To eliminate any errors due to randomness and model layers, five models with each method were trained by changing random initialization and random data sampling.

Following are the metrics considered for evaluating the above methods for trait extraction:

1) Counting Error. Counting error is the difference between ground truth count and predicted count that can be visualized with an error histogram. As there are five models for each method, the error for single image can be computed by taking mean or median of the five predictions. It was observed that median helps in eliminating outlier predictions compared to mean and thus to plot the error histogram, median of five prediction counts can be considered.

2) Root Mean Squared Error (RMSE). One of the standard regression metric is to compute the root mean squared error for the boll counts. The counting error is a measure of how far the predictions are from actual values while RMSE is a measure of its spread. Ideal values for RMSE depends on the range of values in ground truth. For boll counting, the range for counts is [0,15] and thus a value closer to zero should be considered as a good score.

3) Scatter Plot. To visualize the performance of prediction vs. Ground truth data, scatter plots are widely used. For boll counting the scatter plot takes the form of a bubble plot where the radius of the bubble depicts the amount of samples representing corresponding point. The correlation line can be fitted using linear regression and the equations for the fitted line can be obtained to compare with standard Y=X line along with corresponding R^2 values.

4) Annotation time. One of the motivations behind this study was to obtain accurate and precise boll counts with minimum annotation efforts. While labelling the image tiles the time required was manually recorded for each of the labelling approach by single annotator. These measurements can be used to compare the performance with respect to the efforts required prior to training the models. This will be helpful in future research when there is an insufficiency of labelled or raw data.

3.3.8 Implementation Details

Boll Counting

WS-COUNT and CountSeg methods were implemented with the help of codes provided by the authors (Bellocchio, 2019; G. Sun, 2020). For both the methods the feature map layer was modified to suit the boll data while keeping the basic architecture same. S-COUNT network was trained with the counts obtained from point labels with the feature map size, $N = 6$ to 8. For the classifier PAC and entire WS-COUNT model, $N = 6$ was observed to perform better and thus used for further experiments. At first, the PAC is trained with the help of class labels extracted from the point labels to yield a robust classifier that can be used for supervising WS-COUNT architecture. The WS-COUNT model was trained with both SGD and Adam optimizer with a learning rate of 0.0001. The rest of the parameter were unchanged from original implementation.

For CountSeg, the backbone feature extractor was experimented between ResNet-50 and ResNet-101. It was observed that ResNet-101 did not offer more gain in counting performance and thus ResNet-50 was considered for further experimentation as it makes the model less bulky. Apart from backbone, the size of channels for the 1×1 convolution filter was experimented between 2 to 60 and 60 was chosen as it offered significant counting performance gain. Stage 1 of training was limited to 30 epochs and stage 2 was performed until convergence for a maximum of 200 epochs. The network was optimized using SGD optimizer with a learning rate of 0.001 and the custom loss function explained in section 3.3.4. All the networks used PyTorch based implementation with GPU support.

The basic Mask R-CNN network was built using Matterport's tensorflow implementation of Mask R-CNN (Abdulla, 2017) and the further experiments were performed by tuning the hyperparameters provided by them with a small sample size. To keep the model simpler, ResNet-50 was used as feature extractor with pre-trained weights on COCO dataset. Rest of the parameters are kept unchanged from original implementation. The model was trained using SGD optimizer and a fixed learning rate of 0.001. As stated in 3.3.2 the data used to train this network for boll counting contains 300 training images which were augmented with the help of image augmentation library. The augmentations were randomly selected from flip, rotate, blur, scale operations.

All the above mentioned models were trained on Nvidia Tesla P100-PCIE-16GB devices available at Georgia Advanced Computing Resource Center (GACRC) (of Georgia, 2015) clusters.

Main stralk and Node

To obtain features specific to main stalk and eliminate false branch proposals, the feature extractor and region proposal network (RPN) was improved in the Matterport's implementation. ResNet-101 along with feature pyramidal network was used as a backbone feature extractor. Apart from this, the configurations given with the implementation- such as non-max suppression threshold, percentage of positive examples used for mask training, etc.- were tweaked to attain a stable performance. The models were optimized using SGD with an initial learning rate of 0.001 scheduled to decrease by 1/10th every 200 epochs. The loss weights given with the original implementation focuses more on mask prediction branch while RPN layers were less penalized. In this implementation, more weights were given to the RPN loss that helped improving elimination of false positive branches being identified as main stalk. Rest of the hyperparameters were kept same as of the original implementation.

3.4 Results and Discussion

3.4.1 Boll Counting

The histogram of error distribution revealed that the mean errors of the two supervised counting methods were closer to zero and had less spread of error than those of the unsupervised methods (Fig. 3.6). For example, the majority (99%) of counting errors for the supervised methods were within ±3, whereas only 93% of the images had the counting error within this range for weakly supervised methods. However, the spread of error for CountSeg are comparable to that for both the supervised methods. Errors on positive side signifies the under-counting by the methods within a certain range. This can be attributed to the fact that the images were not directly trained on actual boll counts, but were trained on the presence-absence of bolls and unlike supervised methods, weakly supervised training process was unaware about the exact quantity of bolls present in an image. If the data contain fixed size objects and subitizing range can be increased, then this under-counting can be reduced. But in the case of bolls, the images in this dataset has random variation in boll size, shape, and 2D depth of bolls that limits the subitizing range to be smaller, in this case [0,10] .

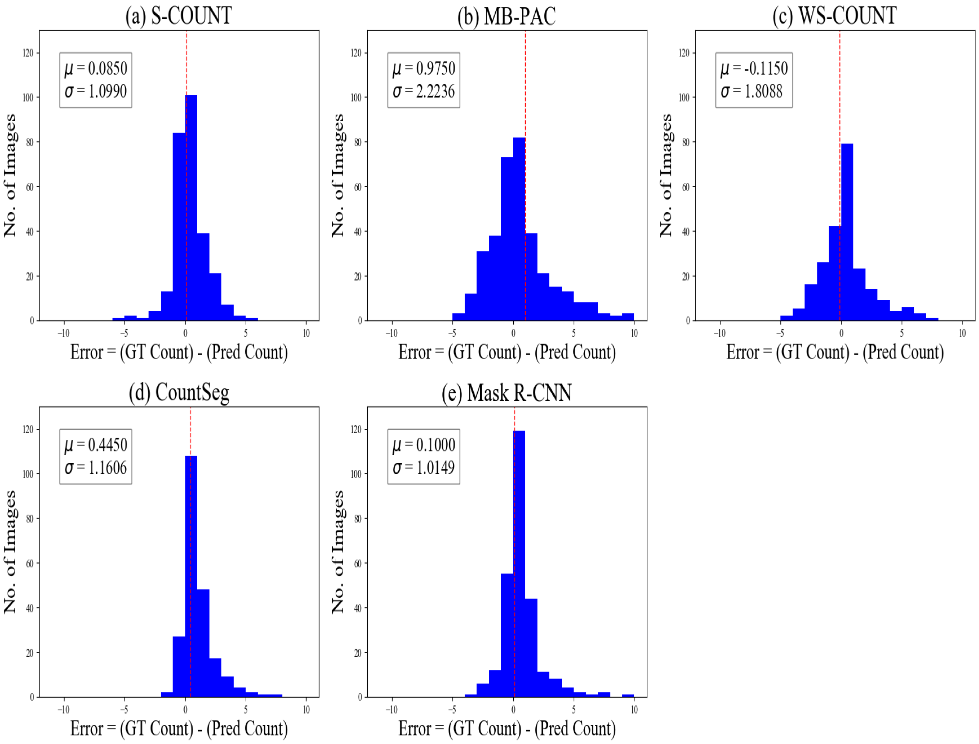

Figure 3.6: **Error histograms from median predictions given by each method**. Error is computed as the difference between GT count and median of predicted counts from five model variations.

The linear regression analyses show that although supervised methods performed better than unsupervised methods, CountSeg achieved a comparable performance as Mask R-CNN and S-Count. For MB-PAC network the under-performance was expected as the model represents the summation of class labels obtained from presence-absence classifier, while for the S-COUNT network, highly accurate predictions can be observed. The WS-COUNT network that combines the two (MB-PAC and S-COUNT) learns the counting with the help of PAC and thus performs as per expectations. The under-counting is improved from MB-PAC to WS-COUNT as the latter consists of a regression network that learns to regress total count.

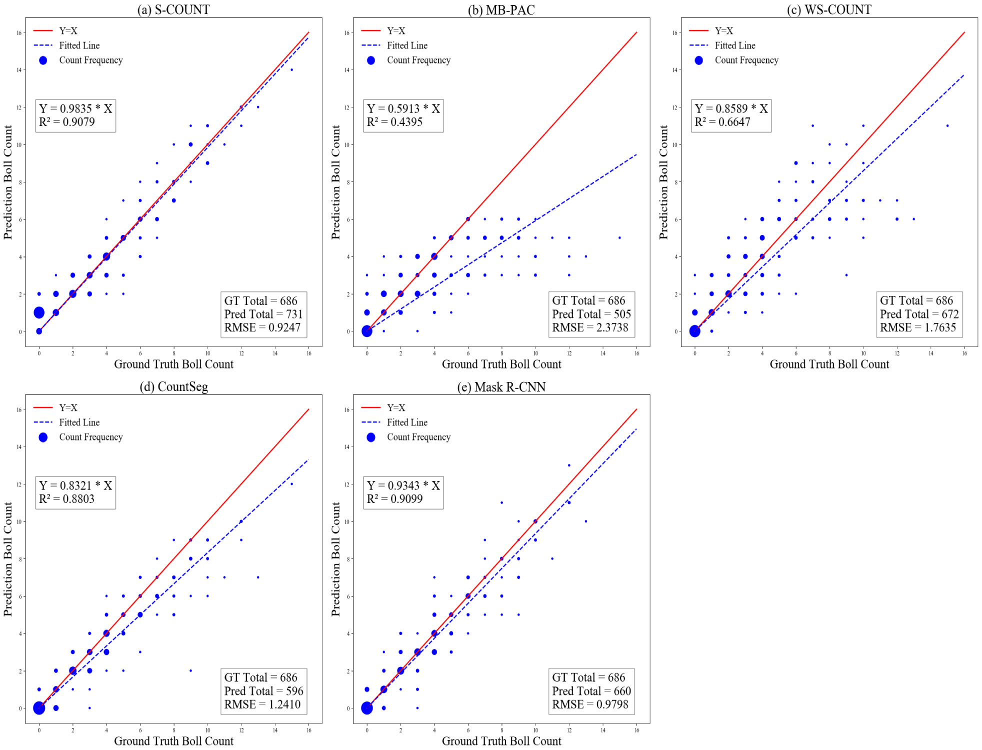

Figure 3.7: **Bubble plots for boll counts**. The performance of best models of each method were plotted in the grond-truth vs. predicted count plot. The figure shows fitted linear regression line and corresponding R^2 value. Also, the total boll count from 200 test (validation) images is shown to demonstrate counting capabilities of the method.

Table 3.4: **Testing set** boll counting mean RMSE along with std. deviation for different methods with respect to boll count per image.

Boll count/image	**0**	**[1-5]**	**[6-10]**	**[11-15]**	**Total**
#Train/Test split	919/50	1852/102	710/42	231/6	**3712/200**
S-COUNT	0.582 ± 0.25	1.069 ± 0.16	1.556 ± 0.26	2.430 ± 1.04	**1.181 ± 0.16**
MB-PAC	1.285 ± 1.04	1.371 ± 0.24	3.705 ± 0.19	8.414 ± 0.31	**2.567 ± 0.21**
WS-COUNT	0.708 ± 0.07	1.431 ± 0.25	2.489 ± 0.23	5.314 ± 0.39	**1.826 ± 0.05**
CountSeg	0.286 ± 0.06	0.869 ± 0.02	1.978 ± 0.14	3.805 ± 0.45	**1.284 ± 0.08**
Mask R-CNN	0.566 ± 0.20	0.982 ± 0.04	1.586 ± 0.42	2.884 ± 1.03	**1.175 ± 0.20**

Comparisons of mean RMSE of five models on test set revealed that the weakly supervised methods performs well within a certain range above the subitizing range ([0,10]) while is not as accurate as supervised methods for boll counts greater than ten (Table 3.4). It can be noted that as the boll number grows, all the methods showed more errors as the images with higher number of bolls are likely to contain occlusions, closely placed bolls or visually challenging instances that are hard to detect without any depth information. Overall, the mean RMSE values of weakly supervised methods were close to the supervised methods despite being trained with minimum amount of supervision.

To verify the performance of this end-to-end counting approach, these models were compared with respect to the total boll count on full scale images (Table 3.5). The outputs of one weakly supervised method (CountSeg) and one supervised method (Mask R-CNN) were illustrated on images representing the dataset of in-field and potted plants taken under different conditions (Figure 3.8). For example, image Boll_008 was taken from a tilted angle such that the plant casts shadows of the bolls on the ground creating a false boll instance in 2D image. In this case all the methods show random performance as the boll count varies largely for each model. Supervised methods such as Mask R-CNN overestimated the count due to false predictions of masks to shadows while CountSeg underestimated the number due to occluded bolls forming a single peak. Image Boll_022 depicts an irregular shaped cotton plant that has visually distinguishable cotton bolls and every patch in the image has a different background texture. The top part of image has bolls on the far away background, which produce noisy input and all the models mis-detected few of those bolls in the background as true bolls belonging to the plant in final predictions. Nevertheless, due to the clear separation between bolls, both fully and weakly supervised models gives predictions close to the actual ground truth. In fact, CountSeg yields better counts on all of its 5 model variations than S-COUNT and Mask R-CNN.

One of the challenges for this end-to-end processing pipeline was to handle crowded bolls in a single image as can be seen in image Boll_041. All the methods underestimated the total count for the image because this image is visually difficult for counting in 2D view-point. Nevertheless, the counts obtained from WS-COUNT show consistency and are close to the predictions of supervised counts. Extending the application to indoor potted plants (Boll_116), counting performance shows highly precise results for supervised and weakly supervised methods. The accuracy is slightly less as the background and illumination contribute to the noise in the image, which led to reduced spatial context that can be used to separate boll instances. On the other hand, potted images in outdoor conditions (Boll_127) are slightly easier to process as the boll instances are less occluded by background and illumination. To sum up, supervised methods performed slightly better in adverse image conditions but the weakly supervised methods without adequate spatial context during training were able to localize most of the boll instances even in adverse conditions such as noisy background, large boll population, and changing illuminations.

Table 3.5: **Hold-out Test set** average boll count with std. deviation for the five models of each method. These images represents entire plants under various conditions.

Image	**Boll_008**	**Boll_022**	**Boll_041**	**Boll_116**	**Boll_127**
Actual Count	41	34	100	20	22
S-COUNT	43.6 ± 4.22	41.8 ± 11.12	87.8 ± 7.50	16.4 ± 1.67	21.4 ± 1.67
MB-PAC	**41.2 ± 4.87**	42.0 ± 12.39	59.2 ± 8.23	18.8 ± 5.26	13.6 ± 1.95
WS-COUNT	48.2 ± 1.31	40.2 ± 3.49	86.4 ± 2.79	23.2 ± 1.48	18.8 ± 1.30
CountSeg	37.8 ± 2.49	**34.2 ± 0.84**	81.8 ± 3.56	17.0 ± 0.00	18.0 ± 1.414
Mask R-CNN	46.0 ± 6.16	34.8 ± 1.90	**89.0 ± 3.94**	**17.2 ± 0.84**	**21.2 ± 0.45**

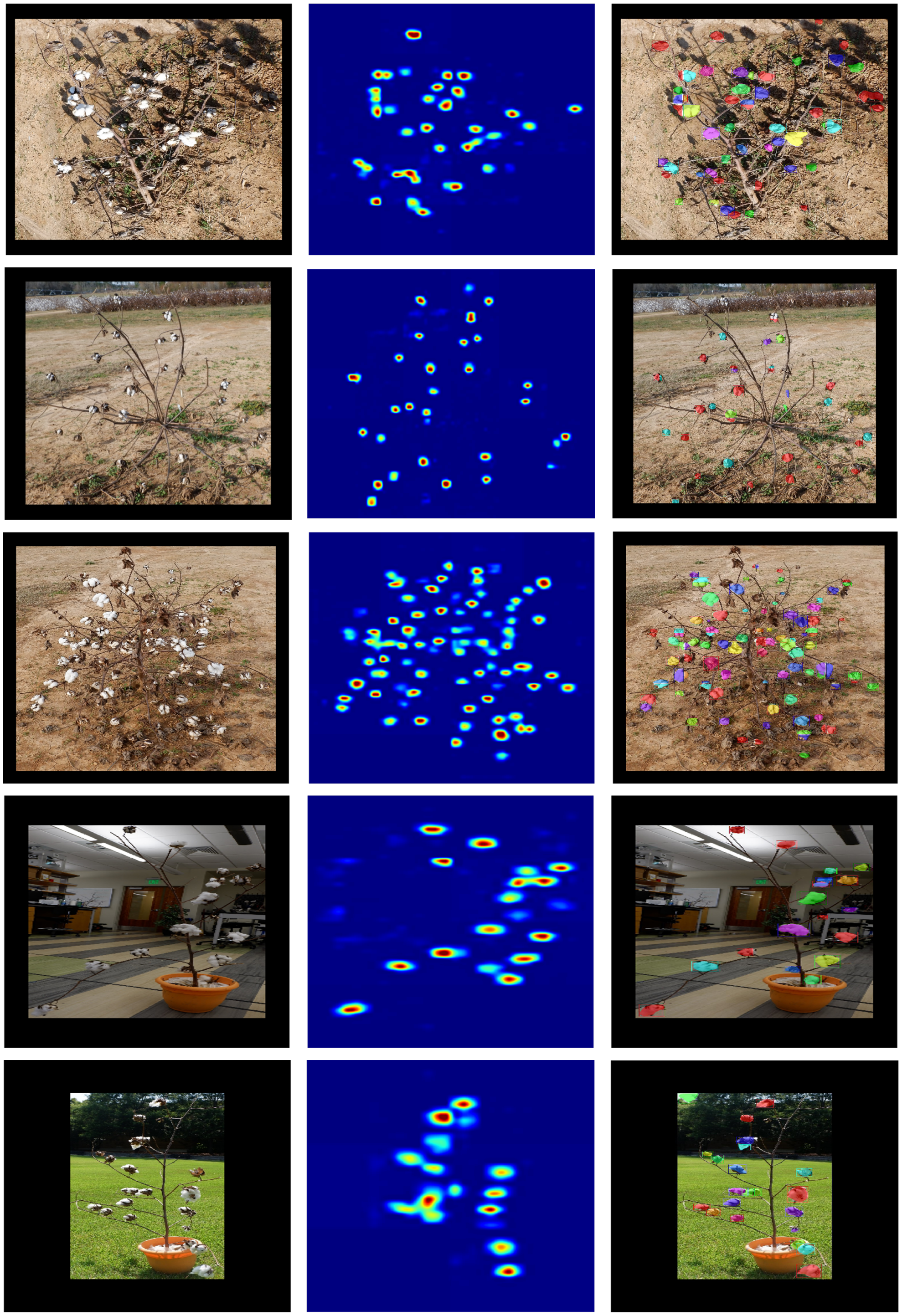

Figure 3.8: **Comparison of CountSeg and Mask R-CNN**. This shows the output from CountSeg density maps and prediction instance masks from Mask R-CNN for 5 held-out test samples: (starting from top row) Boll_008, Boll_022, Boll_041, Boll_116, Boll_127. It can be observed that even with lower supervision, CountSeg was able to retain the spatial contexts for most of the bolls.

The CountSeg method produces highly accurate density maps which can be combined with a proposal ranking method to achieve weakly supervised instance segmentation (Figure 3.8). The instance masks produced by CountSeg can be compared with the masks obtained from Mask R-CNN, thus in turn, replacing the intensive supervised method with a low-cost weakly supervised method.

Average annotation time was compared among three labeling methods (mask, point, and class labels) and class labeling method showed a clear advantage (Figure 3.9). For labelling instance masks, the annotator has to draw exact polygons that cover all the pixels in that instance. As a result, the time required to annotate a single instance varies according to the size, shape, and visual separability from other boll instances (which is difficult to achieve in 2D images). In case of point labels, annotator is required to click/draw a single point for a single instance which can be done in comparably same amount of time for a fixed number of bolls per image. Thus, the average time required for a single count value was considered for representing point labels. The class labels are the simplest form of labelling in which annotator can simply select "yes" or "no" for the presence or absence of the object instance, respectively. Irrespective of the boll count per image, on an average, class labels take around 2 seconds per image.

It can be seen that point labels are at least 10 times faster for the images with boll counts in subitizing range ([0,10]) and at least 15 times faster for images beyond counting range. This essentially allows researchers to use more raw data in the training process given the fixed amount of time. In terms of boll counting, the performance gap between fully supervised and weakly supervised seems acceptable considering the huge advantage with regard to annotation costs. Experimentation with current weakly supervised methods to improve the spatial consistency such as the use of Generative Adversarial Networks (GAN) along with WS-COUNT architecture may result in an even better performance in the future (Bellocchio et al., 2020).

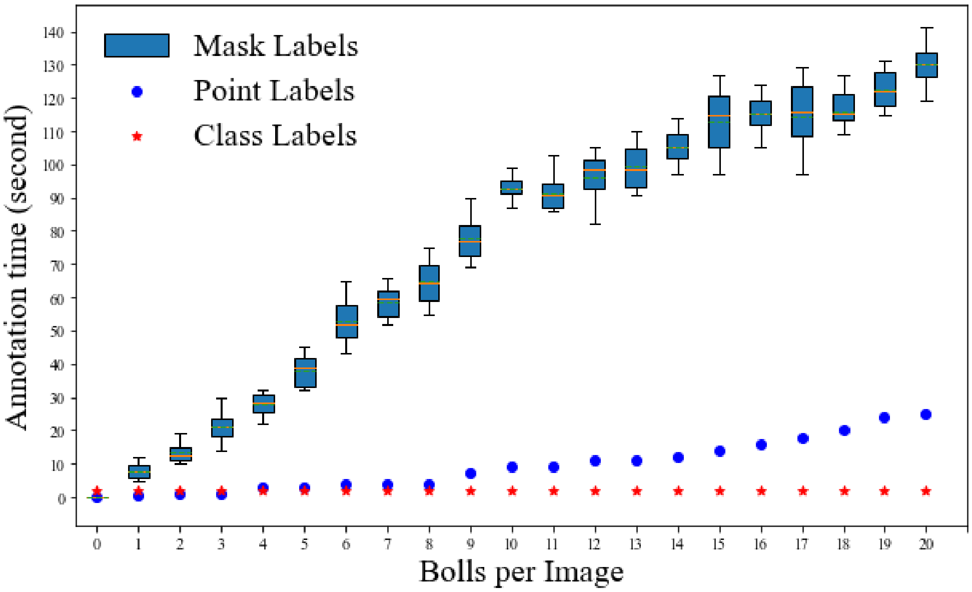

Figure 3.9: **Annotation times comparison by label type**. The time taken for labelling an image tile was measured with respect to the boll count in that image tile. A sample set of 10 images per boll count was considered and the average times are reported for point labels while the box plots represents the range of time taken by mask labels for the same boll count.

3.4.2 Main Stalk and Node Segmentation

For the boll counting task it was observed that instance segmentation using Mask R-CNN yields sharp masks and differentiates the boll instances accurately. In case of main stalk and nodes, identifying a single instance is difficult in a 2-dimensional image due various factors such as lack of depth information, overlapping parts, highly dense and similar features. However, with the help of a small annotated data the models discussed above performed significantly better with respect to the segmentation. Figure 3.10 shows representative plant samples from the test set after the segmentation and trait extraction pipeline. The main stalk segmentation, despite of facing the problem of occlusion, works sufficient enough to identify the longest stalk in most of the cases. On the other hand, identifying nodes present on the main stalk faces some challenges. Despite of identifying nodes at the lower level, the node segmentation model fails to identify most of the nodes at higher level of main stalk. However, for a cotton plant, most of the fruiting branches emerges from the lower part of the main stalk and for researchers, these lower level nodes are of high interest.

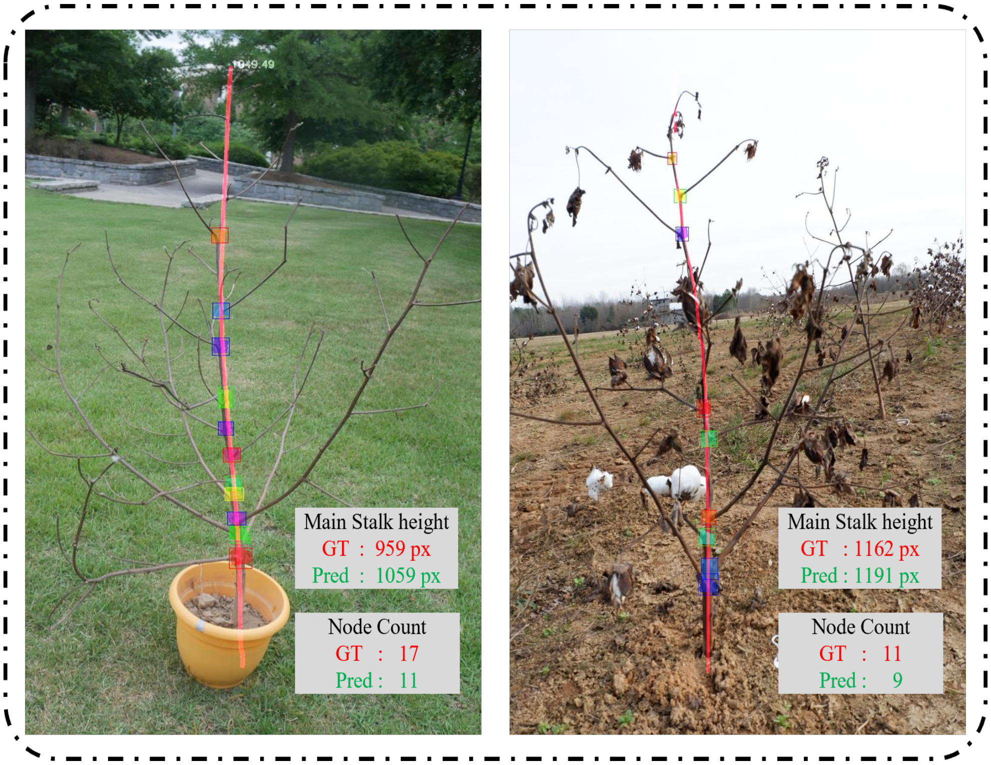

Figure 3.10: **Segmentation of main stalk and nodes**. Main stalk and nodes are segmented separately with the help of trained models and the final skeleton of main stalk is obtained after post-processing individual instance masks. The models can be used for the plants in potted as well as in-field conditions.

The failure in segmentation mostly occurred due to the ambiguity of main stalk and nodes in the image. During the main stalk segmentation, it was observed that the mask for top part of main stalk was not segmented properly or missed totally. There can be several reasons for this, primarily, the top part was thinner than the rest of the main stalk which contributes to less pixels in the ground truth mask itself, sometimes just a line of connecting pixels. Thus, at the time of learning the model's loss function does not yield a higher loss for missing that part. Another reason was the inconsistent background at the top part that can confuse the model from associating the top part to the main stalk instance due to the slight obfuscation of the image. Figure 3.11 shows some of the segmentation outlier cases of potted plants in outdoor conditions that failed to segment main stalk figure 3.11(a,b,d) or all the nodes figure 3.11(b,c,d) accurately.

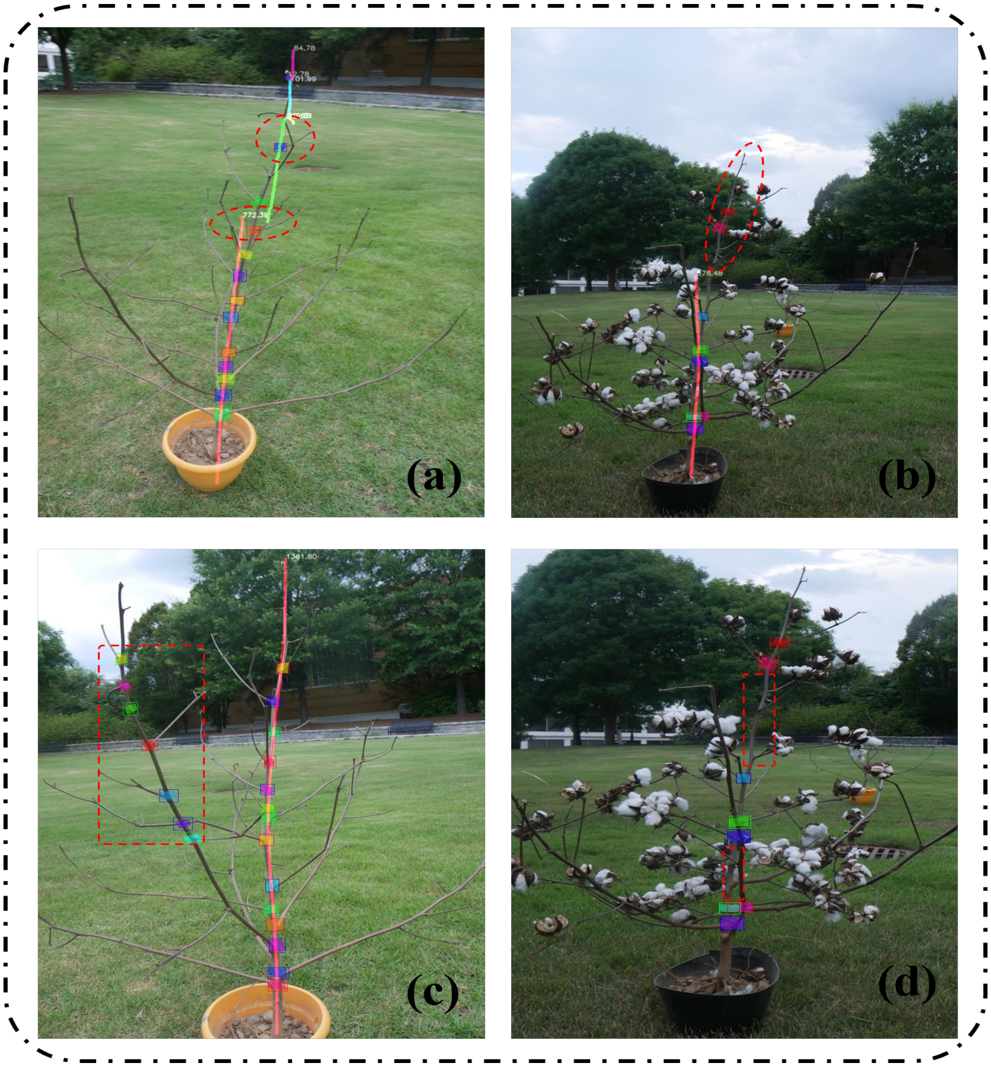

Figure 3.11: **Outlier cases in segmentation and detection**.The main stalk segmentation model suffered from loss or discontinuity in output (a,b,d) that are marked with red ovals. The discontinuities occur mostly in case of curved main stalk and was addressed in post-processing. While node segmentation model produced extra nodes (c), or missed intermittent nodes (b,d) in case of ambiguous main stalk.

For the segmentation of nodes, the ground truth masks were labelled with a polygonal shape approximately around the area at which a primary branch starts at main stalk. Furthermore, the separation between two nodes varies over the length of main stalk as there are more nodes present at the bottom than top part of a cotton plant (Ritchie et al., 2007). The node segmentation model, thus, failed to detect all the nodes, missing some of them or identifying the nodes that are not on main stalk. Therefore, based on observations, the images with node counting error greater than or equal to 6 (missing or extra) were considered as outliers. Although the additional nodes segmented on other branches share similar feature representation as of a primary node, this study was to identify nodes on main stalk only and hence those should be eliminated. It can be said that in the 2-D images with densely spaced nodes distinguishing separate nodes becomes challenging when the plant is not well-captured or there are occlusions that breaks the continuity of the main stalk. Some of the issues in identifying main stalk and nodes can be solved with the help of high-quality image data with precise annotations. The issue of occlusion cannot be addressed directly with the help of 2-D image segmentation model, however, pre-processing and modern imaging techniques such as subject removal, gap filling may alleviate it.

3.4.3 Main Stalk Height and Node Count

Once all the instance masks of main stalk are post-processed, the length of final line segment of the main stalk was measured in terms of pixels. Figure 3.12 shows the scatter plot and error histogram for the measured height from instance masks with respect to height measured from post processing ground truth masks. The predicted heights were close to the ground truth within a range of $\pm 45\%$ of pixels without considering the outliers. The relative pixel error was less than 20% in 20 out of 28 test images. with a mean of $19.4\% \pm 16.5$, showing that the main stalk segmentation can produce quality results if the instances are identified without complete segmentation failures. Additionally, a linear fit model is obtained from the scatterplot of predicted vs. ground truth height that shows a good correlation with almost overlapping best fit line. Considering the average image dimensions were in the range of a 2000 pixels with main stalk occupying 1200 pixels in height at an average, the RMSE of 284 pixels can be considered significantly better. These measurements prove the effectiveness of individual mask segmentation and the post-processing to obtain the main stalk skeleton.

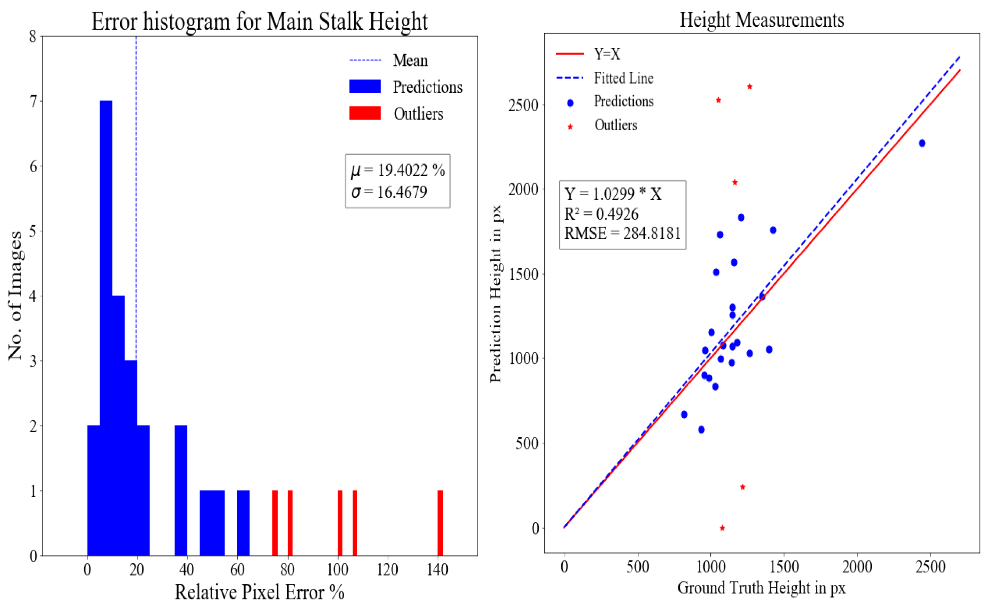

Figure 3.12: **Scatter plot for main stalk height calculation**. The model's prediction versus ground truth pixel count is plotted and a regression line is obtained. The outliers due to the main stalk segmentation failure were not considered in the regression analysis. Corresponding error histogram is shown with the relative pixel error in percentage with outliers.

In terms of node segmentation, the model underperformed while segmenting all of the nodes present. Figure 3.13 shows the node counting error histogram along with the scatterplot for the predicted vs. ground truth node count. As the average node count for test dataset was 15 nodes, the predictions beyond ±5 nodes than ground truth were considered as segmentation failures and was not considered in modelling the regression. The RMSE of 3.34 shows that even if there are many missing nodes, at a higher level of inspection - such as 2-D imaging- this model can be used to get an approximate estimate of the node counts. Furthermore, the internode distance will not differ much considering the average number of nodes and total main stalk height.

Overall, the two models trained separately to extract main stalk and nodes can segment acceptable instances even when trained with a smaller labelled dataset of 175 training images. To get accurate traits from these segmented instance, post-processing is must as the model cannot associate individual instances to the plant skeleton. Hence, the role of plant phenotyping libraries such as PlantCV is crucial to obtain several morphological traits with the help of well-organized and generalized functions.

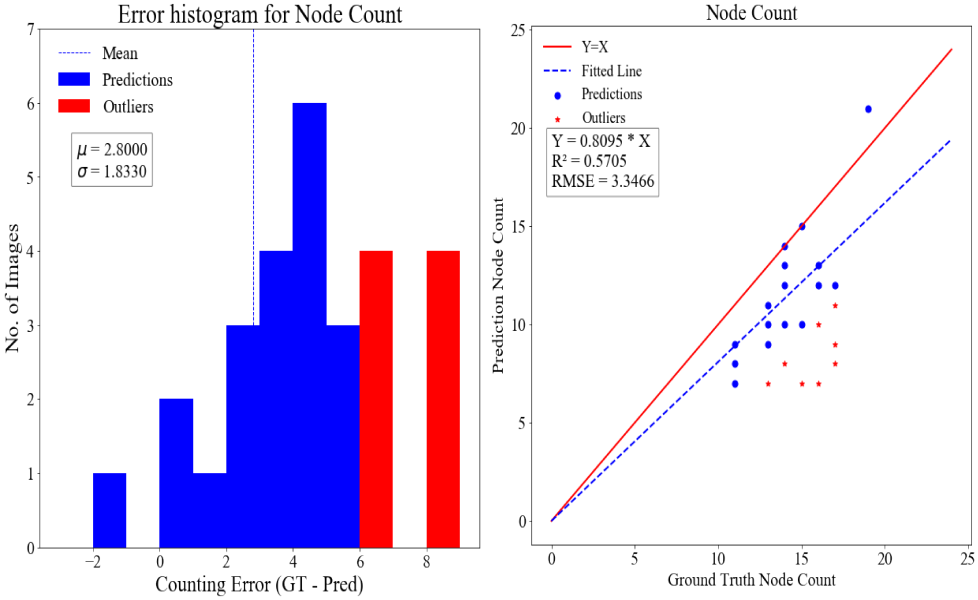

Figure 3.13: **Scatter plot for node counting.** The model's prediction versus ground truth node count is plotted and a regression line is obtained. The outliers due to the node segmentation failure were not considered in the regression analysis. Corresponding error histogram is shown with the error ranging from [-3, 10] nodes missed along with outliers (above 6).

3.4.4 Discussion

Since last decade, supervised learning approaches are studied widely across the agricultural domain (Hasan et al., 2020; Loey et al., 2020) but they still pose a question of how much data is sufficient enough for the training (Adke et al., 2020; Ni et al., 2020). The proposed trait extraction pipeline for main stalk and node segmentation suffers from the same problem of data scarcity and needs to be addressed with more training data and precise annotations. Nevertheless, the main stalk height measurements can be performed at a real scale with the help of calibrated data which could certainly estimate the desired traits with high accuracy. Additionally, with the help of branch annotations, a third model to detect secondary fruiting branches can be developed. These segmented branches could be subjected to post-processing for skeletonization with custom algorithms similar to sorghum skeletonization(Gaillard et al., 2020). The proposed methods for instance segmentation of main stalk and nodes are studied for special cases of in-field and potted plants, whereas, in the cotton farm the conditions are not guaranteed to be similar. There are several challenges such as closed spacing between plants, the high volume of background plants and their branches, irregularly shaped main stalks, etc. To perform accurate and precise instance segmentation in such conditions still poses a problem to the proposed methods.

On the other hand, using 3-D imaging methods to collect the plant data is piquing interest in plant phenotyping. 3-D point cloud encompasses the missing depth data from 2-D images, and thus, are considered widely to avoid the 2-D overlapping structures and occlusions by other parts (S. Sun et al., 2018). A 3-D point cloud obtained from a cotton plant can be segmented with the help of deep learning to accurately segment the organs to perform more intricate trait extraction as demonstrated by Saeed and Li, 2021. The output of this point-voxel-based method can be considered as a superior alternative to perform the individual instance segmentation compared to a 2-D Mask R-CNN-based pipeline. However, data collection in such scenarios demands extra efforts than a 2-D supervised approach. Furthermore, the pre-processing and post-processing of such data is a highly skilled and laborious task. Therefore, to perform the initial field analysis and predict the approximate plant traits, the proposed method is best suitable.

Traditional approaches to extract the plant phenotypic traits without supervised machine learning involves approximate estimations due to lack of trait labels (boll count in this case) and instead of settling for an approximate estimation, weakly supervised learning could be used to solve the ill-posed problem of data annotations. Recently, weakly supervised methods are going through major developments such as new learning paradigms that improves peak response maps to yield instance activation maps (Y. Zhu et al., 2019), combining the old approaches(CAM)with traditional machine learning (PCA) (Ibrahem et al., 2021; Muhammad & Yeasin, 2020), or combining supervised and weakly supervised approaches (Otálora et al., 2021). In this study, the example learning task of counting cotton bolls was performed using both supervised and weakly supervised approaches along with their annotation costs. With minimum efforts spent in data annotations, weakly supervised methods performed as well as supervised approach under similar inference conditions. Also, the intermediate results obtained from the peak response maps could easily be used for semantic segmentation problem, and if provided with object proposals, could perform instance segmentation at the cost of classification. Finally, while working with weakly supervised methods for boll detection, few questions remain open for discussion such as occlusion handling, identifying instances in high density of cotton bolls, and real-time in-field detection of bolls. Supervised segmentation of cotton bolls shown great awareness of occlusions and counting densely populated cotton bolls. However, the proposed weakly supervised methods do not learn on the instance boundaries, instead they aggregate the peaks observed in the boll feature map, which makes them prone to under-count in such scenarios. Therefore, to apply these techniques in real-time detection needs more improvements such as isolated image acquisition platforms that captures the images with improved focus on a single plant Jiang et al., 2018. The advancements in current WS-COUNT architecture (Bellocchio et al., 2020) to handle high density and occlusions can help answering those questions with a few modifications in the current experimental setup. Furthermore, since the data annotated for weakly supervised methods is easy to reproduce and can be reused across a variety of different algorithms, one can try an ensemble of these methods to obtain best match for their counting task. For example, boosting the boll counting weak detector with a pre-trained source using knowledge transfer (Zhong et al., 2020). This may help to solve the problem of dense population of cotton bolls and the occlusions caused by other organs by incorporating the instance boundary knowledge of a strong base learner.

3.5 Conclusions

In this work, the applications of supervised and weakly supervised learning methods were studied to obtain phenotypic traits of cotton plant. With the help of manually annotated datasets for segmenting cotton plant main stalk and nodes, a supervised trait extraction pipeline was developed that achieved sufficient segmentation performance. Furthermore, to demonstrate the use of weakly supervised methods in boll counting two different approaches: WS-COUNT and CountSeg, were studied and compared with supervised counting results. The weakly supervised frameworks can be used for boll segmentation with the help of object proposals and could be used for semantic segmentation of cotton plant skeleton . Future work will be directed at improving the imaging methods to train models with high resolution data and collecting calibrated data for trait extraction.

Chapter 4

Conclusion

In this thesis we demonstrated the use of deep learning and computer vision in plant phenotyping with the examples of corn image instance segmentation and cotton plant trait extraction. A Mask R-CNN based corn consumption estimation framework was presented and evaluated to quantify the consumption of corn with a relatively small number of images collected by community scientists. This thesis demonstrated that the Mask R-CNN model can be used to produce high quality results of pixel-wise segmentation for the challenging task of multi-label segmentation of consumed corn. We have proposed two approaches for labelling the ground truth and found that segmenting only the whole corn and its consumed part is sufficient for estimating consumption. The best results were obtained when the training data were sufficient and labelled with high accuracy. We also studied the effect of varying light conditions and background and found that our model, which was not trained with such images, was able to identify certain segmentation instances accurately, and can be improved upon by including such images for further training. The framework developed in this study can be used to predict more samples collected in the GMO Corn Experiment and will produce reliable results more efficiently. Future work will be directed at improving the variation in accuracy as well as testing the visually challenging images.

On the other hand, the applications of supervised and weakly supervised learning methods were studied to obtain phenotypic traits of cotton plant. With the help of manually annotated datasets for segmenting cotton plant main stalk and nodes, a supervised trait extraction pipeline was developed that achieved sufficient segmentation performance. Furthermore, to demonstrate the use of weakly supervised methods in boll counting two different approaches: WS-COUNT and CountSeg, were studied and compared with supervised counting results. The weakly supervised frameworks can be used for boll segmentation with the help of object proposals and could be used for semantic segmentation of cotton plant skeleton . Future work will be directed at improving the imaging methods to train models with high resolution data and collecting calibrated data for trait extraction along with additional traits such as branch angle, internode distance, etc.

Appendix A

Supplementary Materials for Corn Image Instance Segmentation

A.1 Supplimentary Data

All the annotated images, labels, and source code are archieved in the GitHub repository (https://github.com/UGA-BSAIL/Corn-Segmentation) along with the instruction to run pretrained models and train new models for consumption estimation.

A.2 Supplementary Tables and Figures

A.2.1 Statistical analysis of the model performance with various sample sizes

With two-sided t-tests (Table Table A.1-Table A.3), the performance metrics were statistically analyzed to find the significant improvements corresponding to sample size increase.

Table A.1: **Statistical Analysis using t-test on overall segmentation mAP with various sample size (t-value, P-value)**

Sample size	50	100	150	200	250	300
50	0,1	**1.93,0.09**	1.12,0.29	3.46,0.009	2.38,0.04	4.3,0.002
100	-	0,1	**0.34,0.7**	2,0.08	0.53,0.60	2.9,0.02
150	-	-	0,1	**1.7,0.12**	0.7,0.5	2,0.07
200	-	-	-	0,1	**1.54,0.1**	0.18,0.8
250	-	-	-	-	0,1	**2.34,0.04**
300	-	-	-	-	-	0,1

Table A.2: **Statistical Analysis using t-test on segmentation IoU for Whole corn with various sample size (t-value, P-value)**

Sample size	50	100	150	200	250	300
50	0,1	**0.97,0.35**	2.36,0.04	2.55,0.03	2.94,0.01	3.71,0.005
100	-	0,1	1.46,0.18	1.7,0.12	2.12,0.06	3.02,0.01
150	-	-	0,1	**0.35,0.73**	0.85,0.4	1.95,0.08
200	-	-	-	0,1	**0.45,0.6**	1.44,0.18
250	-	-	-	-	0,1	**1,0.34**
300	-	-	-	-	-	0,1

Table A.3: **Statistical Analysis using t-test on segmentation IoU for Bare cob with various sample size (t-value, P-value)**

Sample size	50	100	150	200	250	300
50	0,1	**1.93,0.09**	1.12,0.29	3.46,0.009	2.38,0.04	4.3,0.002
100	-	0,1	**0.34,0.7**	2,0.08	0.53,0.60	2.9,0.02
150	-	-	0,1	**1.7,0.12**	0.7,0.5	2,0.07
200	-	-	-	0,1	**1.54,0.1**	0.18,0.8
250	-	-	-	-	0,1	**2.34,0.04**
300	-	-	-	-	-	0,1

A.2.2 Performance comparision between Mask R-CNN and SoloV2

To explore other segmentation approaches that are more recent and are lighter than Mask R-CNN, we performed experiments on available dataset using SoloV2. To train a SoloV2 model, we had to convert our data (images and annotation masks) into COCO labelling format. For the sake of simplicity and in the interest of time, instead of converting the entire dataset, we converted a smaller sample size (up to 150).

We compared the performance of the two methods with regard to the mean average precision, and IoU for Whole Corn and Bare Cob classes.

Table A.4: **Effect of training sample size on segmentation performance of Mask R-CNN and SoloV2 model**

Metric	**mAP**		**Whole Corn IoU**		**Bare cob IoU**	
Method	**Mask R-CNN**	**SoloV2**	**Mask R-CNN**	**SoloV2**	**Mask R-CNN**	**SoloV2**
50	0.574	0.09	0.868	0.5792	0.606	0.435
100	0.592	0.24	0.876	0.7412	0.636	0.436
150	0.588	0.42	0.885	0.815	0.626	0.561

Bibliography

Abdulla, W. (2017). Mask R-CNN for object detection and instance segmentation on keras and tensorflow [https://github.com/matterport/Mask_RCNN [Online; accessed: 01/11/2018]]. *GitHub repository*. https://github.com/matterport/Mask_RCNN.

Adke, S., Haro Von Mogel, K., Jiang, Y., Li, C., et al. (2020). Instance segmentation to estimate consumption of corn ears by wild animals for gmo preference tests. *Frontiers in artificial intelligence*, *3*, 119.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 577–584.

Ashapure, A., Jung, J., Chang, A., Oh, S., Maeda, M., & Landivar, J. (2019). A comparative study of rgb and multispectral sensor-based cotton canopy cover modelling using multi-temporal uas data. *Remote Sensing*, *11*(23), 2757.

Atanbori, J., Montoya-P, M. E., Selvaraj, M. G., French, A. P., & Pridmore, T. P. (2019). Convolutional neural net-based cassava storage root counting using real and synthetic images. *Frontiers in plant science*, *10*, 1516.

Bellocchio, E. (2019). Ws-count [https://github.com/isarlab-department-engineering/WS-COUNT [Online]]. *GitHub repository*. https://github.com/isarlab-department-engineering/WS-COUNT.

Bellocchio, E., Ciarfuglia, T. A., Costante, G., & Valigi, P. (2019). Weakly supervised fruit counting for yield estimation using spatial consistency. *IEEE Robotics and Automation Letters*, *4*(3), 2348–2355.

Bellocchio, E., Costante, G., Cascianelli, S., Fravolini, M. L., & Valigi, P. (2020). Combining domain adaptation and spatial consistency for unseen fruits counting: A quasi-unsupervised approach. *IEEE Robotics and Automation Letters*, *5*(2), 1079–1086.

Bollis, E., Pedrini, H., & Avila, S. (2020). Weakly supervised learning guided by activation mapping applied to a novel citrus pest benchmark. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019a). Yolact: Real-time instance segmentation. *Proceedings of the IEEE international conference on computer vision*, 9157–9166.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019b). Yolact++: Better real-time instance segmentation. *arXiv preprint arXiv:1912.06218*.

Butler, D. (2012). Rat study sparks GM furore. [https://www.nature.com/news/rat-study-sparks-gm-furore-1.11471 [Online; accessed 30-Oct-2018]]. *Nature*, *486*(7417), 484–485.

Ceccarelli, S. (2015). Efficiency of plant breeding. *Crop Science*, *55*(1), 87–97.

Chamanzar, A., & Nie, Y. (2020). Weakly supervised multi-task learning for cell detection and segmentation. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 513–516.

Chassy, B., & Tribe, D. (2010). Animals graze where there's feed. [Academics Review, http://academicsreview.org/reviewed-content/genetic-roulette/section-1/1-19animals-can%E2%80%99t-identify-gm-crops/ [Online; accessed 30-Oct-2018]]. http://academicsreview.org/reviewed-content/genetic-roulette/section-1/1-19animals-can%E2%80%99t-identify-gm-crops/.

Chen, L., Fu, Y., You, S., & Liu, H. (2021). Efficient hybrid supervision for instance segmentation in aerial images. *Remote Sensing*, *13*(2), 252.

Chen, Y., Lee, W. S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., & He, Y. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. *Remote Sensing*, *11*(13), 1584.

Cheng, B., Parkhi, O., & Kirillov, A. (2021). Pointly-supervised instance segmentation. *arXiv preprint arXiv:2104.06404*.

Cholakkal, H., Sun, G., Khan, F. S., & Shao, L. (2019). Object counting and instance segmentation with image-level supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12397–12405.

Cholakkal, H., Sun, G., Khan, S., Khan, F. S., Shao, L., & Van Gool, L. (2020). Towards partial supervision for generic object counting in natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Durand, T., Mehrasa, N., & Mori, G. (2019). Learning a deep convnet for multi-label classification with partial labels. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 647–657.

Durand, T., Mordan, T., Thome, N., & Cord, M. (2017). Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 642–651.

Dutta, A., Gupta, A., & Zissermann, A. (2016). VGG image annotator (VIA) [http://www.robots.ox.ac.uk/~vgg/software/via/ Version: 2.0.1, [Online; accessed: 01/11/2018]]. http://www.robots.ox.ac.uk/~vgg/software/via/.

Fahlgren, N., Gehan, M. A., & Baxter, I. (2015). Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current opinion in plant biology*, *24*, 93–99.

FAOSTAT. (2018). FAOSTAT statistical database. *Publisher: FAO (Food and Agriculture Organization of the United Nations), Rome, Italy*.

FAOSTAT. (2019). FAOSTAT statistical database. *Publisher: FAO (Food and Agriculture Organization of the United Nations), Rome, Italy*.

Gaillard, M., Miao, C., Schnable, J., & Benes, B. (2020). Sorghum segmentation by skeleton extraction. *European Conference on Computer Vision*, 296–311.

Gehan, M. A., Fahlgren, N., Abbasi, A., Berry, J. C., Callen, S. T., Chavez, L., Doust, A. N., Feldman, M. J., Gilbert, K. B., Hodge, J. G., et al. (2017). Plantcv v2: Image analysis software for high-throughput plant phenotyping. *PeerJ*, *5*, e4088.

Gewin, V. (2003). Genetically modified corn—environmental benefits and risks. *PLoS Biol*, *1*(1), e8.

Ghosal, S., Zheng, B., Chapman, S. C., Potgieter, A. B., Jordan, D. R., Wang, X., Singh, A. K., Singh, A., Hirafuji, M., Ninomiya, S., et al. (2019). A weakly supervised deep learning framework for sorghum head detection and counting. *Plant Phenomics*, *2019*.

Hafiz, A. M., & Bhat, G. M. (2020). A survey on instance segmentation: State of the art. *International Journal of Multimedia Information Retrieval*, 1–19.

Hamuda, E., Glavin, M., & Jones, E. (2016). A survey of image processing techniques for plant extraction and segmentation in the field. *Computers and Electronics in Agriculture*, *125*, 184–199.

Haro von Mogel, K., & Bodnar, A. (2015). The GMO Corn Experiment. [Biology Fortified Inc. https://biofortified.org/experiment/]. https://biofortified.org/experiment/.

Hasan, R. I., Yusuf, S. M., & Alzubaidi, L. (2020). Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion. *Plants*, *9*(10), 1302.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7310–7311.

Huang, Z., Huang, L., Gong, Y., Huang, C., & Wang, X. (2019). Mask scoring r-cnn. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6409–6418.

Huang, Z., Li, Y., & Wang, H. (2020). In-field cotton boll counting based on a deep neural network of density level classification. *Journal of Electronic Imaging*, *29*(5), 053009.

Ibrahem, H., Salem, A. D. A., & Kang, H.-S. (2021). Real-time weakly supervised object detection using center-of-features localization. *IEEE Access*, *9*, 38742–38756.

Jiang, Y., Li, C. et al. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: A review. *Plant Phenomics*, *2020*, 4152816.

Jiang, Y., Li, C., Paterson, A. H., & Robertson, J. S. (2019). Deepseedling: Deep convolutional network and kalman filter for plant seedling detection and counting in the field. *Plant methods*, *15*(1), 1–19.

Jiang, Y., Li, C., Robertson, J. S., Sun, S., Xu, R., & Paterson, A. H. (2018). Gphenovision: A ground mobile system with multi-modal imaging for field-based high throughput phenotyping of cotton. *Scientific reports*, *8*(1), 1–15.

Jiang, Y., Li, C., Xu, R., Sun, S., Robertson, J. S., & Paterson, A. H. (2020). Deepflower: A deep learning-based approach to characterize flowering patterns of cotton plants in the field. *Plant methods*, *16*(1), 1–17.

Johnson, J. W. (2018). Adapting Mask-RCNN for automatic nucleus segmentation. *arXiv preprint arXiv:1805.00500*.

Jung, A. B. (2018). Imgaug. [https://github.com/aleju/imgaug [Online; accessed 30-Oct-2018]]. *GitHub repository*. https://github.com/aleju/imgaug.

Kalluri, T., Varma, G., Chandraker, M., & Jawahar, C. (2019). Universal semi-supervised semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5259–5270.

Koirala, A., Walsh, K. B., Wang, Z., & McCarthy, C. (2019). Deep learning–method overview and review of use for fruit detection and yield estimation. *Computers and Electronics in Agriculture*, *162*, 219–234.

Kumar, K. V., & Jayasankar, T. (2019). An identification of crop disease using image segmentation. *Int. J. Pharm. Sci. Res*, *10*(3), 1054–1064.

Laradji, I. H., Rostamzadeh, N., Pinheiro, P. O., Vazquez, D., & Schmidt, M. (2018). Where are the blobs: Counting by localization with point supervision. *Proceedings of the European Conference on Computer Vision (ECCV)*, 547–562.

Laradji, I. H., Saleh, A., Rodriguez, P., Nowrouzezahrai, D., Azghadi, M. R., & Vazquez, D. (2021). Weakly supervised underwater fish segmentation using affinity lcfcn. *Scientific Reports*, *11*(1), 1–10.

Laradji, I. H., Vazquez, D., & Schmidt, M. (2019). Where are the masks: Instance segmentation with image-level supervision. *arXiv preprint arXiv:1907.01430*.

Li, Y., Cao, Z., Lu, H., Xiao, Y., Zhu, Y., & Cremers, A. B. (2016). In-field cotton detection via region-based semantic image segmentation. *Computers and Electronics in Agriculture*, *127*, 475–486.

Li, Y., Cao, Z., Lu, H., & Xu, W. (2020). Unsupervised domain adaptation for in-field cotton boll status identification. *Computers and Electronics in Agriculture*, *178*, 105745.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Lin, C., Wang, S., Xu, D., Lu, Y., & Zhang, W. (2020). Object instance mining for weakly supervised object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 11482–11489.

Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature pyramid networks for object detection. *CVPR*, *1*(2), 4.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, 740–755.

Liu, J., Lai, H., Jia, Z., et al. (2011). Image segmentation of cotton based on ycbccr color space and fisher discrimination analysis. *Acta Agronomica Sinica*, *37*(7), 1274–1279.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768.

Loey, M., ElSawy, A., & Afify, M. (2020). Deep learning in plant diseases detection for agricultural crops: A survey. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, *11*(2), 41–58.

Lu, H., Cao, Z., Xiao, Y., Zhuang, B., & Shen, C. (2017). Tasselnet: Counting maize tassels in the wild via local counts regression network. *Plant methods*, *13*(1), 1–17.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. (2020). Soybean yield prediction from uav using multimodal data fusion and deep learning. *Remote Sensing of Environment*, *237*, 111599.

McCarthy, C., Hancock, N., & Raine, S. (2009). Automated internode length measurement of cotton plants under field conditions. *Transactions of the ASABE*, *52*(6), 2093–2103.

Muhammad, M. B., & Yeasin, M. (2020). Eigen-cam: Class activation map using principal components. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7.

Naylor, P., Laé, M., Reyal, F., & Walter, T. (2018). Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*.

Ni, X., Li, C., Jiang, H., & Takeda, F. (2020). Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield. *Horticulture Research*, *7*(1), 1–14.

Ni, X., Li, C., Jiang, H., & Takeda, F. (2021). Three-dimensional photogrammetry with deep learning instance segmentation to extract berry fruit harvestability traits. *ISPRS Journal of Photogrammetry and Remote Sensing*, *171*, 297–309.

Normanly, J. (2012). *High-throughput phenotyping in plants: Methods and protocols*. Springer.

Oberti, R., & Shapiro, A. (2016). Advances in robotic agriculture for crops. *Biosystems Engineering*, *100*(146), 1–2.

of Georgia, U. (2015). The georgia advanced computing resource center (GACRC) [https://gacrc.uga.edu/ [Online]].

Otálora, S., Marini, N., Müller, H., & Atzori, M. (2021). Combining weakly and strongly supervised learning improves strong supervision in gleason pattern classification. *BMC Medical Imaging*, *21*(1), 1–14.

Ou, J.-R., Deng, S.-L., & Yu, J.-G. (2021). Ws-rcnn: Learning to score proposals for weakly supervised instance segmentation. *Sensors*, *21*(10), 3475.

Pabuayon, I. L. B., Kelly, B. R., Mitchell-McCallister, D., Coldren, C. L., & Ritchie, G. L. (2021). Cotton boll distribution: A review. *Agronomy Journal*.

Pabuayon, I. L. B., Yazhou, S., Wenxuan, G., & Ritchie, G. L. (2019). High-throughput phenotyping in cotton: A review. *Journal of Cotton Research*, *2*(1), 1–9.

Petti, D. J., & Li, C. (2021). Graph neural networks for plant organ tracking. *2021 ASABE Annual International Virtual Meeting*, 1.

Qu, H., Wu, P., Huang, Q., Yi, J., Riedlinger, G. M., De, S., & Metaxas, D. N. (2019). Weakly supervised deep nuclei segmentation using points annotation in histopathology images. *International Conference on Medical Imaging with Deep Learning*, 390–400.

Ritchie, G. L., Bednarz, C. W., Jost, P. H., & Brown, S. M. (2007). Georgia cotton growth and development. [Georgia Cotton Growth and Development,http://cotton.tamu.edu/General%20Production/Georgia%20Cotton%20Growth%20and%20Development%20B1252-1.pdf [Online; accessed 30-Apr-2021]].

Roseboro, K. (2008). Mice eat farmer's non-GM corn, ignore GM. [The Organic & Non-GMO Report, https://www.non-gmoreport.com/articles/may08/farmers_non-GM_corn.php [Online; accessed 30-Oct-2018]].

Roseboro, K. (2013). Farmer's experiment finds that squirrels prefer organic over GMO corn. [The Organic & Non-GMO Report, https://www.non-gmoreport.com/articles/june2013/farmer-experiment-squirrels-prefer-organic-corn.php [Online; accessed 30-Oct-2018]]. https://www.non-gmoreport.com/articles/june2013/farmer-experiment-squirrels-prefer-organic-corn.php.

Saeed, F., & Li, C. (2021). Plant organ segmentation from point clouds using point-voxel cnn. *2021 ASABE Annual International Virtual Meeting*, 1.

Saleem, M. H., Potgieter, J., & Arif, K. M. (2019). Plant disease detection and classification by deep learning. *Plants*, *8*(11), 468.

Séralini, G.-E., Clair, E., Mesnage, R., Gress, S., Defarge, N., Malatesta, M., Hennequin, D., & Vendômois, J. S. D. (2012). RETRACTED: Long term toxicity of a roundup herbicide and a roundup-tolerant genetically modified maize. *Elsevier*.

Shen, Y., Ji, R., Chen, Z., Wu, Y., & Huang, F. (2020). Uwsod: Toward fully-supervised-level capacity weakly supervised object detection. *Advances in Neural Information Processing Systems*, *33*.

Sibiya, M., & Sumbwanyambe, M. (2019). A computational procedure for the recognition and classification of maize leaf diseases out of healthy leaves using convolutional neural networks. *AgriEngineering*, *1*(1), 119–131.

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in plant science*, *21*(2), 110–124.

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience*, *2016*.

Steinke, J., van Etten, J., & Zelan, P. M. (2017). The accuracy of farmer-generated data in an agricultural citizen science methodology. *Agronomy for Sustainable Development*, *37*(4), 32.

Sukmana, S. E., & Rahmanti, F. Z. (2017). Blight segmentation on corn crop leaf using connected component extraction and CIELAB color space transformation. *2017 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 205–208.

Sun, G. (2020). Object counting and instance segmentation with image-level supervision [https://github.com/GuoleiSun/CountSeg [Online]]. *GitHub repository*. https://github.com/GuoleiSun/CountSeg.

Sun, S., Li, C., Chee, P. W., Paterson, A. H., Meng, C., Zhang, J., Ma, P., Robertson, J. S., & Adhikari, J. (2021). High resolution 3d terrestrial lidar for cotton plant main stalk and node detection. *Computers and Electronics in Agriculture*, *187*, 106276.

Sun, S., Li, C., Paterson, A. H., Chee, P. W., & Robertson, J. S. (2019). Image processing algorithms for infield single cotton boll counting and yield prediction. *Computers and electronics in agriculture*, *166*, 104976.

Sun, S., Li, C., Paterson, A. H., Jiang, Y., Xu, R., Robertson, J. S., Snider, J. L., & Chee, P. W. (2018). In-field high throughput phenotyping and cotton plant growth analysis using lidar. *Frontiers in Plant Science*, *9*, 16. https://doi.org/10.3389/fpls.2018.00016

Tong, P., Zhang, X., Han, P., & Bu, S. (2021). Point in: Counting trees with weakly supervised segmentation network. *2020 25th International Conference on Pattern Recognition (ICPR)*, 9546–9552.

Twine, A., & Redfern, R. (2021). Australian cotton production manual 2021.

Uddin, M. S., & Bansal, J. C. (2021). *Computer vision and machine learning in agriculture*. Springer.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.

Wang, H., Li, H., Qian, W., Diao, W., Zhao, L., Zhang, J., & Zhang, D. (2021). Dynamic pseudo-label generation for weakly supervised object detection in remote sensing images. *Remote Sensing*, *13*(8), 1461.

Wang, J., Yao, J., Zhang, Y., & Zhang, R. (2018). Collaborative learning for weakly supervised object detection. *arXiv preprint arXiv:1802.03531*.

Wang, X., Kong, T., Shen, C., Jiang, Y., & Li, L. (2019). Solo: Segmenting objects by locations. *arXiv preprint arXiv:1912.04488*.

Wang, X., Zhang, R., Kong, T., Li, L., & Shen, C. (2020). Solov2: Dynamic, faster and stronger. *arXiv preprint arXiv:2003.10152*.

Ward, D., Moghadam, P., & Hudson, N. (2018). Deep leaf segmentation using synthetic data. *arXiv preprint arXiv:1807.10931*.

Wei, J.-d., Fei, S.-m., Wang, M.-l., & Yuan, J.-n. (2008). Research on the segmentation strategy of the cotton images on the natural condition based upon the hsv color-space model. *Cotton Sci*, *20*(1), 34–38.

Xia, M., Li, W., Fu, H., Yu, L., Dong, R., & Zheng, J. (2019). Fast and robust detection of oil palm trees using high-resolution remote sensing images. *Automatic Target Recognition XXIX*, *10988*, 109880C.

Xiong, H., Cao, Z., Lu, H., Madec, S., Liu, L., & Shen, C. (2019). Tasselnetv2: In-field counting of wheat spikes with context-augmented local regression networks. *Plant Methods*, *15*(1), 1–14.

Yu, G., Zare, A., Xu, W., Matamala, R., Reyes-Cabrera, J., Fritschi, F. B., & Juenger, T. E. (2020). Weakly supervised minirhizotron image segmentation with mil-cam. *European Conference on Computer Vision*, 433–449.

Zhang, D., Han, J., Cheng, G., & Yang, M.-H. (2021). Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Zhong, Y., Wang, J., Peng, J., & Zhang, L. (2020). Boosting weakly supervised object detection with progressive knowledge transfer. *European conference on computer vision*, 615–631.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., & Jiao, J. (2018). Weakly supervised instance segmentation using class peak response. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3791–3800.

Zhu, N., Liu, X., Liu, Z., Hu, K., Wang, Y., Tan, J., Huang, M., Zhu, Q., Ji, X., Jiang, Y., et al. (2018). Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *International Journal of Agricultural and Biological Engineering*, *11*(4), 32–44.

Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D., & Jiao, J. (2019). Learning instance activation maps for weakly supervised instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3116–3125.