Project report (first quarter Jan 1 2023 to March 31, 2023)

Project funded by North Central Soybean Research Program

Project tile - Field phenotyping using machine learning tools integrated with genetic mapping to address heat and drought induced flower abortion in soybean

Participating institutions – Texas Tech University, Kansas State University, University of Missouri, and University of Tennessee

Goals & Objectives

Long-term Goal – Develop soybean cultivars with 20 to 30% lower flower abortion under favorable to challenging environmental conditions, leading to about 10-15% increase in yield potential

Objectives (Year 1)

• Explore the genetic diversity in flower abortion under different soil moisture and climatic conditions using a large diversity panel

• Develop an image-based field phenotyping system and deep-learning tools to precisely document temporal dynamics in flower abortion and pod retention in genetically diverse soybeans

• Discover environmentally stable and region-specific genomic regions controlling flower abortion in diverse soil types, moisture, and climatic conditions

Progress achieved

Objective 1 - Explore the genetic diversity in flower abortion under different soil moisture and climatic conditions using a large diversity panel

A total of 350 diverse soybean lines were sent for winter nursery seed increase at Costa Rica in December 2022. They were planted in foundation seed increase plot (total of 150 ft row length for each line) to make sure enough seeds (5 lbs) is available for field planting at multiple locations in summer 2023. Among the 350 lines, 310 lines had good germination and plant stand in the seed multiplication field. We expect to receive sufficient seeds for these lines in late April for 2023 summer planting.

Figure 1. Winter nursery seed increase for the project materials at Costa Rica.

Genetic diversity among the group 3 and 4s are targeted in terms of genetic structure

The 310 lines represents genetic diversity of the USDA soybean germplasm collection in maturity group III and IV. We have whole genome sequencing data for this set with an average sequencing coverage of 20x. Approximately, 0.6 million high quality SNPs and 0.5 million In/Del are available for robust GWAS to identify genetic loci and genes regulation of stress resilience and flower abortion in soybean. The average SNP and In/Del density together is about 1 marker/Kbp.

Preparation of field trails at multiple participating locations

The experimental site at the ***University of Missouri*** for this project is located in the Bradford Research Center (Columbia, MO). Three-acre field was reserved in the farm for this project. We will collect soil samples to identify basic soil properties. The field will be prepared for planting in April. The proposed ~310 diverse lines will be planted in mid-May to early-June, depending on the local weather.

The experimental site to evaluate the diversity panel under rain-fed conditions at ***Kansas State University*** will be located at the Agronomy North Farm near Manhattan, KS. Three and one-half acres have been reserved for planting the experiment. Field preparation for planting is underway and soil samples will be taken following planting. We expect to receive seed of the panel from the winter nursery in April or early May (shared by University of Missouri colleagues) with an expected planting date in May.

The experimental site in ***University of Tennessee*** that the experiment will be carried out will be located in West TN Research and Education Center (WTREC) under rainfed condition. We have secured a little over 2.5 acres for this study in 2023. The soil samples collection is in progress and detailed information will be documented about the field. The burndown will be done in a couple of weeks. We will be receiving 310 soybean lines seeds in from University of Missouri colleagues and planting will be done in early May.

The experiment will be conducted on the Quaker Avenue Research Farm at ***Texas Tech University*** in Lubbock, TX. The experiment will be carried out under sub-surface drip irrigation (SDI). Multiple irrigation zones have been obtained for this trail, which total to an area of 3 acres. Soil samples will be collected and analyzed along with documentation of the field history over prior years. Herbicide applications for burndown will be completed in April followed by a pre-emerge herbicide application in mid-May prior to planting of the ~310 soybean lines thereafter.

Objective 2 - Develop an image-based field phenotyping system and deep-learning tools to precisely document temporal dynamics in flower abortion and pod retention in genetically diverse soybeans

Before the field season begins the team has taken good advantage of greenhouse grown soybean plants and other existing datasets to develop a robust machine learning tool to detect flower number and rate of abortion under field conditions.

The team is implementing two general strategies for enumerating aborted flowers and has begun to apply them to greenhouse grown soybean plants.

1. Pre-abortion: Counting flowers on the plant and comparing the counts over time
2. Post-abortion: Collecting and counting aborted flowers over time

Strategy 1: We have developed a preliminary imaging protocol by which images of greenhouse plants are collected from multiple views and with high enough resolution (e.g., 4K x 6K) such that the smallest flowers are comprised of a minimum of 30 pixels. Our proposed strategy would then detect the flowers in two stages.

a) Subsample acquired image and feed it to a node-detection network. Subsampling the original high-resolution image would make it possible for the detection network to ingest it without compromising image fidelity.
b) Having the nodes localized from the previous step, crop the original image, and feed the resulting high-resolution sub-images to a flower-detection network. This ensures that even the smallest flowers are comprised of a sufficiently large number of pixels and yet, the cropped input images are small enough for the network to ingest.

Figure 2. Strategy 1: Flowers are detected by first localizing the plant nodes.

Node-Detection Network: As an initial approach to detecting nodes, we have employed the Faster R-CNN architecture. We started by pre-training our model with a dataset provided by the study in

2023 that focuses on detecting nodes on Eggplant, Chili, and Tomato plants.[1] A summary of and examples images from this dataset can be seen in *Table 1* and *Figure 3*, respectively.

Table 1: Summary of the external dataset used for pre-training the Faster R-CNN model.

Title	Number of Images	Total Number of Nodes
Chili	50	304
Tomato	350	2403
Eggplant	180	1748
All	580	4455

Figure 3: Examples of nodes in the dataset used for pre-training the Faster R-CNN model. (a) Chili; (b) Tomato; (c) Eggplant[1].

Moreover, we constructed a dataset of 154 images that were captured from our greenhouse soybean plants before March 1st. These were subsequently annotated and divided into training and test sets for model development. Notably, during our annotation process, we separated the nodes into two distinct categories: nodes with flowers and nodes without flowers. Further details about this dataset can be found in *Table 2*.

Table 2: Summary of the first dataset captured from our plants and used for model development.

Image Set	Number of Images	Nodes with Flowers	Nodes Without Flowers	Total Number of Nodes
Training	123	817	106	923
Test	31	175	25	200

[1] https://doi.org/10.3390/agriculture13020473

All	154	994	129	1123

Initial results of the pre-trained model on our test set, without any further training on our training set, reveal that the model has a relatively good understanding of nodes. However, it still struggles in densely populated scenes. Moreover, the majority of the predicted bounding boxes by the model fail to enclose the flowers, which may be critical to our flower detection strategy; see *Figure 3*.

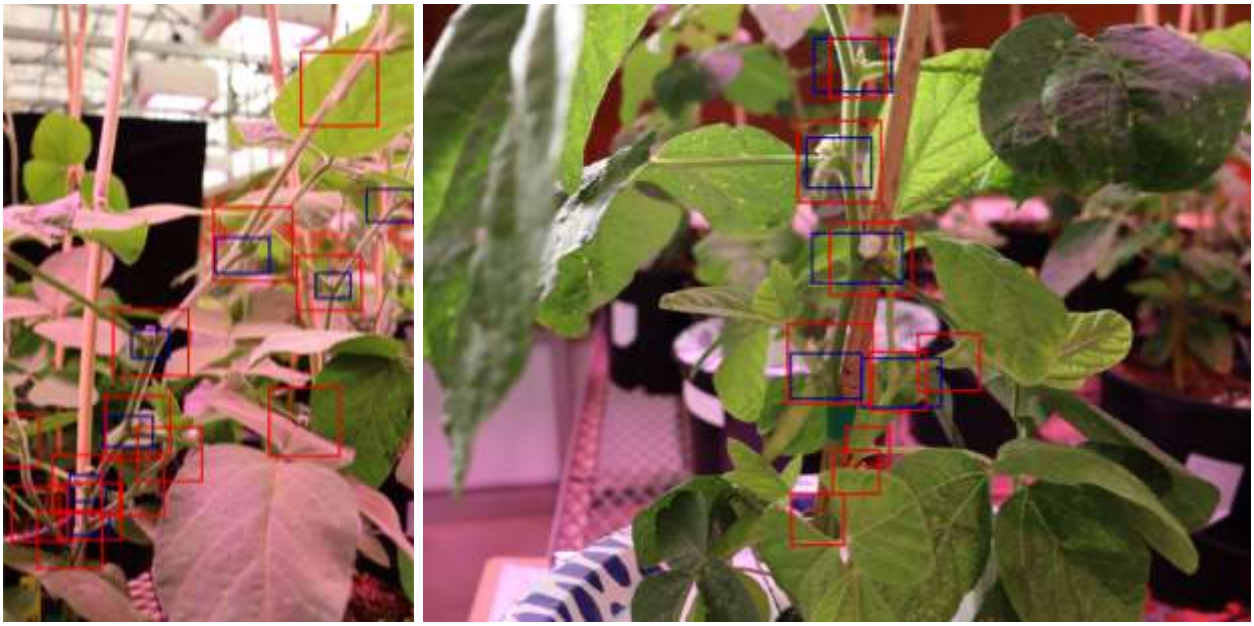


Figure 3: Initial results of the pre-trained model on some of the images from our test set. The red and blue bounding boxes indicate ground-truth and model's predictions, respectively.

We then proceeded to fine-tune the pre-trained model for our application by training it on our training set. As a result, the model exhibited improved performance on our test set, particularly in densely populated scenes. Furthermore, the predicted bounding boxes were more suitable for our flower detection strategy. *Figure 4* demonstrates the performance of the model on the same images depicted in *Figure 3*.

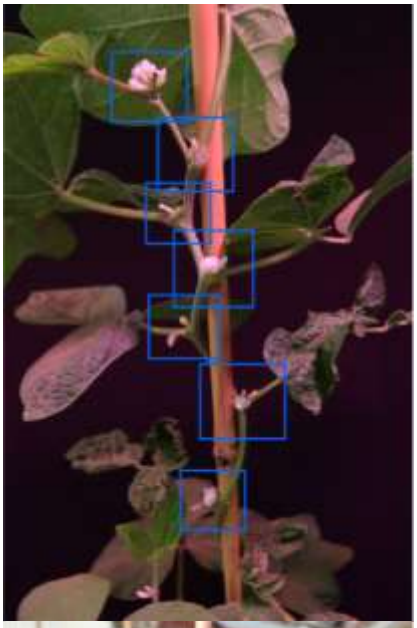


Figure 4: Results of the fine-tuned model on our test images. To prevent overcrowding, only the model's predictions are displayed. Ground-truths are displayed in Figure 3. The blue and purple bounding boxes are indicators of nodes with and without flowers, respectively.

By March 16th, a new dataset of 466 images was compiled to be annotated for further model development. This dataset includes several characteristics that the model had not encountered before:

1) Images from a different greenhouse with more congested backgrounds; *Figure 5A*.
2) New soybean plant varieties; *Figure 5B*.
3) Nodes with several matured soybean pods; *Figure 5C*.

The existing model's inference on this dataset indicates that the model's ability to generalize is reasonably good as it is able to locate most of the visible nodes in the new images; see *Figure 5*. However, as seen in *Figure 6*, the model also outputs several more False-Positives (FP) and False-Negatives (FN) in the new images.

Figure 5: Results of the existing model on the new dataset. The blue and purple bounding boxes indicate nodes with and without flowers, respectively. A) Images from a new location; B) New soybean plant varieties and C) Nodes with several developed soybean pods.

Figure 6: Examples of false positives and false negatives (two for each) from left to right.

Future work includes:

1) Simplifying the annotation process for the new dataset, which contains three times more images than the previous one, by using the existing model's predictions (as shown in Figure 5) as preliminary annotations. Therefore, the annotators will primarily focus on refining the predicted bounding boxes and occasionally making additions or deletions. This approach will significantly accelerate the annotation process, which is essential for efficient model development.
2) Exploring and implementing other state-of-the-art network architectures that may be better suited and capable of achieving superior performance for our application.
3) Associating the model predictions with the ground truth flower and node data to ascertain the efficiency of the model predictions and the extent of refinement needed for models to be precise to allow for deployment under field conditions.

Flower Detection Network: Similar to the node detection network, the flower detection network is also based on the Faster R-CNN architecture. Specifically, we used the Faster R-CNN implementation available in Detectron2 (a library containing state-of-the-art detection and segmentation algorithms made publicly available by Facebook AI Research). We trained an initial model based on a dataset published by Zhu et al. (2022). A summary of the dataset is shown in *Table 3*, while the statistics on the training/validation/test subsets are shown in *Table 4*. and some sample images from the dataset are shown in *Figure 7*.

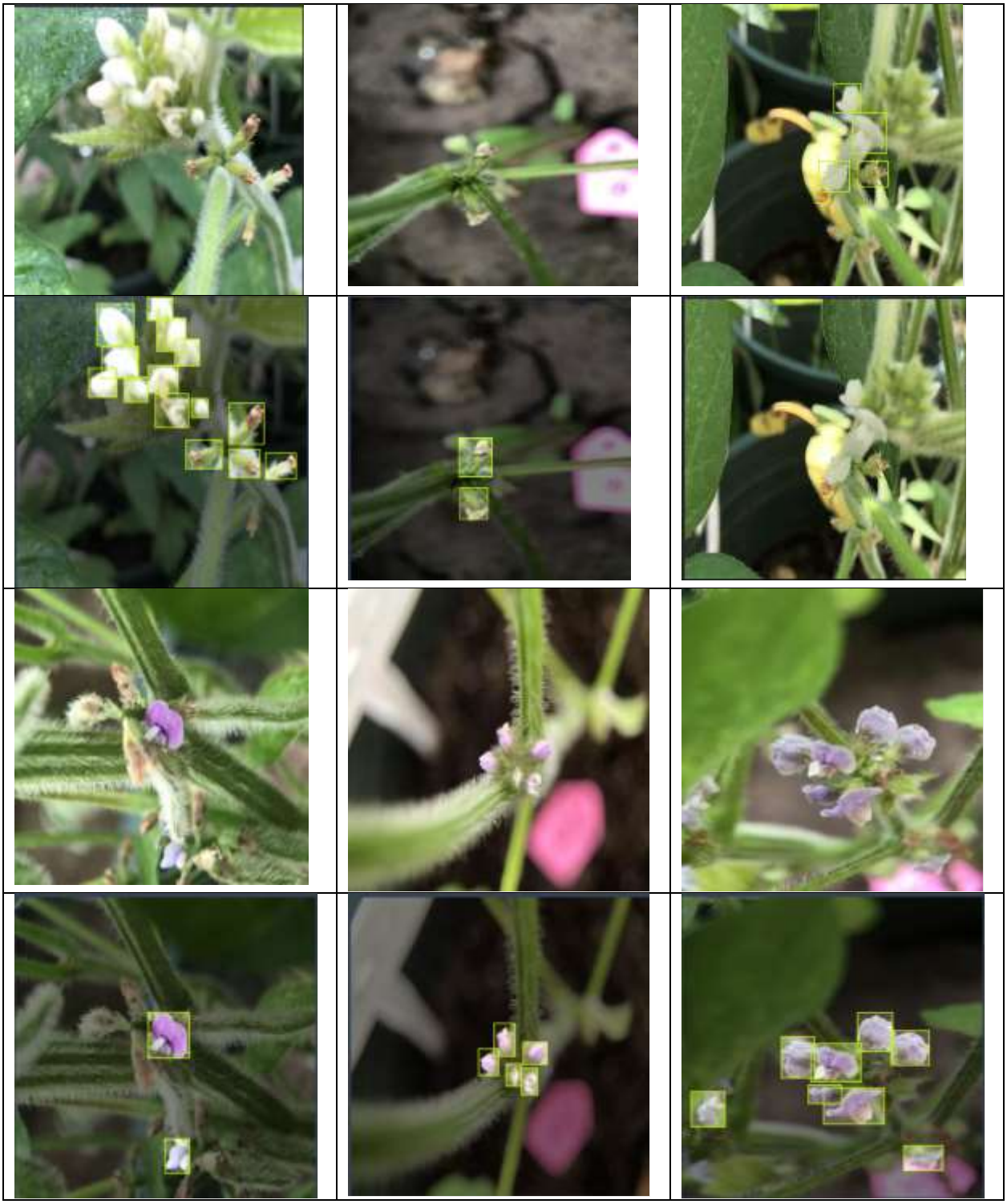
Table 3. Summary of the dataset for soybean flower detection (Zhu et al., 2022)

Variety	Image acquisition time	Podding habit	Number of images	Color of flower
DN252	2019	Sub-limited podding habit	568	White
ChunFengZao	2020	Sub-limited podding habit	545	White
ZheNong NO. 6	2020	Limited podding habit	266	Purple
HN51	2019	Limited podding habit	516	Purple

Table 4. Training/validation/test subsets used to train Faster R-CNN model for flower detection (Zhu et al., 2022)

Subset	Number of Images
Training	1364
Validation	152
Testing	379
Total	1895

Figure 7. Examples of purple and white flowers in the soybean flower dataset used for training [Zhu et al., 2022], together with their manual ground truth annotations (shown below each original image)

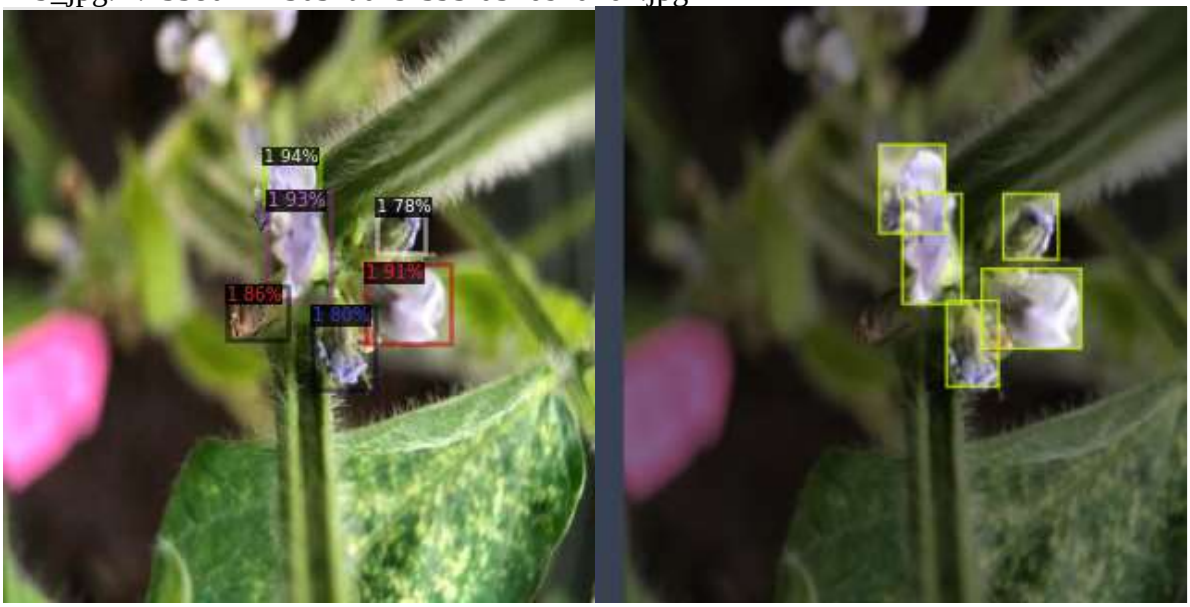


The Faster R-CNN model was trained for 1000 epochs (i.e., 1000 passes through the training data). The Average Precision for detections whose bounding boxes overlap by at least 50% with the ground truth bounding boxes (denoted as AP50) was 82.036 on the test images (and 82.767 on the training images). Some sample predictions on test images are shown below, together with their corresponding ground truth annotations (the predicted and ground truth counts are also shown underneath each image).

174_jpg.rf.5378b9d2c081dd05ff89526219b17ce3

Predicted: 5 Ground Truth: 4

479_jpg.rf.4338041f45054ad19f395fd546b2d261.jpg



Predicted: 6 Ground Truth: 5

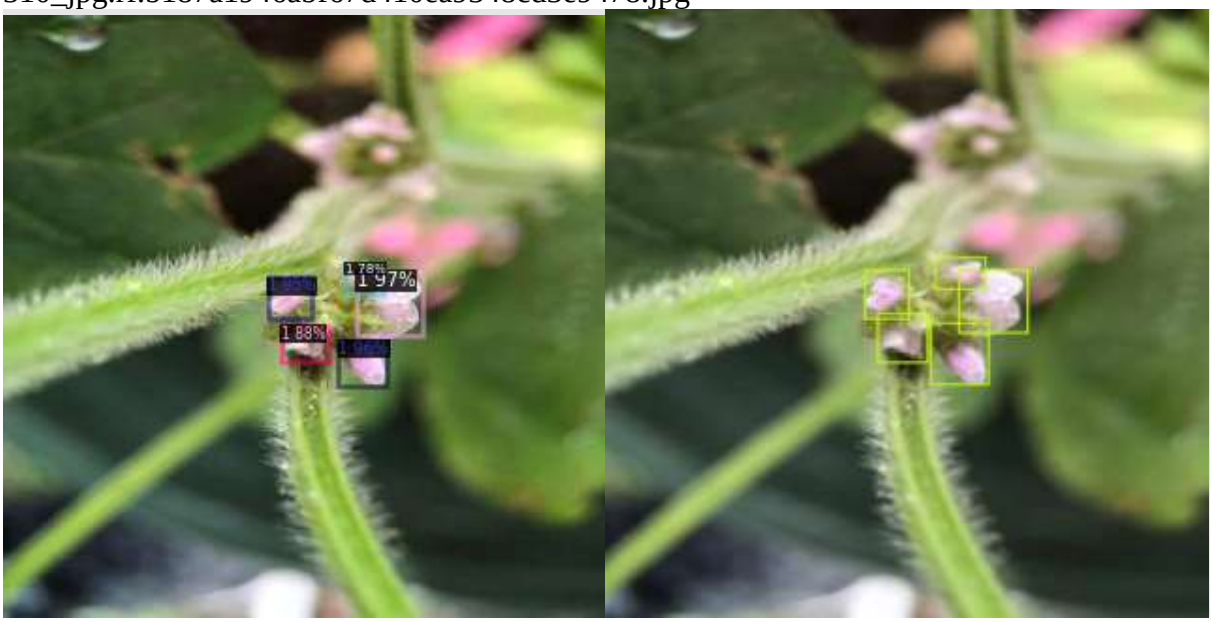
218_jpg.rf.27de6be46a007b42bc7b1eebdffa4e96.jpg

Predicted: 6 Ground Truth: 8

246_jpg.rf.678e7cafd8806935305e2e0a907ca2d8.jpg

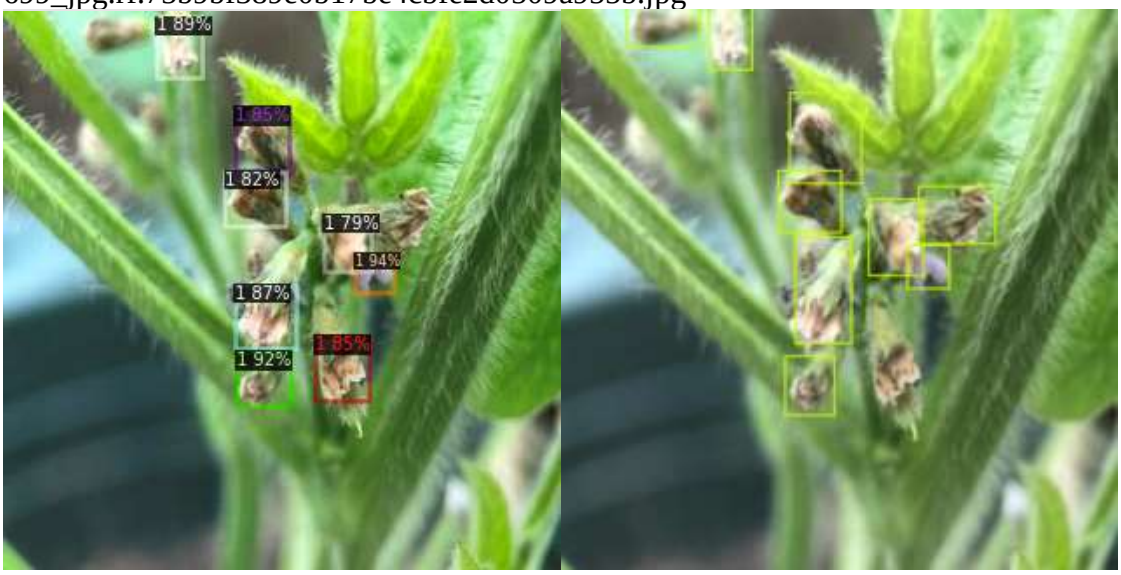
Predicted: 4 Ground Truth: 5

310_jpg.rf.b187a1946abf67d410ea9348cd3e9478.jpg



Predicted: 5 Ground Truth: 5

699_jpg.rf.75b9bf389e0b17be4cbfe2d0509a955b.jpg



Predicted: 8 Ground Truth: 9

904_jpg.rf.9474ccfa417ab167fe2ad0a883bc8e0e.jpg

As can be seen, the predictions closely match the ground truth annotations, although the model can predict both false positives and false negatives. However, a closer analysis of the results is needed as the annotations themselves may not always be consistent (e.g., some buds are counted, while others are not counted, or some dried flowers are counted, while others are not counted). For example, in the last image above, the model correctly identifies 5 flowers, while the annotation includes only 4 flowers. That is the case for other images.

We used the model trained on the data from Zhu et al. (2022) on images of nodes with flowers extracted from images that the team collected, which have varying resolutions compared to the images in the dataset the model was trained on. As can be seen in the figure below (which shows two examples of predictions on the left and the corresponding original images on the right), the model identifies some flowers but fails to identify other flowers, suggesting that we will need to fine-tune the current model on a variety of images of different resolutions.

Future work includes:

1) Fine-tuning the original model trained on images from Zhu et al. (2022) to images selected from our images to ensure the model performs well on our images and is robust to variations in image resolution and other image variations (e.g., images with smaller or larger number of flowers, images with more or less leaves, etc.)
2) Exploring and implementing other state-of-the-art network architectures (e.g., YOLOv7) that may be better suited and capable of achieving superior performance for our application.

Strategy 2: We have developed a preliminary imaging protocol by which the aborted flowers from greenhouse plants are collected, imaged, and annotated as shown below.

Figure. Strategy 2: Aborted flowers are collected, imaged and counted.

Annotated images are used to train a network for aborted flower detection and counting. The network used is also a Faster R-CNN network available in Detectron2. To gain an understanding of what plate color may lead to best predicted counts for aborted flowers, we imaged aborted flowers on plates of three colors: Sky Blue (2 images), Deep Blue (2 images), and Black (3 images), and we trained a model for each plate color (we used one image for training and one for test). Furthermore, we trained a model based on all imaged plates regardless of the color (three images of three different colors were used for training and three images for testing). A total of 168 aborted flowers were annotated on the 7 plate images. Statistics about the number of aborted flowers on each plate color are shown in the table below:

Plate Color	Aborted Flowers Annotated
Deep Blue	50
Sky Blue	50
Black	68
Total	168

Sample predictions of the models on train/test images are shown below.

Deep Blue Model

Train **Test**

Sky Blue Model

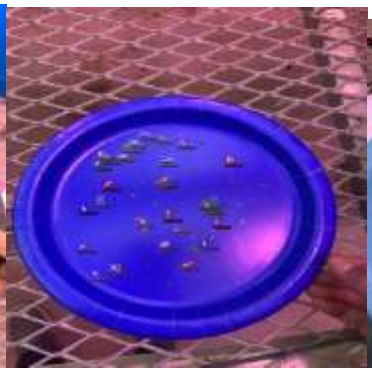
Train **Test**

Black Model

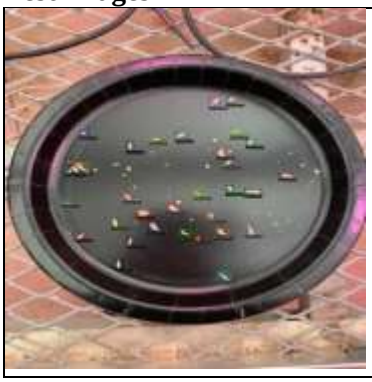
Train **Test**

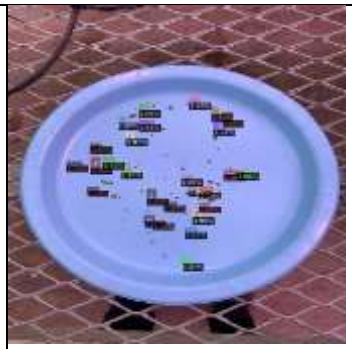

All plates/colors Model

Training images


Test images





Overall, the models perform well considering the very limited data they are trained on, but a close examination of the results shows some false positives and false negatives as can be seen in the example below:

Comparing the four models we trained, we see better results for the models that are trained with a single-color plate image. However, the final goal is to train models that are robust to variations in plate color and work for a variety of backgrounds that resemble what one may observe on the ground in the greenhouse or potentially in the field. Given the promising results we got with only 7 annotated images (including 168 aborted flowers), we expect very accurate results, highly correlated with the ground truth annotations with more training data.

Future work includes:

1) Annotating more image plates and training a model that is robust to plate color/background.
2) Exploring transfer learning from a model that the team has trained in prior work for detecting sorghum seeds spread on a piece of paper.
3) Exploring and implementing other state-of-the-art network architectures (e.g., YOLOv7) that may be better suited and capable of achieving superior performance for our application.

Objective 3 - Discover environmentally stable and region-specific genomic regions controlling flower abortion in diverse soil types, moisture, and climatic conditions

Organ abscission (in this case pistil and flower) is an important process that regulates the detachment of flower from the stem. However, the underlying genetic mechanism of flower abscission is largely unknown in plants. To understand the flower abscission in soybean we surveyed the key determinant genes involved in flower and flower organ abscission in Arabidopsis and identified orthologs in soybean genome. The majority of genes expressed in abscission layer in the model organisms are associated with hormone biosynthesis/transport and nutrient uptake. We have selected a subset of these genes (mainly transcription factors) involved in hormone regulation. We will conduct a gene-based haplotype analysis to select the group of

lines and correlated the large effect variants with the phenotypic data. The confounding effect (if any) (if any) of flowering QTLs will be compared for the selected genes.