# GenSeg: Generating and Segmenting Plant Images for Sustainable Agriculture

Can Eren Derman

## Abstract

Natural weeds growing alongside crops is a tedious problem since they damage the crops by competing for natural resources such as sunlight, water and nutrients. Current methods of eliminating these weeds rest upon the usage of different chemical compositions called herbicides. However, herbicides are most of the time toxic to not only the unwanted weeds but also the crops and the harmless weeds. The technical approach described in this paper aims to tackle this problem by achieving pixel-wise classification of 9 main different classes of plants in a single image using a convolutional neural network trained on artificial training images. CNN's are powerful tools to use in object classification tasks when there are thousands of training images. However, in this case there were only 22 real world images and most of the training data had to be generated artificially. What distinguishes the approach described in this paper is the generation of the artificial training data. Hopefully in the future, robots can be trained to distinguish and localize different classes of weeds with this method and can intervene without using chemicals.

## Introduction

Crops such as wheat, corn and rice are the main food sources of the world. Corn is accepted as the world's most important grain. According to Statista, 1,036.76 million metric tons of corn was produced in 2017-2018. [5] However, weeds are one of the main setbacks in corn production. Weeds, which grow around corns and other crops utilize the same resources available to the crops such as water, light and nutrients in soil. Thus weeds prevent crop production to be maximized. Manually detaching weeds from soil require a lot of time and manpower. There are many other methods used to reduce weeds in crop fields such as usage of herbicides. However, only some herbicides are selective in the plants they affect. It is very hard to detect different kinds of weeds automatically since they are mostly the same color and in similar shapes. In most cases herbicides are used on the entire field, thus harm the environment as much as they harm the unwanted weeds. In addition, it requires time and manpower to observe the effects of different herbicides in very large fields of crop production.

There are a fair number of papers on weed classification; however, most of the time the methods used rely mainly on large datasets and are capable of distinguishing less than 4 classes.

Several different technical approaches were tested in our solution; such as a two step pixel-resolution image segmentation of soil vs. non-soil and low-resolution classification of plant species. Yet, our method which yielded the most accurate pixel-wise classifications composed of a U-net implementation from the Ronneberger et al. paper [1] trained on a composition of artificial and real images.

Our implementation of the U-net was not complete due to computational restraints. The real one described in the 2015 Ronneberger et al. paper uses 64 channels in its initial layer, while our implementation had 24 channels on its first layer. The main reasons why we chose the U-net architecture were its capability to yield precise segmentations even with few training images and its ability to work well with data augmentation.

This project was funded by the BASF chemical company and the real world field images were provided by them. In addition to the 22 real world images, we were provided with birds eye view images of different weeds and soil samples. These images were used to create composite

artificial training images. The weeds and crops were rotated and randomly placed on different soil samples. Then, shadows were added at random directions and the color balance was altered to make the generated images look more realistic.

After the U-net was trained with a composite dataset of artificial and real images, the network was used to generate pixelated masks for each class in 249 real world images. These masks were submitted to BASF and compared to ground truths. BASF responded to each prediction set with F1 scores for the prediction of each class.

Pixel-level object classification is a hard task to achieve even with a large number of training images and small number of object classes. In our case both of these conditions were the opposite of optimal. In addition, pixelated segmentation of 10 different classes of crop, soil and weeds is a hard task for even a human to accomplish accurately. Given the amount of time, which was a single month, our average results were much better than mere estimates and pretty accurate on detecting certain classes such as the crop and soil.

**Related Work**

Previous work mainly relies on multispectral images of input data rather than single RGB images. These additional inputs consist of NIR and NDVI images of the land in addition to RGB images. They also mainly focus on segmentation of three main classes such as soil, crop and weed. Sa et al. [2] and Potena et al. [3] are good examples of such papers. On the other hand, papers which are purely on RGB images such as the Milioto et al. [4] could only accurately classify three classes which are valued crops, weeds and soil. They also rely on large datasets, which consist of around 10,000 real world images.

There are two main aspects which make our approach unique. Firstly, we were not provided with a large dataset of several thousand RGB images or any other kind of multispectral data. However, three datasets provided by BASF were used in the experiment. Dataset A consisted of green house isolated plant images of 6 different classes of weed and 1 class of crop. Dataset B consisted of 31 different soil samples. Dataset C consisted of 271 real world images from which we were provided with the ground truths of only 22 images. So by using datasets A and B we constructed composite images and masks to train on.

Secondly, rather than just classifying three different classes our work focused on classifying a total of 10 different types of classes. This, in addition to the fact that we had access to only a limited amount of data made the process extremely hard to get results above a certain threshold.

**Technical Approach**

Classification of 10 different classes with limited data is a very hard task. In this paper we have aimed to distinguish 8 different classes of weed, corn and soil given an RGB image, without having an actual dataset of real world training data. There were two main aspects of our technical approach. First one was generating our own dataset from the provided segmented BASF green house and segmented images. Secondly, implementing a powerful CNN architecture to accomplish pixel-level classification of 10 different classes.

The BASF dataset was composed of three different parts. Dataset A had isolated birds eye view images of specific plants with resolutions around 0.75 mm/px. Each picture contained a single plant and included information about its age. Table 1 consists of the EPPO codes and the scientific names of the plants. In addition to these 7 different types of plants, our work focused

on detecting other broad leafed and non-broad leafed weeds. Dataset B consisted of 31 different soil samples, which had resolutions around 0.40 mm/px.

| EPPO Code | Scientific name | Number of images |
|---|---|---|
| **zeamx** | Zea mays | 1753 |
| **setve** | Setaria verticillata | 538 |
| **digsa** | Digitaria sanguinalis | 544 |
| **echcg** | Echinochloa cruz-galli | 528 |
| **amare** | Amaranthus retroflexus | 540 |
| **abuth** | Abutilon theophrasti | 535 |
| **cheal** | Chenopodium album | 288 |

*Table 1: Species Included in Dataset A*

*Figure 1: Example images from Dataset A*

*Figure 2: Example images from Dataset B*

Dataset C included original field images, which contained most of the individual classes we aimed to classify. In addition segmentet ground-truths of each targeted species were provided for 22 of the images in dataset C. An example of an original field image in addition to its corresponding masks is exemplified in Figure 3. In addition, dataset D consisted of field images of individual plants. The images in this dataset were much more realistic than the ones provided in dataset A.

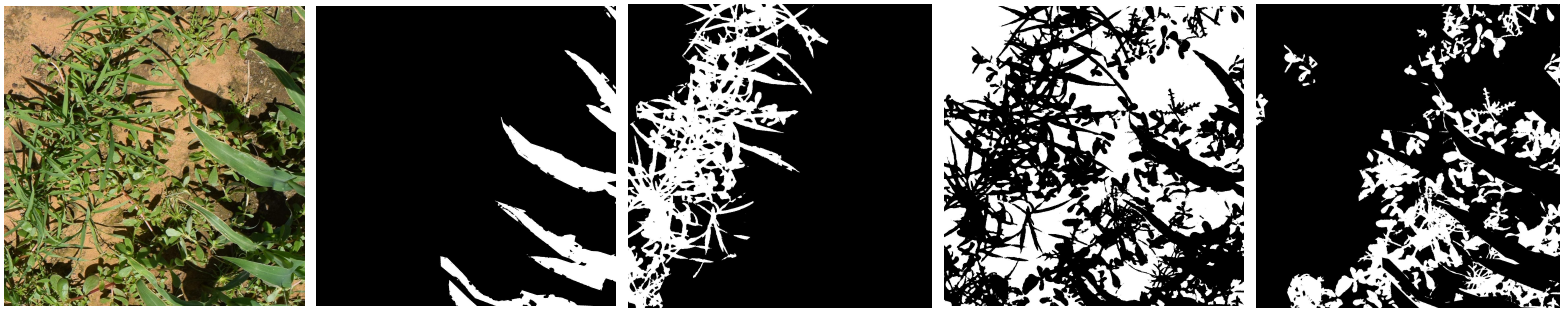


*Figure 3: Example image from Dataset C and its Corresponding Masks*

In order to generate training images, images on dataset A were layered on top of soil samples provided in dataset B. The process of generating training images was semi random. It consisted of three main steps.

First step was placing the weeds and crops on the soil background. Initially, different species of weed were sorted in ascending order of their age. Then, a random number of weeds from each different species of weed was placed on the soil background layer by layer. Before being placed on the soil each plant was rotated randomly. The plants were placed on the soil semi-randomly to avoid overpopulating one specific area in a constructed image. In addition the mm/px ratio of each plant and soil image were taken into account. Each constructed image was

realistic in its proportions. After the last layer of weeds are stacked, crops were added since they are physically the tallest species among the seven plants. Corn of similar age was used in a single picture to increase the liability of the constructed dataset.

Second step was to create the masks for the generated training data. After each layer was added, the pixel values corresponding to the weeds were stored in a mask file. With each new layer added, if there is a overlap between the different weeds and crops, masks of lower layers were updated to exclude those pixel values from their masks. Figure 4 shows an example of a generated image and its associated masks for soil and corn.

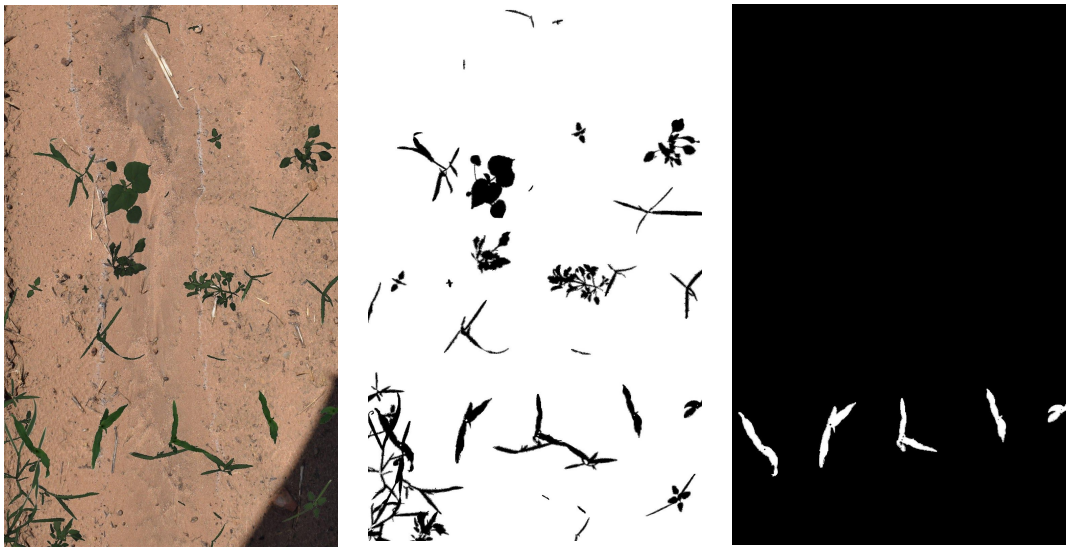


*Figure 4: Example of a Generated Image and its Masks*

The last step of generating training images was on trying to make the images more realistic. The images from dataset A, were very dull compared to soil images. In order to make them more realistic so, we have tried several approaches. The initial approach was to implement the simGAN described in the paper of Shrivastava et al. However, it was not a successful implementation due to computational limitations. After many hours of slow training with small batch sizes, our implementation was only learning about color. We think the literature networks might not be well suited for our generated images. In addition, the actual SimGAN paper was only used for grayscale images. Although there is probably a better architecture for GAN to preserve general color information while allowing for small alterations to brightness we recommend future work to look into implementing a functional style transfer. However, we have successfully added shadows and altered the color balance of the constructed images to make them look more realistic these changes are shown in Figure 6.



*Figure 5: Before and After Results of the Failed GAN Attempt*

*Figure 6: Generated Images with Shadows and Color Balance*

After successfully generating our the dataset, a CNN was implemented to train on the generated images. The U-Net architecture described in the Ronneberger et al. paper was chosen because it is a successful semantic segmentation encoder-decoder network and it does not rely on several thousand annotated images and worked well with augmented data. It has placed first in the ISBI cell tracking challenge by segmenting the 2D transmitted light datasets most accurately. The U-Net architecture used in the Ronneberger et al. paper is illustrated in Figure 6.

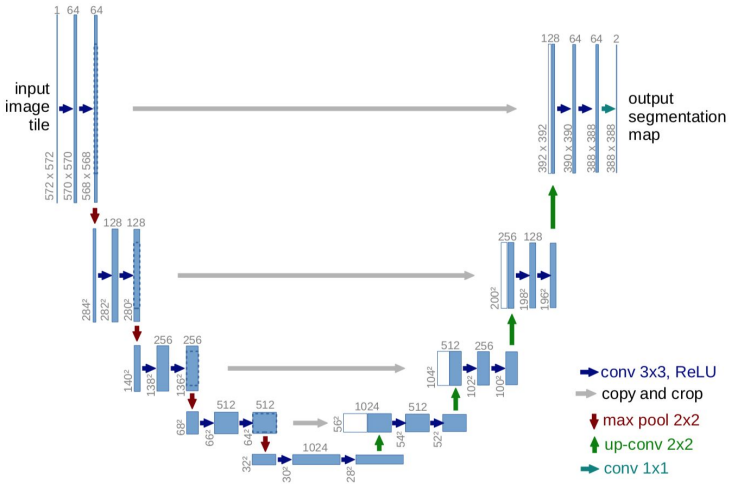



*Figure 6: U-net Architecture in Ronneberger et al. Paper*

The implementation described in this paper was different from the original one in the number of channels it used. Each blue box in Figure 6 shows a multi-channel feature map. The initial channel size is 64 and the rest are proportional to the preceding one by a factor of two. Due to computational limitations the implementation used in this paper had 24 channels in its initial layer. In addition, after the training of the U-Net was completed, it was called to generate prediction masks on 249 real world field images. The field images were resized due to

computational limitations. They were originally 1500x1500 pixels and were downsized to 1000x1000 pixels. After prediction masks for each of 10 classes were constructed for the 249 real world images, they were sent to BASF for validation. At each prediction set different parameters were changed and new techniques were tested.

The altered parameters were the number of training images, training images being real or generated, number of layers on the U-Net, number of channels of the initial layer of the U-Net and number of training steps. Different techniques used were adding rotations, shadows and color balance techniques to make generated images look more realistic.

**Evaluation**

Different prediction sets were sent to BASF and F1 scores were compared for evaluation. The specifications of each set is described on Table 2. Each column describes an aspect of the prediction set. Number of layers is the total number of layers on the U-Net, the original implementation uses 5 layers, so that was the maximum number of layers used. Initial channel size is the u number of channels on the first multi-channel feature map. The rest of the five layers have two times more channels compared to the preceding layer. Number of training images is the total number of training images used for training. If there were rotations used in that dataset, the number of training images is multiplied by the number of rotations. Adjusted weights is a boolean parameter used to reduce the overfitting of soil by reducing its weight by a factor of 10 in the cross entropy loss function. Rotations column denotes if the dataset was augmented with random rotations, if so the corresponding cell specifies how many rotations were used. Number of steps is the total number of steps used in training. Training images column describes the type of training images such as real world images, generated images or a composition of real world and composite images.

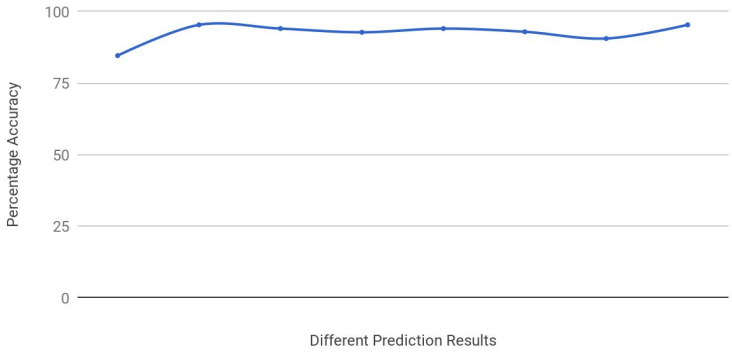
| Prediction Set | Number of Layers on U-Net | Initial Channel Size | Number of Training Images | Adjusted Weights | Rotations | Number of Steps | Training Images |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 8 | 22 | No | No | 2,000 | Real world |
| 2 | 2 | 8 | 400 | No | No | 2,000 | Generated |
| 3 | 5 | 24 | 400 | No | No | 2,000 | Generated |
| 4 | 5 | 16 | 400 | No | No | 20,000 | Generated |
| 5 | 5 | 16 | 1600 (400x4) | No | 4 rotations | 8,000 | Generated |
| 6 | 5 | 16 | 1600 (400x4) | No | 4 rotations | 20,000 | Generated |
| 7 | 5 | 16 | 1600 (400x4) | Yes | 4 rotations | 8,000 | Generated |
| 8 | 5 | 24 | 2442 | No | No | 8,000 | Composite |

*Table 2: Description of Each Prediction Set*

Each prediction set resulted in a different F1 score for each of the classes. While evaluating the results the average F1 scores were weighted the most. However the following graphs show how the accuracy of each class' detection changed over different prediction sets.

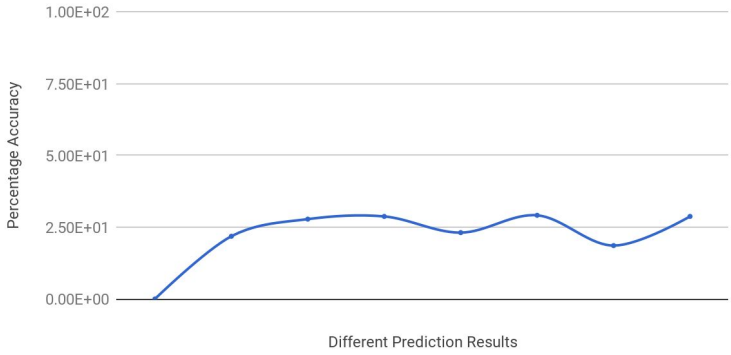
Percentage Accuracy of Soil Detection

Figure 8



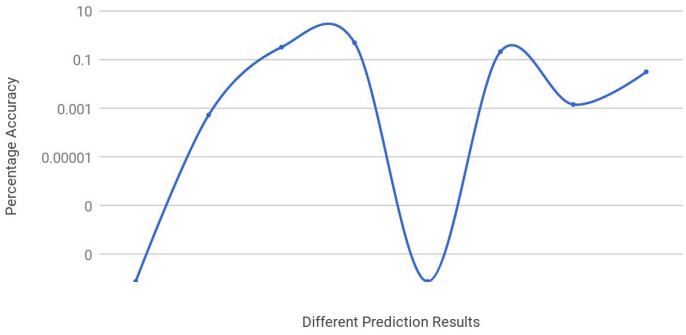

Percentage Accuracy of Zeamx Detection

Figure 9



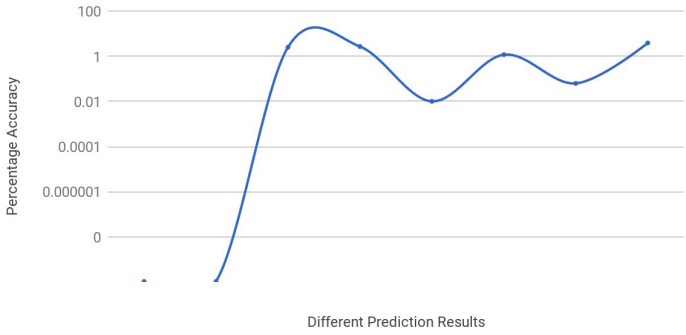

Percentage Accuracy of Setve Detection

Figure 10



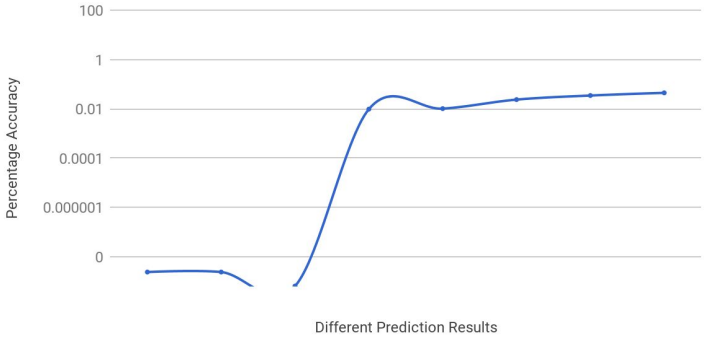

Percentage Accuracy of Digsa Detection

Figure 11



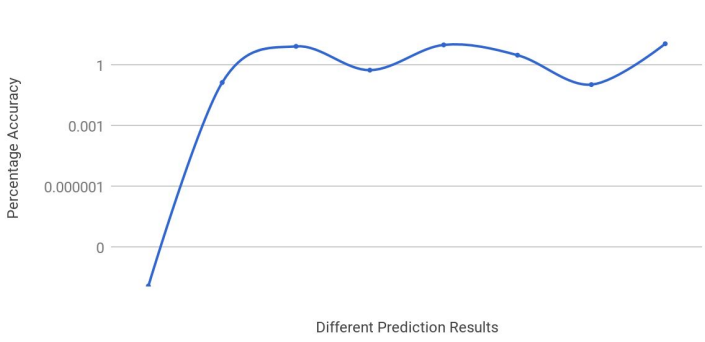

Percentage Accuracy of Cheal Detection

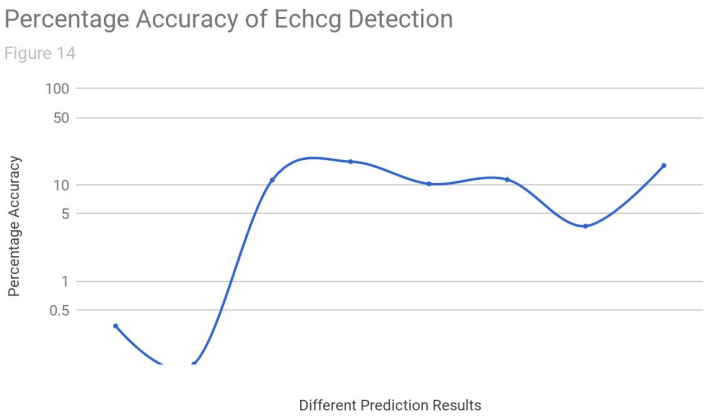
Figure 12

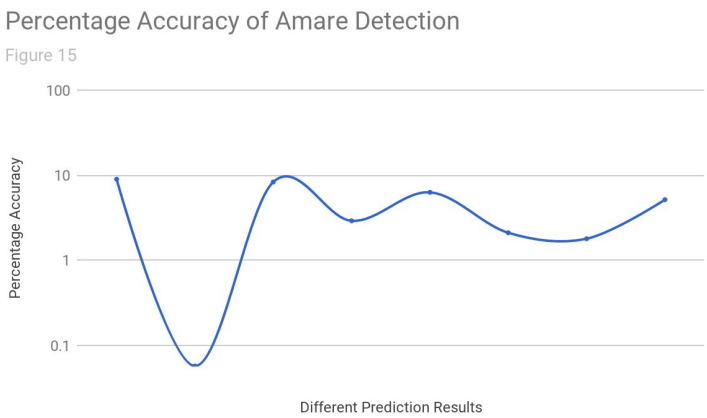



Percentage Accuracy of Abuth Detection

Figure 13





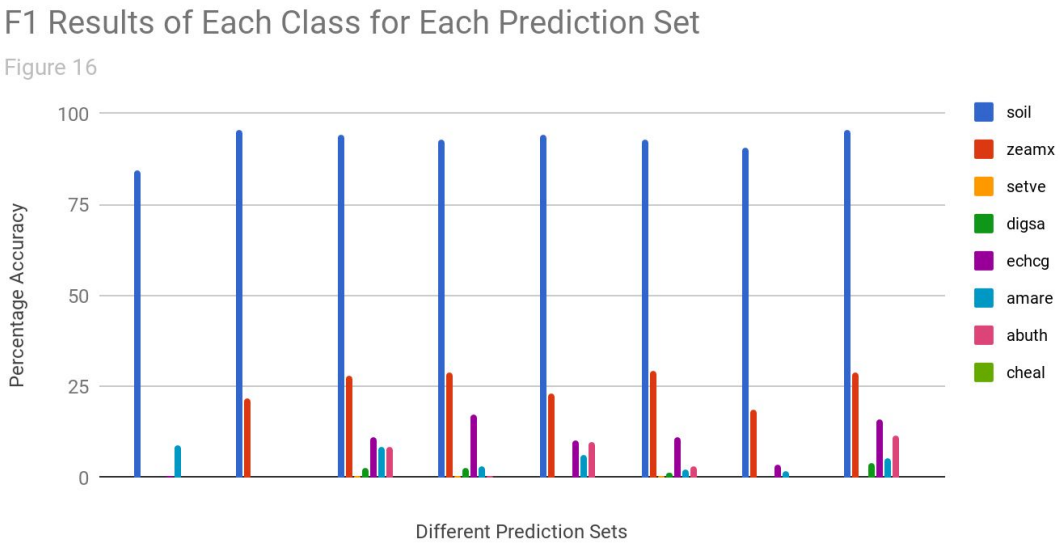



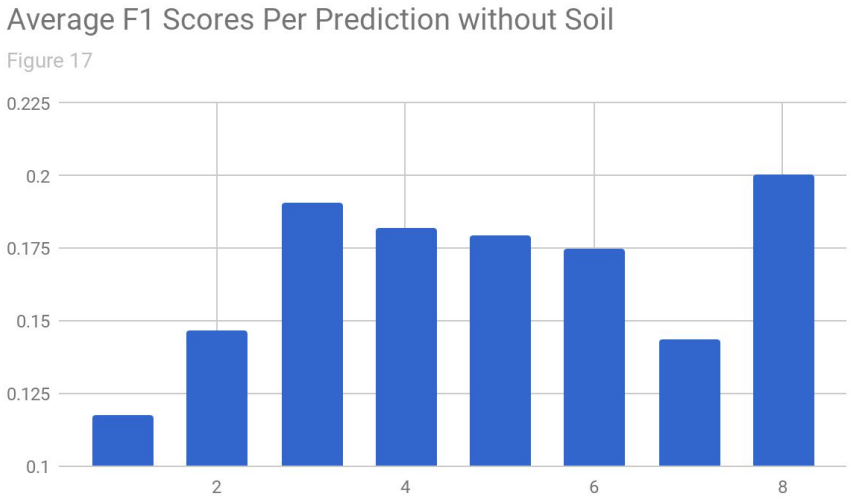

As evident in the graphs it was very hard to decide which parameters to keep since individual weeds responded differently to each alteration. Thus, the average F1 scores were valued the most in our quantitative evaluations. Figure 15 illustrates the difference of the average F1 scores on each prediction set while Figure 16 provides a general view of how each class reacted differently on each prediction set. Overall, number of training images and the number of channels on each layer proved to be the most important parameters. Although rotations were useful, adding them reduced the number of channels on each layer due to the limited computational power we had access to. Thus for the final prediction set rotations were excluded.

Quantitatively the final results were the best with an average F1 score of 0.20 out of 1. F1 scores are calculated by taking into account both precision and recall. Precision (p) is the number of true positives divided by the total number of positive detections. Recall (r) is the number of true positives divided by the number of positive results that should have been detected. F1 score is then calculated with the following formula:

$$F_1 = (p * r)/(p + r)$$

Thus, it is a reliable description of the accuracy of detection results. All in all, although the average F1 percentage was %20, %95 accuracy in soil segmentation and %28 in corn segmentation was achieved. The results were not optimal, but given that the network has trained mainly on generated images the results are still impressive.






**Conclusion**

In this paper, we have generated our own training images and implemented a CNN with the U-Net architecture to achieve pixel level semantic segmentation of an input image into 10 different classes. We have generated layered images using the BASF dataset and trained a CNN on it to detect different classes of weed, corn and soil in a given real field image. On %20 accuracy was achieved in accurately detecting each class.

For future work, on the data generation side we would recommend successfully implementing a GAN or Pix2Pix style transfer to make images more realistic. In addition, our images were constructed by adding them layer on top of layer on a 2D surface, we would recommend reconstructing the training images in a 3D rendering engine and adding artificial light and shadows.

If data generation is not needed then an interesting approach could be trying to add different spectral inputs in addition to RGB images. On the other hand, it would be interesting to try a low resolution logistic classifier with several layers to distinguish soil versus plant and then different classes of plants.

**References**

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox *U-Net: Convolutional Networks for Biomedical Image Segmentation*

[2] Inkyu Sa1, Zetao Chen2, Marija Popovic´1, Raghav Khanna1, Frank Liebisch3, Juan Nieto1, Roland Siegwart1 *Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs*

[3] Ciro Potena, Daniele Nardi, and Alberto Pretto *Fast and Accurate Crop and Weed Identification with Summarized Train Sets for Precision Agriculture*

[4] Andres Milioto, Philipp Lottes, Cyrill Stachniss *Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs*

[5] Grain production worldwide 2017/18, by type
*https://www.statista.com/statistics/263977/world-grain-production-by-type/*