Semantic Segmentation of Agricultural Anomalies in Aerial Imagery: Analysis of Segmentation Models

Arunima Srivastav
Stanford University
aru04@stanford.edu

Devan Agrawal
Stanford University
agrawald@stanford.edu

Malvyn Lai
Stanford University
malvlai@stanford.edu

Abstract

Agricultural crop anomalies have a direct impact on the health and efficacy of a crop yield. To explore possible algorithms in order to detect such patterns, we utilize a Agriculture-Vision dataset from 2021 to compare various architectures' effectiveness in performing a semantic segmentation task onto satellite imagery of farmlands. We attempt to utilize a direct semantic segmentation encoder-decoder structure based around a U-Net as a baseline, and improve upon that architecture by first running a classifier on the dataset.

Having uncovered the U-Net's general significant challenges in this setting of segmentation, our project overall highlights the difficulties and potential mismatch of architectures like the U-Net to large spatial data such as the Agriculture-Vision dataset. In addition to uncovering knowledge around the U-Net's applicability in the task of aerial image segmentation, we gathered insights from the applicability of other model architectures—such as Segment Everything Everywhere all at Once (SEEM) to the Agriculture-Vision dataset and segmentation task.

1. Introduction

The Agriculture-Vision dataset (2021) is a large-scale aerial farmland image dataset for semantic segmentation of agricultural anomalies. Such instances include anomalies like the presence of weed clusters, storm damage, farmland drydown, nutrient deficiency, water floods, among more. Discovering such patterns is important for farmers: the recognition and subsequent management is ultimately critical to protecting yields. Algorithms that can detect these field anomalies and conditions can provide a timely mechanism to prevent major losses and increase yields of farmer crops. Our work aims to use pre-existing models trained on the Agriculture-Vision dataset to understand the effect of certain feature adjustment operations on the ultimate evaluation metrics, such as the loss between predicted segmentation mask and the ground truth segmentation mask as well as the metric of mean Intersection-over-Union (mIoU). Specifically, our project aimed to do two things. First, we wanted to start with a pre-existing baseline U-Net implementation that performs binary segmentation (i.e. discerning between background and a certain class of image told to the model) on the Agriculture-Vision dataset to understand where the results lie at baseline with just a simple architecture, such as a U-Net. We used the implementation from Mark Lisi [1]. In this first part we wanted to understand how the baseline U-Net performed. Second, we sought to observe the impact of combining a classifier with nine binary classifying U-Net models that we trained on each of the classes of the dataset. In particular, we designed the structure of the classifier to incorporate a Residual Neural Network (ResNet) with a MultiOutputClassifier object from scikit-learn, which allowed for packaging of the classifiers to run them in parallel.

2. Related Works

2.1. Foundational Off-the-Shelf Models

An important piece of related work was a prefacing publication from Chiu *et al.* [2] that provided depth of background specifically around the dataset we are using. In this paper we are introduced to popular models for semantic segmentation: U-Net, which leverages an encoder-decoder architecture for pixel-wise classification; PSPNet which uses spatial pooling; and foundational off-the-shelf models from DeepLab. DeepLab, in comparison to these other aforementioned methods, has many strengths. It is a very deep pre-trained network that can be fine-tuned to the specific application—and it showed fairly decent performance when implemented by the Agriculture-Vision dataset team, getting mIoU scores that compete with the team's best models yet. Heffels and Vanschoren [3], on a different aerial imagery dataset, share similar findings, achieving high mIoU metrics using a model architecture based on DeepLabv3. Learning

about DeepLab led us to explore other state-of-the-art foundational models, SEEM and YOLOv4. SEEM, proposed by Zou *et al.* [4], as suggested by the name, claims to be an all-around robust model in semantically segmenting any given image prompt. Further, its strengths lie in its versatility, it being able to take in a generalized range of images; compositionality, as it has the capability to learn complex spatial relationships between text and visual prompts; interactivity, as incorporated are learnable memory prompts into the decoder that maintains segmentation history; and semantic-awareness and vocabulary segmentation capability. A weakness of SEEM, as we discovered and will discuss, is its lack of applicability outside of general resolution images—i.e., it is not as applicable in cases similar to trying to semantically segment aerial imagery. We further explored YOLOv4, a popular off-the-shelf model for semantic segmentation. One source in particular that provided insight on YOLOv4's specific strengths in its use cases was Samyal *et al.* [5]. The authors emphasized that in this case of object detection in aerial imagery, YOLOv4 performed better in comparison to many models. Ruan et al. [11], in "A precise crop row detection algorithm in complex farmland for unmanned agricultural machines" use YOLO as well and also have strong performance in this context of aerial imagery.

2.2. Pre-Existing Models on our Dataset

Lisi [1] looks at the performance of U-Net in the context of the Agriculture-Vision dataset. U-Net is a much simpler approach to the task at hand. However a significant weakness is that its simpler nature does not allow it to pick up on the more complex features that deeper models might allow. Training a deeper U-Net could help this, though in Lisi's implementation, their model often predicted grainy mask label predictions, as well as a significant amount of blank label predictions. However papers like Baheti *et al.* [12], who had high mIoU scores from spatial imagery using a U-Net, provided arguments for the effectiveness. The U-Net success seems to be very context-dependent. Liu *et al.* [6] motivate our work with a more complex model approach to the Agriculture-Vision dataset. Using a novel architecture, the Multi-view Self-Constructing Graph Convolutional Network (MSCG-Net), showcase in their paper the effectiveness of the MSCG-Net in the context of the aerial imagery from the Agriculture-Vision dataset. The MSCG-Net is a composition of Self-Constructing Graphs (SCGs) and Graph Convolutional Networks (GCNs). The approach of having a CNN backbone allowed the model to learn high-level representations from images while gaining the semantic meaning and rotational invariance benefits from the SCGs. Their model was among the top scorers in the actual Agriculture-Vision competition.

Yang *et al*. [7] in the Agriculture-Vision Competition use DeepLabV3+ with what they call Switchable Normalization blocks led to very successful mIoU scores. The strength here is that the switchable normalization block allowed the model to alleviate feature divergence, and they also had a versatile hybrid loss function. Another team, Park *et al*. [8] used a Residual DenseNet—a very strong approach due to the network depth, and its unique approach when it comes to several steps of post-processing.

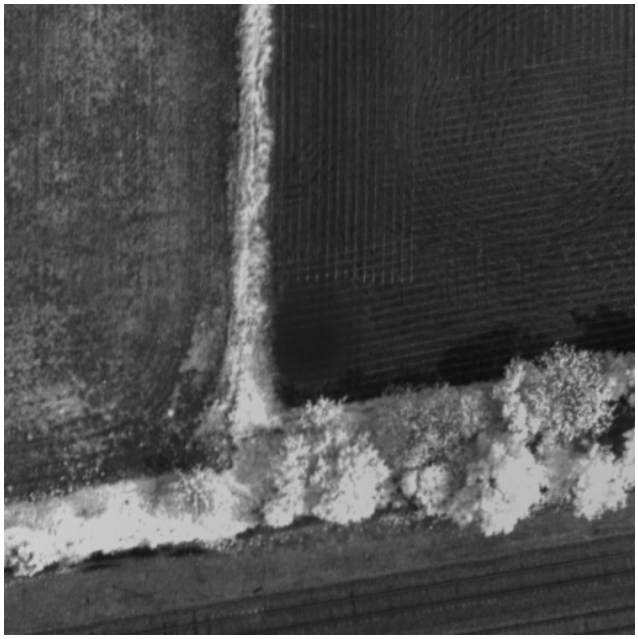

Figure 1. An example NIR satellite imagery, taken from the validation set.

3. Dataset & Features

The 2021 Agriculture-Vision dataset consists of 94,986 RGB and near-infrared (NIR) images from 3,432 different farmlands. These farmland images were captured between 2017 and 2019 across multiple growing seasons in various farming locations in the US. Each training image is hand-processed by agronomic experts, who provide boundaries (e.g., drawn boxes) and labels to identify the anomalies in each farmland image. The full set of possible labels and their corresponding numerical labels for segmentation in the dataset are: double plant (1), drydown (2), endrow (3), nutrition deficient (4), planter skip (5), storm damage (6), water (7), waterway (8), and weed cluster (9). As a form of data handling, the dataset does not include images from the same farmland in more than one subset from the training, test, and validation set to preserve the integrity of our model.

Due to the large size of the dataset, we created our own train-val-test split consisting of 18,334 images for training, 3,667 images for validation, and 1,833 images for testing. To preprocess the data for our ResNet-based model, we ap-

Figure 2. The corresponding RGB image for the above figure.

plied a series of transformations using the PyTorch transformer module. First, the pixel values of the images were rescaled to the range [0.0, 1.0]. Then, the images were normalized using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225] across the color channels. These preprocessing steps are necessary to ensure compatibility with the pre-trained ResNet model. The labels were converted to tensors without any further transformations.

By utilizing both the RGB and NIR images during training, our model aims to learn from a broader spectrum of information, potentially improving its ability to identify and segment the various anomalies present in the farmland images. The preprocessing steps of rescaling and normalization align the input data with the requirements of the ResNet model, enabling effective training and inference. We chose to normalize the images in the same way across all experiments to ensure consistency.

4. Methods

In this study, we aimed to develop an effective segmentation model for the 2021 Agriculture-Vision dataset, which consists of farmland images captured across various locations in the US. Our primary metric for evaluating model performance was the mean Intersection over Union (mIoU), which measures the overlap between the predicted and ground truth masks for each class.

4.1. MSCG-Net : Baseline

To establish a baseline, we selected a state-of-the-art model from a competition that demonstrated low loss and high accuracy. The chosen model, Multi-view Self-Constructing Graph Convolutional Networks with Adaptive Class Weighting Loss (MSCG-Net), is a deep learning architecture designed for semantic segmentation tasks. MSCG-Net leverages multi-view information by processing the input image from different perspectives (e.g., original, rotated, and flipped views) and constructing graph convolutional networks (GCNs) to capture long-range dependencies within each view. The model also employs an adaptive class weighting loss function to address class imbalance issues commonly found in semantic segmentation datasets. By integrating multi-view information and adaptive class weighting, MSCG-Net aims to improve segmentation accuracy and robustness. The code for MSCG-Net is provided in the attached document.

4.2. Segment Everything Everywhere All at Once (SEEM)

In addition to MSCG-Net, we explored the use of an off-the-shelf segmentation model called Segment Everything Everywhere All at Once (SEEM). SEEM is a prompt-based model that can segment images based on user-provided textual descriptions. We conducted experiments to evaluate the effectiveness of SEEM in segmenting the Agriculture-Vision dataset images.

4.3. Baseline U-Net Implementation

We explored the use of a U-Net architecture for semantic segmentation of the Agriculture-Vision dataset. U-Net, proposed by Ronneberger et al. [9], is a convolutional neural network designed for biomedical image segmentation. The architecture consists of an encoder path that captures context and a symmetric decoder path that enables precise localization. The skip connections between the encoder and decoder paths allow for the propagation of spatial information, making U-Net well-suited for segmentation tasks. Inspired by the architecture from github.com/marklisi1/ag-vision-segmentation/blob/main/ag-vision.ipynb, we made modifications to the original U-Net to adapt it to our specific task. Our U-Net implementation consists of an encoder with four convolutional blocks, each followed by max pooling for downsampling. The decoder path includes four upsampling blocks, each followed by convolutional layers. Skip connections are added between the corresponding encoder and decoder blocks to preserve spatial information. To address potential over-fitting issues, we introduced dropout layers between the convolutional layers in the encoder path. Dropout is a regularization technique that randomly sets a fraction of input units to zero during training, which helps

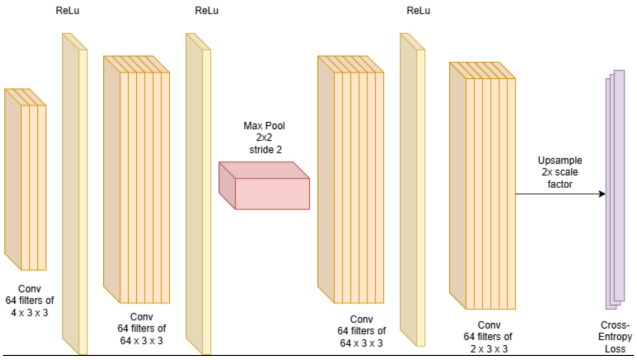

Figure 3. Architecture of our baseline U-Net, which contains four conv layers, three relu layers, and a pooling layer.

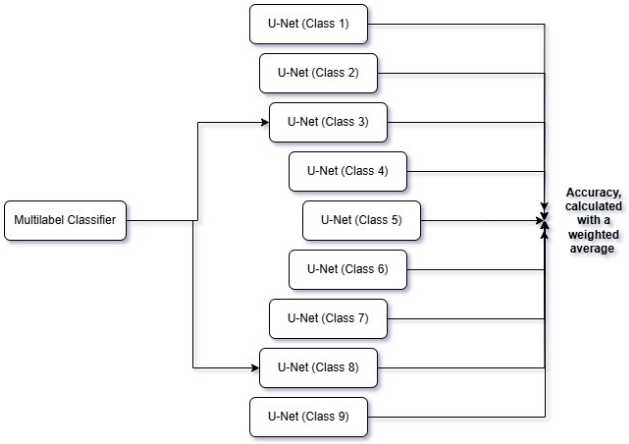

Figure 4. Proposed architecture for the classifier/U-Net pipeline. The above figure is an example with an image where classes 3 and 8 are found present by the multilabel classifier. The U-Nets trained on classes 3 and 8, endrow and waterway respectively, are then ran on the image to detect the pixels.

prevent the network from relying too heavily on specific features and promotes generalization.

4.4. Classifier & U-Net Pipeline Implementation

4.4.1 Classifier

Another architecture we experimented with was to first utilize a multilabel classifier to output the existing classes for each image. The advantage of such a structure was to allow 9 separately trained U-Nets to focus on binary segmentation. In theory, this would allow us to first detect which classes existed in each image, and run each image in a more finely tuned U-Net to then find the pixels where for each of the detected classes. Since our dataset contained only ground truth masks, as a segmentation dataset, we first extracted the necessary class labels for each image by checking the 9 ground truth masks for each image. The length 9 label vector for each image was then stored in a csv file. Our multilabel classifier was structured using the MultiOutputClassifier object from scikit-learn. We implemented ResNet-50 with pretrained weights as our estimator by defining a ResNet class with methods fit and predict, corresponding to train and testing. ResNet, introduced by He et al. [10], has become a popular choice for many computer vision tasks due to its ability to train deep networks effectively by mitigating the vanishing gradient problem through the introduction of residual connections. This ResNet-50 was then fed into the MultiOutputClassifier, which finetuned 9 separate classifiers, each with a pretrained ResNet-50 as the backbone.

4.4.2 U-Net

We then aimed to create a pipeline between the classifier and the U-Net models. Again, the goal was to have a separate U-Net for each of the classes, providing a more specialized model to segment and learn the features of each class more effectively. To implement this, we created a U-Net with the same decoder architecture as 4.3. We chose to use ResNet-18 as the encoder for our U-Net models instead of the original encoder architecture due to its proven effectiveness in computer vision tasks. Our hypothesis for using ResNet-18 as the encoder was that it would provide a more powerful and efficient feature extraction mechanism compared to the original U-Net encoder. ResNet-18 has been pre-trained on large-scale datasets like ImageNet, which enables it to learn robust and transferable features. By leveraging the pre-trained weights of ResNet-18, we aimed to benefit from transfer learning, where the knowledge gained from image classification tasks can be applied to semantic segmentation. Furthermore, the deeper architecture and residual connections of ResNet-18 allow for the capture of more complex and hierarchical features while facilitating the flow of information and gradients throughout the network. This enables the training of deeper models without performance degradation. We hypothesized that this enhanced feature extraction capability, combined with the benefits of transfer learning, would lead to improved segmentation accuracy and generalization compared to the original U-Net encoder. ResNet-18 is also a relatively lightweight model compared to its deeper counterparts (e.g., ResNet-50, ResNet-101), making it computationally efficient and faster to train. This is particularly advantageous when training separate U-Net models for each class, as it reduces the overall training time and resource requirements.

5. Experiments, Results & Discussion

5.1. MSCG-Net

We used the checkpoint in the codebase for the MSCG-Net on the Agriculture-Vision dataset. The model check-

point was stored after 20 epochs of training. We ran the model on our testing dataset, and achieved a loss of 1.12 and an accuracy (as defined by the paper itself) of 63.9%. The mIoU metric returned a value of 43.

5.2. Segment Everything Everywhere All At Once (SEEM)

Prompt-based Segmentation with SEEM: To explore the potential of prompt-based segmentation, we conducted 10 trials using the SEEM model. In each trial, we uploaded a batch of RGB images from the Agriculture-Vision dataset and provided the corresponding label class as the prompt. Unfortunately, despite our efforts, SEEM was unable to effectively segment the images based on the provided prompts. The model's performance in this task was unsatisfactory, suggesting that prompt-based segmentation may not be directly applicable to the specific characteristics and challenges of the Agriculture-Vision dataset, focusing on a single class with non-empty labels. However, despite the inclusion of dropout, we observed an exponential reduction in training loss, with the final training loss reaching around 0.018 and an mIoU of 1.00. These results suggested that the model was overfitting to the training data, and the performance on the test data confirmed this assumption with our model not generalizing well to these unseen samples, as our loss spiked to 1.997. Furthermore, when looking at individual samples, the IoU for most samples was 0, as our model suffered from having a bias towards producing blank masks (as confirmed by printing out the predicted masks).

5.3. Classifier & U-Net Pipeline

We were unable to run inference on our classifier model as we kept running out of memory on our GPU. To fix this, we tried several approaches including decreasing the size of our dataset - but to no avail. In further testing, we would have switched the ResNet-50 model for a lighter model with comparable performance.

To evaluate the effectiveness of our ResNet-18 U-Net approach for semantic segmentation of the Agriculture-Vision dataset, we conducted a series of experiments. We initially trained the model on the entire set of non-empty labels and corresponding images for a single class. However, upon testing, we observed that the model consistently produced blank segmentation masks. To investigate whether this issue was specific to a particular class, we repeated the experiment with various classes. Unfortunately, the model generated blank masks across all classes, indicating a more systematic problem.

Throughout the training process, we monitored the loss function, which exhibited a decreasing trend. The final loss values ranged between 0.5 and 1.5, suggesting that the model was learning to minimize the objective function. We employed the Binary Cross-Entropy (BCE) loss with logits, a commonly used loss function for binary segmentation tasks, implemented in the PyTorch framework.

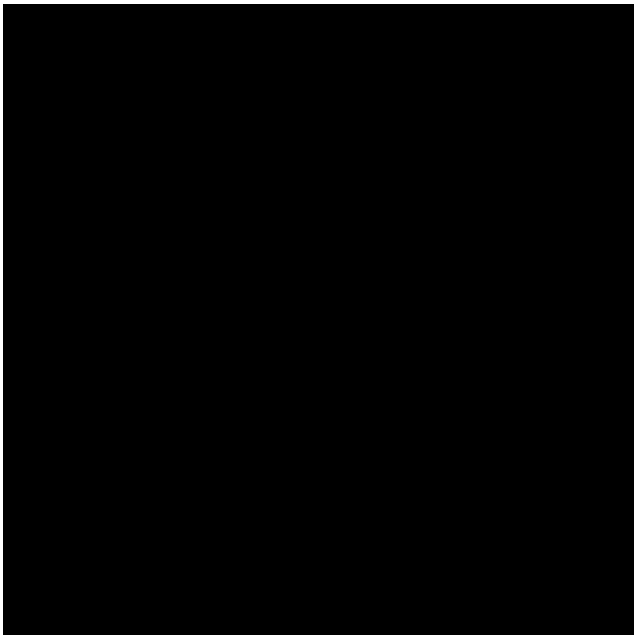

Figure 5. An example of the blank masks generated by the U-Nets. The black pixels in this output mask indicate the lack of any crop anomalies detected by the model in the image.

To further investigate the model's behavior, we created a small dataset consisting of approximately 300 non-empty images and trained the model on this subset for 20 epochs. The goal was to overfit the model on this limited dataset and assess its performance. However, even when tested on the same set of images used for training, the model still produced blank segmentation masks.

Our hypothesis is that the model learned to minimize the loss by producing blank masks as an average representation of the entire dataset. This behavior suggests that the model struggled to capture the discriminative features necessary for accurate segmentation. The consistent decrease in loss values during training, despite the generation of blank masks, supports this hypothesis.

6. Conclusion

In this study, we explored various approaches for semantic segmentation of agricultural anomalies in aerial imagery using the Agriculture-Vision dataset. Our primary goal was to compare the effectiveness of foundational segmentation models and investigate the impact of architectural modifications and pipeline designs on the segmentation performance.

We established a baseline using the state-of-the-art MSCG-Net model, which leveraged multi-view information and adaptive class weighting to address the challenges of aerial

image segmentation. MSCG-Net achieved promising results on the Agriculture-Vision dataset, demonstrating the effectiveness of its architecture in capturing long-range dependencies and handling class imbalance.
We also explored the applicability of off-the-shelf models, such as SEEM, to our task. However, despite its success in other domains, SEEM struggled to effectively segment the aerial images based on the provided prompts. This highlights the specific challenges posed by the Agriculture-Vision dataset and the need for domain-specific adaptations.
Our experiments with the U-Net architecture revealed significant challenges in achieving accurate segmentation results. Despite modifications such as the introduction of dropout regularization and the use of ResNet-18 as the encoder, the U-Net models consistently produced blank segmentation masks. This behavior suggests a mismatch between the U-Net architecture and the complex spatial patterns present in the aerial imagery.
We attempted to address these challenges by designing a pipeline that combined a multilabel classifier with specialized U-Net models for each class. The classifier aimed to detect the presence of different anomalies in each image, while the U-Net models focused on binary segmentation. However, memory optimization issues hindered the successful implementation of this pipeline.

6.1. Future Work

To further advance the field of agricultural anomaly segmentation in aerial imagery, several avenues for future research can be explored:

1. Investigating alternative architectures: While U-Net and its variants faced challenges in this task, exploring other architectures specifically designed for aerial image segmentation, such as DeepLabV3+ or PSPNet, may yield better results. These architectures have shown success in capturing long-range contextual information and handling the unique characteristics of aerial imagery.

2. Efficient memory management: Developing efficient memory management strategies is crucial for training deep learning models on large-scale datasets like Agriculture-Vision. Exploring techniques such as gradient checkpointing, mixed-precision training, or distributed computing could enable the successful implementation of more complex pipeline designs.

By addressing these challenges and exploring novel approaches, future research can contribute to the development of robust and reliable segmentation models for agricultural anomaly detection in aerial imagery. Such advancements have the potential to support precision agriculture practices, improve crop yield, and enhance food security.

7. Contributions

1. Malvyn : Implemented the Classifier as well as edited the MSCG-Net it get it to run.

2. Arunima : Modified the implementation of the U-Net to add a Res-Net decoder as well as ran the trials for the Classifier & U-Net pipeline model. Helped edit MSCG-Net to get it to run

3. Devan : Implemented the baseline U-Net and ran all trials for it.

All three members contributed to writing the project proposal, milestone and report.

8. References

1. M. Lisi. *Crop Anomaly Detection with Semantic Segmentation*. 2023. https://github.com/marklisi1/ag-vision-segmentation/blob/main/CV_Final_Project.pdf

2. M. T. Chiu, X. Xu, Y. Wei, Z. Huang, A. G. Schwing, R. Brunner, H. Khachatrian, H. Karapetyan, I. Dozier, G. Rose, D. Wilson, A. Tudor, N. Hovakimyan, T. S. Huang, and H. Shi. *Agriculture-Vision: A large aerial image database for agricultural pattern analysis*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2828–2838, 2020.

3. M. R. Heffels and J. Vanschoren. *Aerial Imagery Pixel-level Segmentation*. 2020.

4. X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, Y. J. Lee. *Segment Everything Everywhere All at Once*. Submitted on 13 Apr 2023 (v1), last revised 11 Jul 2023 (this version, v4).

5. A. S. Samyal, A. K. R, S. Hans, K. A. K, S. S. B. *Analysis and Adaptation of YOLOv4 for Object Detection in Aerial Images*. 2021.

6. Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg. *Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 7, 2020.

7. S. Yang, S. Yu, B. Zhao, Y. Wang. *Reducing the feature divergence of RGB and near-infrared images using Switchable Normalization*. 2020.

8. M. T. Chiu, X. Xu, K. Wang, J. Hobbs, N. Hovakimyan, T. S. Huang, H. Shi, Y. Wei, Z. Huang,

A. Schwing, R. Brunner, I. Dozier, W. Dozier, K. Ghandilyan, D. Wilson, H. Park, J. Kim, S. Kim, Q. Liu, M. C. Kampffmeyer, R. Jenssen, A. B. Salberg, A. Barbosa, R. Trevisan, B. Zhao, S. Yu, S. Yang, Y. Wang, H. Sheng, X. Chen, J. Su, R. Rajagopal, A. Ng, V. T. Huynh, S.-H. Kim, I.-S. Na, U. Baid, S. Innani, P. Dutande, B. Baheti, S. Talbar, J. Tang. *The 1st Agriculture-Vision Challenge: Methods and Results*. 2020.

9. O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.

10. 10. K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. Submitted on 10 Dec 2015.

11. Z. Ruan, P. Chang, S. Cui, J. Luo, R. Gao, Z. Su. A precise crop row detection algorithm in complex farmland for unmanned agricultural machines. Biosystems Engineering, Volume 232, August 2023, Pages 1-12.

12. 12. B. Baheti, S. Innani, S. Gajre, S. Talbar. Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 358-359.