Agronav: Autonomous Navigation Framework for Agricultural Robots and Vehicles using Semantic Segmentation and Semantic Line Detection

Shivam K Panda* Yongkyu Lee* M Khalid Jawed
Dept. of Mechanical and Aerospace Engineering, University of California Los Angeles
{shivamkp, yongkyulee}@g.ucla.edu, khalidjm@seas.ucla.edu

Abstract

The successful implementation of vision-based navigation in agricultural fields hinges upon two critical components: 1) the accurate identification of key components within the scene, and 2) the identification of lanes through the detection of boundary lines that separate the crops from the traversable ground. We propose Agronav, an end-to-end vision-based autonomous navigation framework, which outputs the centerline from the input image by sequentially processing it through semantic segmentation and semantic line detection models. We also present Agroscapes, a pixel-level annotated dataset collected across six different crops, captured from varying heights and angles. This ensures that the framework trained on Agroscapes is generalizable across both ground and aerial robotic platforms. Codes, models and dataset will be released at github.com/shivamkumarpanda/agronav.

1. Introduction

According to World Food Programme, the population of those experiencing food insecurity is projected to be 342.5 million in 2023, which is more than double the same population in 2020. This concerning trend can be attributed to population growth [11], climate change [45], labor shortage [27], and food affordability driven by high fertilizer prices [10]. To address these challenges, the concept of Precision Agriculture has emerged as a sustainable solution, leveraging modern technology to optimize crop and livestock management practices [15, 37, 40]. A key aspect of Precision Agriculture is automation technology, which aims to minimize resource expenditure while maximizing efficiency. Central to automation technology is autonomous navigation, which enables robotic platforms to operate in the field without human intervention.

Some existing methods of achieving autonomous navigation in agricultural fields rely on real-time kinematic GNSS (RTK-GNSS) [8, 43], LiDAR [25] and depth cameras [1]. Some limitations of RTK-GNSS equipment include its high cost, vulnerability to region-specific outages, reliance on geo-referenced auto-seeding, and signal attenuation problems for smaller mobile robots designed for under-canopy tasks [3]. In addition, the trajectories planned using waypoints with the RTK-GNSS system does not account for the dynamic, changing environment of the agricultural field [19]. This necessitates the use of onboard sensors to observe the changes in the environment in closer proximity. While LiDARs are an integral part of autonomous driving in urban environments, where most obstacles are defined by hard and simple surfaces, their use in the agricultural field is limited due to the natural differences in the environment. The complex and soft nature of the surrounding obstacles, such as leaves and stems and their close proximity to the UGV make it a hostile environment for using LiDARs. In addition, LiDARs entail complex tasks such as point cloud classification and multimodal fusion to extract meaning from the acquired data [32].

To this end, we propose an end-to-end pipeline for autonomous navigation in the agriculture field centered around efficiency and simplicity. Our method avoids the use of expensive equipment such as RTK-GNSS and LiDAR and is entirely vision-based, which requires nothing but a single RGB camera (see Figure 1a). We frame the problem as a series of two downstream tasks: 1) semantic segmentation, which labels each pixel of the input RGB image as one of the predefined classes; and 2) semantic line detection, which extracts the two boundary lines from the overlayed image, an equally weighted blend of the raw image and the color mask, which is obtained as a result of the first task (see Figure 1b). The main contributions of our paper are as follows:

- We propose a simple and efficient end-to-end pipeline that extracts the centerline from an RGB image from a series of two operations: semantic segmentation and semantic line detection.
- We utilize domain adaptation with minimal annotated

*These authors contributed equally to this work

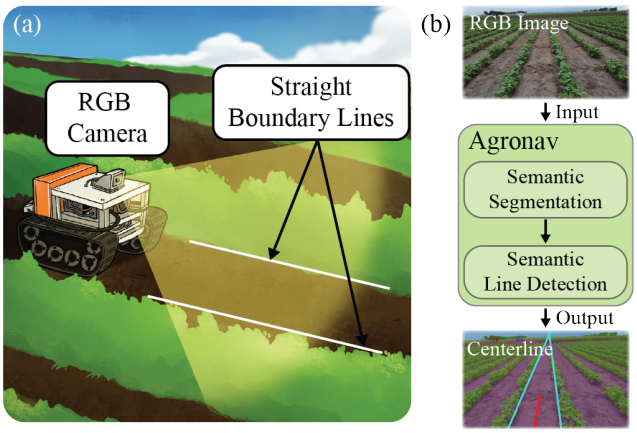


Figure 1. Overview of Agronav. (a) In an agricultural field, lanes are represented as straight boundary lines. A vision-based navigation framework can effectively capture these lines. (b) High-level pipeline of Agronav. Centerlines are extracted from input RGB images through a series of operations.

data for the semantic segmentation model by reorganizing the labels of a publicly available dataset, Cityscapes, and using accurate inferences from a large vision model, ViT-Adapter.

- We demonstrate that semantic line detection, which capitalizes on the structured environment of an agriculture field where crops are planted in straight rows, can successfully extract two boundary lines for various test cases.

- We provide an open-source dataset, Agroscapes, which can be used as a benchmark for scene understanding in agricultural fields for different crops.

The rest of the paper is organized as follows. Section 2 provides a detailed review of semantic segmentation, semantic line detection, and autonomous navigation in agriculture, which are core related topics of our work. Section 3 introduces the data collection and annotation for the training of the semantic segmentation model and the line detection model. Section 4 covers the detailed methodologies related to our autonomous navigation pipeline. Section 5 provides the result that quantifies the performance of our navigation pipeline. Lastly, Section 6 discusses the future direction of this research.

2. Related Works

2.1. Semantic Segmentation in Autonomous Driving

Semantic segmentation is the process of assigning each pixel in an image to a particular object class. In autonomous driving, semantic segmentation is used to identify objects in the environment, such as roads, pedestrians, vehicles, traffic signs, etc. and to generate a detailed map of the scene. This information is then used to plan the vehicle's trajectory and ensure safe and efficient driving.

Over the years, several state-of-the-art models have been developed for semantic segmentation in autonomous driving. These models leverage deep learning techniques to achieve high accuracy and robustness in a variety of driving scenarios. One of the most popular models is the Fully Convolutional Network (FCN) proposed by Long *et al.* [33], which replaces the fully connected layers of a traditional Convolutional Neural Network (CNN) with convolutional layers to enable pixel-wise predictions. Other popular models such as the U-Net model [39] uses a U-shaped architecture to capture both local and global features and has been shown to achieve high accuracy on medical image segmentation tasks. Deeplab model [13] on the other hand, uses atrous convolution to increase the receptive field of the network and improve segmentation accuracy.

These models have been trained end-to-end on annotated datasets such as the Cityscapes [16], which contains high-resolution images of urban environments with pixel-level annotations for 30 object classes. Other 3D dataset have also been used in the autonomous driving community such as the KITTI dataset [22], which contains images as well as point cloud data obtained from a moving vehicle.

In recent years, transformer-based models such as Vision Transformer (ViT) [18] and Swin Transformer [31] have shown remarkable performance on a variety of computer vision tasks including semantic segmentation. ViT utilizes the self-attention mechanism to capture global dependencies and is designed to work well on large-scale datasets. Swin Transformer introduces hierarchical structures with shifted windows to further improve the performance. Some other state-of-the-art models in this context are ViT-Adapter [14], HRNet [42], SegFormer [48], and ResNeSt [49].

2.2. Semantic Line Detection

Semantic lines are characteristic straight lines that capture the essence of a scene. Identifying these lines, which are often implied than obvious, can enhance understanding of an image. Besides the most prominent application in photographic composition to improve aesthetics, semantic lines are also critical in autonomous driving. In autonomous driving, the boundaries of road lanes and key road features serve as important semantic lines. Though many road features are represented as curved lines in urban settings, a typical agricultural field is more structured, where crops are planted in straight rows. This organized structure of the agricultural field simplifies the task of detecting lanes. Lanes can be approximated with straight lines, which are the boundaries that divide the crops from the traversable ground.

The practice of detecting straight lines from images dates back to an image processing technique called the Hough

Transform. More recently, success of CNNs in computer vision led to numerous deep learning-based approaches. A CNN-based semantic line detector named SLNet and open-source dataset SEL was proposed in [28]. This work treated the identification of semantic lines as a combination of classification and regression tasks. An improvement of this study, which used the attention mechanism as well as matching and ranking, was introduced by Jin *et al.* [26]. The proposed line detection algorithm DRM consists of three neural networks: D-Net, R-Net and M-net. While the D-Net extracts semantic lines through the mirror attention module, R-Net, and M-net are Siamese networks that select the most meaningful lines and remove redundant lines. Most recently, Zhao *et al.* introduced the Deep Hough Transform method, which is an end-to-end framework that combines convolutional layers for feature extraction and the Hough Transform to detect semantic lines [51].

2.3. Autonomous Navigation in Agriculture

Although RTK-GNSS based solutions had been popular in field robots, robust navigation in complex agricultural environments requires perception information from local field structures. In this direction, LiDAR was used for perception by Barawid *et al.* [7] on orchards and by Malavazi *et al.* [34] in a simulated environment. Winterhalter *et al.* [46] used both LiDAR and RGB images to extract single lines in row-crop fields with equal spacing.

Most of the earlier vision-based techniques use segmentation between plants and soil by applying some variation of greenness identification *e.g.* excess green index (ExG) [47]. The lines required to extract paths from the segmented image have been estimated using techniques like Hough Transform [4, 35] or least squares fit [6, 21, 50]. However, such image processing based methods might fail in several situation *e.g.*, plants covered with dirt after rainfall, ground covered with weeds or offshoots, different seasons, etc. Some other approaches involve using multi-spectral images [23] and plant stem emerging point (PSEP) using hand-crafted features [36] for crop row location, however these methods cannot be generalized to mutliple crops. In a recent study, Ahmadi *et al.* [2] devised an image processing technique for crop center extraction using ExG followed by detecting individual crop-rows, achieving good performance on crops with both sparse and normal intensities. However, all the above techniques lack the ability to provide an overall scene understanding to facilitate multiple decision-making processes in a robot.

Supervised deep learning models for scene understanding have been successfully applied and popular in the autonomous navigation community as discussed in section 2.1. However there have been very limited work in agriculture. In a recent study, Song *et al.* [41] used semantic segmentation, based on FCN, on wheat fields. However they only provide three classes (wheat, ground & background), and do not provide fine pixel-wise annotations. Apart from wheat, it has been applied and tested for tea plantation [30] and strawberry plantation [38], which are relatively easier crops compared to wheat, corn, rice, canola, flax etc. In another study, Cao *et al.* [12] proposed an improved ENet semantic segmentation network, followed by the random sampling consensus (RANSAC) algorithm to extract navigation lines. They evaluated the method on Crop Row Detection Lincoln Dataset (CRDLD) [17] - a UAV dataset for sugar beet crop. However, the method does not provide pixel-wise labels for scene understanding, and the evaluations are limited to sparse and normal density crop rows. Bai *et al.* [5] conducted a detailed and comprehensive review of vision-based navigation for agricultural autonomous vehicles and robots. DNN-based methods largely relies on annotated dataset but there are no good open-source benchmark datasets across multiple crops that can be used for development of semantic segmentation models for autonomous navigation in agriculture, the equivalent of Cityscapes dataset [16] on roads and streets. This motivates collection of our dataset, Agroscapes along with our transfer learning approach, in order to greatly reduce the amount of data required.

Crop	Row-row [m]	Crop Density	Camera Placement
Canola	0.4	normal, dense	UGV, UAV, HH
Flax	0.45	normal, dense	UGV, UAV, HH
Strawberry	0.3	dense	UAV, HH
Bean	0.5	dense	UAV, HH
Corn	0.4	normal	UAV
Cucumber	0.5	sparse	UAV

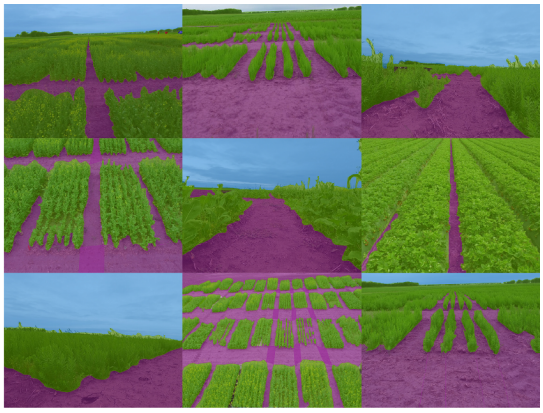
Table 1. Distribution of data across crops with different row-to-row widths, crop density and camera placement. *HH = Handheld.

3. Dataset

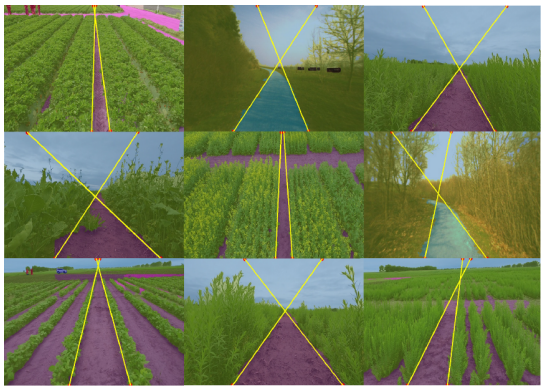
3.1. Data Collection

The data were collected in multiple locations across the United States for six different types of crops: strawberry, flax, canola, bean, corn and cucumber. Strawberry and cucumber data were collected in Oxnard, CA, while the data of the other crops were acquired in Fargo and Carrington, ND (see Figure 2a). The total data amounts to approximately 2,000 seconds of high-resolution video. Data were collected by cameras on three types of platforms: UAV, UGV and handheld (see Table 1). In order to enhance the richness of the data, the viewpoint of the camera were varied in heights and angles for different iterations. This ensured that the collected data suited our initial purpose of developing an integrated framework applicable to different

(a) Agroscapes images from different crops and camera heights



(b) Fine pixel-wise annotations for scene understanding



(c) Line annotations for semantic line detection

Figure 2. An overview of Agroscapes dataset comprising different crops at different heights and subsequent annotations.

types of autonomous systems - mobile robots, drones and vehicles (e.g. autonomous tractors). The variations also provided our machine learning model more robustness for higher accuracy in varied environments.

3.2. Pixel-wise Annotation

Since one of the major goals of our work is to provide an open-source benchmark dataset in agriculture navigation for scene understanding, fine pixel-level annotations were carried out on the collected dataset (see Figure 2b). The classes of the annotations were selected based on domain adaptation and the requirements of an agriculture scene. Hence a total of 120 images were finely annotated in 9 classes - soil, crop, weed, sky, human, vehicle, building, fence and other. Segmentation of soil and crop is the most important to determine the traversable and non-traversable regions. It is also important for the autonomous robot to identify and understand the presence of humans in its vicinity in an agricultural field. This is crucial for a safe-work environment with the workers on the field [9]. Other obstacles such as vehicles (or robots) and fences are also common to an agriculture environment. All images were annotated by a single annotator to ensure precise boundaries for each class and consistency among all annotations. CVAT (Computer Vision Annotation Tool) from Intel was used to execute all annotations. Among the annotated images, roughly 50% are single-row images and the rest multiple-row images. Annotation of single-row images took approximately 20 minutes per image, whereas annotating multiple-row images took between 30 to 100 minutes per image. Therefore, the total annotation effort for all images amounted to approximately 60 hours.

3.3. Line Annotation

The dataset used to train the semantic line detection model is a combination of our own data and the annotated Freiburg Forest dataset [44]. To ensure that the model is trained on images similar to that of our test environment, we filter images from the Freiburg Forest dataset that are semantically similar to those taken from the agriculture field. The two main considerations for filtering were: 1) whether the boundary line that separates the crops from the traversable ground can be approximated using straight line, and 2) whether there were comparable amount of vegetation on the both sides of the two boundary lines. Once the RGB images have been selected, they were overlayed with their ground truth color masks, where the RGB images and the color masks have been weighed equally. Lastly, each weighted image was labeled with two semantic lines that mark the boundary between the crops and the ground (see Figure 2c). Each line was represented with the pixel coordinates of the endpoints that lie on the edges of the image. In total, 400 images were annotated and used for training. The ratio of ground images to aerial images was 1:1.

4. Methodology

In this section we discuss the individual downstream tasks in the overall Agronav pipeline (see Figure 3).

4.1. Semantic Segmentation

4.1.1 Models Selection

Various semantic segmentation models were explored in the direction of transfer learning, using different pretrained checkpoints, to ensure good performance on Agroscapes dataset. Unsupervised domain adaptation inferences were executed on two kinds of models - large (high parameter) models with limited real-time performance, and relatively smaller (low parameter) models, which were real-time capable. For the pretrained checkpoints, the models were experimented with checkpoints trained on ADE20K [52], Cityscapes [16], COCO [29] and Pascal VOC [20]. The results showed that models pretrained on Cityscapes dataset achieved the best performance, which is reasonable, given the greater domain relevance of the Agroscapes dataset to the Cityscapes dataset compared to the other datasets. Among the large models, ViT-Adapter [14] was selected, which also has the state-of-the-art performance on the Cityscapes currently. Among the real-time models, three models - HRNet [42], MobileNetV3 [24] and ResNeSt [49] were selected for an ablation study on our domain adaptation. ViT-Adapter (ViT-A) had the best performance among the four models based on the inference study.

HRNet [42] is a high-resolution network that is designed to preserve high-resolution representations throughout the network while maintaining a low computational cost. It employs a multi-resolution fusion approach that combines high-resolution and low-resolution representations to achieve both high accuracy and efficiency. MobileNet [24] is a lightweight network designed for mobile devices, which utilizes depthwise separable convolutions to reduce the number of parameters and computations. It has a small memory footprint and can achieve real-time performance on mobile devices. ResNeSt [49] is a recent advancement of the ResNet architecture that introduces nested and scale-specific feature aggregation to improve the model's ability to capture fine-grained patterns. It uses a split-attention mechanism to capture information from different feature maps and scales, resulting in improved accuracy on various computer vision tasks.

4.1.2 Supervised Domain Adaptation

The classes of the pretrained checkpoints (on Cityscapes dataset) were reorganized for all the models. The Cityscapes checkpoints contain 19 classes, which were reorganized into 8 classes - soil, vegetation, sky, human, vehicle, building, fence and other. Now, the checkpoints were trained again on the source domain *i.e.* Cityscapes based on these 8 labels. However, prior to retraining, the Cityscapes annotations had to be reorganized based on our domain relevance. Table 2 explains the reorganization from Cityscapes labels to the Agronav labels. This reorganization scheme guarantees inheritance of human recognition and obstacles avoidance capabilities of Cityscapes.

Cityscapes	Agroscapes
Road, Sidewalk	Soil
Vegetation, Terrain	Vegetation
Sky	Sky
Person, Rider	Human
Building	Building
Wall, Fence	Fence
Car, Truck, Train, Bus, Motorcycle, Bicycle	Vehicle
Pole, Traffic Light, Traffic Signal	Other

Table 2. Reorganization of Cityscapes labels for Agronav domain adaptation

The final objective here is to achieve accurate, real-time semantic segmentation on the Agroscapes dataset using minimal number of annotated images (120 images). Considering the superior performance of the ViT-Adapter for zero-shot learning, given its large size, we first trained this large model on the annotated Agroscapes dataset. The resulting checkpoint was used to generate labels from 3850 unlabelled Agronav images. We visually inspected the generated labels and selected 1165 labels with mIoU of approximately 90% or higher. These 1165 labels then served as the training data for the real-time models - HRNet, MobileNetV3 and ResNeSt. The models were trained in two stages: first, they were trained on the ViT-Adapter generated labels; then, fine-tuned by adding the manually annotated labels. In summary, our domain adaptation strategy leverages the high accuracy of the large model to improve the results of the real-time models, despite the limited availability of manually annotated data.

4.2. Semantic Line Detection

Our semantic line detection model is trained using the Deep Hough Transform method [51]. The pipeline includes four core components to detect semantic lines: 1) the feature pyramid network (FPN) extracts pixel-wise deep representations; 2) deep representations are converted from the spatial domain to the parameteric domain via Deep Hough Transform; 3) the line detector module to detect lines in the parametric space; 4) Reverse Hough Transform which converts the detected lines back into image space. In contrast to classical line detection algorithms which detect countless straight edges in an image, the Deep Hough Transform model can be explictly trained to output few most semantically meaningful lines. In addition, the lightweight of the trained semantic line detection model make it suitable for real-time applications.

For autonomous crop-row navigation, every image con-

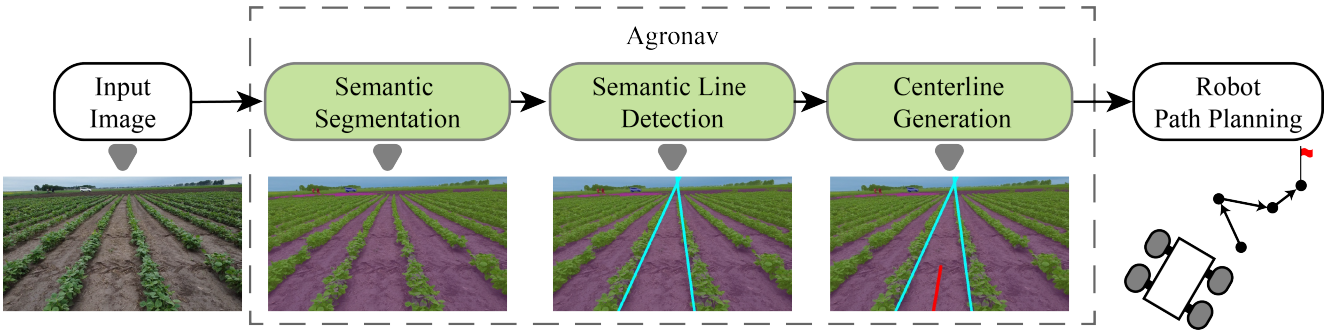


Figure 3. Overview of the Agronav pipeline with semantic segmentation, semantic line detection and centerline generation.

sists of two most semantically meaningful lines. These lines mark the left-side and right-side boundaries of the traversable ground and the crops. The centerline, which serves as the ultimate guideline for navigation, can be extracted from these two lines. In an effort to develop a pipeline which is applicable for different scenarios, both ground and aerial robotic platforms, both single and multi-row images are annotated with two lines that mark the boundaries. As mentioned in Section 3, line annotations are performed on the overlayed image, which is an equally weighted blend of the RGB image and the color mask. The latter is obtained as an output of the semantic segmentation model.

4.3. Centerline Generation

The final piece of our pipeline involves generating the centerline from the two output lines l_1, l_2 predicted by the semantic line detection model. Within the Deep Hough Transform framework, each line is parameterized by two parameters r and θ, where the distance parameter r measures the distance between the line l and the center of image, and the orientation parameter θ represents the angle between the line l and the horizontal axis of the image. Following simple calculation, the entire set of pixels that correspond to each semantic line can be computed. We define the centerline as a set of midpoints of the two pixels $(x^{l_1}_{y_i}, y_i)$ and $(x^{l_2}_{y_i}, y_i)$, which reside on the same horizontal axis y_i. For simplicity, the centerlines were visualized for the bottom third of the image.

5. Results

In this section we individually report the accuracy results from semantic segmentation and semantic line detection. We also report an ablation study of different models and techniques. Finally we provide visual assessment of the overall pipeline on unlabelled data (see Figure 4).

5.1. Semantic Segmentation

The training and validation procedures for all models were conducted on NVIDIA A100 GPUs. However,

Method	mIoU	soil	vegetation	sky
ViT-Adapter	**96.43**	**94.24**	**96.72**	**98.23**
HRNet	95.28	92.39	95.48	97.92
ResNeSt	95.34	92.84	95.7	97.35
MobileNetV3	94.57	91.27	95.01	97.29

Table 3. Comparison of mIoU scores on Agroscapes dataset

Method	FPS
ViT-Adapter	0.34
HRNet	9.12
ResNeSt	4.35
MobileNetV3	**14.25**

Table 4. Real time performance across all models

NVIDIA GeForce RTX 2080 GPUs were used for inference or testing on unlabelled images. This was done to evaluate the real-time performance of our models for deployment on mobile robots and other platforms. We used SGD as the optimizer for all models, with a learning rate of 0.01. Given the balanced distribution among the important classes - soil, vegetation and sky in most of the images, cross entropy loss was selected as our training loss. As for the evaluation metric, we used the most widely used mIoU (mean Intersection-over-Union) score.

The mIoU scores of all models are reported in Table 3. As the soil, vegetation, and sky classes represent the majority of the dataset and are the most important ones for our framework, we only report the mIoU scores for these classes. Our results show that the ViT-Adapter achieved the highest mIoU scores and is the most accurate model. The other real-time models are first trained on the ViT-A inferences and then fine-tuned on the manual labeled dataset. Among these models, ResNeSt achieved the best performance for soil, vegetation, and overall mIoU, while HRNet and MobileNetV3 also achieved equally good performance.

We further evaluated the real-time performance of all models on NVIDIA GeForce RTX 2080 GPUs using video feeds from our dataset. The results in frames per second (FPS) are reported in Table 4. Our results indicate that ViT-

Method	mIoU	soil	vegetation	sky
w/o ViT-A inferences				
HRNet	92.62	92.86	89.43	96.64
ResNeSt	92.11	91.85	89.07	96.42
MobileNetV3	92.48	92.45	89.35	96.7
with ViT-A inferences				
HRNet	95.28	92.39	95.48	**97.92**
ResNeSt	**95.34**	**92.84**	**95.7**	97.35
MobileNetV3	94.57	91.27	95.0	97.29

Table 5. Comparison of mIoU scores on the dataset with and without ViT-inferences

Method	mIoU	soil	vegetation	sky
Ground				
HRNet	96.54	95.77	96.31	97.63
ResNeSt	96.52	**95.82**	96.26	97.57
MobileNetV3	**96.66**	95.74	**96.43**	**97.88**
Aerial				
HRNet	**93.08**	88.52	**94.92**	**95.18**
ResNeSt	93.05	**88.58**	94.86	95.1
MobileNetV3	93.04	88.45	94.9	95.15

Table 6. Comparison of mIoU scores on ground vs. aerial images.

Adapter can not achieve a real-time performance, as it takes approximately 3 seconds to process a single frame. On the other hand, HRNet, ResNeSt, and MobileNetV3 all achieve good real-time performance, with MobileNetV3 achieving the highest FPS. However, the real-time performance for all models can be potentially improved by reducing the crop-size of the frames at the cost of accuracy.

To assess the impact of ViT-A inferences on model accuracy, we evaluated the three models with and without adding the inferences. The results are reported in Table 5, which clearly show that the inferences significantly improved the mIoU scores for all three models. With ViT-A inferences we achieve superior performance, particularly an improved performance in segmenting vegetation. These findings demonstrate the success of our domain adaptation strategy. We also compare the performance of Agronav between ground and aerial images in Table 6. It shows better performance with ground images primarily because it is more challenging to segment multiple soil boundaries from a height in aerial imagery. Nevertheless in a practical aerial operation, we can compromise some accuracy in segmenting soil and crop boundary.

5.2. Semantic Line Detection

We adopt the metrics proposed in [51] to evaluate the similarity between a pair of predicted and ground truth lines, where the similarity between two lines is a function of the Euclidean distance and angular distance between the two lines. By representing the predicted lines and ground truth lines as each set of a Bipartite graph, the matching between the lines of two sets is done by solving the Maximum Bipartite Matching problem.

Following the matching process, true positives, false positives, and false negatives are identified, after which the Precision, Recall, and F-measure scores are evaluated.

To assess the benefits of training the semantic line detection model on the overlayed images (an equally weighted blend of the color mask and raw RGB image), we compare it against a model trained on raw RGB images, as summarized in Table 7. For fair comparison, all other training, optimizer, and dataset parameters were controlled. The semantic model trained on overlayed images trained more effectively, validating the reasoning behind the intended synergy between semantic segmentation and semantic line detection.

The results shown in Table 8 highlight good performance of the semantic line detection model on both ground and aerial images. The relatively inferior performance on aerial imagery can be attributed to the fact that aerial images include images with multiple crop rows, where multiple sets of boundary lines can be drawn, hence making the task more difficult.

Method	Precision	Recall	F-measure
Raw Images	0.9021	0.7442	0.8156
Overlayed Images	**0.9513**	**0.8984**	**0.9241**

Table 7. Comparison of the line detection model trained on raw and overlayed images.

Image type	Precision	Recall	F-measure
Ground images	0.9692	0.9692	0.9692
Aerial Images	0.7833	0.8616	0.8206

Table 8. Performance evaluation of the line detection model on ground and aerial images.

6. Conclusion

In this study, an end-to-end framework for vision-based autonomous navigation in agricultural fields was proposed. By adopting domain adaptation for semantic segmentation, we have trained a robust segmentation model that can successfully segment an image into 8 classes including soil, vegetation, sky, human, vehicle, building, fence, and other. The original 19 classes of the Cityscapes dataset were reorganized and then merged with our own data, collected across six different crops. Additionally, we have demonstrated the effectiveness of the semantic line detection model for detecting boundary lines, which capitalizes

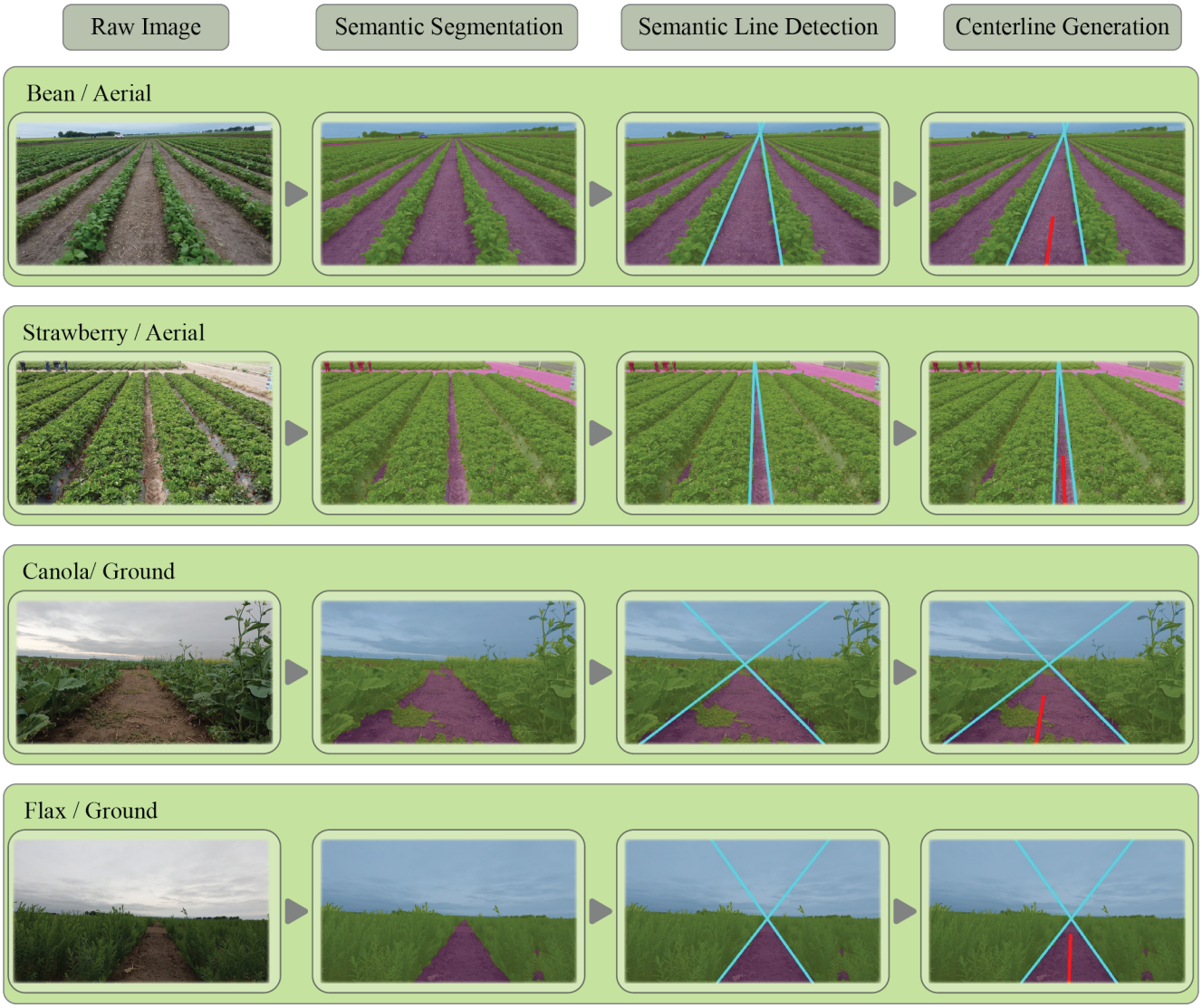


Figure 4. Demonstration of the end-to-end Agronav pipeline, tested on various crops.

on the inherent structure of the field where crops are planted in straight rows. Furthermore, the study validates the synergy between semantic segmentation and semantic line detection by showing that the semantic line detection model trains better on overlayed images. We are also releasing Agroscapes - an open-source dataset for scene understanding in agriculture. This dataset is expected to serve as a valuable resource for studying autonomous navigation in agricultural fields and as a benchmark for future research.

In future work, the results of this study will be applied to a physical robotic platform with dimensions of 75 in (L) x 63 in (W) x 40 in (H), which was initially designed for weed management and data collection in flax and canola fields. The use of this platform will allow for the collection of field trial data. In addition, we hope to release more labeled and unlabeled images in our open-source dataset, Agroscapes, collected across more crops. We also aim to incorporate segmentation of weeds from the crops in future semantic segmentation models.

Acknowledgement

We acknowledge the funding and support from United States Department of Agriculture (USDA Award No. 2021-67022-34200 & 2022-67022-37021) and National Science Foundation (NSF Award No. IIS-2047663 & CNS-2213839).

References

[1] Diego Aghi, Vittorio Mazzia, and Marcello Chiaberge. Local motion planner for autonomous navigation in vineyards with a rgb-d camera-based algorithm and deep learning synergy. *Machines*, 8(2), 2020. 1

[2] Alireza Ahmadi, Michael Halstead, and Chris McCool. Towards autonomous visual navigation in arable fields. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6585–6592. IEEE, 2022. 3

[3] Alireza Ahmadi, Lorenzo Nardi, Nived Chebrolu, and Cyrill Stachniss. Visual servoing-based navigation for monitoring row-crop fields. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4920–4926. IEEE, 2020. 1

[4] Björn Åstrand and Albert-Jan Baerveldt. A vision based row-following system for agricultural field machinery. *Mechatronics*, 15(2):251–269, 2005. 3

[5] Yuhao Bai, Baohua Zhang, Naimin Xu, Jun Zhou, Jiayou Shi, and Zhihua Diao. Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review. *Computers and Electronics in Agriculture*, 205:107584, 2023. 3

[6] Marianne Bakken, Richard JD Moore, and Pål From. End-to-end learning for autonomous crop row-following. *IFAC-PapersOnLine*, 52(30):102–107, 2019. 3

[7] Oscar C Barawid Jr, Akira Mizushima, Kazunobu Ishii, and Noboru Noguchi. Development of an autonomous navigation system using a two-dimensional laser scanner in an orchard application. *Biosystems Engineering*, 96(2):139–149, 2007. 3

[8] Owen Bawden, Jason Kulk, Ray Russell, Chris McCool, Andrew English, Feras Dayoub, Chris Lehnert, and Tristan Perez. Robot for weed species plant-specific management. *Journal of Field Robotics*, 34(6):1179–1199, 2017. 1

[9] Lefteris Benos, Avital Bechar, and Dionysis Bochtis. Safety and ergonomics in human-robot interactive agricultural operations. *Biosystems Engineering*, 200:55–72, 2020. 4

[10] Vibeke Bjornlund, Henning Bjornlund, and André van Rooyen. Why food insecurity persists in sub-saharan africa: A review of existing evidence. *Food Security*, 14(4):845–864, 2022. 1

[11] Thomas C Brown, Vinod Mahat, and Jorge A Ramirez. Adaptation to future water shortages in the united states caused by population growth and climate change. *Earth's Future*, 7(3):219–234, 2019. 1

[12] Maoyong Cao, Fangfang Tang, Peng Ji, and Fengying Ma. Improved real-time semantic segmentation network model for crop vision navigation line detection. *Frontiers in Plant Science*, 13, 2022. 3

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[14] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2, 5

[15] Isabel Cisternas, Ignacio Velásquez, Angélica Caro, and Alfonso Rodríguez. Systematic literature review of implementations of precision agriculture. *Computers and Electronics in Agriculture*, 176:105626, 2020. 1

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 3, 5

[17] Rajitha de Silva, Grzegorz Cielniak, and Junfeng Gao. Towards agricultural autonomy: crop row detection under varying field conditions using deep learning. *arXiv preprint arXiv:2109.08247*, 2021. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[19] L Emmi, E Le Flécher, V Cadenat, and M Devy. A hybrid representation of the environment to improve autonomous navigation of mobile robots in agriculture. *Precision Agriculture*, 22:524–549, 2021. 1

[20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 5

[21] Iván García-Santillán, José Miguel Guerrero, Martín Montalvo, and Gonzalo Pajares. Curved and straight crop row detection by accumulation of green pixels from images in maize fields. *Precision Agriculture*, 19(1):18–41, 2018. 3

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[23] Sebastian Haug, Peter Biber, Andreas Michaels, and Jörn Ostermann. Plant stem detection and position estimation using machine vision. In *Workshop Proc. of Conf. on Intelligent Autonomous Systems (IAS)*, pages 483–490, 2014. 3

[24] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 5

[25] Jawad Iqbal, Rui Xu, Shangpeng Sun, and Changying Li. Simulation of an autonomous mobile robot for lidar-based in-field phenotyping and navigation. *Robotics*, 9(2), 2020. 1

[26] Dongkwon Jin, Jun-Tae Lee, and Chang-Su Kim. Semantic line detection using mirror attention and comparative ranking and matching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 119–135. Springer, 2020. 3

[27] David Laborde, Will Martin, Johan Swinnen, and Rob Vos. Covid-19 risks to global food security. *Science*, 369(6503):500–502, 2020. 1

[28] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Semantic line detection and its applications. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3249–3257, 2017. 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[30] Yu-Kai Lin and Shih-Fang Chen. Development of navigation system for tea field machine using semantic segmentation. *IFAC-PapersOnLine*, 52(30):108–113, 2019. 3

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[32] Shreya Lohar, Lei Zhu, Stanley Young, Peter Graf, and Michael Blanton. Sensing technology survey for obstacle detection in vegetation. *Future Transportation*, 1(3):672–685, 2021. 1

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[34] Flavio BP Malavazi, Remy Guyonneau, Jean-Baptiste Fasquel, Sebastien Lagrange, and Franck Mercier. Lidar-only based navigation algorithm for an autonomous agricultural robot. *Computers and electronics in agriculture*, 154:71–79, 2018. 3

[35] John A Marchant and Renaud Brivot. Real-time tracking of plant rows using a hough transform. *Real-time imaging*, 1(5):363–371, 1995. 3

[36] Henrik S Midtiby, Thomas M Giselsson, and Rasmus N Jørgensen. Estimating the plant stem emerging points (pseps) of sugar beets at early growth stages. *Biosystems engineering*, 111(1):83–90, 2012. 3

[37] António Monteiro, Sérgio Santos, and Pedro Gonçalves. Precision agriculture for crop and livestock farming—brief review. *Animals*, 11(8), 2021. 1

[38] Vignesh Raja Ponnambalam, Marianne Bakken, Richard JD Moore, Jon Glenn Omholt Gjevestad, and Pål Johan From. Autonomous crop row guidance using adaptive multi-roi in strawberry fields. *Sensors*, 20(18):5249, 2020. 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[40] Uferah Shafi, Rafia Mumtaz, José García-Nieto, Syed Ali Hassan, Syed Ali Raza Zaidi, and Naveed Iqbal. Precision agriculture techniques and practices: From considerations to applications. *Sensors*, 19(17):3796, 2019. 1

[41] Yan Song, Feiyang Xu, Qi Yao, Jialin Liu, and Shuai Yang. Navigation algorithm based on semantic segmentation in wheat fields using an rgb-d camera. *Information Processing in Agriculture*, 2022. 3

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2, 5

[43] Benoît Thuilot, Christophe Cariou, Philippe Martinet, and Michel Berducat. Automatic guidance of a farm tractor relying on a single cp-dgps. *Autonomous robots*, 13(1):53–71, 2002. 1

[44] Abhinav Valada, Gabriel Oliveira, Thomas Brox, and Wolfram Burgard. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *International Symposium on Experimental Robotics (ISER)*, 2016. 4

[45] Jasper Verschuur, Sihan Li, Piotr Wolski, and Friederike EL Otto. Climate change as a driver of food insecurity in the 2007 lesotho-south africa drought. *Scientific reports*, 11(1):3852, 2021. 1

[46] Wera Winterhalter, Freya Veronika Fleckenstein, Christian Dornhege, and Wolfram Burgard. Crop row detection on tiny plants with the pattern hough transform. *IEEE Robotics and Automation Letters*, 3(4):3394–3401, 2018. 3

[47] David M Woebbecke, George E Meyer, Kenneth Von Bargen, and David A Mortensen. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995. 3

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[49] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 2, 5

[50] Xiya Zhang, Xiaona Li, Baohua Zhang, Jun Zhou, Guangzhao Tian, Yingjun Xiong, and Baoxing Gu. Automated robust crop-row detection in maize fields based on position clustering algorithm and shortest path method. *Computers and electronics in agriculture*, 154:165–175, 2018. 3

[51] Kai Zhao, Qi Han, Chang-Bin Zhang, Jun Xu, and Ming-Ming Cheng. Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4793–4806, 2021. 3, 5, 7

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5