1 Image analysis and segmentation for agricultural applications

1.1 Objective and relevance to the project

As the project intends to cover a wide spectrum of research and development in drone technology, agriculture stands out as a major application area. Moreover, building an ecosystem specifically for drone-based applications will lead to improved results and provide opportunities for further innovation. Utilizing drone-captured images of crop fields and analyzing them to draw meaningful conclusions in various domains is the primary objective. This includes applications in precision agriculture, such as smart irrigation, plant and soil health monitoring, pest and disease prediction and control, among others.

Image segmentation plays a pivotal role in agriculture by enabling precise identification of different regions within a field, such as distinguishing crops from soil or weeds. This is critical for tasks like monitoring crop health, estimating yields, and identifying stressed or diseased plants at an early stage. Accurate segmentation allows for efficient resource allocation, such as targeted irrigation and pesticide application, which reduces costs and improves overall sustainability. By automating these processes through drone-based imagery and advanced segmentation models like SAM, farmers can make data-driven decisions, ultimately leading to higher productivity and better crop management.

Furthermore, the project investigates image segmentation algorithms for agricultural data for the calculation of the Normalized Difference Vegetation Index (NDVI), a critical indicator for assessing soil treatment and quantifying vegetation health and density. We applied the Segment Anything Model (SAM) to effectively segment crop regions from non-crop areas, experimenting with various combinations of spectral bands to find the optimal combination.

1.2 Background

1.3 Literature Review

1.3.1 Segment Anything Model

The Segment Anything Model (SAM),[2] developed by Meta AI, is a state-of-the-art image segmentation model designed to be highly generalize and versatile across diverse segmentation tasks. SAM is built to "segment anything" meaning it can handle a wide variety of segmentation challenges without the need for task-specific fine-tuning.

Architecture:
Image Encoder: SAM uses a ViT (Vision Transformer) as the backbone to extract feature representations from the input image
Prompt Encoders: SAM introduces prompt encoders, which allow it to take different types of user inputs (prompts). These prompts guide the segmentation

process. The prompts can be of various forms:

- Points: Users can mark specific points in the image to indicate the areas of interest (foreground/background)
- Bounding Boxes: Users can draw bounding boxes around objects to segment them
- Masks: Users can provide an initial mask, and SAM can refine it.

Further, the model is also trained to segment automatically without any prompts. Each type of prompt is encoded and combined with the image features to influence the segmentation process

Mask Decoder: The mask decoder is responsible for generating the final segmented output based on the image features and the encoded prompt. This decoder generates segmentation masks for the regions of interest specified by the user prompts. The decoder predicts a binary mask, which identifies which pixels belong to the object(s) of interest.

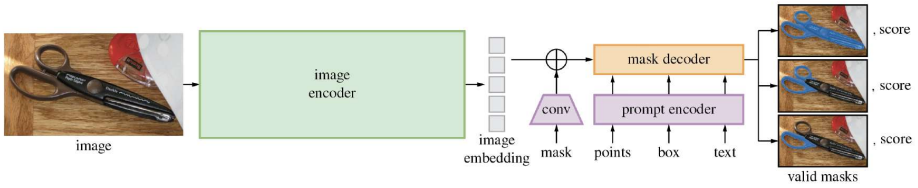

Figure 1: SAM architecture. Image courtesy: [2]

1.3.2 Segment Anything Model 2

SAM 2 [7], the next iteration of Meta's Segment Anything Model (SAM), is an advanced tool for thorough object segmentation in images and videos. It effectively manages complex visual data using a unified, promptable model architecture, enabling real-time processing and zero-shot generalization capabilities.

- Unified Model Architecture: It integrates image and video segmentation into a single model, streamlining deployment and ensuring consistent performance across various media formats. It utilizes a versatile prompt-based interface, allowing users to define objects of interest using different types of prompts, including points, bounding boxes, or masks.
- Real-Time Performance: The model achieves real-time inference speeds, processing approximately 44 frames per second. This makes SAM 2 suitable for applications requiring immediate feedback, such as video editing and augmented reality.

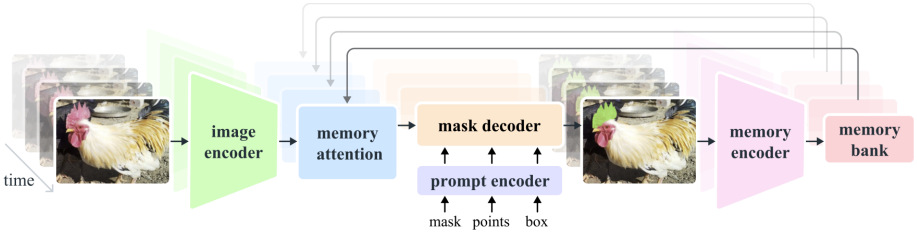

Figure 2: The SAM 2 architecture. For a given frame, the segmentation prediction is conditioned on the current prompt and/or on previously observed memories. Videos are processed in a streaming fashion with frames being consumed one at a time by the image encoder, and cross-attended to memories of the target object from previous frames. The mask decoder, which optionally also takes input prompts, predicts the segmentation mask for that frame. Finally, a memory encoder transforms the prediction and image encoder embeddings (not shown in the figure) for use in future frames.

- Zero-Shot Generalization: It segment objects it has never encountered before, demonstrating strong zero-shot generalization. This is particularly useful in diverse or evolving visual domains where pre-defined categories may not cover all possible objects.
- Interactive Refinement: Users can iteratively refine the segmentation results by providing additional prompts, allowing for precise control over the output. This interactivity is essential for fine-tuning results in applications like video annotation or medical imaging.

1.3.3 KAN-based Segmentation

Kolmogorov–Arnold Networks (KAN)[4] are alternatives to Multi-Layer Perceptrons (MLPs), where the activation functions are learnable which enhance the ability to decompose high-dimensional complex functions into univariate transformations, enabling efficient and flexible representation of intricate relationships in data. Xianping Ma et al. [5] have demonstrated that the KAN-enhanced segmentation model achieves superior performance in terms of accuracy compared to state-of-the-art methods. A KAN based encoder-decoder architecture is employed to capture complex spatial and rich semantic relationships from high-dimensional features. The encoder outputs are projected into low-dimensional space(using dimensionality reduction techniques like PCA and UMAP etc.) and the segmentation is obtained by clustering these low dimensional features.

1.3.4 DeeplabV3

DeepLabv3 is used as the segmentation model to perform semantic segmentation on the phenobench dataset . It uses ResNet-101 backbone, which is pre-trained

on ImageNet and fine-tuned for the specific task of segmenting agricultural images (e.g., soil/background, crop, weed) from the PhenoBench dataset . The model enhances segmentation accuracy using atrous convolution for dense feature extraction. CNNs typically use repeated combination of max-pooling and striding at consecutive layers which significantly reduces the spatial resolution of the resulting feature maps .

DeepLabv3 employs atrous (dilated) convolutions, allowing the model to capture multi-scale contextual information without increasing the computational cost or losing resolution. Consider two-dimensional signals, for each location i on the output y and a filter w, atrous convolution is applied over the input feature map x:

$$y[i] = \sum_k x[i + r \cdot k]w[k] \tag{1}$$

The model incorporates Atrous Spatial Pyramid Pooling (ASPP), which applies multiple parallel atrous convolutions with different dilation rates (e.g., 4, 8, 16) to capture information at various scales.

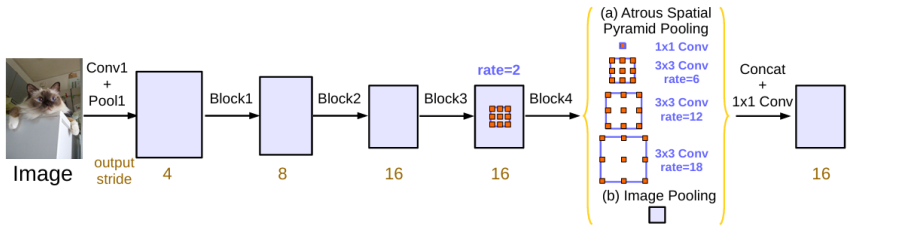

Figure 3: Parallel modules with atrous convolution (ASPP), augmented with image-level features [1]

1.3.5 MSCG-Net

Multi-view SCG-Net (MSCG) [3] for semantic labeling tasks with the proposed adaptive class weighting loss was used semantic segmentation on phenobench dataset . It utilizes Self-Constructing Graph module (SCG) to exploit the rotation invariance in airborne images, by extending it to consider multiple views . SCG is used to consider undirected graph to capture relations across the image . The feature map . GCN (Graph Convolutional Network) is applied after the SCG module . The SCG-GCN module then outputs various predictions with different degrees of rotation . Se_ResNext101_32x4d was used as backbone CNN to learn the high level representations of the feature map of the images . To handle the imbalance in classes in the dataset - the Adaptive Class Weighting (ACW) loss function was used . It updates class weights based on pixel-frequency statistics at each training step instead of using fixed weights. The pixel-frequency for

class j at training step t is computed as:

$$f_j^t = \frac{\hat{f}_j^t + (t-1) \cdot f_j^{t-1}}{t}$$

The adaptive class weight is then calculated using median frequency balancing:

$$w_j^t = \frac{\text{median}(\{f_j^t \mid j \in C\})}{f_j^t + \epsilon}$$

These weights are normalized with pixel-wise predictions $\tilde{y}_{ij}$ and ground-truth y_{ij}:

$$\tilde{w}_{ij} = \frac{w_j^t}{\sum_{j \in C} w_j^t} \cdot (1 + y_{ij} + \tilde{y}_{ij})$$

Finally, a positive-negative contrastive (PNC) function is used with this weighting in the ACW loss [3]:

$$\mathcal{L}_{acw} = \frac{1}{|Y|} \sum_{i \in Y} \sum_{j \in C} \tilde{w}_{ij} \cdot p_{ij} - \log\left(\text{MEAN}\{d_j \mid j \in C\}\right)$$

This model is typically designed for the agrivision dataset , the model was tuned to handle RGB images instead of accepting 4 channels (NIR-RGB) . The images were resized due to the high computational complexity and to match the dimensions of the original model.

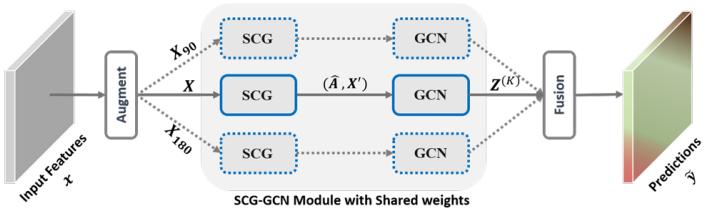

Figure 4: MSCG-Net [3]

1.3.6 Segmentation of unlabeled data

SAM Segment Anything Model (SAM)[2] is a foundation model which segments images based on the prompts, such as points and bounding boxes. Even though it can segment objects without prior knowledge of the object type, it struggle to detect small objects of early-stage crops. Since SAM is not specifically trained on multi-spectral images, it fails to segment canopies of an agricultural land.

1.4 Achievements

The above-mentioned work develops techniques for unsupervised segmentation of agricultural crops, enabling more efficient NDVI calculation without manual intervention. It successfully segmented the crop field areas from the soil while comparing the NDVI values obtained from the model with those from a black-box software, Pix4D. Moreover, RMSE error observed was as low as 0.07.

1.5 Challenges and Solutions

- Absence of labelled data prevented us from validating the results we obtained from segmentation - Agri-Vision Dataset [6]
- It is required to fine-tune the model for specific agricultural crop fields to provide better results; proper dataset collection required.
- For more refined and detailed individual plant segmentation, high resolution images are required.
- Started interacting with other groups to use the collected sugarcane field data.

1.6 Future Work

- SAM is a segmentation model. We aim to use this model to classify the age of sugarcane and determine the optimal maturity period. We are looking forward to using the SAM model as a classification tool.
- We aim to identify and differentiate between healthy and unhealthy plants (and the reasons behind their condition) so they can be treated accordingly.
- We plan to integrate crop yield data with NDVI values and find the correlation between these two parameters.
- Furthermore, we aim to apply various segmentation models to the sugarcane dataset and draw meaningful inferences from it.
- Additionally, we will explore better segmentation techniques available for agricultural fields.

1.7 Outcomes till date

1.7.1 Results from SAM

The Agri-vision dataset [6] used was annotated with nine types of field anomaly patterns that are most important to farmers. This dataset contains six types of annotations: Cloud shadow, Double plant, Planter skip, Standing Water,

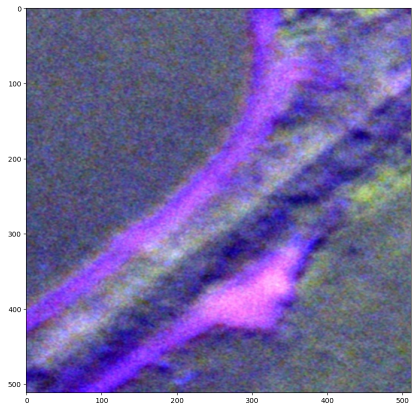

Figure 5: Input RGB image

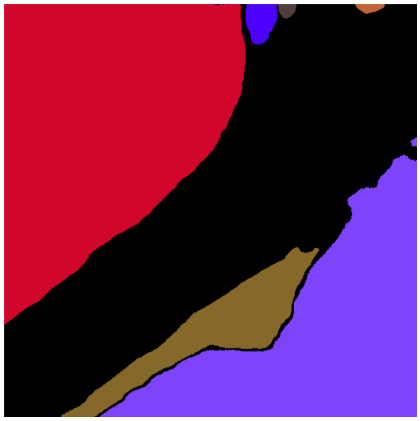

Figure 6: Masks from SAM

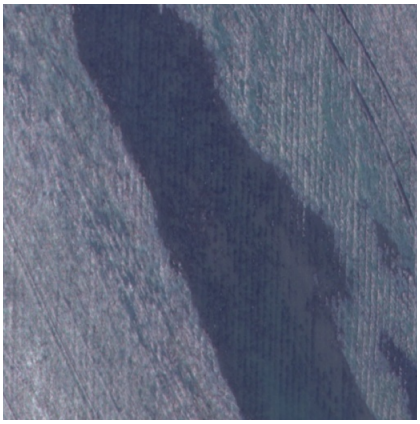

Figure 7: Composite RGB image

Figure 8: Masks from SAM

Waterway and Weed cluster. These types of field anomalies have great impacts on the potential yield of farmlands, therefore it is extremely important to accurately locate them.

The above results from SAM depicts the segmented image from the Agri-Vision Dataset when a three channel RGB image specifically Waterway annotated image was fed into the model, making each colored patch distinct from the others.

1.7.2 Results from SAM2

Similar results as from SAM were obtained from SAM 2 also. Since, it also perfroms segmentation in similar way as SAM where each colored patch is distinct from the others.

1.7.3 Results from KAN

The multi-spectral images were obtained at a resolution of 5cm/pixel by flying the drone at a height of 53 meters. The multi-spectral image comprises of four bands stacked in the order Green, Red, RedEdge and NIR. For better segmentation, the orthomosaic image obtained from Parrot Sequoia camera is converted into tiles of size 512×512 pixels.

The results of KAN-based self-supervised segmentation on the multi-spectral image are shown in Fig.10. It depicts the superior performance of KAN-based networks over SAM in segmenting small objects or early-stage crops.

Figure 9: First three bands of Multi-spectral Image(vineyard)

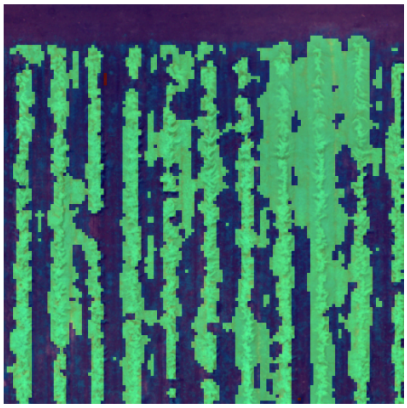

Figure 10: Overlay of segmentation masks on the input image

1.7.4 Results from DeepLabV3 and MSCG-Net

Phenobench Dataset Dataset consists of RGB images (1024 x 1024) in real field conditions recorded by a UAV equipped with a high-resolution camera that captures imagery of the field. For recording the data, a DJI M600 was employed and PhaseOne iXM-100 camera with an 80 mm RSM prime lens mounted on a gimbal to obtain motion-stabilized RGB images at a resolution of 11664×8750 pixels.

The UAV was flying at a height of approximately 21, resulting in a ground sampling distance (GSD) of 1mm/pixel. To cover the entire field, we used the DJI Ground Station Pro app to plan a flight that covers the field row-wise.

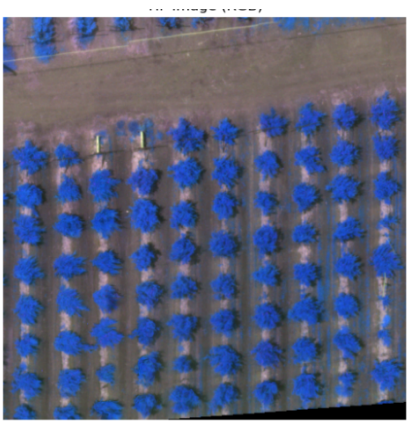

Figure 11: First three bands of Multi-spectral Image

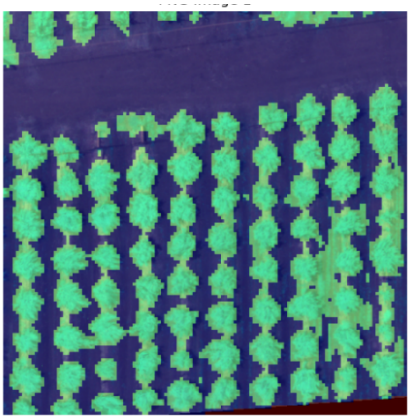

Figure 12: Overlay of segmentation masks on the input image (pomegranate)

We set the forward overlap between consecutive images (by motion vector) at 75%, and the side overlap between images placed in neighboring rows at 50%. Each image is geo-referenced using the on-board GNSS.

The PhenoBench dataset provides semantic annotations for vegetation segmentation tasks with the following label classes: **background** (label 0), **crop** (label 1), **weed** (label 2), **partial crop** (label 3), and **partial weed** (label 4). To reduce ambiguity at vegetation boundaries and simplify the classification task, we merge the partial classes into their respective primary categories. Specifically, label 3 (partial crop) is reassigned to label 1 (crop), and label 4 (partial weed) is reassigned to label 2 (weed). After this remapping, the dataset effectively contains three semantic classes: background, crop, and weed. This merging strategy ensures more robust model training and cleaner separation between meaningful vegetation classes in the field imagery [8].

Figure 13: Image from phenobench dataset [8]

Table 1: IoU Results on PhenoBench Dataset: DeepLabV3 vs MSCG-net

Class	**DeepLabV3 IoU**	**MSCG-net IoU**
Soil	0.99	0.90
Crop	0.91	0.85
Weed	0.47	0.35
Mean IoU	**0.79**	**0.70**

MSCG-net performed slightly worse than deeplabv3 because the images were resized due to high computational complexity . Also , the images in phenobench don't suffer from the problem of random orientations in images like in agrivision dataset so it didn't bring significant novelty in performance .

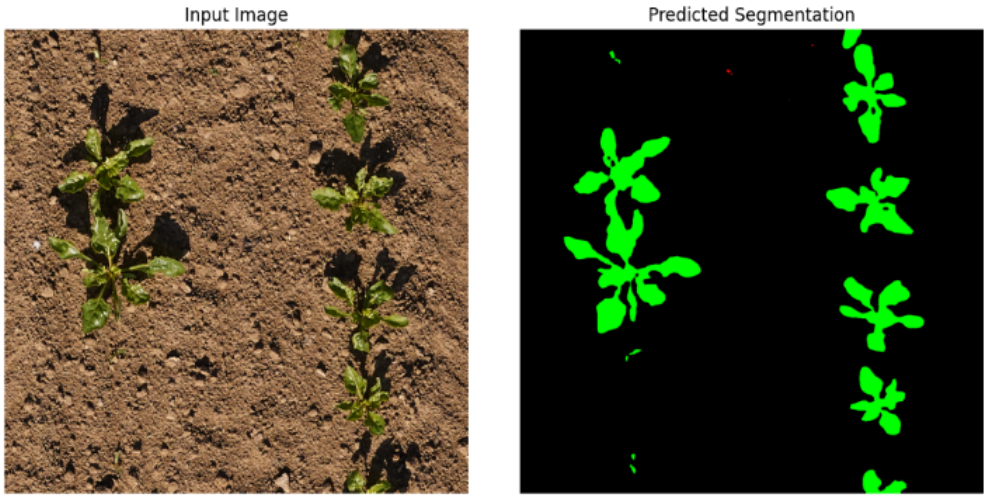

Figure 14: Segmentation result on phenobench dataset by DeepLabV3

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[3] Qinghui Liu, Michael Kampffmeyer, Robert Jenssen, and Arnt-Børre Salberg. Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation. pages 199–205, 06 2020.

[4] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[5] Xianping Ma, Ziyao Wang, Yin Hu, Xiaokang Zhang, and Man-On Pun. Kolmogorov-arnold network for remote sensing image semantic segmentation. *arXiv preprint arXiv:2501.07390*, 2025.

[6] Yunchao Wei Zilong Huang1 Alexander Schwing Robert Brunner Hrant Khachatrian Hovnatan Karapetyan Ivan Dozier Greg Rose David Wilson Adrian Tudor Naira Hovakimyan Thomas S. Huang Honghui Shi Mang Tik Chiu, Xingqian Xu. Agriculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis. *CoRR*, abs/2001.01306(x), 2020.

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[8] Jan Weyler, Federico Magistri, Elias Marks, Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley. Phenobench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain. *IEEE transactions on pattern analysis and machine intelligence*, PP, 06 2024.