PlantSAM: Towards Real-Time Plant Segmentation with Efficient Vision-Language Foundation Models

Daniel Petti (daniel.petti@ufl.edu), Charlie Changying Li (cli2@ufl.edu), Alina Zare (azare@eng.ufl.edu)

Paper 2500547

Introduction

- Locating crops and weeds in images (segmentation) is a needed capability for agricultural robots.
- AI foundation models such as Segment Anything Model (SAM) could potentially be used for this.
- However, SAM is a generalist model and not especially good at segmenting plants.
- BioCLIP is an existing model that associates plant images with scientific species names.
- **Goals are: (1) Generate a high-quality plant segmentation dataset by leveraging SAM and BioCLIP.**
- **(2) Train a version of SAM that segments plant species based on text prompts.**

Methods

- SAM segments objects based on spatial prompts.
- A large Vision Transformer (ViT) is used to extract features.
  - Masks are then generated based on features and prompts.
- SAM can segment based on text prompts by leveraging CLIP.
- We replace CLIP with BioCLIP to segment plant species.

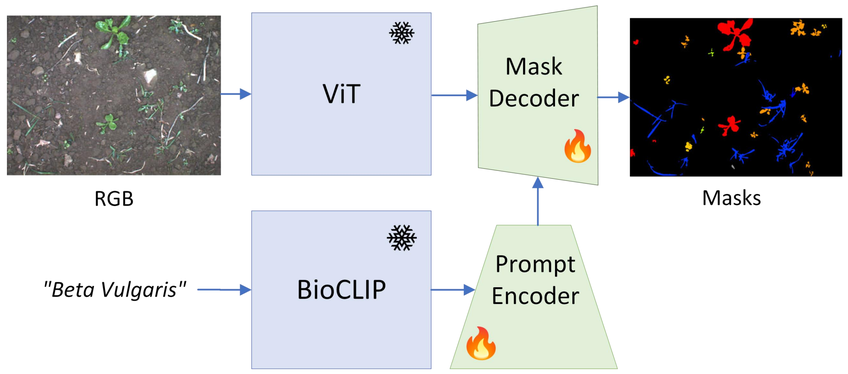

Figure 1: The basic architecture of SAM with BioCLIP prompts.

- We use some public datasets to train this model.
  - In total, this yields ~53,000 images from crop/weed species.
- We also leverage ~20,000 images from our private collection.
  - Encompass crops like blueberry, brassica, and cotton.

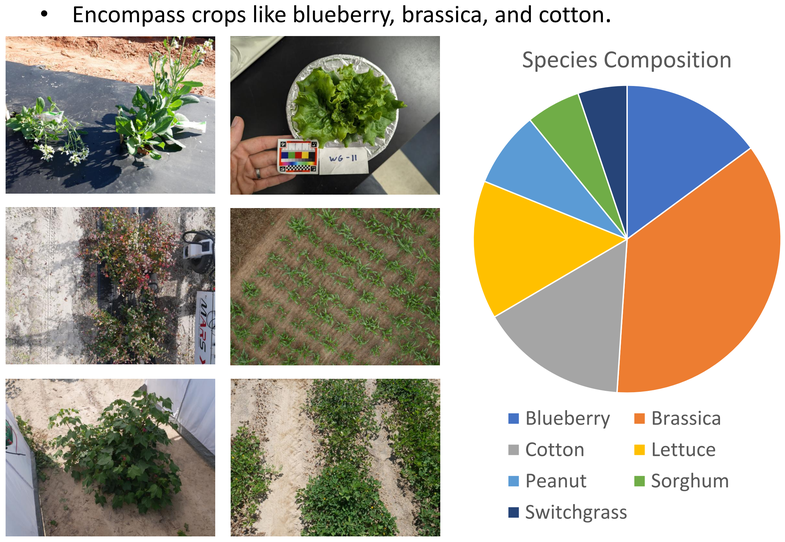

Figure 2: Species composition and images from our private dataset.

Methods (cont.)

- Most images in our dataset don't have mask annotations.
- We propose a pipeline to generate these automatically.
- We use SAM Automatic Mask Generation (AMG) mode to generate candidate masks.
- Then we embed the masks using BioCLIP.
  - Masks with a low probability of being plants are removed.

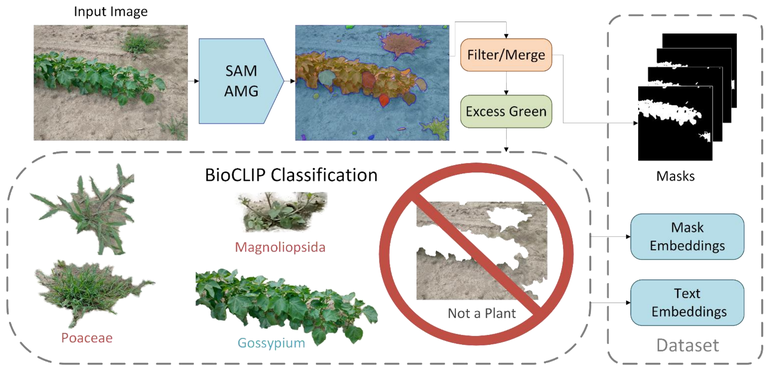

Figure 3: We generate "ground-truth" masks automatically with SAM and use BioCLIP to filter and embed the results. Excess green is used as an additional cue for determining plant-ness.

- We fine-tune the mask decoder on our data while keeping the image encoder frozen.

Results

- The model is capable of segmenting individual species when prompted.
- However, performance varies widely between species
  - Some species that are poorly represented in the dataset exhibit poor performance.

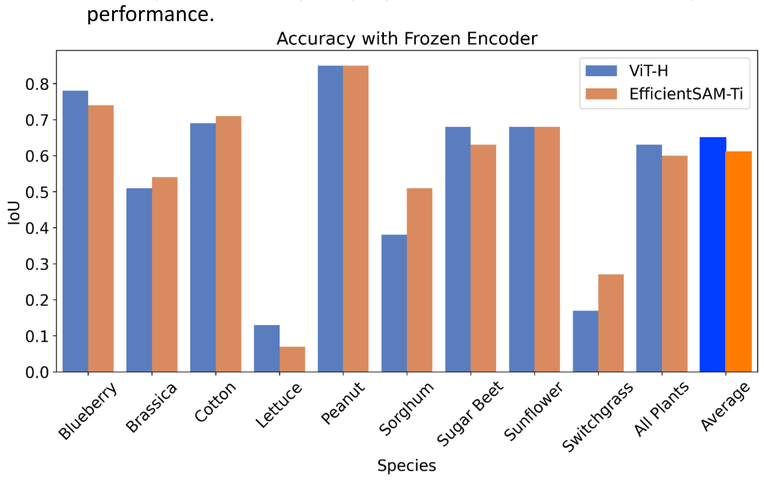

Figure 4: Segmentation accuracy of SAM models with fine-tuned decoders by species on a hand-annotated validation dataset.

- Using smaller EfficientSAM model and fine-tuning the encoder improves performance.
  - Some species also benefit from a mask refinement step.
- Qualitatively, we can observe that the model learns to differentiate between species.
  - This can be useful for segmenting crops vs. weeds

Results (cont.)

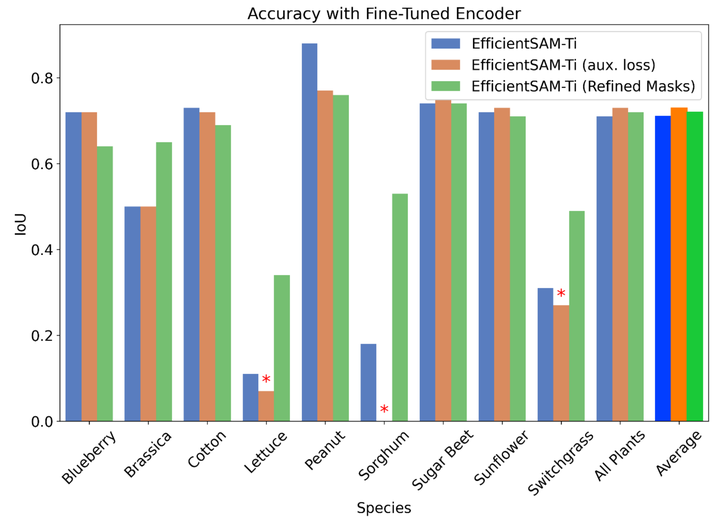

Figure 5: Performance of EfficientSAM models with fine-tuned encoders. One variant uses an auxiliary loss to ensure that the image embeddings are similar to those from the full-sized model.

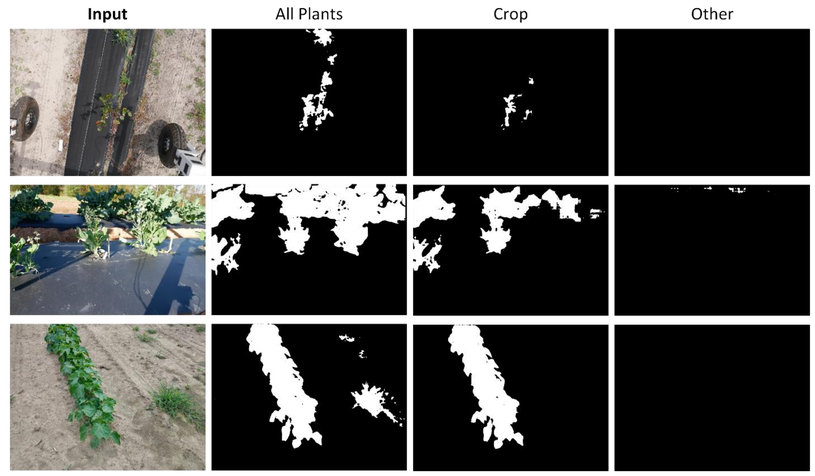

Figure 6: Qualitative segmentation results. Segmenting all plants produces a combined mask for crops and weeds. Segmenting just the crop removes the weeds. Segmenting a species that is not in the image produces a blank mask.

- EfficientSAM model can achieve 12 FPS on an edge device (Nvidia Jetson Orin AGX)

Conclusion

- SAM can be adapted for crop/weed segmentation.
- This process can be done with little annotated data
- Results are promising, but underscore limitations of the datasets used.

https://doi.org/10.13031/aim.202500547