American Society of Agricultural and Biological Engineers
2950 Niles Road | St. Joseph MI 49085-9659 | USA
269.429.0300 | fax 269.429.3852 | hq@asabe.org | www.asabe.org

An ASABE Meeting Presentation
DOI: https://doi.org/10.13031/aim.202500547
Paper Number: 2500547

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

Daniel Petti[1], Changying Li[1], Alina Zare[2]
[1]Bio-Sensing, Automation and Intelligence Laboratory, University of Florida, USA
[2]Machine Learning and Sensing Laboratory, University of Florida, USA

Written for presentation at the
2025 ASABE Annual International Meeting
Sponsored by ASABE
Toronto, Ontario
July 13-16, 2025

Abstract. Deep-learning-based semantic segmentation algorithms are very powerful and could be useful for advancing precision farming with row and specialty crops. However, this use case is limited by the difficulty of building large, general datasets. Consequently, new models are expensive to develop and may not generalize well to novel conditions. Recently, there has been a spate of interest in foundation models: large models trained on internet-scale data that have shown remarkable generalization capacity. One example is Segment Anything, which is able to segment objects based on simple visual prompts. However, in their most common form, these models may not be well-suited for agricultural tasks, as data scraped from the internet may not be applicable to agriculture, and the resulting models may be too large to run on edge devices. To address these issues, we develop an agriculture-specific foundation model. Using a large, unlabeled dataset of plant images, we extract embeddings with BioCLIP and use these to train a customized version of Segment Anything that can segment particular plant species. We then use knowledge distillation to create a small segmentation model that can run in real-time on an edge device and investigate its effectiveness for crop/weed segmentation. This model can achieve 73% average IOU while running at 12 frames per second. Overall, we demonstrate the feasibility of customized foundation models for precision agriculture tasks.

Keywords. *Foundation Model, Semantic Segmentation, Robotic Weeding, Weakly-Supervised Learning, Knowledge Distillation.*

Introduction

Accurate semantic segmentation of crops and weeds is an important problem in agricultural robotics. For instance, crop row segmentation plays an important role in visual navigation algorithms that are often developed to allow a robot to automatically follow a crop row. Typical approaches start by segmenting the crops in the image from a camera on the robot, then analyzing the resulting masks to adjust the robot's velocity (de Silva et al., 2024; Cerrato et al., 2024). Obviously, the success of this algorithm leans heavily on the accuracy of the segmentation, which should ideally be robust to crop selection, weed pressure, and illumination conditions. Similarly, robotic weeding approaches rely on the accurate segmentation of both crops and weeds (Ahmadi et al., 2024). Poor performance of this segmentation approach will lead to either weeds going unaddressed or the destruction of crops.

Though traditional threshold-based segmentation can differentiate between plants and soil, such approaches are typically not able to differentiate between crops and weeds (Weyler et al., 2023), necessitating the use of more advanced deep learning models. However, deep learning-based approaches traditionally require large training data sets, which can be cumbersome to collect and annotate. A common issue is that the resulting models won't generalize well to conditions that are different from the training data. The typical response is either to increase the diversity of the training data (Ilyas et al., 2025; Steininger et al., 2023) or the generalizability of the model (Weyler et al., 2023).

These considerations changed with the introduction of the Segment Anything Model (SAM) (Kirillov et al., 2023). SAM is part of a new crop of vision foundation models that arose in the past few years, spurred by advances in model design (Dosovitskiy et al., 2021) which bring with them the ability to leverage internet-scale data. Since the advent of modern deep learning techniques (Long et al., 2015; Ronneberger et al., 2015), semantic segmentation has generally been implemented as a fully-supervised pixel-level classifier. Instead, the evolution of semantic segmentation over the years has typically been a story of larger, better models and larger datasets (Cheng et al., 2022). SAM departs from this paradigm in its ability to perform *promptable* segmentation. Instead of being limited to a closed set of pre-defined classes, SAM will generate segmentation masks for any object specified by the user through point selections, bounding boxes, or even text descriptions.

In agriculture, SAM has been used to perform segmentation at diverse scales, from entire fields (Gurav et al., 2023; Kovačević et al., 2024; Ferreira et al., 2025) down to individual plant organs (Tan et al., 2025; Li et al., 2025; Nguyen et al., 2023). However, SAM is a generalist model trained with data scraped from the internet and therefore sometimes underperforms more specialized segmentation approaches on specific tasks. Furthermore, the promptable, interactive nature of SAM makes it challenging to use as a drop-in replacement for standard segmentation algorithms. Therefore, most successful deployments of SAM in agriculture use it either as part of a larger pipeline (Tan et al., 2025; Li et al., 2025; Nguyen et al., 2023), or modify the model in some way to enhance its capabilities (Li et al., 2023). This necessity undermines the inherent advantages of using SAM compared to a bespoke model.

Additionally, SAM is disadvantaged in robotic applications by its large and cumbersome model architecture. The original SAM implementation relies on a ViT-H (Dosovitskiy et al., 2021) backbone in order to extract as much information as possible from its large training dataset. However, such a large model (635 million parameters) is difficult to fit in the memory of edge devices. Furthermore, inference with SAM is relatively slow, and robotic weeding and navigation are both sensitive to inference latency. Therefore, a faster segmentation approach is needed.

Both problems can be addressed by using SAM merely to generate a large training dataset with minimal supervision and then training a more efficient model with the results (Li et al., 2025; Cao et al., 2025). DepthCropSeg (Cao et al., 2025) is a good recent example of this approach, although it uses Depth Anything v2 (Yang et al., 2024) instead of SAM, which is a vision foundation model designed for monocular depth estimation. DepthCropSeg is able to generate pseudo-masks for a large dataset of crop images with very little human supervision. Though this approach is promising, getting it to work in practice can often be difficult. Complex and carefully-engineered post-processing steps are usually necessary to convert the raw output from the foundation model into a reasonable ground truth example. Furthermore, the final trained model produced from this dataset is still limited by the strictures of traditional closed-set segmentation, meaning that it might need to be re-trained to accommodate changes in the crop type. It would be preferable to have a model that is usable for inference while retaining as much of the inherent flexibility of SAM as possible. This, after all, is the primary advantage of a foundation model.

An alternative method is to reduce the size of the SAM model in order to speed up inference (Sun et al., 2024). Typically, this involves replacing the large SAM encoder with a lighter-weight model and then using some sort of knowledge distillation (Hinton et al., 2015) to reconstruct the performance of the larger model. Specifically, this study focuses on EfficientSAM (Xiong et al., 2024), which uses a novel self-supervised reconstruction loss on the image embeddings. EfficientSAM is small enough to achieve good inference speed on a robot, so it was adopted as the basis for the proposed approach.

Precision agriculture applications require the ability to not only segment plants but also differentiate plant species, which vanilla SAM, being class-agnostic, cannot do on its own. SAM's text prompt mode can provide a method of segmenting plants, but this relies on CLIP (Radford et al., 2021) embeddings which are not sufficient for fine-grained differentiation between different plant species. BioCLIP (Stevens et al., 2023) is a vision-language foundation model that is able to align the embeddings of scientific clade names to images of organisms within that clade. This makes it a useful drop-in replacement for standard CLIP in situations where one is working with such data. Though the original BioCLIP is quite capable, it has also been extended several times, including with a broader training dataset (Yang et al., 2025), and with support for additional modalities (Gong et al., 2024; Sastry et al., 2024).

Although various efficient, high-performing crop/weed segmentation models have been proposed (Ahmadi et al., 2022; Dang et al., 2023), to our knowledge, all of them lack the ability to generalize significantly beyond the data they were trained on. This study explores an alternative to simply scaling up the dataset size. Mainly, generalization is achieved by leveraging existing vision foundation models (SAM and BioCLIP). To alleviate data annotation requirements, the proposed model can be trained using automatically generated masks, unlocking access to much larger and more diverse training datasets.

In short, the overarching goal of this study is to create a modified version of SAM that is capable of segmenting specific plant species when prompted. This is achieved by leveraging the BioCLIP foundation model in order to make SAM aware of the taxonomic hierarchy. By leveraging BioCLIP, it is possible to generate a large and diverse dataset of plant segmentation examples in a self-supervised manner (Figure 1). This dataset can then be used to train a modified version of SAM for segmenting plants. The resulting model, however, is too large to reasonably be deployed on a robot, so we also leverage the EfficientSAM approach to produce a tiny version. This version is sufficiently fast to be used for visual navigation or robotic weeding applications.

The specific objectives of this study are to:

- Generate a high-quality plant segmentation dataset by leveraging SAM and BioCLIP.
- Create a modified version of SAM that can segment specific plant species based on text prompts.
- Distill an efficient plant segmentation model that can run in real time on an edge device.

Methods

In the proposed approach, a modified version of SAM is trained (which we call PlantSAM) to segment particular plant species based on text prompts (Figure 1). This is enabled through the use of the BioCLIP (Stevens et al., 2023) foundation model, which allows for the addition of semantic information (the species classification) to the class-agnostic generated masks from SAM. In order to implement this, we propose an automated data generation pipeline that can produce semantic plant masks for unlabeled field imagery.

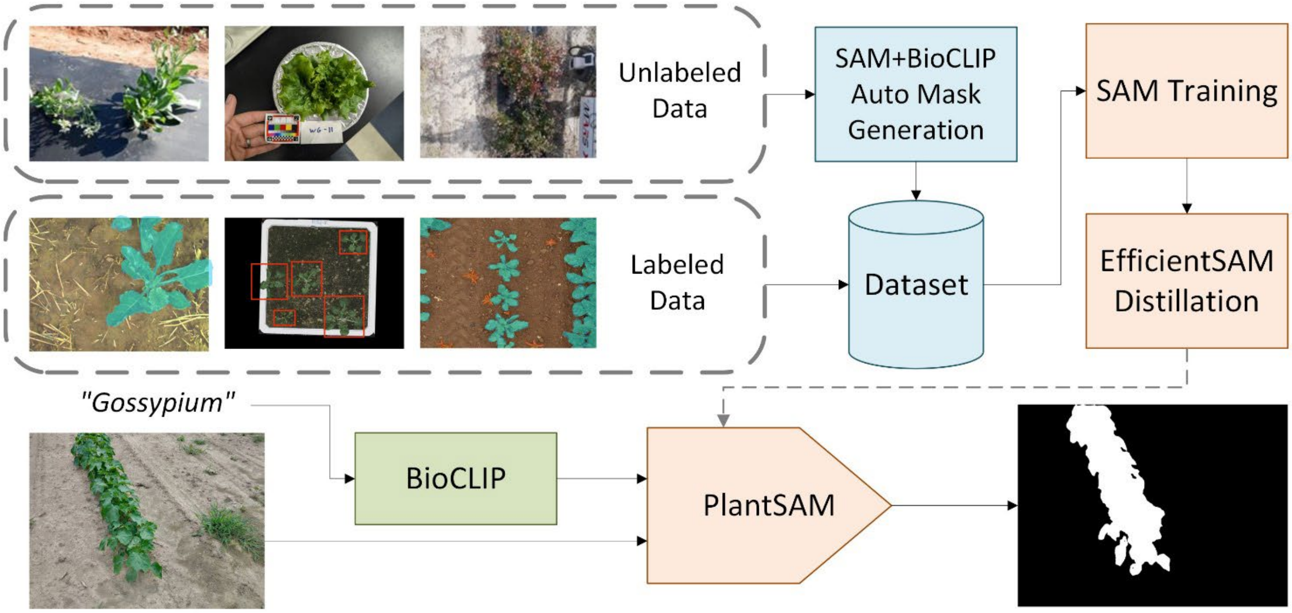

Figure 1: An overview of the proposed system. First, an unsupervised mask generation process, which leverages pre-trained SAM and BioCLIP models, is used to generate masks for unlabeled data. The resulting pseudo-labels are combined with public labeled datasets to produce a combined training dataset for the PlantSAM model. This large model can be distilled to a smaller model (based on EfficientSAM) that runs on edge devices. During inference, the model processes an image along with the BioCLIP embedding of the clade to segment as a prompt, and outputs a mask for that clade.

Dataset Construction

A comprehensive dataset of plant images with associated mask annotations was required for training the model. This dataset incorporates data from various public datasets, along with private data that we collected over several years (Table 1). Overall, the data encompass a variety of different crop species (with an additional, miscellaneous "weeds" category) (Figure 2). However, the data itself is fairly imbalanced, with some crops being much more heavily represented.

Table 1: Sources of the data used to construct the dataset. Note that "size" denotes the number of images used; not every image from every source dataset was used.

Source Dataset	Size	Annotation Type

SugarBeets (Chebrolu et al., 2017)	12,324	Mask
PhenoBench (Weyler et al., 2023)	2,875	Mask
OPPD (Madsen et al., 2020)	7,624	Bounding Box
Deep Weeds (Olsen et al., 2019)	17,509	None
Leaf Counting (Teimouri et al., 2018)	9,372	None
TerraRef (LeBauer et al., 2021)	542	None
LettuceMOT (Hu et al., 2022)	2457	Bounding Box
Private Data	20,345	None

Some of the dataset contained annotations which could be leveraged. For datasets that contained full mask annotations, these were used directly. For datasets that contained bounding box annotations, SAM was used with bounding box prompts to convert the bounding boxes into masks. (The annotations for LettuceMOT were not used at all.) All together, these public annotations provided a core of high-quality mask annotations for the dataset. However, there were still a significant number of images that had no annotations at all.

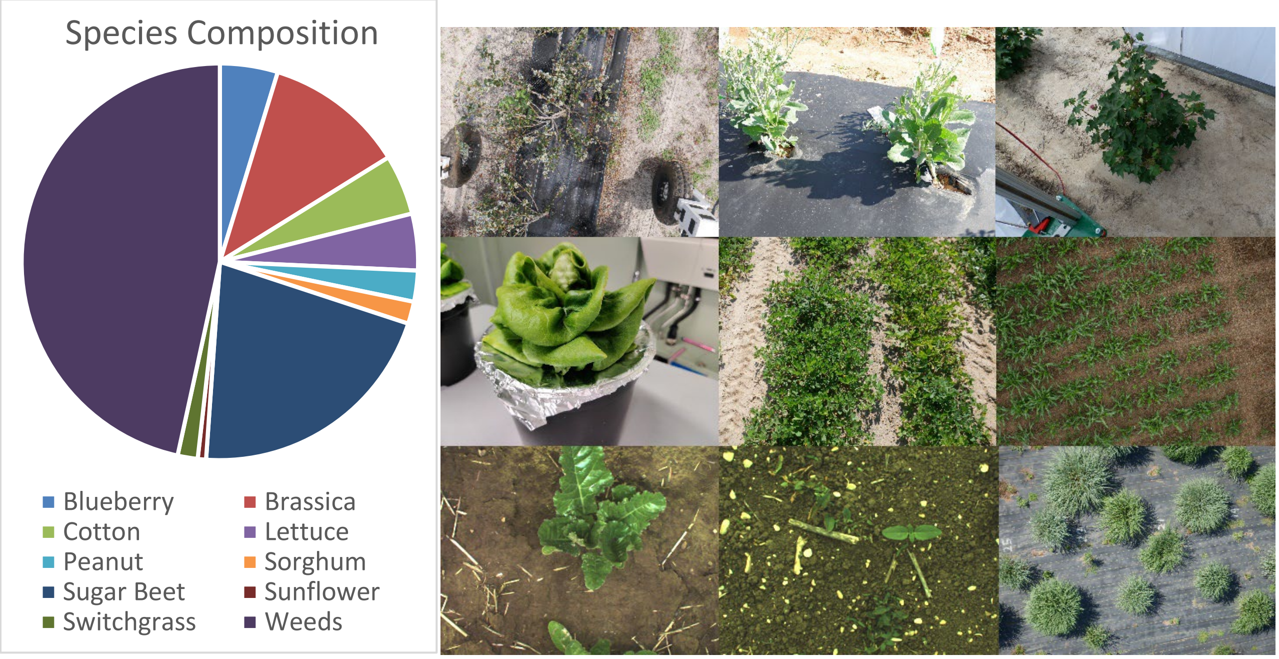

Figure 2: The composition of the dataset by species. On the right are several example images from the dataset.

Automatic Mask Generation

A pipeline to generate high quality masks of plants from input images without any manual labeling (Figure 3) is now required. The pipeline takes as input raw, unlabeled images, and outputs a set of masks for each image with associated ground-truth mask embeddings. The mask embeddings are generated by BioCLIP, which is applied only to the masked pixels of the image for each mask. These embeddings will be used as input prompts later when training the PlantSAM model.

The process leverages the automatic mask generation mode of SAM (Kirillov et al., 2023), in which the model is prompted with a regularly spaced grid of points, and all segmented masks are returned. For the proposed method, AMG mode is used with default parameters. Though SAM does perform some filtering on the generated masks, we find it helpful to perform additional filtering. Specifically, we detect masks that have >80% overlap and merge them into a single mask.

This process tends to produce masks that together cover most of the image. However, only some of these masks will encompass plants, and the rest will be part of the background. To determine which is which, the innate classification capabilities of BioCLIP (Stevens et al., 2023) are leveraged. Specifically, BioCLIP is applied to each mask and the model's hierarchical classification capability is used to determine whether that mask falls into the phylum "Plantae Tracheophyta" (vascular plants). Any masks that BioCLIP does not associate strongly with plant clades are eliminated at this stage.

Additionally, the excess green index is used as a cue that influences the mask classification step. If more than 50% of the pixels in the mask have an excess green value above a certain threshold, a lower confidence threshold is used for the BioCLIP clade prediction step. Otherwise, a higher confidence threshold is used. We found empirically that this strategy tended to increase the quality of BioCLIP predictions.

Excess green thresholding on a pixel level is also used in order to enhance mask quality. Specifically, SAM can struggle when segmenting highly complex plant structures with many small leaves and branches. To work around this, a mask enhancement step is introduced, where any pixels from a generated mask that do not rise above a certain excess green threshold are removed. The new mask is then applied to the image and BioCLIP is re-run. If the BioCLIP prediction does not change, we save this thresholded mask, otherwise, we discard it and use the original.

BioCLIP clade predictions are performed hierarchically, using a modified version of the classification strategy in the original BioCLIP code. After embedding a masked image, the similarity between this embedding and the text embedding is first calculated for every species in the TreeOfLife-1M dataset (Stevens et al., 2023). If no species meets the necessary similarity threshold, the similarity is then computed for each genus, summing the similarities for the species in that genus. In this way, we work our way up the taxonomy until the similarity for a clade reaches the threshold. Both the image and text embeddings from BioCLIP are then saved as part of the dataset, to be used as prompts during training. In this way, the most specific clade possible is assigned to each mask, even when BioCLIP exhibits a high degree of uncertainty.

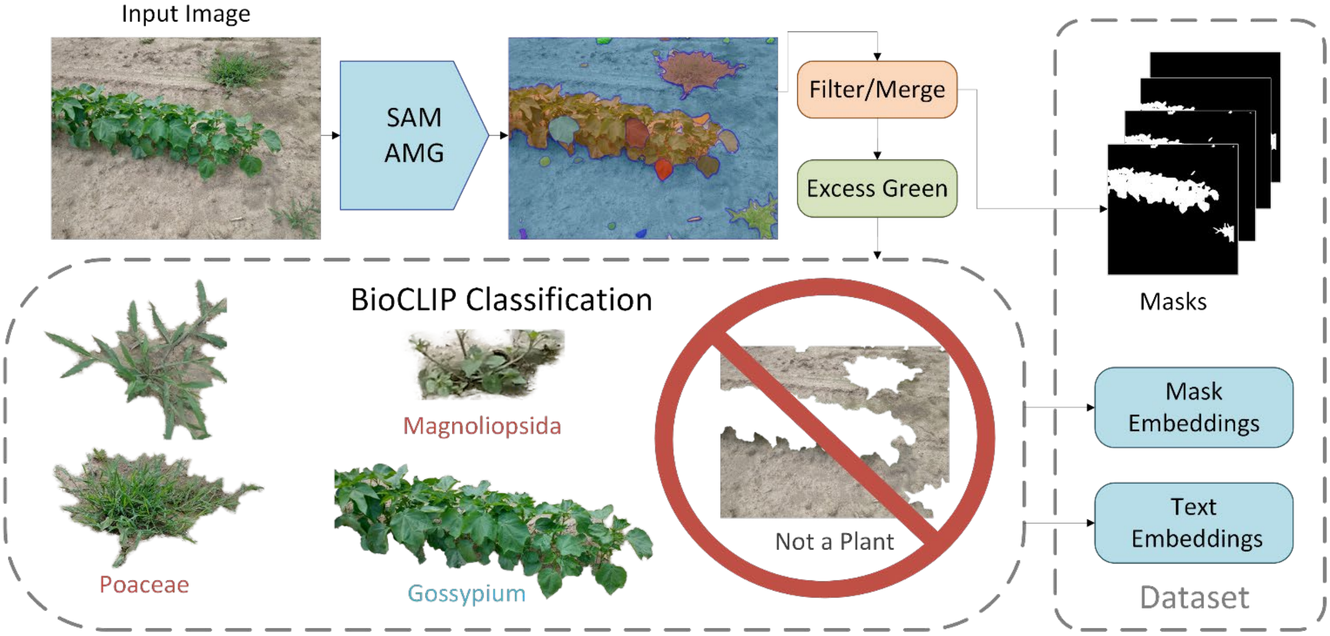

Figure: The mask generation workflow. SAM is first used in Automatic Mask Generation mode to generate candidate masks. These are filtered to remove overlaps, and BioCLIP is used to classify masks into clades. Non-plant masks are eliminated. Excess green is also calculated and used as an additional cue when determining whether a mask is a plant. The dataset contains generated plant masks for each image, as well as the associated image and text embeddings for each mask.

Even with these strategies, BioCLIP still produces sub-optimal results. Specifically, it will often fail to correctly ascertain the species for individual plants, only being able to classify them at a higher taxonomic level. This can be a problem when generating a training dataset, because our ultimate use-case involves prompting the model to segment a specific crop species. It was found that if too little of the dataset contains species-specific labels, the model has trouble learning how to respond to these granular prompts.

With an eye towards increasing the percentage of masks that include species-level annotations, prior knowledge about our dataset can be leveraged. Specifically, it is known which images in the dataset contain specific crops. With this knowledge, the similarity between each mask embedding and the text embeddings for the specific crop species in our dataset can then be computed. The softmax of the resulting scores can then be taken, essentially taking a difficult classification problem and turning it into a much easier one. If a very high probability is observed for a species that we know *a priori* is in the image, we go ahead and assign that species to this mask.

A final post-processing pass for all of the generated masks is also performed in order to ensure that the taxonomic hierarchy is respected. In particular, any masks of a higher-level clade should be a superset of the masks of its subclades. Therefore, we go through all of the generated masks and manually merge masks for low-level clades into the masks for their parent clades. A top-level mask for the “Plantae Tracheophyta” clade is also added if it doesn’t exist already. Once each mask is assigned to a clade, the BioCLIP text embedding are generated for the assigned clade and saved to our dataset as well.

When trained on this dataset, the model was found to successfully segment plants in general but struggle to produce correct species-level masks. Instead, it tends to produce the same results regardless of the prompt. In order to force the model to pay attention to species-level prompts, “negative examples” are added for each image in the dataset. These consist of a blank mask that is added to the dataset for every image. This mask has an associated BioCLIP embedding for a randomly selected crop species that is known to not be in the image.

Model Training

PlantSAM closely follows the original SAM architecture, with modifications to the prompting approach. Kirillov et al., (2023) specify how SAM may be modified to segment objects based on text prompts. Importantly, this approach does not require training SAM with an image-text dataset, instead, CLIP (Radford et al., 2021) is leveraged to enable text prompting. Specifically, the CLIP image embeddings are extracted for each ground-truth mask in the dataset, and SAM is prompted with these embeddings during training. Because CLIP aligns image and text embeddings, SAM can then be prompted with CLIP *text* embeddings during inference.

The proposed approach is similar but substitutes BioCLIP (Stevens et al., 2023) for CLIP. BioCLIP is trained on millions of images of different organisms in order to align their embeddings with the names of the taxonomic clade that they fall into. When SAM is trained with BioCLIP embeddings, it becomes possible to prompt the model with the name of the species to segment (Figure 4).

The SAM model consists of three primary components: the image encoder, prompt encoder, and mask decoder. The image encoder encodes the input image, the prompt encoder encodes the prompt, and the mask decoder takes both and produces mask predictions (Kirillov et al., 2023). When training PlantSAM, the image encoder is kept frozen and only the prompt encoder and mask decoder are fine-tuned (Figure 4). Because the ViT-H-based image encoder is enormous, with 635 million parameters, (Dosovitskiy et al., 2021) keeping it frozen saves a significant amount of computation.

The model is trained for 60 epochs using the Adam optimizer. All input images are resized to 1024x1024, following the procedure used for the original SAM model. Furthermore, the same losses and hyperparameters were used as in the original implementation. The image embeddings for the entire dataset were pre-generated using the pre-trained image encoder and save to the disk beforehand. This avoids having to run the image encoder during training and significantly speeds up the training process.

During training, three masks were predicted for every input prompt, but back-propagation was only performed through the one with the lowest loss. The original SAM model does this to address the inherent ambiguity in the prompting mechanism. By contrast, prompting with species names theoretically has no ambiguity. It was found, however, that allowing the model to predict multiple masks still produces better results overall. This might be because it allows the model to more easily address the inevitable noise present in the auto-generated target masks.

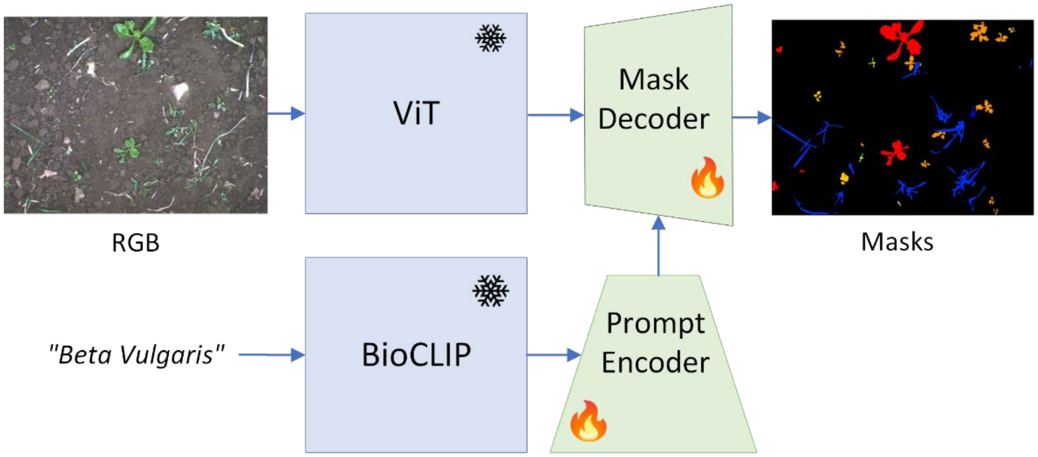

Figure 3: Architecture of the PlantSAM model. The ViT image encoder and BioCLIP model are frozen and used as-is. The remaining components are fine-tuned for plant segmentation.

Mask Refinement

Once the initial training is complete, the trained model is used to further refine the masks in the dataset. This is not a wholly novel idea. A similar strategy is used in DepthCropSeg (Cao et al., 2025). More broadly, active learning pipelines also use partially trained models for data annotation (Ren et al., 2021), but ultimately rely on human supervision to correct self-annotation errors. We employ no such supervision.

At the most basic level, mask refinement works by using an initial trained PlantSAM model to predict masks, then combining these masks with the original automatically extracted masks and re-running the mask embedding pipeline. (Figure 5). Predicted masks that overlap significantly with what's already in the dataset should be assigned to the same clade and therefore handled by the final mask post-processing pass. In the event that they are assigned to unrelated clades, the post-processing algorithm will select whichever clade is more specific (e.g. prioritizing a species-level classification over a class-level one).

During the mask refinement process, the model is prompted to segment all plants and include all three generated masks, with a very low IOU prediction threshold. The goal is to produce as many masks as possible; they can be refined later using BioCLIP. Since the model performs semantic and not instance segmentation, each mask is further broken up into connected components and each component is processed individually, as a basic heuristic for separating different plants.

Additionally, a second pass is performed in which the model is prompted to segment the specific crop species that we know is visible in the input image. The goal of this step is to further increase the portion of the target crop that is correctly segmented. Each mask generated with this process is simply checked with BioCLIP to ensure that it is, in fact, a plant. Assuming that it is, it is automatically assigned to the specific crop species that the model was prompted with, regardless of what BioCLIP says.

DepthCropSeg uses a simple refinement process in which only the mask regions that appear in both the initial generated masks and the predicted masks are kept. Obviously, this is a good strategy for removing false positive pixels. However, it provides no mechanism for correcting false negatives and could even introduce new false negatives. By contrast, the proposed refinement strategy is well-equipped to address false positives, because predicted masks are added directly to the dataset as long as they pass some basic checks. To address false negatives, another heuristic is introduced that penalizes regions of the mask in which the original generated masks and model predictions do not agree. Simply put, we check the amount of overlap between each original generated mask and the predicted mask. If this overlap is less the 50%, we use a higher threshold when performing the BioCLIP embedding step. In other words, the assumption is that this mask is less likely to be real and thus BioCLIP is required to be more confident that it is, in fact, a plant in order to keep it.

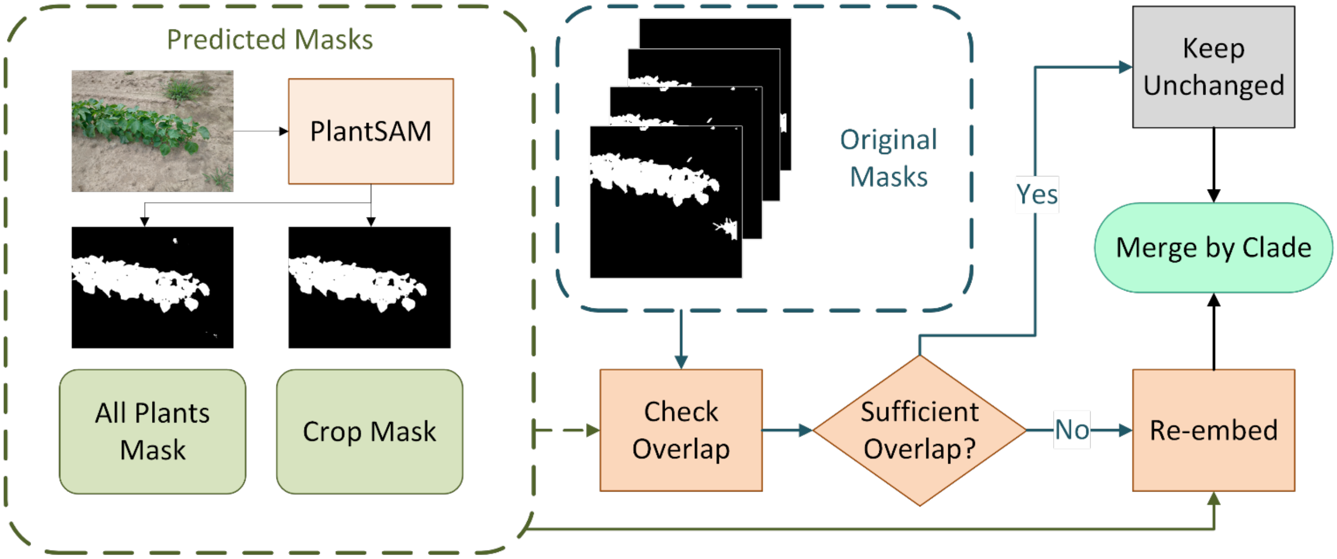

Figure 4: The mask refinement process. An initial trained version of the model is first used to predict masks. These predicted masks are then intelligently merged with the original masks from the dataset, and re-embedded if necessary using BioCLIP.

Knowledge Distillation

The original SAM model is based on ViT-H and therefore requires a large amount of computation during training and inference. Although computation is saved in our approach by only training the decoder and leaving the encoder frozen, we cannot escape the need to run the encoder during inference time. This makes the original PlantSAM model unsuitable for real-time deployment on a robot, where computational power is limited, and decisions need to be made in fractions of a second.

This is circumvented by leveraging the EfficientSAM-Ti encoder (Xiong et al., 2024), which is much smaller (10m parameters vs. 636m). Initially, this model was trained using the same method as the larger ones: freezing the encoder and fine-tuning the decoder. However, the model is small enough that it is also practical to un-freeze the encoder and train the entire thing. For both strategies, we begin with the pre-trained EfficientSAM weights and fine-tune on our dataset for 60 epochs.

The original EfficientSAM is trained using a knowledge distillation approach that imposes an MSE loss between the image embeddings generated by the EfficientSAM encoder, and those generated by the original pre-trained ViT-H encoder. A variation of this technique is tested on our own model. When training EfficientPlantSAM, a similar auxiliary MSE loss is imposed between the original and the new image embeddings. The total loss is a weighted sum of the image embedding loss and the original SAM losses.

The inference speed of our models was tested on an Nvidia A100 GPU. For this test, the decoder was run once for each run of the image encoder (predicting one mask for each input image). The models are optimized using TensorRT for the best possible inference performance. EfficientPlantSAM was also tested separately on an Nvidia Jetson Orin AGX, which more closely approximates the hardware that might be available on a robotic platform. Unfortuantely, we were unable to test the full-sized PlantSAM model on the Jetson because TensorRT conversion failed, likely due to memory constraints.

Evaluation

The automatically generated dataset poses a problem when evaluating model performance, because the quality of the masks cannot be guaranteed. To ensure meaningful evaluation results, we therefore create a small, hand-annotated dataset of 100 images. These images are selected randomly from the original private data collection and encompass all the crop species in that collection.

For the public datasets, if hand-annotated masks are available, these are used during evaluation. If a specific validation split is specified, this is also used. Otherwise, one is generated randomly. For OPPD (Madsen et al., 2020), we also use the masks generated from bounding boxes for validation, since they were empirically determined to be of relatively high quality.

Intersection Over Union (IOU) is selected as our primary evaluation metric. During species-specific evaluation, the model is prompted specifically to segment that species. Since the model generates three masks, we select the one with the lowest loss and report the accuracy of that one.

Results

Training the full PlantSAM (with the ViT-H image encoder frozen) on the combined dataset achieves a performance of 65% mIoU across all species (Figure 6). Overall, the results are respectable, but there are also clearly visible deviations in the performance between different species. Several species, such as Sorghum and Switchgrass, are not well-represented in the dataset, potentially contributing to their poor performance. Lettuce is better represented but mainly using images of potted lettuce plants taken in a laboratory setting. By contrast, most of the other images were collected in the field. Due to this discrepancy, the proposed model may have a difficult time operating on data from this setting.

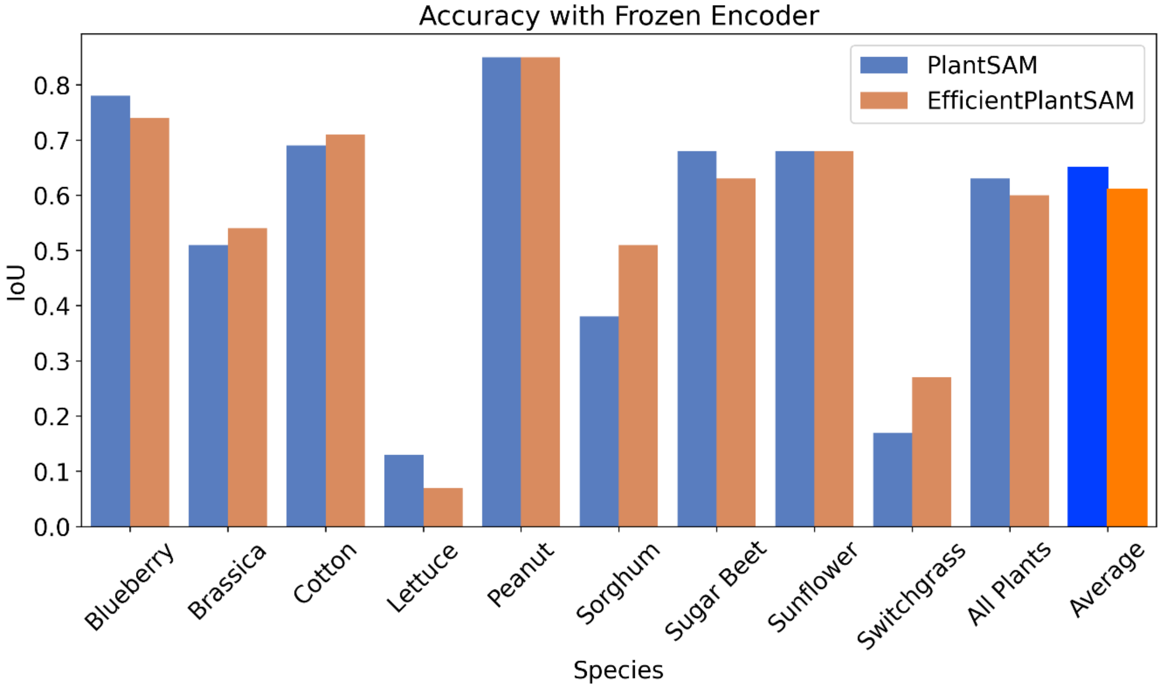

Figure 5: Accuracy of the models trained with a frozen image encoder. Color denotes which pre-trained image encoder was used.

The qualitative results of the PlantSAM approach appear promising. In images containing both crop and weed species, the model is able to segment either the crops and weeds together, or just the crops, depending on the prompt given (Figure 7). Also, the model produces empty masks when prompted with a species that is not in the image. It should be noted that the resulting masks are generally fairly high quality.

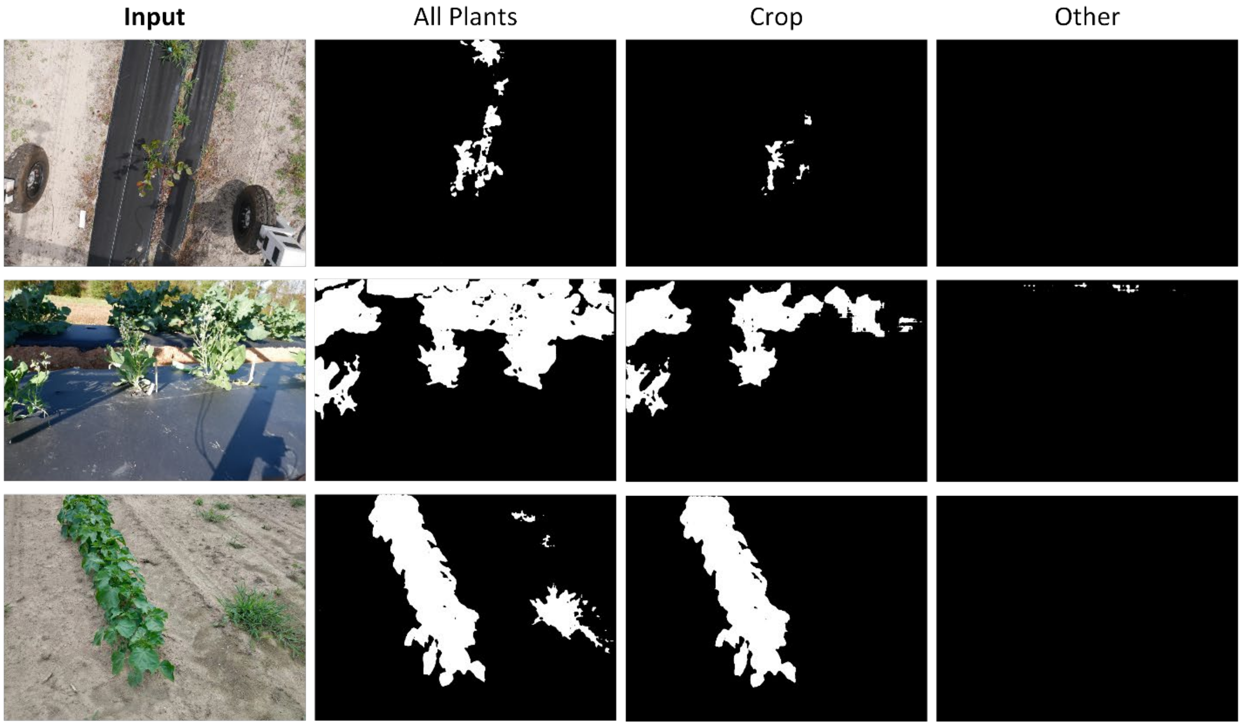

Figure 6: Some interesting examples of the segmentation results of our model. The "All Plants" column shows the results when prompted to segment all the plants. The "Crop" column shows the results when prompted to segment just the crop species. The "Other" column shows the results when prompted to segment a species that is not in the image.

PlantSAM also exhibits some interesting failure modes that might be an artifact of the dataset generation process. In general, it seems much more prone to false negatives than false positives (Figure 8, a, b, and c). When a false negative does occur, it usually fails to segment an entire plant. In particular, it often seems to not know how to handle blurred plants in the background of an image (Figure 8c). It is relatively rare that it successfully segments a plant but merely generates a low-quality mask. (When this does happen, it tends to be confused by shadows and the like, which is likely a tendency inherited from the base SAM through the automated mask generation process.)

The tendency towards false negatives is almost certainly a limitation of BioCLIP. We observed during the dataset generation phase that BioCLIP is not perfectly reliable at determining the species of a plant, especially when the original mask happens to be of low quality. BioCLIP, it should be noted, was trained mostly on data from iNaturalist (Stevens et al., 2023), and the private dataset contains images that are significantly out-of-domain for iNaturalist photos. Most significantly, these include images captured by UAVs and gantry systems. We note that BioCLIP seems to struggle particularly with these types of images. (The switchgrass and sorghum categories in our dataset are heavily dominated by UAV imagery, which is likely another contributing factor to their poor performance.)

Even in ideal conditions, BioCLIP sometimes fails for inexplicable reasons. This includes both false negatives (where BioCLIP fails to recognize a plant) and false positives (where BioCLIP tags something that isn't a plant as a plant). This situation becomes more acute for the much more difficult problem of accurately identifying a species, instead of merely ascertaining plant-ness. This is why we go to great lengths to improve BioCLIP's species recognition capabilities by leveraging prior knowledge. It is also why performance still tends to degrade somewhat when segmenting specific species, compared to when segmenting all plants.

Given that the model generates three separate output masks for each prompt, we note that in most cases, at least one of the generated masks is high quality. The standard practice during inference with SAM is to use the IOU prediction as a proxy for mask quality in order to decide which of the masks to select. Given the importance of the IOU metric, then, it is worth paying attention to its accuracy. It was noticed that, though IOU prediction rankings often do correlate well with subjective mask quality, they are occasionally incorrect (Figure 8d, e, and f). In these situations, a lower-quality mask will be predicted with a higher IOU than a higher-quality one.

A particularly pernicious situation arises when predicting empty masks (as when one prompts for a species that isn't present). Even if an image contains no plants at all, some noisy predictions are usually generated (Figure 8d). Furthermore, all of the IOU predictions will be near zero, even for the correct empty mask. This is due to a quirk of the IOU metric, in which the IOU of two empty masks is typically defined to be zero. Therefore, there is no easy way of distinguishing the correct mask in this case. Though it is not clear how frequently this situation would arise in the field, it might be worth considering modified versions of the IOU metric that exhibit different behavior in the empty mask case.

Knowledge Distillation

When keeping the encoders frozen and training only the decoder, replacing the ViT-H encoder with the EfficientSAM encoder reduces performance slightly, with a 61% mIOU score (Figure 6). This is an entirely expected result, due to the massive decrease in model size. What is less expected is that, if the EfficientSAM encoder is unfrozen, the resulting performance is *significantly better* than the larger model, with all versions achieving >70% mIOU (Figure 9). Obviously, the ability to modify the encoder parameters is likely making up for the much more powerful, but frozen, encoder of the larger model.

When it comes to training the smaller model, it was found that the auxiliary image embedding loss makes little difference (Figure 9). At 73% mIOU, the model with it does end up performing marginally better than the model without (at 71%), but this is not enough to make much of a difference in the real world. Auxiliary losses tend to have unpredictable effects on accuracy, because they force the model to trade off between optimizing the primary and auxiliary losses. This is underscored by the fact that performance actually decreases slightly on some species with the auxiliary loss. Still, we continued to use the loss in further experiments, seeing as it may help slightly and imposed negligible computational cost.

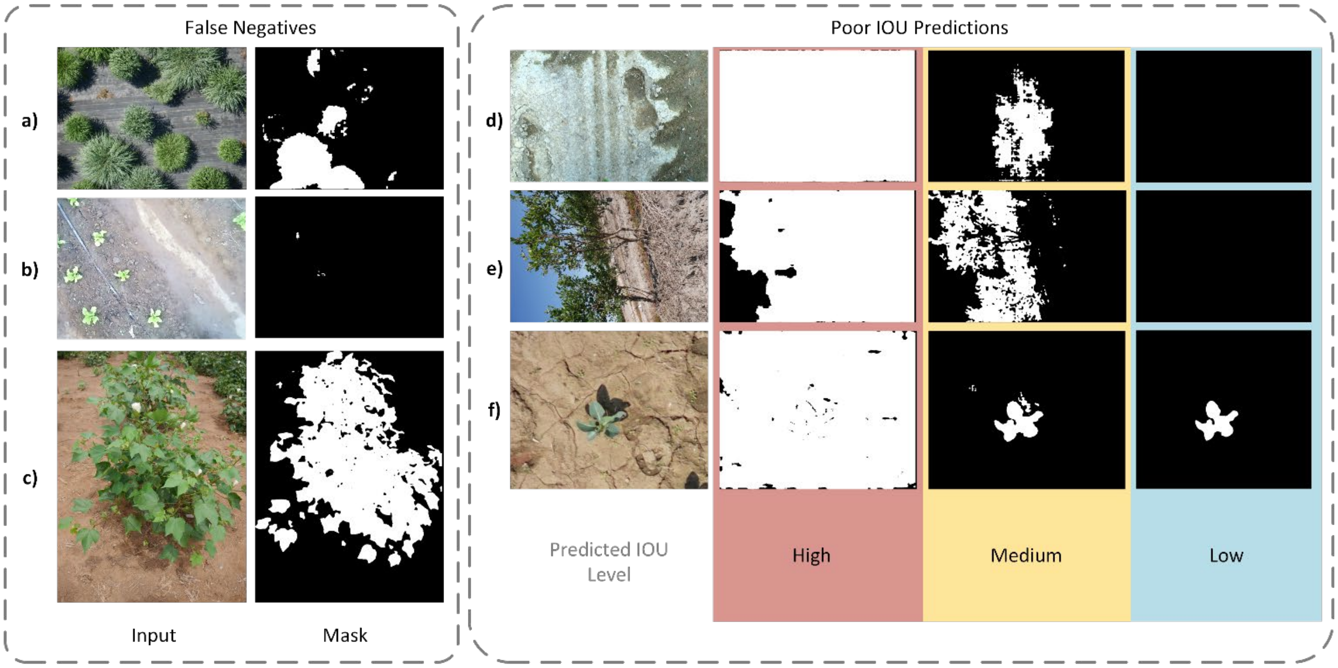

Figure 7: Some examples of the model's failure cases. The most common failure case is false negatives, in which the model typically fails to segment an entire plant (a, b, and c). Plants in the background are especially prone to this (c). In other cases, one of the three predicted masks might be correct, but the IOU prediction might be incorrect (d and e).

Mask Refinement

Performing an additional refinement step on the generated masks and training a new model makes little difference in the overall performance but does affect the performance on individual species (Figure 9). Here, we report the results of mask refinement on the EfficentSAM-based model, because it is both better-performing and faster to train than its larger brethren. Specifically, it was found that certain under-performing species, such as lettuce, improve significantly in performance when the model is trained on the refined dataset. (These species are marked with a red star in Figure 9). It should be noted that the models with trainable encoders seem to perform particularly badly on these species compared to those with frozen encoders, perhaps because freezing the encoder provides some inherent regularization that improves accuracy on species that are poorly represented in the dataset. Mask refinement, however, closes that gap. Therefore, is possible that mask refinement offers a possible method of counteracting the inherent challenges of training on an imbalanced, highly multi-modal dataset.

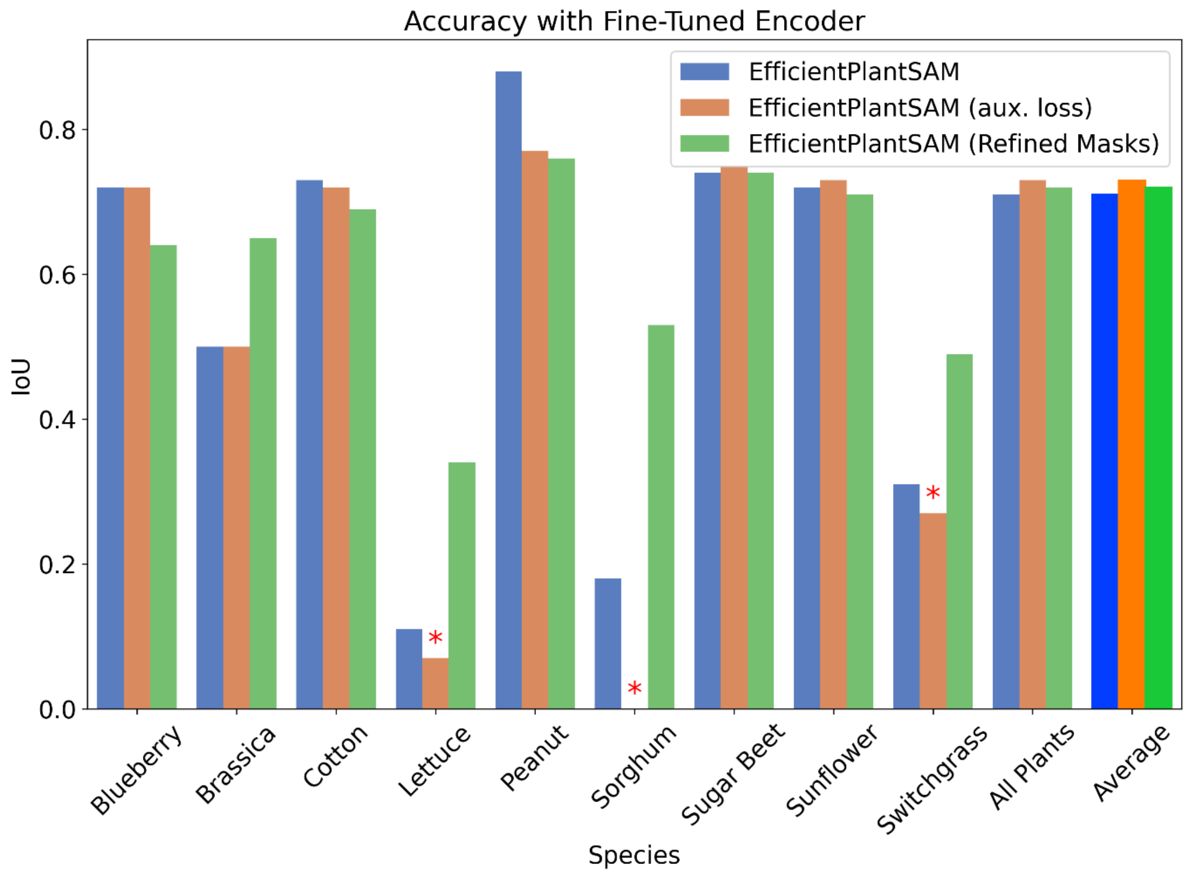

Figure 8: Accuracy of the models trained with an unfrozen image encoder. "Aux. loss" refers to the use of an auxiliary loss on the image embeddings. The red stars indicate species where the model performed unusually poorly.

Overall, the results suggest that mask refinement is a useful tool, but perhaps not applicable in all circumstances. In particular, some species seem to benefit from it, and some see no benefit or even a slight performance degradation. In the future, we can limit refinement to only the species that benefit. We can also investigate whether performing a second iteration of mask refinement provides any additional performance improvement.

Inference Speed

As expected, the inference speed of the vanilla SAM is limited by the size of the ViT-H encoder (Table 2). With TensorRT optimization, the baseline PlantSAM achieves an inference speed of around 14 FPS on an A100 GPU. By contrast, the EfficientSAM-based model runs at around 54 FPS on the same hardware. Even on the much less powerful Jetson AGX, the smaller model is still able to run at 12 FPS. This speed is likely to be sufficient for real-time applications in robotics, such as precision weeding and crop row following.

These experiments shouldn't, however, be taken as a guarantee of "real world" inference speed. Firstly, the implementation of the TensorRT conversion process was rather naïve; with more tweaking and configuration, or aggressive quantization, more favorable performance could possibly be achieved. Secondly, it should be noted that some applications of PlantSAM would require multiple invocations of parts of the model, which we did not test. For instance, if crops and weeds both need to be identified, the mask decoder would have to be invoked twice for each image with two different prompts: one to segment all the plants, and one to segment just the crop. (This could be done with a single invocation of the mask decoder with a batch size of two, which likely wouldn't carry much of a performance penalty on modern hardware.)

Table 2: Inference speeds of the Segment Anything Models when optimized with TensorRT.

Model	**Hardware**	**Speed (FPS)**
PlantSAM (ViT-H)	A100	14
EfficientPlantSAM-Ti	A100	54
EfficientPlantSAM-Ti	Jetson Orin AGX	12

Discussion

Many of the uses of SAM in agriculture appear to be centered around remote sensing, in particular, the detection of field boundaries (Kovačević et al., 2024; Ferreira et al., 2025; Rodriguez Sanchez et al., 2024). These approaches generally use SAM in AMG mode and perform some sort of post-processing. The end-goal of these approaches might be simply delineating field boundaries (Ferreira et al., 2025), or it might be identifying particular crops (Gurav et al., 2023). The overall AMG and post-processing pipeline is similar conceptually to the method we use for generating our dataset. It is also the same method used to generate a portion of the SA-1B dataset used for training the original SAM model (Kirillov et al., 2023). The results on remote sensing data tend to be good, but we suspect that this is because SAM does well when segmenting clearly delineated, relatively convex fields. We found that segmenting individual plants reliably required significant extensions to the basic SAM AMG approach.

In particular, it was observed that SAM can sometimes struggle with the fine structures and complex backgrounds inherent in the plant segmentation task. An attempt was made to ameliorate this through excess green filtering; however, this is not perfectly precise. As an alternative approach, the original SAM model could be replaced with High Quality SAM (Ke et al., 2023), which adapts the SAM model in order to increase the accuracy of complex masks. This could be a potential avenue for improving the generated mask quality of PlantSAM. HQ SAM has already been shown to provide some benefits when segmenting sugar beet images (Nguyen et al., 2023).

The experiments training the full EfficientSAM encoder suggest that decoder-only fine-tuning leaves some performance on the table. One way to alleviate this is to add an adapter module to SAM (Chen et al., 2023), which adds a small number of trainable parameters but allows the bulk of the model itself to remain frozen. This approach has specifically shown promise for challenging agricultural segmentation tasks (Li et al., 2023), and could be explored in the future as a way of enhancing PlantSAM.

PlantSAM is currently limited to semantic segmentation of entire plant species. In certain cases (e.g. stand counting), it is important to be able to distinguish individual instances, even when they overlap. This is one reason why such approaches typically use object detectors (Wang et al., 2021). Additionally, there are a wide variety of tasks in agricultural robotics, such as fruit counting, which require the ability to segment individual plant *organs* (Akiva et al., 2020; Bolouri et al., 2024). With that goal in mind, PlantSAM could perhaps be modified or employed as part of a larger pipeline to perform hierarchical segmentation (Nguyen et al., 2023).

There have been many proposals to reduce the size of the SAM model without sacrificing too much accuracy (Sun et al., 2024). The EfficientSAM-based (Xiong et al., 2024) approach that was used in this study provides a convenient balance of speed and accuracy, but the space of efficient SAM variants has not been fully explored. There remains the possibility that there could be something more suitable for this application. Choosing a proper distillation approach will likely be key to deploying PlantSAM in the field.

Conceptually, the proposed automated dataset generation approach is similar to the data generation strategy used in DepthCropSeg (Cao et al., 2025). However, the latter (as the name implies) relies on the DepthAnything-v2 (Yang et al., 2024) monocular depth estimation foundation model instead of SAM. Furthermore, despite being mostly automated, DepthCropSeg still requires human review of the generated masks, perhaps because there is not always a clear depth plane separation between the foreground crops and the background. Though DepthCropSeg does introduce a mask refinement technique that leverages a trained model to enhance the generated masks, their refinement approach is much simpler than ours and does not allow for the correction of false negatives in the original dataset. The overall idea of using depth cues for segmentation, however, is a promising one, and should be explored in the future, likely as an auxiliary input to PlantSAM in conjunction with RGB.

Conclusion

This study revealed that SAM can be modified to segment particular plant species. The PlantSAM model was able to differentiate between target crop species and weeds based on input text prompts. What is more, by introducing an automated mask generation pipeline that leveraged the pre-trained SAM and BioCLIP models, the proposed model achieved promising results without any manual data annotation. Because one ultimate goal of the study is to deploy PlantSAM on agricultural robots, the EfficientSAM encoder was leveraged in order to drastically reduce the size of the model, such that it can run in real time on an edge computer. There are still many possible avenues to explore that could improve the segmentation accuracy. In particular, certain failure patterns of the model could result from limitations of the dataset generation process. This approach represents a clear proof-of-concept of a general segmentation foundation model for precision agriculture.

References

Ahmadi, A., Halstead, M., & McCool, C. (2022, October). BonnBot-I: A Precise Weed Management and Crop Monitoring Platform. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (pp. 9202-9209). https://doi.org/10.1109/IROS47612.2022.9981304

Ahmadi, A., Halstead, M., Smitt, C., & McCool, C. (2024, July). BonnBot-I Plus: A Bio-Diversity Aware Precise Weed Management Robotic Platform. *IEEE Robotics and Automation Letters, 9*, 6560-6567. https://doi.org/10.1109/LRA.2024.3408080

Akiva, P., Dana, K., Oudemans, P., & Mars, M. (2020, June). Finding Berries: Segmentation and Counting of Cranberries using Point

Supervision and Shape Priors. arXiv:2004.08501. Retrieved from https://openaccess.thecvf.com/content_CVPRW_2020/html/w5/Akiva_Finding_Berries_Segmentation_and_Counting_of_Cranberries_Using_Point_Supervision_CVPRW_2020_paper.html

Bolouri, F., Kocoglu, Y., Pabuayon, I. L., Ritchie, G. L., & Sari-Sarraf, H. (2024). CottonSense: A high-throughput field phenotyping system for cotton fruit segmentation and enumeration on edge devices. *Computers and Electronics in Agriculture, 216*, 108531. https://doi.org/https://doi.org/10.1016/j.compag.2023.108531

Cao, S., Xu, B., Zhou, W., Zhou, L., Zhang, J., Zheng, Y., . . . Lu, H. (2025). The Blessing of Depth Anything: An Almost Unsupervised Approach to Crop Segmentation with Depth-Informed Pseudo Labeling. *Plant Phenomics*, 100005. https://doi.org/https://doi.org/10.1016/j.plaphe.2025.100005

Cerrato, S., Mazzia, V., Salvetti, F., Martini, M., Angarano, S., Navone, A., & Chiaberge, M. (2024). A Deep Learning Driven Algorithmic Pipeline for Autonomous Navigation in Row-Based Crops. *IEEE Access, 12*, 138306-138318. https://doi.org/10.1109/ACCESS.2024.3465873

Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., & Stachniss, C. (2017). Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The International Journal of Robotics Research, 36*, 1045-1052. https://doi.org/10.1177/0278364917720510

Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., . . . Mao, P. (2023, October). SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, (pp. 3367-3375).

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022, June). Masked-Attention Mask Transformer for Universal Image Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1290-1299). Retrieved from https://openaccess.thecvf.com/content/CVPR2022/html/Cheng_Masked-Attention_Mask_Transformer_for_Universal_Image_Segmentation_CVPR_2022_paper.html

Dang, F., Chen, D., Lu, Y., & Li, Z. (2023). YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems. *Computers and Electronics in Agriculture, 205*, 107655. https://doi.org/https://doi.org/10.1016/j.compag.2023.107655

de Silva, R., Cielniak, G., & Gao, J. (2024). Vision based crop row navigation under varying field conditions in arable fields. *Computers and Electronics in Agriculture, 217*, 108581. https://doi.org/https://doi.org/10.1016/j.compag.2023.108581

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*. Retrieved from https://arxiv.org/abs/2010.11929

Fawakherji, M., Potena, C., Pretto, A., Bloisi, D. D., & Nardi, D. (2021). Multi-Spectral Image Synthesis for Crop/Weed Segmentation in Precision Farming. *Robotics and Autonomous Systems, 146*, 103861. https://doi.org/https://doi.org/10.1016/j.robot.2021.103861

Ferreira, L. B., Martins, V. S., Aires, U. R., Wijewardane, N., Zhang, X., & Samiappan, S. (2025). FieldSeg: A scalable agricultural field extraction framework based on the Segment Anything Model and 10-m Sentinel-2 imagery. *Computers and Electronics in Agriculture, 232*, 110086. https://doi.org/https://doi.org/10.1016/j.compag.2025.110086

Gong, Z., Wang, A. T., Haurum, J. B., Lowe, S. C., Taylor, G. W., & Chang, A. X. (2024). BIOSCAN-CLIP: Bridging Vision and Genomics for Biodiversity Monitoring at Scale. *arXiv preprint*.

Gurav, R., Patel, H., Shang, Z., Eldawy, A., Chen, J., Scudiero, E., & Papalexakis, E. (2023). Can SAM recognize crops? Quantifying the zero-shot performance of a semantic segmentation foundation model on generating crop-type maps using satellite imagery for precision agriculture. *Can SAM recognize crops? Quantifying the zero-shot performance of a semantic segmentation foundation model on generating crop-type maps using satellite imagery for precision agriculture*. Retrieved from https://arxiv.org/abs/2311.15138

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *Distilling the Knowledge in a Neural Network*. Retrieved from https://arxiv.org/abs/1503.02531

Hu, N., Wang, S., Wang, X., Cai, Y., Su, D., Nyamsuren, P., . . . Wei, H. (2022). LettuceMOT: A dataset of lettuce detection and tracking with re-identification of re-occurred plants for agricultural robots. *Frontiers in Plant Science, 13*. https://doi.org/10.3389/fpls.2022.1047356

Ilyas, T., Arsa, D. M., Ahmad, K., Lee, J., Won, O., Lee, H., . . . Park, D. S. (2025). CWD30: A new benchmark dataset for crop weed recognition in precision agriculture. *Computers and Electronics in Agriculture, 229*, 109737. https://doi.org/https://doi.org/10.1016/j.compag.2024.109737

Ke, L., Ye, M., Danelljan, M., liu, Y., Tai, Y.-W., Tang, C.-K., & Yu, F. (2023). Segment Anything in High Quality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Ed.), *Advances in Neural Information Processing Systems. 36*, pp. 29914–29934. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/5f828e38160f31935cfe9f67503ad17c-Paper-Conference.pdf

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., . . . Girshick, R. (2023, October). Segment Anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (pp. 4015-4026). Retrieved from https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html

Kovačević, V., Pejak, B., & Marko, O. (2024, July). Enhancing Machine Learning Crop Classification Models through SAM-Based Field Delineation Based on Satellite Imagery. *2024 12th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, (pp. 1-4). https://doi.org/10.1109/Agro-Geoinformatics262780.2024.10661028

LeBauer, D., Burnette, M., Fahlgren, N., Kooper, R., McHenry, K., & Stylianou, A. (2021, October). What Does TERRA-REF's High Resolution, Multi Sensor Plant Sensing Public Domain Data Offer the Computer Vision Community? *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, (pp. 1409-1415). Retrieved from https://openaccess.thecvf.com/content/ICCV2021W/CVPPA/html/LeBauer_What_Does_TERRA-

REFs_High_Resolution_Multi_Sensor_Plant_Sensing_Public_ICCVW_2021_paper.html

Li, Y., Wang, D., Yuan, C., Li, H., & Hu, J. (2023). Enhancing Agricultural Image Segmentation with an Agricultural Segment Anything Model Adapter. *Sensors, 23*. https://doi.org/10.3390/s23187884

Li, Z., Xu, R., Li, C., Munoz, P., Takeda, F., & Leme, B. (2025). In-field blueberry fruit phenotyping with a MARS-PhenoBot and customized BerryNet. *Computers and Electronics in Agriculture, 232*, 110057. https://doi.org/https://doi.org/10.1016/j.compag.2025.110057

Long, J., Shelhamer, E., & Darrell, T. (2015, June). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Retrieved from https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

Madsen, S. L., Mathiassen, S. K., Dyrmann, M., Laursen, M. S., Paz, L.-C., & Jørgensen, R. N. (2020, April). Open Plant Phenotype Database of Common Weeds in Denmark. *Remote Sensing, 12*, 1246. https://doi.org/10.3390/RS12081246

Nguyen, K. D., Phung, T.-H., & Cao, H.-G. (2023). A SAM-based Solution for Hierarchical Panoptic Segmentation of Crops and Weeds Competition. *A SAM-based Solution for Hierarchical Panoptic Segmentation of Crops and Weeds Competition*. Retrieved from https://arxiv.org/abs/2309.13578

Olsen, A., Konovalov, D. A., Philippa, B., Ridd, P., Wood, J. C., Johns, J., . . . White, R. D. (2019). DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. *Scientific Reports, 9*, 2058. https://doi.org/10.1038/s41598-018-38343-3

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In M. Meila, & T. Zhang (Ed.), *Proceedings of the 38th International Conference on Machine Learning. 139*, pp. 8748–8763. PMLR. Retrieved from https://proceedings.mlr.press/v139/radford21a.html

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., . . . Wang, X. (2021, October). A Survey of Deep Active Learning. *ACM Comput. Surv., 54*. https://doi.org/10.1145/3472291

Rodriguez Sanchez, J., Li, C., & Johnsen, K. (2024). *Automated Unmanned Systems and Data Analytics for In-Field Digital Crop Phenotyping.* Ph.D. dissertation, University of Georgia, United States – Georgia. Retrieved from https://ufl-flvc.primo.exlibrisgroup.com/openurl/01FALSC_UFL/01FALSC_UFL:UFL?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations&sid=ProQ:ProQuest+Dissertations+%26+Theses+Global&atitle=&title=Automated+Unmanned+Systems+and+Data+Analytics+for+In-Field+Digital+Crop+Phenotyping&issn=&date=2024-01-01&volume=&issue=&spage=&au=Rodriguez+Sanchez%2C+Javier&isbn=9798384014850&jtitle=&btitle=&rft_id=info:eric/&rft_id=info:doi/

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Ed.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing. Retrieved from https://ui.adsabs.harvard.edu/abs/2015arXiv150504597R

Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., & Jacobs, N. (2024). TaxaBind: A Unified Embedding Space for Ecological Applications. *TaxaBind: A Unified Embedding Space for Ecological Applications*. Retrieved from https://arxiv.org/abs/2411.00683

Steininger, D., Trondl, A., Croonen, G., Simon, J., & Widhalm, V. (2023, January). The CropAndWeed Dataset: A Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (pp. 3729-3738). Retrieved from https://openaccess.thecvf.com/content/WACV2023/html/Steininger_The_CropAndWeed_Dataset_A_Multi-Modal_Learning_Approach_for_Efficient_Crop_WACV_2023_paper.html?utm_referrer=https://dzen.ru/media/id/5e048b1b2b616900b081f1d9/63baac0e87e9f80578545e0e

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., . . . Su, Y. (2023, June). BioCLIP: A Vision Foundation Model for the Tree of Life. *BioCLIP: A Vision Foundation Model for the Tree of Life*, 19412-19424.

Sun, X., Liu, J., Shen, H. T., Zhu, X., & Hu, P. (2024). On Efficient Variants of Segment Anything Model: A Survey. *On Efficient Variants of Segment Anything Model: A Survey*. Retrieved from https://arxiv.org/abs/2410.04960

Tan, C., Sun, J., Song, H., & Li, C. (2025). A customized density map model and segment anything model for cotton boll number, size, and yield prediction in aerial images. *Computers and Electronics in Agriculture, 232*, 110065. https://doi.org/https://doi.org/10.1016/j.compag.2025.110065

Teimouri, N., Dyrmann, M., Nielsen, P. R., Mathiassen, S. K., Somerville, G. J., & Jørgensen, R. N. (2018). Weed Growth Stage Estimator Using Deep Convolutional Neural Networks. *Sensors, 18*. Retrieved from http://www.mdpi.com/1424-8220/18/5/1580

Wang, L., Xiang, L., Tang, L., & Jiang, H. (2021). A Convolutional Neural Network-Based Method for Corn Stand Counting in the Field. *Sensors, 21*. https://doi.org/10.3390/s21020507

Weyler, J., Läbe, T., Magistri, F., Behley, J., & Stachniss, C. (2023, June). Towards Domain Generalization in Crop and Weed Segmentation for Precision Farming Robots. *IEEE Robotics and Automation Letters, 8*, 3310-3317. https://doi.org/10.1109/LRA.2023.3262417

Weyler, J., Magistri, F., Marks, E., Chong, Y. L., Sodano, M., Roggiolani, G., . . . Behley, J. (2023). PhenoBench — A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *arXiv preprint*. Retrieved from https://arxiv.org/pdf/2306.04557

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., . . . Chandra, V. (2024, June). EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 16111-16121). Retrieved from https://openaccess.thecvf.com/content/CVPR2024/html/Xiong_EfficientSAM_Leveraged_Masked_Image_Pretraining_for_Efficient_Segment_Anything_CVPR_2024_paper.html

Yang, C.-H., Feuer, B., Jubery, Z., Deng, Z. K., Nakkab, A., Hasan, M. Z., . . . Ganapathysubramanian, B. (2025). BioTrove: A Large Curated Image Dataset Enabling AI for Biodiversity. *BioTrove: A Large Curated Image Dataset Enabling AI for Biodiversity*. Retrieved from https://arxiv.org/abs/2406.17720

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth Anything V2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Ed.), *Advances in Neural Information Processing Systems. 37*, pp. 21875–21911. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2024/file/26cfdcd8fe6fd75cc53e92963a656c58-Paper-Conference.pdf