Is Segment Anything Model a revolution in precision agriculture?

Alberto Carraro[1] and Francesco Marinello[1]

TESAF Department, University of Padova

Abstract. Precision agriculture uses accurate identification and mapping of crop features by automated mechanisms. The use of computer vision techniques implemented by supervised deep learning systems to solve many precision agricultural problems necessitates large-scale data collection and prolonged ground truth annotation by humans. The so-called foundation models in artificial intelligence are therefore becoming more and more significant. Meta is working on a project called Segment Anything to provide a base model for image segmentation. Without extra training, it can accomplish zero-shot generalization to strange objects and images. This study evaluates the performance of Meta's Segment Anything Model (SAM) for the problem of semantic segmentation of objects in the context of precision agriculture.

Keywords: Precision agriculture · Meta's Segment Anything Model (SAM) · Image segmentation · Computer vision

1 Introduction

The main points that the reader should get from this paper are:

- There are large labelled datasets for computer vision tasks but they contain images of easily recognizable subjects (animals, cars, roads, etc.)
- Scientists can take open source architectures pre-trained on those large datasets but then they need to adapt them to their specific case, with additional ground-truth labelling (that requires experts) and additional training
- SAM requires no additional training and the authors claim it has learned the very concept of "object", disregarding any specific semantic context: how well can it perform on non-ordinary images?

Bisogna prendere spunto da

- https://www.sciencedirect.com/science/article/pii/S2214317323000112
- https://www.sciencedirect.com/science/article/pii/S0168169921002593/pdf

Precision agriculture is a broad field that aims to improve agricultural practices' efficiency, productivity, and sustainability through science and technology. One of the critical challenges in precision agriculture is accurately identifying and mapping crop features and conditions such as diseases, plant height, leaf area index,

and crop growth stage. Computer vision-based approaches have driven innovation in the agricultural industry and are nowadays used in scenarios such as pest identification [40], precision livestock farming [26], and many more. Machine vision technologies support farmers and industries, saving time and costs while reaching very good efficacy.

In computer vision, there are four primary tasks that can be combined and handled in different ways. One is image categorization, which arranges images according to a limited number of classes. Another is object detection, which creates bounding boxes, or tiny rectangles, around particular items in photos that fall into specified categories. Semantic segmentation seeks to categorize areas inside the images that have been labeled at the pixel level based on texture, color, and spatial distribution. Instance Segmentation, which combines object recognition and semantic segmentation, recognizes the various instances presented in the image together with their borders at the pixel level.

Since early applications of semantic segmentation, this task has been implemented in various ways. Depending on the lighting conditions and the sharpness of edges, various techniques have been adopted, such as colour space conversion and combination of colour channels [30]. Other implementations use machine-learning-based classification techniques such as decision trees [39] and clustering [41]. Deep Learning is a subset of Machine Learning, where the primary tools are Deep Neural Networks. The use of Deep Neural Networks for the solution of computer vision tasks reached enormous diffusion and success also due to the public availability of models like VGG [32], U-Net [31], SegNet [2], DeepLab [5]. These are Deep Neural Nets of various kinds that have been trained over large datasets. Table 1 summarizes the main deep learning models and the datasets they were trained on, together with the task they are meant to perform.

Open-source deep learning models have revolutionized the field of artificial intelligence by providing accessible and adaptable solutions for various tasks, allowing researchers and developers to utilize them for their own projects. A particularly successful principle in precision agriculture is that of Transfer Learning [45], offering significant advantages in developing robust and accurate models for various agricultural tasks. By leveraging pre-trained deep learning models on large-scale general image datasets, transfer learning enables the transfer of knowledge from the source domain to the target agricultural domain. This approach allows researchers to overcome challenges associated with limited annotated agricultural datasets and improve model performance by initializing the network with learned features. Several studies have demonstrated the effectiveness of transfer learning in precision agriculture tasks, such as crop disease detection [25], crop type mapping [24], and yield prediction [37]. Transfer learning not only accelerates model training but also enhances the generalization and adaptability of the models to different agricultural environments, ultimately contributing to improved crop management and increased agricultural productivity.

Transfer learning still requires some ground-truth labelling and additional training in order to specialize the chosen Deep Learning model for the domain at hand. For this last purpose there are dataset specialised on the agricultural domain like

Model	Dataset used for training	Computer Vision Task
AlexNet [17]	ImageNet [9] (14,197,122 images)	Image Classification
VGGNet [32]	ImageNet	Semantic Image Segmentation
ResNet [13]	ImageNet	Image Classification, Object Detection
U-Net [31]	Various medical imaging datasets: - MICCAI [18] [22], - ISIC [6]	Semantic Image Segmentation
DeepLab [5]	- PASCAL VOC [11], - Cityscapes [7] (5,000 images), - ADE20K [44] (20,210 images), - COCO [20] (118,000 images)	Semantic Image Segmentation
YOLO [28]	COCO	Real-Time Object Detection
SegNet [2]	- CamVid [3], - Cityscapes, - SUN RGB-D [33] (10,335 images), - ADE20K	Semantic Image Segmentation
EfficientNet [34]	ImageNet	Semantic Image Segmentation

Table 1: Most popular open source deep learning models and datasets used in computer vision tasks.

the PlantVillage [23] project containing 54,306 images of 14 crop species with 26 diseases (or healthy) made openly available.

There are also large-scale datasets commonly used for semantic segmentation in precision agriculture. Some of the largest and widely used datasets are:

1. Plant Phenotyping Datasets: These datasets focus on crop and plant analysis, providing pixel-level annotations for various plant structures.
   - PlantCLEF [19]: A collection of plant images with annotations for leaf, stem, and flower structures.
   - Plant Phenotyping Dataset [42]: Contains images of Arabidopsis plants with annotations for leaves and other plant components.
2. Crop Field Datasets: These datasets consist of aerial or satellite images of crop fields with annotations for different crops or objects of interest.
   - CropSeg [21]: A dataset with high-resolution aerial images of different crops, annotated at the pixel level.
   - Crop DeepLab Dataset [43]: Contains crop field images with fine-grained annotations for different crop types and objects.
3. Aerial Imagery Datasets: These datasets focus on aerial images captured by drones or satellites, providing annotations for various objects or classes of interest.
   - ISPRS 2D Semantic Labeling [29]: A large-scale dataset with high-resolution aerial images and annotations for building, road, and vegetation classes.
   - DeepGlobe Land Cover Classification [8]: Consists of satellite imagery with annotations for land cover classes, including agricultural areas.

These datasets offer substantial amounts of labeled data for semantic segmentation tasks in precision agriculture. Researchers often utilize them to develop and evaluate models for crop analysis, disease detection, weed detection, and other applications in precision agriculture.

Recently, segmentation foundation models have seen tremendous advancements in the field of natural image segmentation [38][46], enabling accurate and efficient segmentation of objects in a fully automatic or interactive way. These models are typically based on transformer architectures and leverage pre-trained weights to achieve state-of-the-art performance and unprecedented generalisation ability on a wide range of natural images. Among these the Segment Anything project [15] by Meta® is a task, a model, and a dataset for image segmentation. The Segment Aanything Model (SAM) in particular has learned a general notion of what objects are and this understanding enables *zero-shot* generalization to unfamiliar objects and images without requiring additional training. SAM has been trained on the SA-1B dataset [14], the largest segmentation dataset to date, with over 1 billion masks on 11 million images.

SAM is designed and trained to be promptable, so its segmentation capabilities can be extended and transferred to new image distributions and tasks. We evaluate its capabilities on the specific tasks of segmenting objects inside the pictures of weed and crop of the CWFID dataset [12].

2 Materials and methods

Our goal is to compare the performances of the foundational segmentation model SAM with no training. The chosen task is semantic segmentation i.e. to classify the pixels of each image into two different semantic categories: foliage (foreground) and background. This task could be an important first part in a Machine Learing pipeline whose later stages focus on the analysis of the Region of Interest inside the pictures, which in this case we considered to be the area occupied by foliage. Figure 1 illustrates what we considered as foreground: all entities that were at the distance of within 1m from the camera.

2.1 Segment Anything

The Segment Anything Model (SAM) utilises a transformer-based architecture [36], which has been shown to be highly effective in natural language processing [4] and image recognition tasks [10]. Specifically, SAM uses a vision transformer-based **image encoder** to extract image features and **prompt encoders** to incorporate user interactions, followed by a **mask decoder** to generate segmentation results and confidence scores based on the image embedding, prompt embedding, and output token.

The prompt encoders are tailored for different user inputs. SAM supports four different prompts: points, boxes, texts, and masks. Each point is encoded by Fourier positional encoding [35] and two learnable tokens for specifying foreground and background, respectively. The bounding box is encoded by the point encoding of

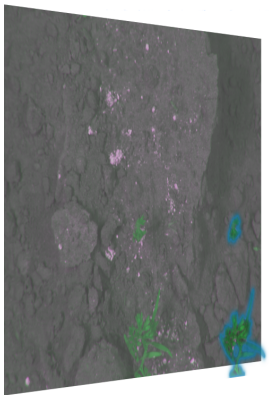


Fig. 1: Foreground vs background.

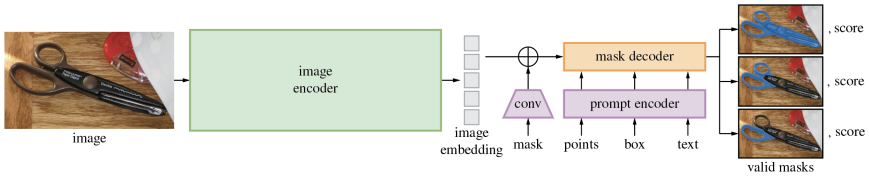
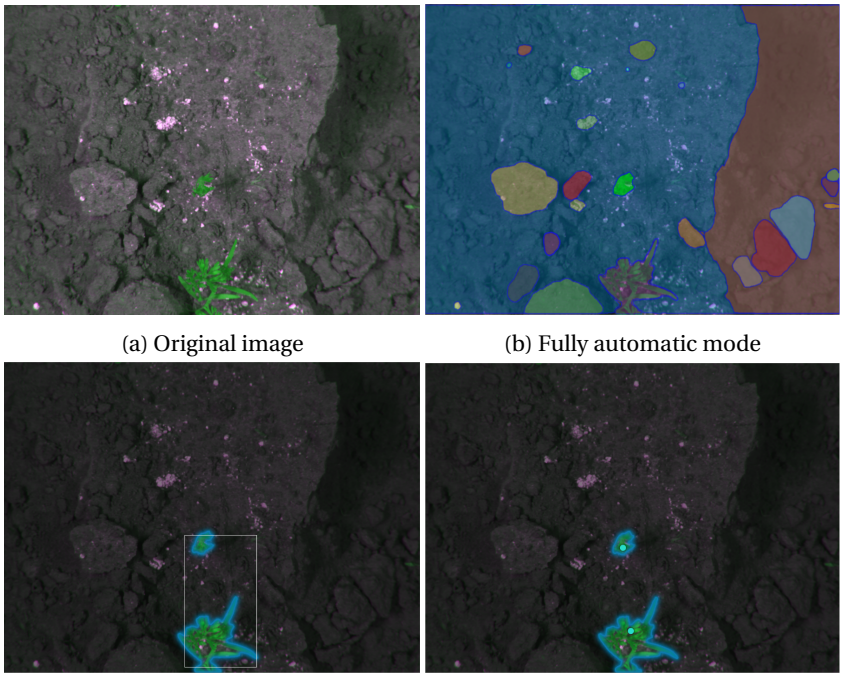


Fig. 2: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks.

its top-left corner and bottom-right corner. The free-form text is encoded by the pre-trained text-encoder in CLIP [27]. The mask prompt has the same spatial resolution as the input image, which is encoded by convolution feature maps. Finally, the mask decoder employs a lightweight design, which consists of two transformer layers with a dynamic mask prediction head and an Intersection-over-Union (IoU) score regression head. The mask prediction head can generate three $4\times$ downscaled masks, which correspond to the whole object, part, and subpart of the object, respectively.

To summarize, when applying SAM for agricultural image segmentation, the segment-everything mode is prone to generate useless region partitions and the point-based mode is ambiguous and requires multiple prediction-correction iterations. In contrast, the bounding box-based mode can clearly specify the ROI and obtain reasonable segmentation results without multiple trials and errors. We argue that the bounding box-based segmentation mode has wider practical values than the segment-everything and point-based mode when using SAM in medical image segmentation tasks.

Since the image encoder can be applied prior to prompting the model, we can pre-compute the image embedding for all training images to avoid replicated computing of the image embedding per prompt, which can significantly improve the



(a) Original image (b) Fully automatic mode

(c) Bounding box mode (d) Additional point modality

Fig. 3: Segmentation results of SAM based on different segmentation modes.

training efficiency. The mask decoder only needs to generate one mask rather than three masks because the bounding box prompt can clearly specify the expected segmentation target in most situations.

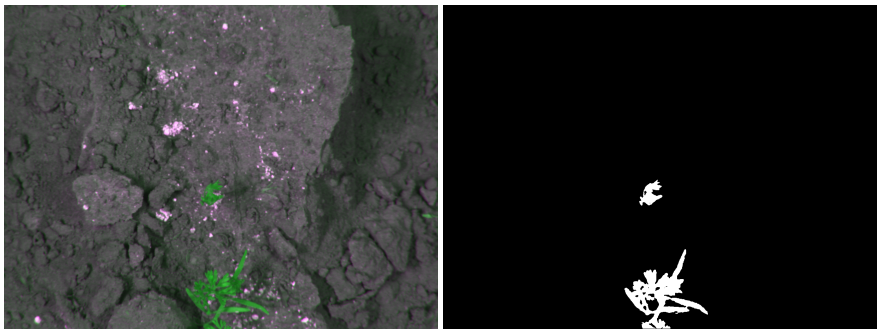
3 Experiments and Results

The CWFID dataset [12] contains 60 RGB images in png format taken at a working distance of approximately 30 cm from plants .....

Deep neural networks require large amounts of training data to tune millions of parameters and develop a learned model for subsequent predictions. While it is possible to crowdsource image annotation for natural scenes and collect large datsets as shown in Table 1, it is difficult to find experts to annotate images for specific scientific domains. Each research topic requires researchers to annotate their own images. Annotation being a very time consuming task, a researcher can end up with a handful of annotated images containing only tens of labeled objects.

Our goal was to evalluate the performances of the foundational segmentation model SAM with no training. The chosen task is semantic segmentation i.e. to classify the pixels of each image into two different semantic categories: plants and background. This task could be an important first part in a Machine Learing pipeline

whose later stages focus on the analysis of the Region of Interest inside the pictures, which in this case we considered to be the area occupied by plants.



(a) (b)

Fig. 4: Ground truth segmentation of foliage (hand-made mask) for the file 050_image.png

The 60 image dataset was captured at a commercial organic carrot farm in Northern Germany in 2013 just before manual weed control was applied. The carrots were grown in single rows on small soil dams.

The 60 images have been manually annotated by the authors of the CWFID dataset ..... The result of the manual segmentation are a total of 60 ".yaml" files containing the coordinates of contour points that define masks delineating weed and crop inside each image.....

In Table **??** we recorded the time used to manually draw precise segmentation masks for 60 images. We keep the so-obtained 60 masks as a reference for the rest of the paper and consider them as ground-truth to measure performance of the two segmentation tools analyzed. In 3 we summarized the classification of pixel predictions in the context of semantic segmentation applied to the task of partitioning an image into its Region Of Interest (ROI) and its background.Frase contorta.

3.1 Segmentation of plant images with SAM

SAM supports three main segmentation modes: segmenting everything in a fully automatic way, bounding box mode, and point mode. The segment-everything mode divides the whole image into six regions based on the image intensity (Fig. 3b).

We segmented 10 images with bounding box mode and other 10 images with point mode. These modes are semi-automatic, i.e. they require a human input for each image.

Figure 3 shows the results of the three segmentation modes on an image of vines from the ESCA dataset [1] and Table **??** reports the exact time necessary to manually draw segmentation masks for the 10 selected images [1].

[1] The results are based on the online demo: `https://segment-anything.com/demo`.

Image name	Number of points on the border	Manual annotation time (seconds)
001_annotation.yaml	233	0
002_annotation.yaml	106	0
003_annotation.yaml	54	0
004_annotation.yaml	124	0
005_annotation.yaml	41	0
006_annotation.yaml	95	0
007_annotation.yaml	217	0
008_annotation.yaml	108	0
009_annotation.yaml	127	0
010_annotation.yaml	106	0
050_annotation.yaml	18	0
051_annotation.yaml	150	0
052_annotation.yaml	155	0
053_annotation.yaml	55	0
054_annotation.yaml	115	0
055_annotation.yaml	105	0
056_annotation.yaml	123	0
057_annotation.yaml	148	0
058_annotation.yaml	209	0
059_annotation.yaml	204	0
060_annotation.yaml	171	0
	Sum	**Sum**
	5904	0
	Average	**Average**
	98.4	0

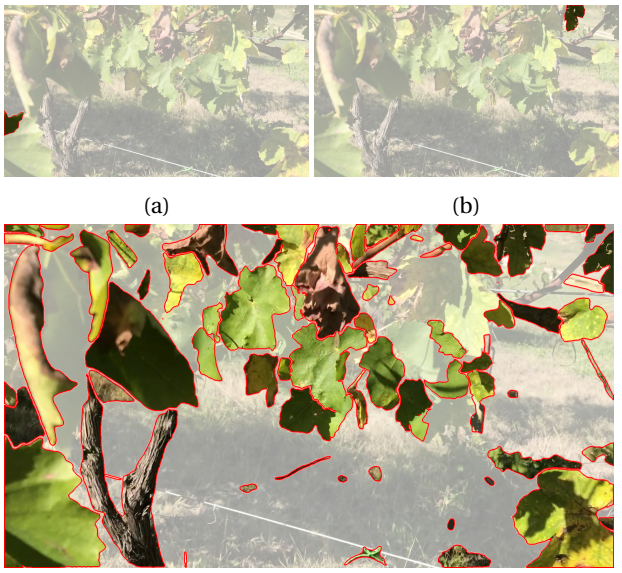
Table 2: Time spent for drawing the masks manually (ground-truth labelling).

Correct predictions	Incorrect predictions
True Positive (TP): the pixel has been predicted as part of the ROI and actually belongs to the ROI	False Positive (FP): the pixel has been predicted as part of the ROI but does not belong to the ROI
True Positive (TN): the pixel has been predicted as part of the background and actually belongs to the background	True Positive (FN): the pixel has been predicted as part of the background but actually belongs to the ROI

Table 3: Definitions of TP, TN, FP, and FN

We launched a batch segmentation job exploiting the fully-automatic mode on the entire set of 1770 images of the Esca dataset, using as hardware platform an NVIDIA RTX™ A6000 GPU. For the instantiation of the model we adopted the `ViT-H`[16] checkpoint. SAM produced for each segmented image a folder containing one mask for each object found in the image (between 100 and 200 objects). In order to pro-

cess 888 images of the esca folder took 55 minutes and 13.8 seconds. In order to process 882 images of the healthy folder took 69 minutes and 19.4 seconds to complete the merge of masks took 48 minutes and 45.4 seconds to complete.



(a) (b)

(c)

Fig. 5: Two masks produced by SAM in automatic mode on healthy1.jpg and the union of all masks produced for the same image.

4 Discussion and Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nos-

trud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Acknowledgements The authors of this paper highly appreciate all the challenge organizers and owners for providing the public dataset to the community. We also thank Meta AI for making the source code of segment anything publicly available to the community.

References

1. Alessandrini, M., Calero Fuentes Rivera, R., Falaschetti, L., Pau, D., Tomaselli, V., Turchetti, C.: Esca-dataset. Mendeley Data, V1, https://data.mendeley.com/datasets/89cnxc58kj/1 (2021). https://doi.org/doi:10.17632/89cnxc58kj.1
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1843–1851 (2016)
3. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. In: IEEE conference on computer vision and pattern recognition. pp. 2366–2373. IEEE (2009)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
5. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: 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)
6. Codella, N.C., Rotemberg, V., Tschandl, P., Visconti, A., Helba, B., Sinz, C., Celebi, M.E., Dusza, S., Gutman, D., Halpern, A., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). IEEE transactions on medical imaging **38**(2), 285–295 (2019)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
8. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., Kress, W., et al.: Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 172–173 (2018)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition pp. 248–255 (2009)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
12. Haug, S., Ostermann, J.: A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In: Computer Vision - ECCV 2014 Workshops. pp. 105–116 (2015). https://doi.org/10.1007/978-3-319-16220-1_8, http://dx.doi.org/10.1007/978-3-319-16220-1_8
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Available at: https://ai.facebook.com/datasets/segment-anything-downloads/

15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Vit-h sam model (May 2023), available at: `https://dl.fbaipublicfiles.com/segment_anything/sam_vit_h_4b8939.pth` (May. 2023)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2012)
18. Landman, B.A., Warfield, S.K.: Miccai 2012 grand challenge and workshop on multi-atlas labeling. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. pp. 451–460. Springer (2012)
19. Ligterink, W., Müller, H., Bonnet, P., Moulin, C., Joly, A.: Plantclef: The fine-grained visual classification of plant species. CLEF (2013)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. European conference on computer vision pp. 740–755 (2014)
21. Liu, K., Wang, Y., Sun, K., Cao, J.: Cropseg: A cropland segmentation dataset and benchmark. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 2536–2540. IEEE (2019)
22. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 317–329. Springer (2015)
23. Mohanty, S.P., Hughes, D.P., Salathé, M.: Plantvillage dataset: A publicly available dataset for deep learning in agriculture. `https://doi.org/10.5281/zenodo.235808` (2016), accessed: 2023-05-17
24. Nowakowski, A., Mrziglod, J., Spiller, D., Bonifacio, R., Ferrari, I., Mathieu, P.P., Garcia-Herranz, M., Kim, D.H.: Crop type mapping by using transfer learning. International Journal of Applied Earth Observation and Geoinformation **98** (2021). `https://doi.org/https://doi.org/10.1016/j.jag.2021.102313`, `https://www.sciencedirect.com/science/article/pii/S0303243421000209`
25. Paymode, A.S., Malode, V.B.: Transfer Learning for Multi-Crop Leaf Disease Image Classification using Convolutional Neural Network VGG. Artificial Intelligence in Agriculture **6**, 23–33 (2022). `https://doi.org/https://doi.org/10.1016/j.aiia.2021.12.002`, `https://www.sciencedirect.com/science/article/pii/S2589721721000416`
26. Qiao, Y., Truman, M., Sukkarieh, S.: Cattle segmentation and contour extraction based on mask r-cnn for precision livestock farming. Computers and Electronics in Agriculture **165**, 104958 (2019)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 779–788 (2016)
29. Richter, S., Vineet, V., Roth, S., Koltun, V.: The isprs 2d semantic labeling contest. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 111–117 (2016)
30. Riehle, D., Reiser, D., Griepentrog, H.W.: Robust index-based semantic plant/background segmentation for rgb- images. Computers and Electronics in

Agriculture **169**, 105201 (2020). https://doi.org/https://doi.org/10.1016/j.compag.2019.105201, https://www.sciencedirect.com/science/article/pii/S0168169919314346
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
33. Song, S., Lichtenberg, S., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: IEEE conference on computer vision and pattern recognition. pp. 567–576. IEEE (2015)
34. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019)
35. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems **33**, 7537–7547 (2020)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural Information Processing Systems **30** (2017)
37. Wang, A.X., Tran, C., Desai, N., Lobell, D., Ermon, S.: Deep transfer learning for crop yield prediction with remote sensing data. In: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. COMPASS '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209811.3212707, https://doi.org/10.1145/3209811.3212707
38. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
39. Yang, W., Wang, S., Zhao, X., Zhang, J., Feng, J.: Greenness identification based on hsv decision tree. Information Processing in Agriculture **2**(3), 149–160 (2015). https://doi.org/https://doi.org/10.1016/j.inpa.2015.07.003, https://www.sciencedirect.com/science/article/pii/S2214317315000347
40. Yuan, Y., Chen, L., Wu, H., Li, L.: Advanced agricultural disease image recognition technologies: A review. Information Processing in Agriculture **9**(1), 48–59 (2022). https://doi.org/https://doi.org/10.1016/j.inpa.2021.01.003, https://www.sciencedirect.com/science/article/pii/S2214317321000032
41. Zhang, H., Peng, Q.: Pso and k-means-based semantic segmentation toward agricultural products. Future Generation Computer Systems **126**, 82–87 (2022). https://doi.org/https://doi.org/10.1016/j.future.2021.06.059, https://www.sciencedirect.com/science/article/pii/S0167739X21002545
42. Zhang, H., Yang, M., Wang, Y., Wang, H., Ma, T., Li, W., Xia, S., Liu, Y.: Plant phenotyping datasets for computer vision. Data **4**(2), 36 (2019)
43. Zhang, S., Tang, H., Zhang, X., Liu, J., Zhang, J., Zhang, J., Lu, H.: Crop deeplab: Large-scale crop field parsing from satellite imagery. Remote Sensing **13**(8), 1460 (2021)
44. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
45. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning (2020)

46. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)

Measure name	Formula	Description
Intersection over Union (IoU), or Jaccard Index	$\frac{\lvert M \cap G \rvert}{\lvert M \cup G \rvert}$	measures the overlap between the predicted segmentation mask and the ground truth mask. It is calculated as the ratio of the intersection area to the union area of the two masks
Dice Coefficient (DC)	$\frac{2\lvert M \cap G \rvert}{\lvert M \rvert + \lvert G \rvert}$	measures the similarity between the predicted boundaries and the ground truth boundaries. It calculates the ratio of twice the intersection of the boundaries to the sum of their areas
Pixel Accuracy (PA)	$\frac{TP+TN}{TP+TN+FP+FN}$	measures the percentage of correctly classified pixels compared to the total number of pixels in the image. It calculates the overall accuracy of the segmentation
Precision (P)	$\frac{TP}{TP+FP}$	measures the proportion of positive predictions relative to total positive predictions
Recall (R)	$\frac{TP}{TP+FN}$	measures the proportion of correct positive predictions relative to total actual positives
F1-score	$\frac{2 \cdot P \cdot R}{P+R}$	combines precision and recall using their harmonic mean; meaningful when there is imbalance between positives and negtives in the dataset.
Hausdorff Distance	$\max\left\{\max_{x \in \partial G} \mathrm{d}(x, \partial M), \max_{y \in \partial M} \mathrm{d}(y, \partial G)\right\}$	measures the maximum distance (in millimeters) between the predicted contour and the ground truth contour. It quantifies the largest discrepancy between the two contours and is sensitive to outliers or large deviations
Average Symmetric Surface Distance (ASSD)	$\frac{\sum_{x \in \partial G} \mathrm{d}(x, \partial M) + \sum_{y \in \partial M} \mathrm{d}(y, \partial G)}{\lvert \partial G \rvert + \lvert \partial M \rvert}$	measures the average distance between the predicted contour and the ground truth contour. Unlike the Hausdorff distance, it takes into account both the under-segmentation (missing parts of the object) and over-segmentation (extraneous regions

Table 4: Definitions of each similarity / dissimilarity measure used in this paper.

Annotation file	# points inside mask	# points outside mask	Annotation time point mode (seconds)
001_annotation.json	41	20	101.696
002_annotation.json	35	21	95.565
003_annotation.json	32	22	106.857
004_annotation.json	46	29	125.596
005_annotation.json	43	49	148.657
006_annotation.json	53	71	210.004
007_annotation.json	72	63	201.255
008_annotation.json	58	68	186.949
009_annotation.json	59	81	193.657
010_annotation.json	51	68	197.857
050_annotation.json	19	138	98.891
051_annotation.json	118	176	388.167
052_annotation.json	98	171	276.081
053_annotation.json	32	140	114.474
054_annotation.json	67	120	254.87
055_annotation.json	75	138	337.862
056_annotation.json	60	180	338.139
057_annotation.json	77	167	293.152
058_annotation.json	95	184	310.982
059_annotation.json	79	166	338.914
060_annotation.json	94	187	276.054
	Sum	**Sum**	**Sum**
	3482	6268	12239.163
	Average	**Average**	**Average**
	58.033	104.467	203.986

Table 5: Time spent to place points for SAM predictor.

		Image segmentation time (minutes)	
Image folder	**Number of images**	**Object segmentation (automatic mode)**	**Object masks merging**
Esca	888	55’ and 13.8”	???
Healthy	882	69” and 19.4”	48’ and 45.4”

Table 6: Image segmentation time with SAM in automatic batch mode.

Image name	IoU	DC	PA	Prec	Rec	F1	ASSD (mm)	HD (mm)
001_image.png	0.804	0.892	0.962	0.157	0.887	0.266	6.456	328.868
002_image.png	0.779	0.876	0.971	0.101	0.89	0.182	5.887	262.574
003_image.png	0.829	0.907	0.985	0.072	0.872	0.134	6.274	329.474
004_image.png	0.804	0.892	0.976	0.098	0.857	0.175	2.832	147.872
005_image.png	0.861	0.926	0.991	0.054	0.931	0.102	3.275	237.306
006_image.png	0.769	0.87	0.976	0.081	0.806	0.147	2.999	127.44
007_image.png	0.819	0.9	0.968	0.145	0.896	0.25	2.595	170.88
008_image.png	0.81	0.895	0.979	0.089	0.923	0.162	3.451	164.657
009_image.png	0.815	0.898	0.981	0.081	0.933	0.15	5.787	361.334
010_image.png	0.792	0.884	0.983	0.064	0.939	0.12	5.45	311.045
050_image.png	0.276	0.433	0.966	0.013	0.805	0.025	150.533	588.205
051_image.png	0.846	0.916	0.975	0.138	0.912	0.24	6.189	499.401
052_image.png	0.803	0.89	0.974	0.106	0.928	0.19	11.023	485.956
053_image.png	0.722	0.839	0.978	0.056	0.944	0.106	32.785	456.054
054_image.png	0.89	0.942	0.988	0.099	0.919	0.178	2.697	469.998
055_image.png	0.802	0.89	0.979	0.085	0.884	0.155	11.072	358.236
056_image.png	0.789	0.882	0.972	0.103	0.882	0.184	7.121	238.481
057_image.png	0.81	0.895	0.977	0.096	0.905	0.174	6.389	218.009
058_image.png	0.867	0.929	0.98	0.132	0.9	0.231	4.626	173.807
059_image.png	0.756	0.861	0.967	0.101	0.88	0.181	7.385	252.634
060_image.png	0.858	0.924	0.983	0.102	0.924	0.184	5.307	319.265
	Avg.	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**
	.....	.....	.....	.....	.....	.....	.....	.....

Table 7: Evaluation of semantic segmentation performed by SAM in predictor mode with input points.

Image name	IoU	DC	PA	Prec	Rec	F1	ASSD (mm)	HD (mm)
001_image.png	0.285	0.444	0.744	0.102	0.579	0.174	29.965	384.033
002_image.png	0.167	0.286	0.666	0.067	0.588	0.12	38.636	274.39
003_image.png	0.234	0.379	0.757	0.074	0.891	0.137	95.89	638.719
004_image.png	0.186	0.314	0.716	0.065	0.571	0.117	41.91	326.686
005_image.png	0.13	0.23	0.694	0.046	0.786	0.086	120.494	597.759
006_image.png	0.168	0.287	0.721	0.056	0.56	0.102	77.322	518.383
007_image.png	0.169	0.29	0.725	0.056	0.346	0.097	50.349	407.836
008_image.png	0.11	0.198	0.729	0.034	0.35	0.061	55.287	399.606
009_image.png	0.129	0.229	0.724	0.041	0.47	0.075	79.684	612.148
010_image.png	0.046	0.088	0.67	0.016	0.232	0.03	112.531	615.325
050_image.png	0.038	0.073	0.61	0.015	0.973	0.03	238.893	790.864
051_image.png	0.15	0.262	0.52	0.085	0.562	0.148	31.957	512.533
052_image.png	0.315	0.479	0.823	0.081	0.712	0.146	59.305	568.401
053_image.png	0.14	0.245	0.646	0.057	0.964	0.108	103.668	558.358
054_image.png	0.238	0.384	0.722	0.087	0.806	0.157	107.342	545.388
055_image.png	0.156	0.27	0.656	0.064	0.662	0.116	84.925	537.276
056_image.png	0.24	0.387	0.76	0.076	0.65	0.136	91.175	519.773
057_image.png	0.184	0.31	0.749	0.057	0.533	0.102	60.21	326.267
058_image.png	0.293	0.453	0.725	0.114	0.774	0.199	40.258	379.606
059_image.png	0.205	0.341	0.729	0.07	0.609	0.125	61.843	513.448
060_image.png	0.268	0.423	0.758	0.088	0.801	0.159	40.327	354.193
	Avg.	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**
	.....	.....	.....	.....	.....	.....	.....	.....

Table 8: Evaluation of semantic segmentation performed by SAM in fully automatic mode.