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. Many challenges in precision agriculture are solved with computer vision techniques implemented by supervised deep learning systems, requiring extensive data collection and extended time for ground truth annotation by humans. For this reason, the so-called *foundation* models in Artificial Intelligence are becoming increasingly important. Segment Anything is a project by Meta to build a starting point for foundation models for image segmentation. It can achieve zero-shot generalisation to unfamiliar objects and images without additional training. This paper analyses the application of Meta's Segment Anything Model (SAM) to the problem of semantic segmentation in the field of precision agriculture and compares its performances to a pre-existing deep learning model called U-Net.

Keywords: Precision agriculture · Meta's Segment Anything Model (SAM) · U-Net · Image segmentation · Computer vision · Disease detection

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?

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 [44], precision livestock farming [28], and many more. Machine vision technologies support farmers and industries, saving time and costs while reaching very good efficacy.

There are four main tasks in computer vision, applied in several combinations and variations. One is image classification, which maps images onto a finite set of predefined labels called classes. Another one is object detection, which computes minimal enclosing rectangles, called bounding boxes, around specific objects within the images belonging to predetermined categories. Semantic Segmentation aims at classifying pixel-level labelling of areas inside the pictures based on texture, colour and spatial distribution. Finally, Instance Segmentation recognises the different instances given in the image with their boundaries at the pixel level, representing a combination of object detection and semantic segmentation.

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 [32]. Other implementations use machine-learning-based classification techniques such as decision trees [43] and clustering [46]. 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 [34], U-Net [33], 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.

Model	**Dataset used for training**	**Computer Vision Task**
AlexNet [16]	ImageNet [9] (14,197,122 images)	Image Classification
VGGNet [34]	ImageNet	Semantic Image Segmentation
ResNet [12]	ImageNet	Image Classification, Object Detection
U-Net [33]	Various medical imaging datasets: - MICCAI [17] [24], - ISIC [6]	Semantic Image Segmentation
DeepLab [5]	- PASCAL VOC [11], - Cityscapes [7] (5,000 images), - ADE20K [49] (20,210 images), - COCO [20] (118,000 images)	Semantic Image Segmentation
YOLO [30]	COCO	Real-Time Object Detection
SegNet [2]	- CamVid [3], - Cityscapes, - SUN RGB-D [36] (10,335 images), - ADE20K	Semantic Image Segmentation
EfficientNet [37]	ImageNet	Semantic Image Segmentation

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

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 [50], 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 [27], crop type mapping [26], and yield prediction [40]. 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 the PlantVillage [25] 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 [47]: 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 [48]: 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 [31]: 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 [42][51], 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 (SA) project [14] 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 [13], 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 vines of the ESCA dataset [1].

2 Materials and methods

Our goal is to compare the performances of the foundational segmentation model SAM with no training to the previously available segmentation model U-Net, which requires manual annotation. 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 [39], 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 [38] and two learnable tokens for specifying foreground and background, respectively. The bounding box is encoded by the point encoding of its top-left corner and bottom-right corner. The free-form text is encoded by the pre-trained text-encoder in CLIP [29]. The mask prompt has the same spatial resolution as the input image, which is encoded by convolution feature maps. Finally,

Fig. 1: Foreground vs background.

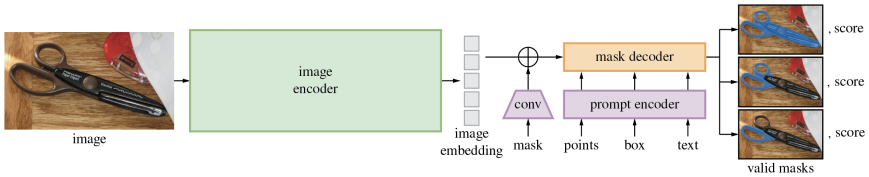
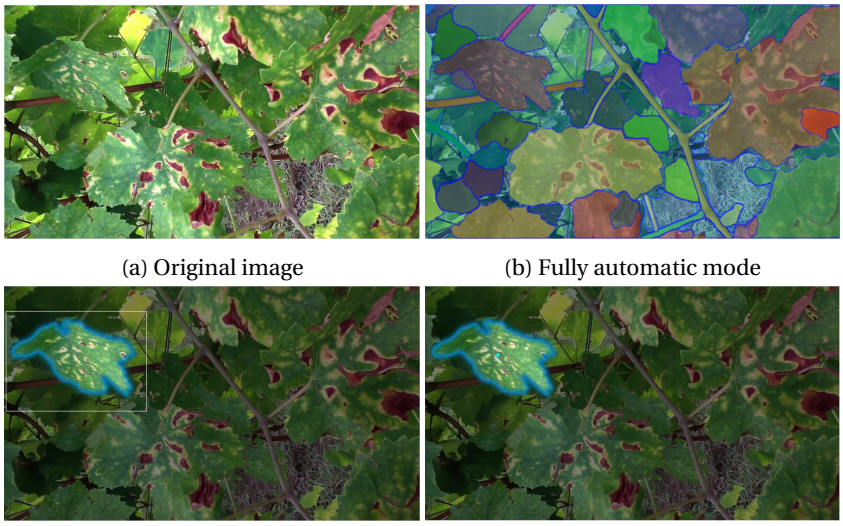


Fig. 2: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks.

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.

2.2 U-Net and EfficientNet

U-Net is a Convolutional Neural Network (CNN) originally proposed for the segmentation of biomedical images [33]. Afterwards the same net has been used for tackling segmentation of Urban Environment surfaces from High Resolution Satellite Imagery [23]. Variations of U-Net have been used in precisoin agriculture, for example to segment cucumber leaves with disease spots [41], or images of Cichorium intybus L. root [35].

The U-Net architecture is, specifically, a Fully Convolutional Neural Network (FCNN) model commonly used for semantics segmentation and instance segmentation tasks. This architecture derives its name from its U-shaped structure (Figure 4) consisting of two main parts, from left to right: the contracting path (encoder) and the expanding path (decoder). The contracting path of U-Net consists of multiple convolutional and pooling layers. These layers gradually reduce the spatial dimensions while increasing the number of feature channels, which helps capture a wide range of features at different scales. Each contracting block typically consists of two convolutional layers followed by a downsampling operation such as max pooling. The expanding path is the decoder part of U-Net and is responsible for generating the segmentation map. It consists of upconvolutional (also known as transposed convolutional) layers, which perform upsampling and increase the

spatial dimensions. The expanding path also includes skip connections that concatenate feature maps from the corresponding contracting path. These skip connections enable the decoder to access the high-resolution feature maps from the encoder, aiding in the precise localization of objects. At the end of the U-Net architecture, a final convolutional layer is typically applied to produce the segmentation map with the desired number of output channels, representing the class probabilities or pixel-wise labels.

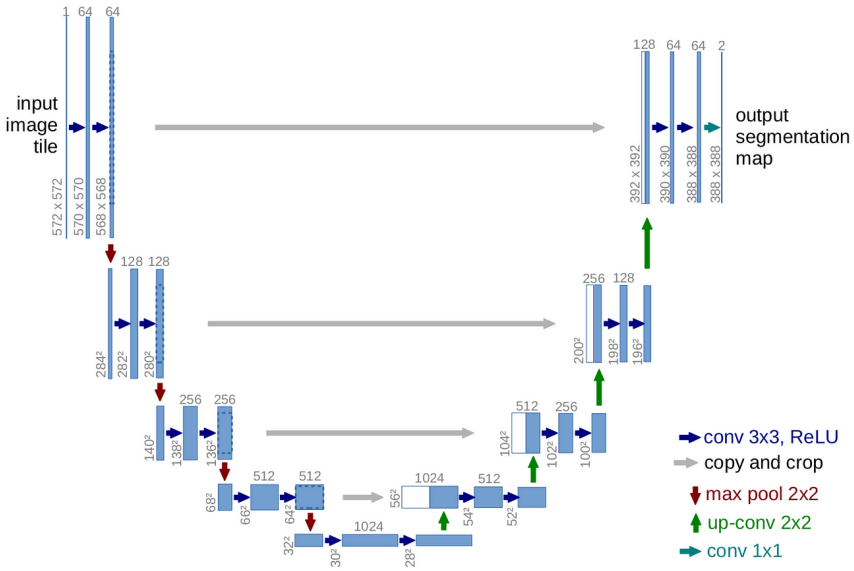


Fig. 4: U-net architecture

EfficientNet [37] is a family of convolutional neural network (CNN) models that are designed to achieve state-of-the-art performance while being highly efficient in terms of computational resources.

EfficientNet models are based on a concept called compound scaling, which optimizes the depth, width, and resolution of the network to achieve a good balance between accuracy and efficiency. The compound scaling technique involves uniformly scaling the network width, depth, and resolution using a coefficient called the compound scaling factor. This factor ensures that all dimensions of the network grow proportionally. By scaling up the network, it captures more complex patterns and features, while scaling down reduces the number of parameters and computations required. EfficientNet models have achieved state-of-the-art performance on various computer vision tasks, including image classification, object detection, and segmentation, across different benchmark datasets. They have demonstrated superior accuracy compared to other popular CNN architectures, such as ResNet,

while maintaining high efficiency in terms of memory usage and computational requirements.

EfficientNet models are often used as a backbone or encoder in various architectures, such as UNet, to improve performance and efficiency in tasks like image segmentation. The UNet architecture with an EfficientNet encoder combines the strengths of both models to achieve highly efficient and accurate image segmentation: it has been used in biomedical applications [18] [**?**] and it is implemented in the APEER ML toolkit by Zeiss® ([45]).

The UNet architecture, known for its U-shaped design, and EfficientNet, a state-of-the-art convolutional neural network (CNN) model, complement each other to improve performance. The EfficientNet encoder serves as the backbone of the network and is responsible for capturing high-level features from the input image. EfficientNet models are known for their superior performance and efficiency by leveraging compound scaling, which optimizes the depth, width, and resolution of the network based on a given resource constraint. In the UNet architecture, the contracting path (encoder) is responsible for capturing context and extracting features at different scales, while the expanding path (decoder) performs precise localization using skip connections. In the UNet with EfficientNet encoder, the contracting path is replaced with the EfficientNet backbone. The EfficientNet encoder provides a powerful feature extraction capability, capturing both low-level and high-level features from the input image. The extracted features are then passed to the expanding path, where upsampling and skip connections are used to recover the spatial resolution and refine the segmentation output.

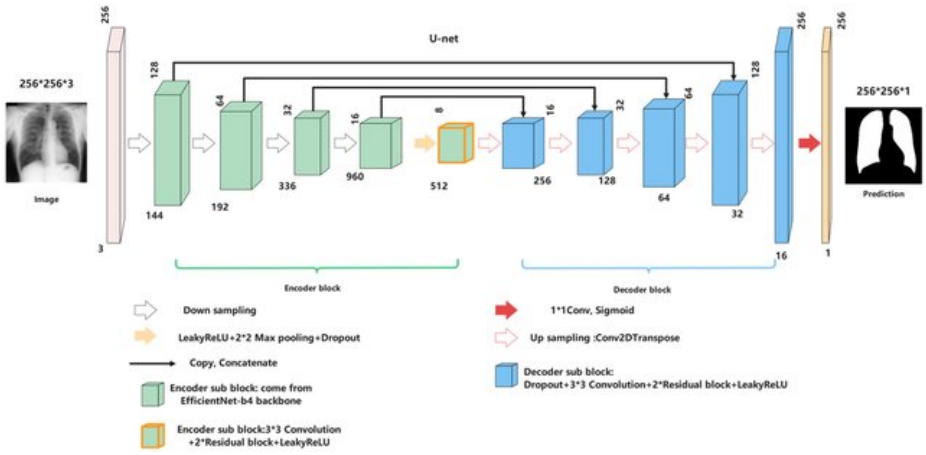


Fig. 5: U-Net architecture with EfficientNet encoder [**?**]

3 Experiments and Results

The ESCA dataset [1] contains 1770 RGB images in jpg format taken at a working distance of approximately 30 cm from grapevine plants during sunny and windy days and considering scenarios with background variety. In 882 of these images, the totality of plants appearing therein is healthy (not affected by Esca). In contrast, the other 888 images contain at least one depiction of shoots with visible foliar symptoms of Esca disease. The ESCA dataset consists of two folders, esca and healthy, each containing images of the relative class.

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 compare the performances of the foundational segmentation model SAM with no training to the previously available segmentation model U-Net, which requires manual annotation. The chosen task is semantic segmentation i.e. to classify the pixels of each image into two different semantic categories: foliage 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.

(a) (b)

Fig. 6: Ground truth segmentation of foliage (hand-made mask) for healthy1.jpg and esca1.jpg

We uploaded 102 images of the healthy folder and 103 of the esca folder to `https://www.apeer.com/annotate` and manually annotated them with the online semantic segmentation tool APEER Annotate [45], which allowed to classify the pixels of each image into two different semantic categories: foliage and background. The reult of the manual sementation are a total of 205 binary masks in which the pixels corresponding to foliage area contain the value 1 and all others contain value 0.

Image name	Manual annotation time (minutes)
healthy1.jpg	xx
healthy2.jpg	xx
healthy3.jpg	xx
healthy4.jpg	xx
healthy5.jpg	xx
esca1.jpg	xx
esca2.jpg	xx
esca3.jpg	xx
esca4.jpg	xx
esca5.jpg	xx

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

In Table 2 we recorded the time used to manually draw precise segmentation masks for 10 images. We keep the so-obtained 10 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.

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

The metrics used in this paper are:

- Intersection over Union (IoU) or Jaccard Index: it 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 (in the case of two classes only this measure is equivalent to Mean Intersection over Union - mIoU).
- Dice Coefficient (DC): It 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): it 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, Recall, and F1-score: these metrics are commonly used in multi-class semantic segmentation evaluation. Precision measures the proportion of correctly classified positive pixels, recall measures the proportion of actual positive pixels that are correctly classified, and F1-score combines precision and recall to provide a balanced measure.
- Normalized Surface Distance (NSD): the average of the Hausdorff distance and the Symmetric Contour Distance. Hausdorff Distance: The Hausdorff distance measures the maximum distance 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. Symmetric Contour Distance: The symmetric contour distance 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).

3.1 Segmentation of vine 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.

Image name	**Annotation time bounding box mode (minutes)**	**Annotation time point mode (minutes)**
healthy1.jpg	xx	xx
healthy2.jpg	xx	xx
healthy3.jpg	xx	xx
healthy4.jpg	xx	xx
healthy5.jpg	xx	xx
esca1.jpg	xx	xx
esca2.jpg	xx	xx
esca3.jpg	xx	xx
esca4.jpg	xx	xx
esca5.jpg	xx	xx

Table 4: Time spent to identify object masks with SAM using its semi-automatic modes.

Figure 3 shows the results of the three segmentation modes on an image of vines from the ESCA dataset [1] and Table 4 reports the exact time necessary to manually draw segmentation masks for the 10 selected images [1].

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`[15] 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 process 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.

		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 5: Image segmentation time with SAM in automatic batch mode.

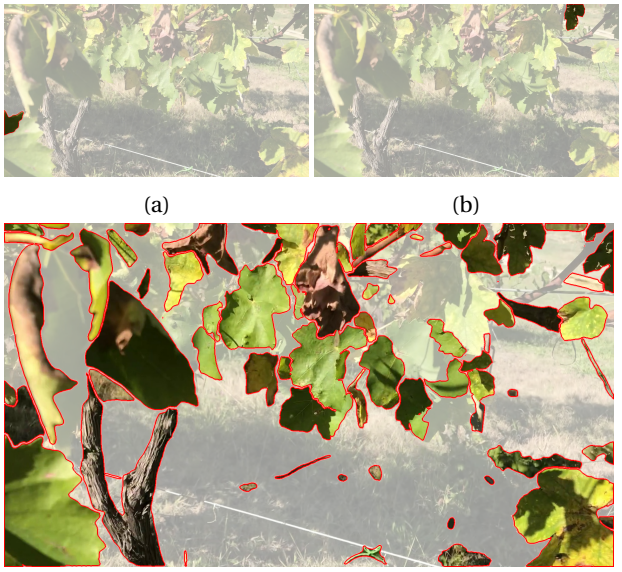
Image Name	**IoU**	**DC**	**PA**	**Prec**	**Rec**	**F1**
healthy1.jpg	...	...	...	...	...	...
healthy2.jpg	...	...	...	...	...	...
healthy3.jpg	...	...	...	...	...	...
healthy4.jpg	...	...	...	...	...	...
healthy5.jpg	...	...	...	...	...	...
esca1.jpg	...	...	...	...	...	...
esca2.jpg	...	...	...	...	...	...
esca3.jpg	...	...	...	...	...	...
esca4.jpg	...	...	...	...	...	...
esca5.jpg	...	...	...	...	...	...

Table 6: Evaluation of semantic segmentation performet by SAM in automatic mode

3.2 Segmentation of vine images with U-Net

In order to test semantic segmentation of vine images with U-Net we used the APEER ML toolkit by Zeiss® ?. The toolkit offers the possibility to train a U-Net with Effi-

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



(a) (b)

(c)

Fig. 7: Two masks produced by SAM in automatic mode on healthy1.jpg and the union of all masks produced for the same image.

Image Name	IoU	DC	PA	Prec	Rec	F1
healthy1.jpg	...	...	...	...	...	...
healthy2.jpg	...	...	...	...	...	...
healthy3.jpg	...	...	...	...	...	...
healthy4.jpg	...	...	...	...	...	...
healthy5.jpg	...	...	...	...	...	...
esca1.jpg	...	...	...	...	...	...
esca2.jpg	...	...	...	...	...	...
esca3.jpg	...	...	...	...	...	...
esca4.jpg	...	...	...	...	...	...
esca5.jpg	...	...	...	...	...	...

Table 7: Evaluation of semantic segmentation performed by SAM in bounding box mode

cientNet encoder that was then used to perform semantic segmentation on all 1770 images of the ESCA dataset. All generated masks were then downloaded.

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 nos-

Image Name	IoU	DC	PA	Prec	Rec	F1
healthy1.jpg	...	...	...	...	...	...
healthy2.jpg	...	...	...	...	...	...
healthy3.jpg	...	...	...	...	...	...
healthy4.jpg	...	...	...	...	...	...
healthy5.jpg	...	...	...	...	...	...
esca1.jpg	...	...	...	...	...	...
esca2.jpg	...	...	...	...	...	...
esca3.jpg	...	...	...	...	...	...
esca4.jpg	...	...	...	...	...	...
esca5.jpg	...	...	...	...	...	...

Table 8: Evaluation of semantic segmentation performed by SAM in point mode

(a) (b)

Fig. 8: Segmentation produced by U-Net on healthy.jpg and esca1.jpg

Image folder	Number of images	Image segmentation time
Esca	888	???
Healthy	882	???

Table 9: Image segmentation time with UNet

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.

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-

Image Name	IoU	DC	PA	Prec	Rec	F1
image1.jpg	...	...	...	...	...	...
image2.jpg	...	...	...	...	...	...
image3.jpg	...	...	...	...	...	...
image4.jpg	...	...	...	...	...	...
image5.jpg	...	...	...	...	...	...
image6.jpg	...	...	...	...	...	...
image7.jpg	...	...	...	...	...	...
image8.jpg	...	...	...	...	...	...
image9.jpg	...	...	...	...	...	...
image10.jpg	...	...	...	...	...	...

Table 10: Evaluation of semantic segmentation performed by UNet.

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. 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)
13. 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/
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.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

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.: Vit-h sam model (May 2023), available at: https://dl.fbaipublicfiles.com/segment_anything/sam_vit_h_4b8939.pth (May. 2023)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2012)
17. 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)
18. Le Dinh, T., Lee, S.H., Kwon, S.G., Kwon, K.R.: Cell nuclei segmentation in cryonuseg dataset using nested unet with efficientnet encoder. In: 2022 International Conference on Electronics, Information, and Communication (ICEIC). pp. 1–4. IEEE (2022)
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. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Büttner, F., et al.: Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv preprint arXiv:2206.01653 (2022)
23. McGlinchy, J., Johnson, B., Muller, B., Joseph, M., Diaz, J.: Application of unet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery. In: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 3915–3918 (2019). https://doi.org/10.1109/IGARSS.2019.8900453
24. 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)
25. 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
26. 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
27. 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
28. 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)
29. 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)

30. 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)
31. 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)
32. 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
33. 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)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
35. Smith, A.G., Petersen, J., Selvan, R., Rasmussen, C.R.: Segmentation of roots in soil with U-Net. Plant Methods **16**(1), 13 (2020). https://doi.org/10.1186/s13007-020-0563-0, https://doi.org/10.1186/s13007-020-0563-0
36. 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)
37. 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)
38. 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)
39. 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)
40. 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
41. Wang, C., Du, P., Wu, H., Li, J., Zhao, C., Zhu, H.: A cucumber leaf disease severity classification method based on the fusion of deeplabv3+ and u-net. Computers and Electronics in Agriculture **189**, 106373 (2021). https://doi.org/https://doi.org/10.1016/j.compag.2021.106373, https://www.sciencedirect.com/science/article/pii/S0168169921003902
42. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
43. 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

44. 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
45. Zeiss: Apeer annotate (2023), https://www.apeer.com/app/
46. 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
47. 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)
48. 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)
49. 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)
50. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning (2020)
51. 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)