Leaf segmentation of seedlings using foundation model on RGB-Depth images

Mathis CORDIER
LARIS, Université d'Angers, France

Hereari METUAREA
INRAe Angers, France

Meriem BENCHEIKH
LARIS, Université d'Angers, France

Cindy TORRES
Vilmorin-Mikado, France

Pejma RASTI
CERADE, ESAIP, France

David ROUSSEAU
LARIS, Université d'Angers, France
david.rousseau@univ-angers.fr

Abstract

We present several ways of integrating the Depth information to the color in order to predict leaf instances via the segment anything model.

1. Introduction

In this communication, we focus on the segment anything model (SAM) [9] that we consider for a segmentation task in plant science with RGB-Depth cameras. These cameras produce color images registered with a Depth map corresponding to the distance from the sensor. The added value of Depth images produced by low-cost cameras has been first demonstrated in [2] for leaf segmentation. The architecture of plants with replicated leaves at different locations along the stems suits the range contrast of Depth cameras and this imaging modality became a standard in plant imaging over the years as illustrated in several reviews [12, 11]. We propose such an investigation here to adapt SAM to RGB-Depth images. As computer vision problem in plant science, we consider instance segmentation of seedling leaves. Seedling is an important stage of development of plant as it includes the deployment of the first leaves which conditions the success of the growth toward adult plants. As most related work, we can point to the numerous adaptations of SAM to specific domains of application [15, 21] and our proposal can be considered as a new contribution in this trend in the scope here of plant sciences. The fusion of RGB and Depth with Deep learning has been mostly investigated with a variety of architectures including CNN [20, 22], LSTM [1, 7], Transformers [18, 14, 7]. As most related work, the monitoring of seedling with RGB-Depth has been investigated for classification of development stage [7] and tracking of individual seedling [3] and we propose an extension of these works to the instance segmentation of individual leaves. As a pilot study on the use of SAM for multimodal data in plant science, several choices are made arbitrarily for a first proof of feasability. These choices could of course be revisited to extend this work.

2. Materials and methods

We settled a grid of low-cost RGB-Depth sensors to acquire sequences of images at a defined sample rate as done in [3, 7, 16]. Different species at seedling stages were observed with different environmental and temporal conditions: Cabbage with one image each 15 minutes in a same a day, Pepper with 1 image per day during 18 days and Komatsuna [19] with 4 images per day during 15 days. Cabbage dataset is useful to study performance trend on short term and Cabbage and Komatsuna for long-term developments. We investigated 3 different ways of injecting the Height information into the RGB image using SAM [9]. A global pipeline of fusion strategies is presented in Fig. 1. Height information can be given as an input of the model. The image encoders ViT-B, ViT-L and ViT-H provided with SAM [5] are built for RGB data and require a 3-channels input format. To be able to use these models despite the addition of the Height channel, we need to have a way of dimension reduction from 4 to 3 channels. As early fusion strategy, we propose here to use Height as a multiplicative filter applied to the 3 color channels. This method is a way of giving more weight to image areas with significant Height from top view. As intermediate fusion strategy, we tested a combination of RGB and Height within the model itself. Rather than operating a dimension reduction of information channels, we propose here to

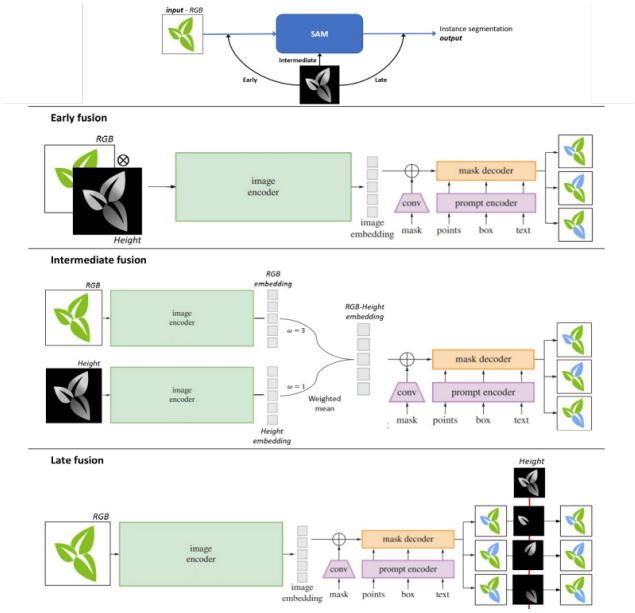

Figure 1. Pipelines of the different proposed strategies of fusion.

	Fusion strategy	AP (%)	AP_{75} (%)	AP_{50} (%)	Dice (%)
Cabbage	RGB - Baseline	10.7±1.4	15.6±2.6	16.7±2.2	20.0±1.0
	Depth	7.1±1.6	18.2±4.5	38.2±7.6	16.2±4.1
	Early	10.4±1.7	13.0±2.8	16.3±2.7	18.3±1.5
	Intermediate	13.9±2.2	26.4±4.0	28.0±3.5	26.8±2.1
	Late	**22.6±2.9**	**47.4±7.1**	**51.9±5.7**	**85.9±1.8**
Pepper	RGB - Baseline	7.9±4.0	15.9±9.6	16.5±9.9	27.7±20.2
	Depth	0.1±0.2	0.1±0.2	0.3±0.5	10.5±9.5
	Early	6.5±3.2	11.5±8.5	12.4±8.9	26.8±21.5
	Intermediate	6.6±4.2	14.6±11.4	14.8±11.5	28.0±22.1
	Late	**17.0±7.2**	**43.0±18.6**	**44.1±19.1**	**78.8±15.6**
Komatsuna	RGB - Baseline	7.1±3.7	10.8±7.1	12.2±7.3	24.3±17.0
	Depth	3.7±4.8	0.4±0.9	14.5±16.0	11.7±11.1
	Early	7.3±4.1	15.1±11.7	16.9±12.2	24.1±16.2
	Intermediate	5.9±4.3	11.5±11.9	12.5±11.9	23.8±16.9
	Late	**9.8±6.2**	**21.4±17.2**	**23.8±17.4**	**52.4±33.3**

Table 1. Overall segmentation results depending on fusion strategies for all datasets using SAM with ViT-B encoder. The average precision (AP) [10] is computed with different thresholds applied to intersection-over-union (IoU) [8] scores. AP_{50} (PASCAL VOC metric [6]) and AP_{75} (strict metric) are obtained with thresholds of 0.5 and 0.75 respectively. AP metric (primary COCO challenge metric [13]) is the area under the Precision-Recall curve with IoU thresholds from 0.50 to 0.95 with a step of 0.05. Dice is the standard Dice-Sorensen coefficient, i.e. twice the IOU between ground truth and leaf instance over the cardinal of the ground truth plus cardinal of leaf instance [4, 17].

encode both RGB and Height images separately. Once the image embedding is obtained for each modality, a local combination is done as shown in Fig. 1. A final RGB-Height embedding is obtained by a weighted mean between RGB and Height embeddings with a $\frac{3}{4}$:$\frac{1}{4}$ ratio respectively, to match the number of channels of each component. As late fusion strategy, we post-process instances produced by SAM with RGB inputs via Height. This simply corresponds to the filtering of instances which appear too low in Height.

3. Results

A first overall outcome is the absence of significant differences of performance obtained with the different encoder ViT-B, ViT-L and ViT-H models provided with SAM. The use of the lightest model in terms of parameters, i.e. ViT-B, therefore seems more appropriate. A complete prediction result for each dataset is provided in Table 1 with Vit-B. For the three datasets, there is a systematic gain in fusing RGB and Depth by comparison with the sole use of Depth or RGB in SAM. The late fusion provides the best results by comparison with early and intermediate ones. We also investigated the combination of late fusion with one of the other fusion. The combination of intermediate and late fusion provides an improvement by comparison with the sole late fusion. The results show a rather large standard deviation on all the segmentation metrics for medium and long-term datasets, e.g. Pepper and Komatsuna. This can be explained by the very different stages of growth within each of these datasets.

For both Pepper and Komatsuna datasets, the timelapse begins at the cotyledon stage, when very few Depth contrast is present and continues through to the adult stage with a large number of leaves and much more contrast in Depth. The evolution of the segmentation performance as a function of plant growth stage is relevant. The performance of late fusion is always better but this superiority is small at the beginning and expands when Height becomes more contrasted. Convergence to a performance plateau is much faster with the injection of Height. This is illustrated in Fig. 2 but was systematically observed for all the tested metrics and datasets. Non surprisingly, this demonstrates that the addition of Height information is beneficial as soon as the canopy height is sufficient.

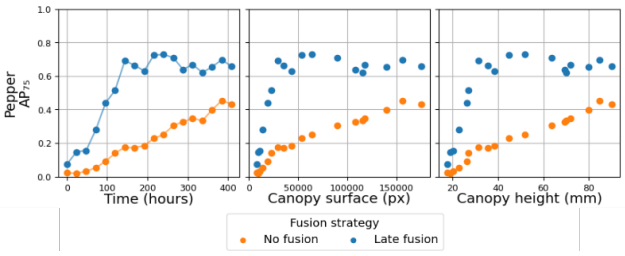

Figure 2. AP_{75} as a function of growth stage represented by time, canopy surface and height.

4. Conclusion

Combination of RGB and Depth in SAM is systematically valuable to predict leaf instances.

References

[1] Ahmed Abobakr, Mohammed Hossny, Hala Abdelkader, and Saeid Nahavandi. RGB-D Fall Detection via Deep Residual Convolutional LSTM Networks. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7, Dec. 2018.

[2] Yann Chéné, David Rousseau, Philippe Lucidarme, Jessica Bertheloot, Valérie Caffier, Philippe Morel, Étienne Belin, and François Chapeau-Blondeau. On the use of depth camera for 3D phenotyping of entire plants. *Computers and Electronics in Agriculture*, 82:122–127, Mar. 2012.

[3] Mathis Cordier, Cindy Torres, Pejman Rasti, and David Rousseau. On the Use of Circadian Cycles to Monitor Individual Young Plants. *Remote Sensing*, 15(11):2704, Jan. 2023. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[4] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[6] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[7] Hadhami Garbouge, Pejman Rasti, and David Rousseau. Enhancing the Tracking of Seedling Growth Using RGB-Depth Fusion and Deep Learning. *Sensors*, 21(24):8425, Jan. 2021. Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Paul Jaccard. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:241–72, Jan. 1901.

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

[10] Kazuaki Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Sept. 2005.

[11] Lei Li, Qin Zhang, and Danfeng Huang. A Review of Imaging Techniques for Plant Phenotyping. *Sensors*, 14(11):20078–20111, Nov. 2014. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[12] Zhenbo Li, Ruohao Guo, Meng Li, Yaru Chen, and Guangyao Li. A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture*, 176:105672, Sept. 2020.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, Feb. 2015. arXiv:1405.0312 [cs].

[14] Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. SwinNet: Swin Transformer Drives Edge-Aware RGB-D and RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, July 2022. Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

[15] Jun Ma and Bo Wang. Segment Anything in Medical Images, Apr. 2023. arXiv:2304.12306 [cs, eess].

[16] Salma Samiei, Pejman Rasti, Joseph Ly Vu, Julia Buitink, and David Rousseau. Deep learning-based detection of seedling development. *Plant Methods*, 16:103, 2020.

[17] Thorvald Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34, 1948.

[18] Georgios Tziafas and Hamidreza Kasaei. Early or Late Fusion Matters: Efficient RGB-D Fusion in Vision Transformers for 3D Object Recognition, Mar. 2023. arXiv:2210.00843 [cs].

[19] Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rin-ichiro Taniguchi. An Easy-to-Setup 3D Phenotyping Platform for KOMATSUNA Dataset. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2038–2045, Venice, Italy, Oct. 2017. IEEE.

[20] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D Segmentation, Mar. 2018. arXiv:1803.06791 [cs] version: 1.

[21] Qiusheng Wu and Lucas Prado Osco. samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM), May 2023.

[22] Zongwei Wu, Guillaume Allibert, Christophe Stolz, and Cedric Demonceaux. Depth-Adapted CNN for RGB-D cameras, Sept. 2020. arXiv:2009.09976 [cs].