Towards Uncertainty Detection in Automated Leaf Tissue Segmentation

Ráchel Grexová[‡], Klara Voggeneder[✎], Danny Tholen[✎], Guillaume Théroux-Rancourt[✎], Walter Kropatsch[‡], and Jiří Hladůvka[‡]
[✎]Institute of Botany, University of Natural Resources and Life Sciences, Vienna
{klara.voggeneder, daniel.tholen, guillaume.theroux-rancourt}@boku.ac.at
[‡]Pattern Recognition and Image Processing Group, Vienna University of Technology
{rachel.grexova, walter.kropatsch, jiri.hladuvka}@tuwien.ac.at

Abstract

In order to use segmented volumetric data for subsequent analyses, it is important to detect and understand, where the segmentation is reliable and where it is uncertain. This is especially critical in deep learning segmentation which relies on manually annotated ground truth. Especially in applications using medical and biological data, ground truth annotations are often sparse, imbalanced, and imprecise.

We propose to utilize 2.5D orthogonal ensembles not only to arrive at dense segmentation but, more importantly, to indicate areas of high prediction fidelity and areas of uncertainty.

Our ensemble achieved accuracy above 95% in the high fidelity areas of a volume of a poplar leaf segment. This accuracy was achieved not only for a fresh leaf sample similar to the training data, but also for a severely dehydrated sample. Well-represented classes contained large areas of high prediction fidelity and exhibited high validation metrics. By contrast, under-represented classes tend to contain large areas of uncertainty.

Indication of uncertainty could be used as a basis to revise the predictions by domain experts. This is in turn expected to improve and/or enlarge the ground truth and allows for training of higher-quality segmentation models.

1. Introduction

Segmentation is crucial step for further biological [17] or biomedical analysis. Traditional approaches of image segmentation rely on homogeneity criteria such as intensity values (threshold) or large gradient magnitude (border line) [12]. Since MRI, CT or µ-CT images are blurred, contain noise or have low contrast, it is more difficult to design such criteria in medical [18] or biological image segmentation. In these fields deep learning is increasingly gaining popularity [8] as the features are learned automatically. The automatic feature learning is beneficial, but the filters important for the segmentation remain unknown, which makes it difficult to interpret and improve the results [15].

Deep-learning approaches rely on large ground truth training sets. Limited annotated data is a remaining challenge in medical imaging [5], but even more in botany and agriculture, where annotated image libraries are missing [13]. Moreover, any manual annotations are subject to inter- and intra-observed variability. In turn, such ground truth annotation often may become unreliable in hard-to-annotate areas.

In 2015 U-Net was introduced [14] and it has become one of the most commonly used architectures in (bio-)medical segmentation [18]. It was originally used for 2D transmitted light microscopy images. Since then it was used for nearly all major imaging modalities such as CT, MRI and X-ray [15]. The drawback of using 2D convolution for 3D data such as MRI, CT or µ-CT is the lack of volumetric context [2]. There have been several extensions of U-Net [15], the 3D U-Net [20] being one of them. Due to high requirements on GPU memory of 3D convolutions [1] volumetric data is usually divided into smaller patches [5]. To overcome the drawbacks of 2D and 3D U-Nets, there have been several attempts to combine these approaches and run the 2D U-Net networks in parallel on several 2D projections of a 3D volume in order to incorporate some volumetric context at computationally efficient cost. This kind of ensemble U-Net is called 2.5D U-Net [15]. Usually the 3D volume is divided into 2D images along three orthogonal axes and then three U-Net models are trained and used for prediction separately. With fusion of the three predictions the final segmentation result is produced [4, 6, 11, 19]. Another possibility is to use random 2.5D U-Net with multiple 2D projections [2].

In this paper we utilize the 2.5D-like approach in order to localize the high fidelity predictions and to flag voxels with uncertain predictions. The aim is on one hand to address the problem with limited ground truth data typical for

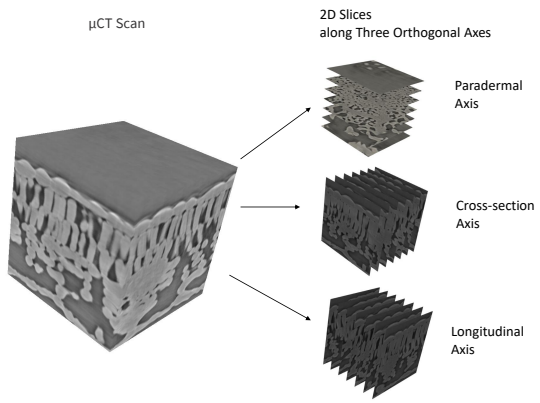


Figure 1. µ-CT scan viewed as three orthogonal stacks of images.

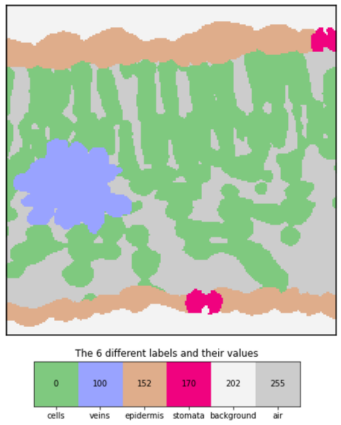


Figure 2. Example of ground truth cross-section slice of scan time 3 showing all 6 available labels.

biological and (bio-)medical image segmentation. On the other hand we aim to build a tool that can enlighten how to fix errors of the predictions. The uncertain regions can be further reviewed by domain experts. This could enlarge the labeled data set, while significantly decreasing the manual labour. We present an approach that serves as the initial step for human-in-the-loop interactive segmentation.

2. Data

µ-CT scans of a hybrid poplar leaf were taken at the TOMCAT beamline at the Swiss Light Source of the Paul Scherrer Institute (Villigen, Switzerland) using acquisition protocols similar to [17]. The leaf was allowed to wilt and scanned in five different scan times. The first scan was done immediately after the leaf strip was prepared and placed into a holder. The other four scans were done after 10, 20, 25, and 30 minutes, while the leaf was dehydrating. While only minor differences in leaf structure were apparent during scan times 1-4, large differences were noticeable at time 5 and the cells were visibly shrunken.

The µ-CT scans were divided into stacks of 2D slices along the three orthogonal axes (Fig. 1). Sparse set of 2D images were manually segmented into 6 classes, i.e., cells, veins, epidermis, stomata, background air, and intercellular airspaces (inner air).

This resulted in 10 to 25 segmented slices for each scan time and each axis. Two of the six classes have been heavily underrepresented: veins (5%) and the small pores on the surface, called stomata ($\approx 1\%$).

3. Methodology

In this section we summarize the methodology of segmenting 2D slices along three orthogonal axes, orthogonal axes ensemble used for the selection of 3-consistent voxels and their evaluation.

3.1. 2D Segmentation Using U-Net

For 2D segmentation we divided the data into the training and validation sets. For the training set, we used scanning times 1, 2, and 4. For the validation set, the time 3 and (the challenging) time 5 were used with the aim to validate the models on a slightly different-looking dataset.

The models were trained and predicted using 3 different resolutions, i.e. 1024×1024, 512×512, 256×256. The models were trained using U-Net [14] architecture. As shown in Fig. 3 one model was trained for the paradermal axis and one for the cross- and longitudinal-section.

In order to address the problem with limited labeled ground truth data-set we used data augmentation [3]. We applied transformation functions such as random crop, flip, rotation both on the µ-CT slices and their corresponding labeled ground truth slices simultaneously.

3.2. Orthogonal Axes Ensemble

The outputs of the three 2D predictions are aggregated in one 3-channel volume with 3 label predictions per voxel (see Figure 3). The number of unique labels per voxel splits the voxels into three categories:

1. all three models predicted consistently (3-consistent voxels);
2. two models were consistent, but inconsistent with the remaining one;
3. all three models were mutually inconsistent.

As no clear consensus is found for voxels of categories 2 and 3, we declare them as uncertain and call for a manual inspection. We'll discuss this later in section 5.

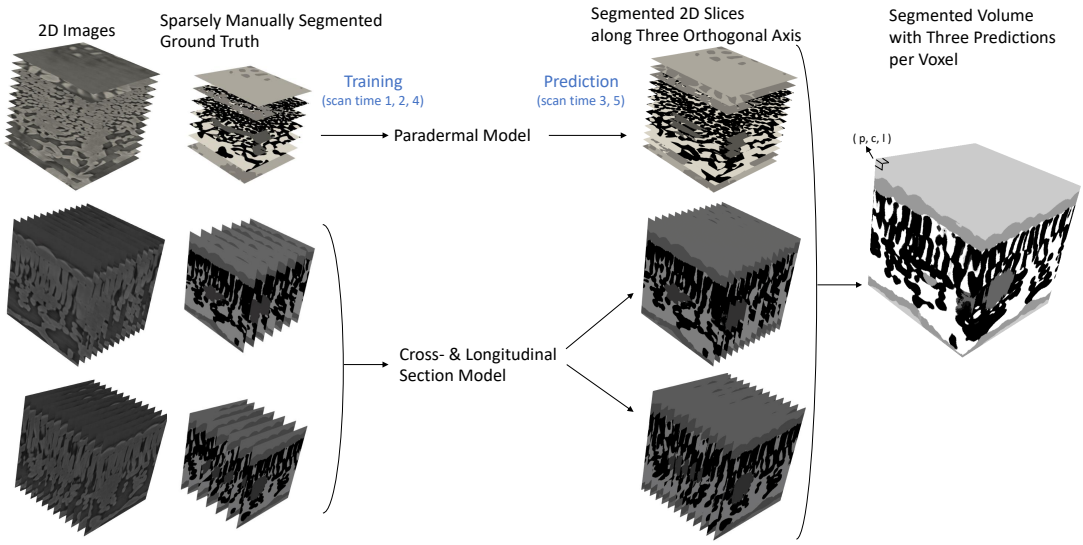


Figure 3. Training and prediction along the 3 orthogonal axes and their aggregation.

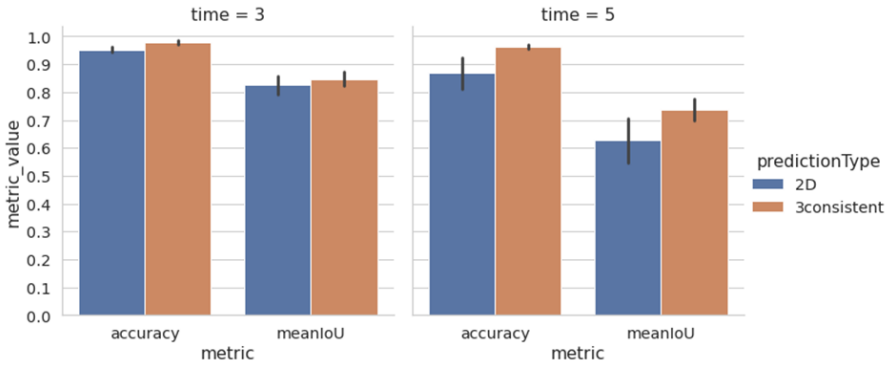


Figure 4. Mean IoU and accuracy for the test set (times 3 and 5). 2D predictions (averaged over all orthogonal axes and resolutions) compared to average of 3-consistent predictions. Black bars represent standard deviation.

In the following we are interested in how reliable are the predictions of category 1 with respect to the ground truth. To do so we compute and compare several metrics for both the ensemble and the three axis-wise 2D predictions.

3.3. Validation Metrics

Five spatial overlap-based metrics [16] are used for validation.

Pixel Accuracy (PA) is a basic metric used for segmentation evaluation. It is the ratio of correctly predicted pixels to the total number of pixels. [9]

Precision is used only for each label class separately:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

where TP is the true positive fraction and FP is the false positive fraction [9]. Precision values indicate whether over-segmentation occurs [10].

Recall Similar to precision, recall is used only for each label class separately:

$$recall = \frac{TP}{TP + FN} \tag{2}$$

where TP is the true positive fraction and FN is the false negative fraction [9]. Recall values indicate whether under-segmentation occurs [10].

Intersection over Union (IoU, a.k.a the Jacard index [7]) is used both in the per-class and the image-mean variants.

IoU for individual class is defined as

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \tag{3}$$

where A is the mask of the class label in the ground truth image and B is the mask of the class label in the predicted image. $|A \cap B|$ is the intersection and $|A \cup B|$ is the union [9] [16].

Mean IoU is defined as the mean for the IoUs of the individual classes [9]. The mean IoU was calculated as

$$meanIoU = \frac{1}{n} \sum_{c=1}^{n} (IoU)_c \tag{4}$$

where $n = 6$ is number of classes and $(IoU)_c$ is the IoU for class c.

4. Results and Discussion

The metrics in 2D predictions were sufficiently high for the well represented classes, i.e. cells, epidermis, background air and intercellular airspace. In scan time 3, IoU, precision, and recall values were usually higher than 90%. In scan time 5 metrics were usually higher than 70% (Fig. 5). Since in scan time 5, the poplar leaf was much more dehydrated than during other scan times, it was expected that the predictions would be less accurate. Indeed the accuracy and mean IoU were lower for scan time 5 than for scan time 3 (Fig. 4).

For the under-represented classes, i.e., stomata and veins the precision was higher than recall. The recall was especially low for stomata for both scan times. This indicates under-segmentation. Therefore IoU was also low for stomata. The low IoU for underrepresented classes can explain why mean IoU is lower than accuracy for the 2D predictions.

After selection of 3-consistent voxels both accuracy and mean IoU increased in both scan times. The amount of voxels of this category was lower in time 5 ($\approx 80\%$) than in time 3 ($\approx 90\%$). The increase of the metrics values was higher for the scan time 5 than for scan time 3. Even though the average accuracy was 95.26% for scan time 3 and 86.82% for scan time 5, after selection of 3-consistent voxels the average accuracy was comparable, i.e. 97.83%, 96.16%, respectively (see Fig. 4). Mean IoU remained lower for scan time 5 than for scan time 3. The difference in the amount of uncertainty voxels for scan time 3 and 5 is demonstrated on an example in Figures 7b and 7d by yellow color.

Except for stomata in time 3 and both stomata and veins in time 5 the metrics increased class-wise. For the under-represented classes the recall was low and it got even lower for the 3-consistent voxels (Fig. 5). Therefore also IoU was lower.

The low recall for stomata for 2D predictions can be observed in Figure 6 (f)-(h). Only some of the stomata were predicted by the particular models, but along each of the orthogonal axis it was predicted differently. In paradermal axis (f) only around half of the stomata were predicted, but when they were predicted it usually corresponded to the ground truth. This corresponds to small recall, but higher precision (see Fig. 5a). Additionally one of the stomata was predicted around hole visible in µ-CT scan (a) and ground truth (e) near the stoma. Such an air gap between stoma and epidermis is highly unusual. In cross- (g) and long- (h) sections the number of predicted stomata is higher, but the shapes are slightly deformed. Therefore, as it is visible in Fig. 6 (b) - (d) the uncertainty depicted with yellow is high in stomata regions and 3-consistent voxels forms only small portion of stomata voxels in ground truth. Additionally around the uncertainty the 3-consistent voxels differs from the ground truth. This explains the decrease of recall after orthogonal axes ensemble. A similar pattern is visible for veins and stomata in scan time 5 (see Fig. 7c and Fig. 5b).

For well represented classes the metric values were for 3-consistent voxels usually above 90% for both scan times. In scan time 3 most voxels labeled as cells, intercellular airspace and background were labeled as their corresponding class in all 2D predictions. This is illustrated in Fig. 7b. In scan time 5 precision was significantly lower for inner air and stomata than in scan time 3 (see Fig. 5b). This indicates over-segmentation of these classes. For the 3-consistent voxels the precision significantly increased.

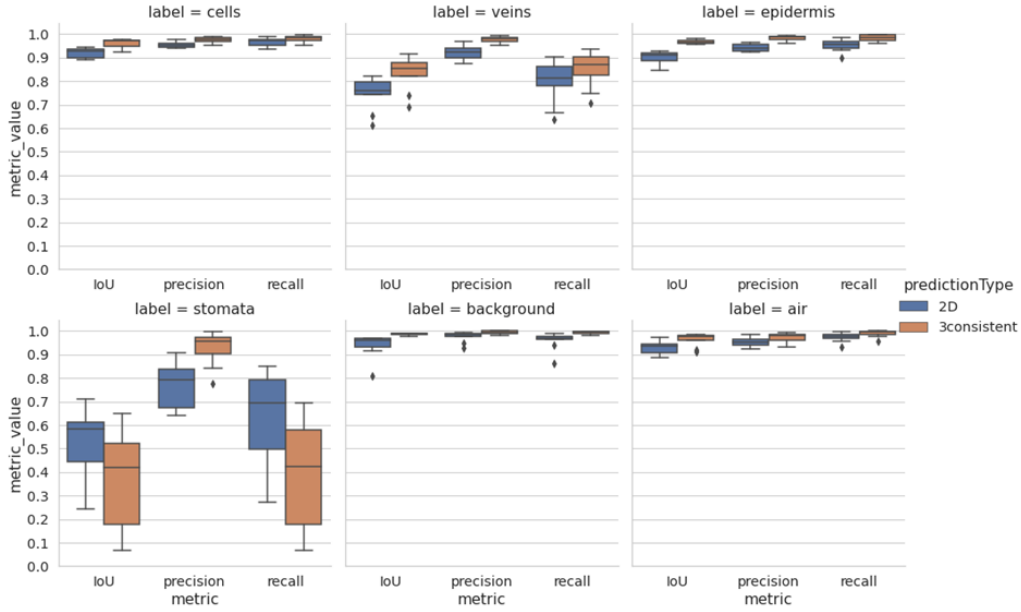
5. Conclusion and Future Work

We presented an approach that utilizes 2.5D orthogonal axis ensembles and detects areas of confidence and uncertainty. The validation metrics were higher for the 3-consistent voxels in comparison to the 2D predictions. For well represented classes, i.e. cells, epidermis, background and inner air, they were usually above 90% even for scan time 5, that was significantly more dehydrated in comparison to the training data-set.

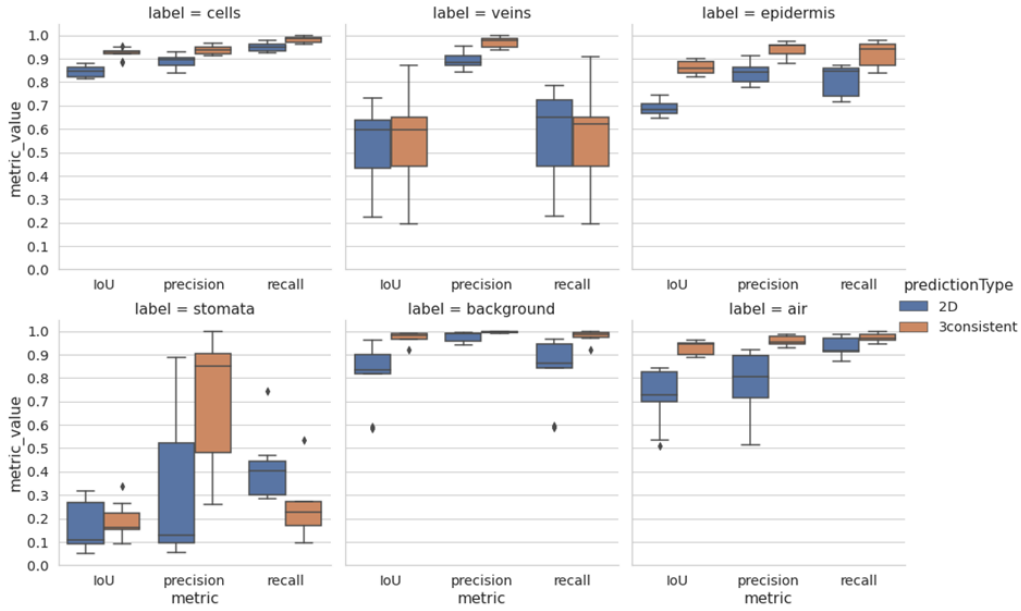
Uncertainty areas tend to correlate with the underrepresented classes, i.e. stomata and veins. Here, small recall was typical in 2D predictions, indicating under-segmentation of these classes. For the classes with large uncertainty areas, under-segmentation remained also for the orthogonal axes ensemble.

In future work this approach could be used as an initial step in a human-in-the-loop segmentation, where the uncertainty areas can be revised.

In Figure 8 we show an example of prediction by orthogonal axes ensemble overlaid by uncertainty (yellow)



(a) scan time 3



(b) scan time 5

Figure 5. Label classes comparison of 2D predictions and 3-consistent voxels for scan time 3 (a) and scan time 5 (b).

for a slice *without* the ground truth. An increasing opacity can become a part of an interactive tool for revision of predictions irrespective of absence/presence of a ground truth. Such a revision can in turn enrich the training set.

In Figure 7a shows 3-consistent voxels of the veins surrounded by yellow uncertainty area and several orange spikes. Because it is hard even for a human expert to distinguish cells closely appressed to the veins, such cells were annotated as veins. Our approach actually correctly annotated these cells, leading to the orange spikes. This shows our approach can help to identify areas that are hard to manually label to improve the ground truth data.

Acknowledgments We acknowledge the Paul Scherrer Institut, Villigen, Switzerland for provision of beamtime at the TOM-

CAT beamline of the Swiss Light Source. The computational results have been partially achieved using the Vienna Scientific Cluster (VSC). This work was supported by the Vienna Science and Technology Fund (WWTF) project LS19-013 and by the Austrian Science Fund (FWF) projects M2245 and P30275.

References

[1] Christoph Angermann and Markus Haltmeier. Random 2.5D U-net for Fully 3D Segmentation. In Hongen Liao, Simone Balocco, Guijin Wang, Feng Zhang, Yongpan Liu, Zijian Ding, Luc Duong, Renzo Phellan, Guillaume Zahnd, Katharina Breininger, Shadi Albarqouni, Stefano Moriconi, Su-Lin Lee, and Stefanie Demirci, editors, *Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting*, Lecture Notes in Computer Science, pages 158–166, Cham, 2019. Springer International Publishing.

[2] Christoph Angermann, Markus Haltmeier, Ruth Steiger, Sergiy Pereverzyev, and Elke Gizewski. Projection-Based 2.5D U-net Architecture for Fast Volumetric Segmentation. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5, July 2019.

[3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, Feb. 2020. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[4] Lin Han, Yuanhao Chen, Jiaming Li, Bowei Zhong, Yuzhu Lei, and Minghui Sun. Liver segmentation with 2.5D perpendicular UNets. *Computers & Electrical Engineering*, 91, May 2021.

[5] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4):582–596, Aug. 2019.

[6] Ke Hu, Chang Liu, Xi Yu, Jian Zhang, Yu He, and Hongchao Zhu. A 2.5D Cancer Segmentation for MRI Images Based on U-Net. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, pages 6–10, July 2018.

[7] Paul Jaccard. The Distribution of the Flora in the Alpine Zone.1. *New Phytologist*, 11(2):37–50, 1912.

[8] Priyanka Malhotra, Sheifali Gupta, Deepika Koundal, Atef Zaguia, and Wegayehu Enbeyle. Deep Neural Networks for Medical Image Segmentation. *Journal of Healthcare Engineering*, 2022:e9580991, Mar. 2022. Publisher: Hindawi.

[9] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[10] Fernando C. Monteiro and Aurélio C. Campilho. Performance Evaluation of Image Segmentation. In Aurélio Campilho and Mohamed S. Kamel, editors, *Image Analysis and Recognition*, Lecture Notes in Computer Science, pages 248–259, Berlin, Heidelberg, 2006. Springer.

[11] Gabriele Piantadosi, Mario Sansone, Roberta Fusco, and Carlo Sansone. Multi-planar 3D breast segmentation in MRI via deep convolutional neural networks. *Artificial Intelligence in Medicine*, 103:101781, Mar. 2020.

[12] Bernhard Preim and Charl P Botha. *Visual computing for medicine: theory, algorithms, and applications*. Newnes, 2013.

[13] Devin A. Rippner, Pranav V. Raja, J. Mason Earles, Mina Momayyezi, Alexander Buchko, Fiona V. Duong, Elizabeth J. Forrestel, Dilworth Y. Parkinson, Kenneth A. Shackel, Jeffrey L. Neyhart, and Andrew J. McElrone. A workflow for segmenting soil and plant X-ray computed tomography images with deep learning in Google's Colaboratory. *Frontiers in Plant Science*, 13, 2022.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing.

[15] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9:82031–82057, 2021.

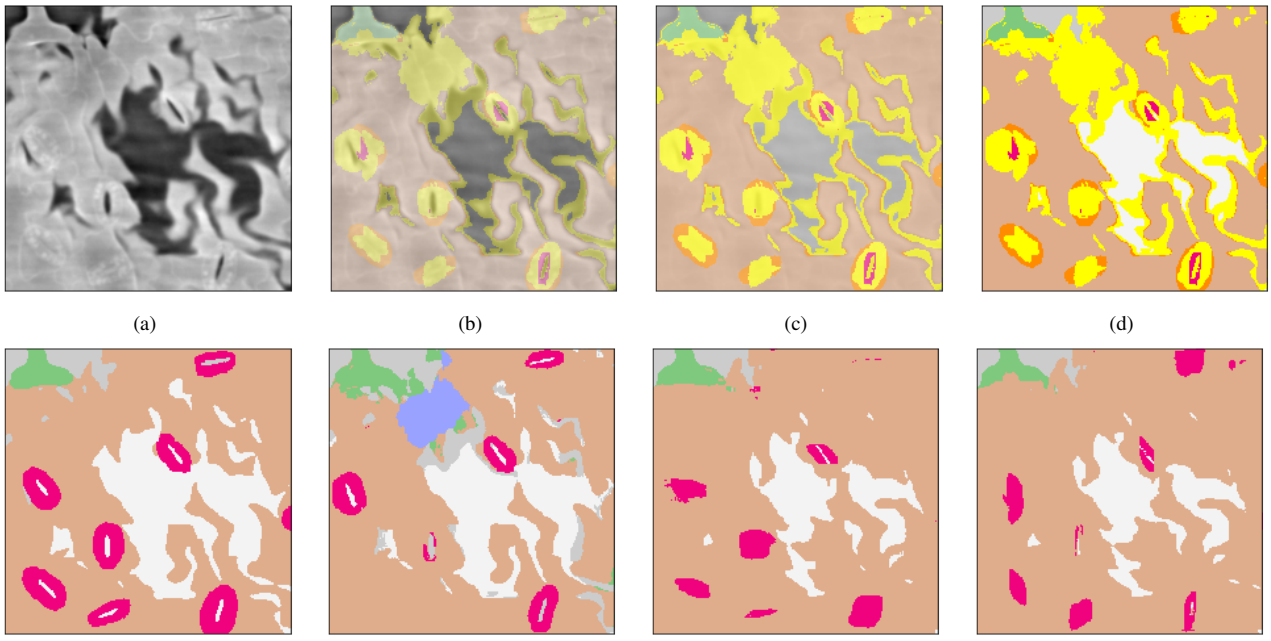
[16] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Aug. 2015.

[17] Guillaume Théroux-Rancourt, Matthew R. Jenkins, Craig R. Brodersen, Andrew McElrone, Elisabeth J. Forrestel, and J. Mason Earles. Digitally deconstructing leaves in 3D using X-ray microcomputed tomography and machine learning. *Applications in Plant Sciences*, 8(7):e11380, 2020.

[18] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 2022.

[19] Jie Wei, Yong Xia, and Yanning Zhang. M3Net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation. *Pattern Recognition*, 91:366–378, July 2019.

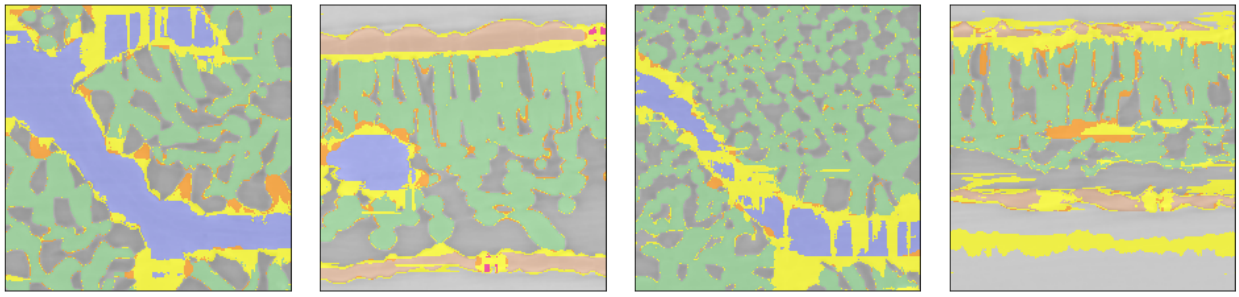
[20] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture Notes in Computer Science, pages 424–432, Cham, 2016. Springer International Publishing.



(a) (b) (c) (d)

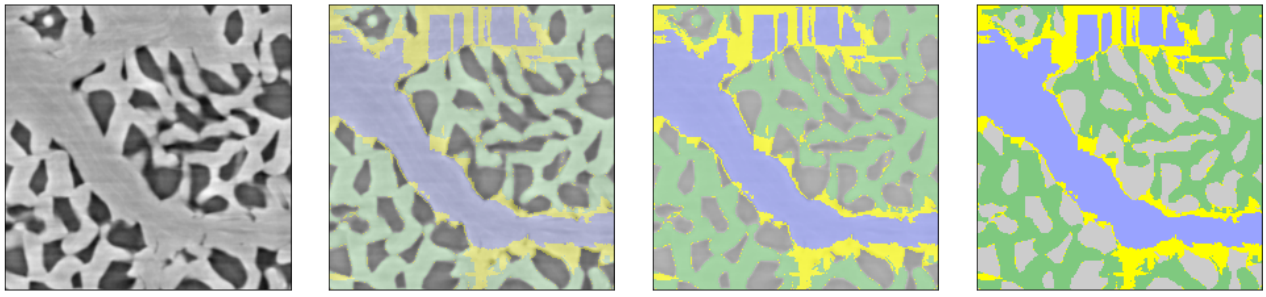
(e) (f) (g) (h)

Figure 6. Top row: slice of scan (a) overlaid by ensemble predictions, using increasing opacity (b)-(d). Yellow indicates uncertainty and requests human revision. Bright orange indicates mismatch between predictions and ground truth. Bottom row: Labels by human expert (e) and predictions along the three orthogonal axes (f)-(h).



(a) (b) (c) (d)

Figure 7. Selected slices from scan time 3 (a),(b) and 5 (c),(d) overlapped by ensemble predictions and uncertainty.



(a) (b) (c) (d)

Figure 8. Slice of scan (a) overlaid by ensemble predictions, using increasing opacity (b)-(d).