Multi-Task Learning for Simultaneous Speed-of-Sound Mapping and Image Reconstruction Using Non-Contact Thermoacoustics

Ajay Singhvi*, Max L. Wang*, Aidan Fitzpatrick*, and Amin Arbabian
Department of Electrical Engineering, Stanford University, Stanford, CA, USA

***Abstract*—Multi-modal imaging via thermoacoustic (TA) approaches provides contrast mechanisms differing from conventional ultrasound (US) imaging – opening up new applications like non-invasive, non-contact below-ground sensing. Due to the high correlation between soil moisture content and speed-of-sound (SoS), knowledge about the SoS in soil can be utilized to improve below-ground image reconstruction and soil moisture mapping at depth. In this work, we present multi-task deep learning networks to accurately predict arbitrarily varying SoS distributions in soil while concurrently reconstructing high-fidelity TA images of root structures. We deploy multi-task U-Net based fully convolutional neural networks trained using US data generated through TA simulations on a wheat root dataset. A multi-input, multi-output architecture performed best – achieving the highest root image contrast-to-noise ratio and lowest SoS mean absolute error.**

***Index Terms*—capacitive micromachined ultrasonic transducer, CMUT, deep learning, thermoacoustics, photoacoustics, ultrasound, speed-of-sound, U-Net, non-contact, soil moisture**

I. Introduction

Intelligent agriculture through the deployment of sophisticated sensors has allowed plant biologists, geneticists, breeders, and farmers to make data-driven decisions – providing benefits like efficient resource usage, improved crop yield, and identification of better plant cultivars. However, while above-ground phenotyping tools have been successfully deployed at scale, most approaches for below-ground sensing today are either high-resolution, lab-based systems that do not translate well to field settings or field-based techniques that are destructive and invasive as well as labor, cost, and infrastructure intensive [1].

A system that enables non-invasive, high-throughput, dynamic measurements of below-ground traits could prove to be the next big advance in plant phenotyping [2]. Large volumes of high fidelity images of root-system architectures could facilitate the development of new root-focused cultivars that can be designed to withstand increasing environmental stresses due to climate change as well as help in carbon sequestration efforts to combat global warming [3]. In addition, field-scale monitoring of the spatial profile of the water content in soil could provide actionable insights for many precision agriculture and resource management applications.

This work was supported by Advanced Research Projects Agency-Energy Grant DE-AR0000825. Code is available at https://github.com/maxlwang/mtl-sos-recon.git.

*A. Singhvi, M. L. Wang and A. Fitzpatrick contributed equally to this work.

Towards that end, we present a non-contact, multi-physics based imaging system that leverages the thermoacoustic (TA) effect to provide high resolution below-ground information at scale [4]–[6]. A conceptual view of the operation of the proposed system is shown in Fig. 1, wherein a hybrid microwave excitation and ultrasound detection approach provides good imaging contrast as well as high resolution. However, the rhizosphere is a highly heterogeneous environment making high-fidelity below-ground TA imaging challenging. One such cause of heterogeneity is the varying below-ground soil moisture distribution which results in spatially varying speed-of-sound (SoS) profiles [7], [8].

Previously, we've developed an iterative SoS reconstruction algorithm for TA sensing systems in [6], but unlike tomographic SoS reconstruction approaches in biomedical imaging [9], the inverse problem formulated in [6] is ill-posed because all below-ground targets emit ultrasound (US) simultaneously. Thus, the algorithm proposed in [6] used known reference targets and did not capture lateral variations in the SoS. To overcome these limitations, in this work we design deep learning networks that can achieve better accuracy and robustness in reconstructing arbitrary SoS distributions for precise soil moisture mapping and concurrently enable high-fidelity below-ground image reconstruction.

II. Deep Learning in Ultrasound Imaging

Deep learning techniques have been implemented in a wide range of ultrasound imaging applications [10]. For example, the U-Net architecture has proved effective in image segmentation and denoising applications [11], [12]. More recently, [13]–[15] have demonstrated that encoder-decoder models using raw

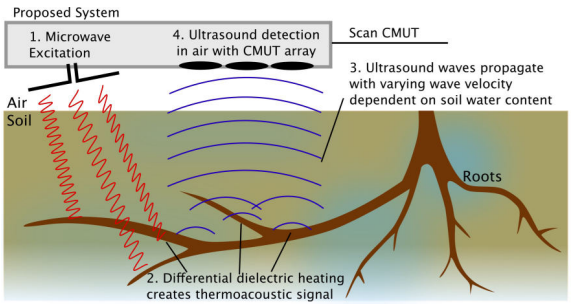

Fig. 1. Conceptual view of the non-contact thermoacoustic system for soil moisture sensing and below-ground imaging in heterogeneous underground environments.

2021 IEEE International Ultrasonics Symposium (IUS) | 978-1-6654-0355-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/IUS52206.2021.9593748

978-0-7381-1209-1/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: Stanford University. Downloaded on March 27,2022 at 16:52:20 UTC from IEEE Xplore. Restrictions apply.

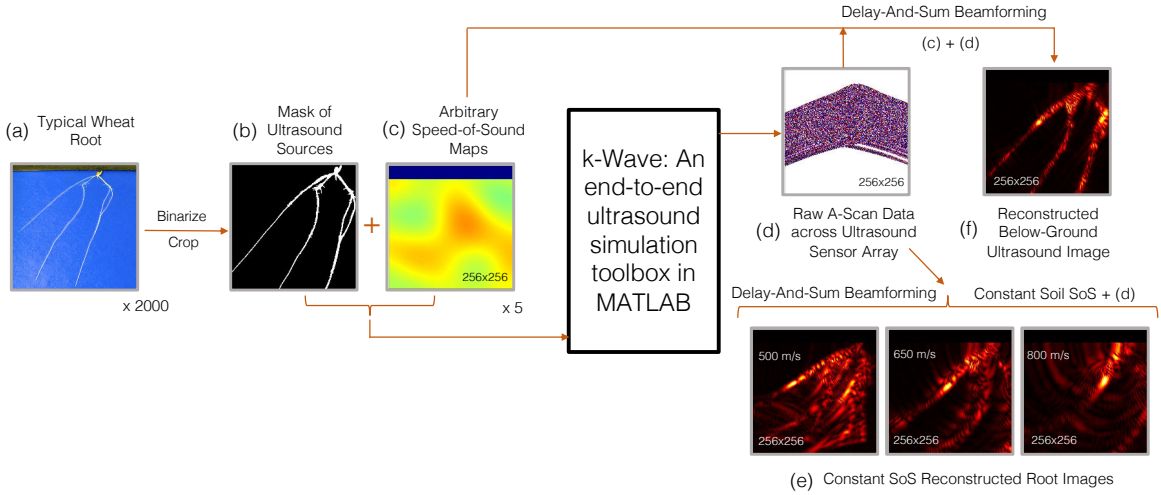

Fig. 2. Dataset generation and processing pipeline using the k-Wave toolbox with example components of the dataset displayed for each stage.

received US signals as image inputs can be deployed directly for biomedical SoS and US image reconstruction.

In [13] and [14], the authors use US imaging signals from a synthetic dataset containing ellipsoid targets with different SoS corresponding to different organs. These signals are fed into a U-Net based architecture and trained to reconstruct the SoS map. Note that although the inputs here are US signals, conventional US involves two-way propagation with US waves reflecting off of interfaces with different SoS while TA imaging involves one-way propagation of US waves emanating from regions with dielectric contrast and traveling through a heterogeneous SoS medium. As a result, in conventional US imaging, the existence of reflected US signals is highly correlated with changes in the SoS, whereas in TA imaging, the SoS only affects the variation in US time-of-flight across different receivers of the array, making it a more challenging inverse problem.

To increase segmentation accuracy and provide more information, [15] used a multi-task learning (MTL) network to output both a reconstructed US image as well as a segmentation mask from received US signals. By combining the image reconstruction and segmentation into an MTL network, the mutual information can help to improve both segmentation and image reconstruction accuracy. In a similar vein, here, we explore how MTL networks for simultaneous SoS mapping and TA image reconstruction can increase accuracy in both tasks compared to a conventional U-Net architecture, with details about specific architectures discussed in Section IV.

III. Custom Dataset Generation

While datasets for SoS reconstruction using tomographic and pulse-echo approaches for biomedical imaging are readily available, given the novelty of the proposed application, large volumes of US data in soil and soil moisture maps do not currently exist. Thus, we synthesize a custom dataset that models realistic scenarios in order to train and test the neural networks we explore in this work. We use the k-Wave toolbox [16] to perform end-to-end acoustic simulations in MATLAB.

For our k-Wave setup, we assume an imaging area of 45 cm $\times$ 45 cm. Within this region, as illustrated in Fig. 1, we insert a root structure to act as an in-soil TA source. We use online root image datasets [17], specifically choosing 2,000 wheat root images captured from wheat grown in an artificial growth medium [18], and photographed as seen in Fig. 2(a).

These images are passed through a data processing pipeline to crop, resize, and binarize them to fit in a 45 cm $\times$ 45 cm window. The resulting image is shown in Fig. 2(b), with the white pixels acting as in-soil TA sources in the k-Wave simulation. Each source is driven with a coded pulse excitation similar to [19]. Five different arbitrary SoS distributions with SoS ranging from 500 m/s to 800 m/s (Fig. 2(c)) are generated for each of the 2,000 root structures, resulting in an augmented dataset containing 10,000 samples in total.

The capacitive micromachined ultrasonic transducer (CMUT) sensor array, shown in Fig. 1, is modeled to have 256 receive elements, placed at a 5 cm standoff in air, operating at a 100 kHz US frequency with a 20 kHz bandwidth [20]. A-Scan data from the k-Wave simulation is captured by each of the 256 CMUTs after which Gaussian white noise is added such that the SNR is 20 dB. The A-Scan data is then time-gain compensated and then stacked together to form a 2D image as shown in Fig. 2(d). Finally, the A-Scan data along with the SoS map is used to reconstruct a ground-truth image of the root structure as shown in Fig. 2(f). As an additional part of our data processing pipeline, we also generate a set of reconstructed root images assuming various constant underground SoS, as shown in Fig. 2(e) which are then stacked to form a multi-channel image. The image reconstruction for both Fig. 2(e),(f) is done using an efficient Fast Marching Method based delay-and-sum algorithm [6].

Next, the reconstructed images, SoS maps as well as A-Scan data are further cropped to a 30 cm $\times$ 30 cm window to avoid large anomalies present at the simulation boundaries

Authorized licensed use limited to: Stanford University. Downloaded on March 27,2022 at 16:52:20 UTC from IEEE Xplore. Restrictions apply.

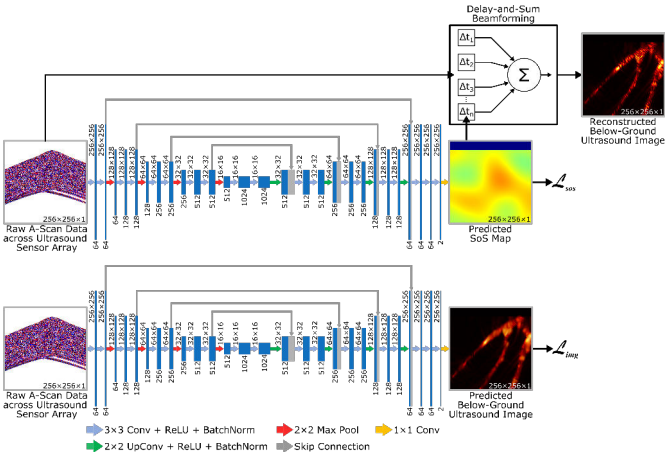

Fig. 3. U-Net architecture used as a baseline for comparison, either with SoS map output (top) or root image output (bottom).

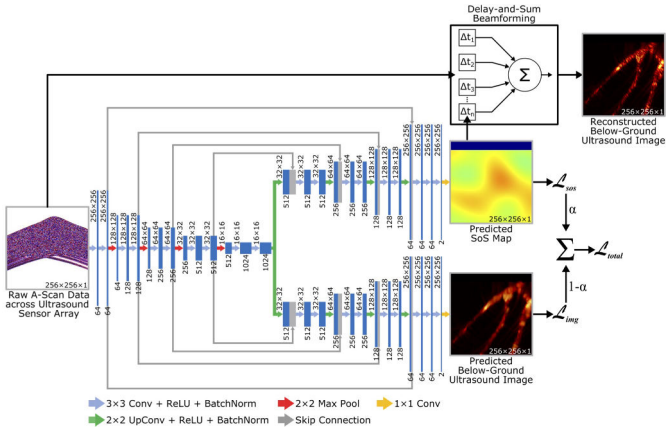

Fig. 4. SIMO U-Net architecture for simultaneous SoS map and root image outputs.

and to make the inverse problem more tractable. Before it is entered into the neural network, the A-Scan data is normalized and down-sampled to a sampling frequency of $\approx$ 250 kHz, resulting in a 256 $\times$ 256 A-Scan image. The outputs of the neural network include a normalized SoS map and reconstructed image of the root target, both of which are also 256 $\times$ 256 images corresponding to a pixel size of $\approx$ 1.5 mm^2.

IV. Network Architectures

A. Baseline

As a baseline for comparison, we use the basic U-Net architecture shown in Fig. 3. An image representation of the raw A-Scan data is input to an encoder which consists of multiple 3×3 convolutional layers, batch normalization, and 2×2 max pooling layers for downsampling. After the encoder, a decoder which consists of multiple 2×2 upsampling convolutional layers and skip connection concatenations from the encoder, outputs either a SoS map or a predicted root image (Fig. 3). For an additional comparison to the predicted image, the A-Scan data and predicted SoS map can be input to a delay-and-sum beamformer to generate a reconstructed root image using a more conventional reconstruction pipeline.

B. Single-Input, Multi-Output Network

Due to the relationship between SoS and image reconstruction, we hypothesized that simultaneously predicting the SoS and image could benefit both tasks. To modify the U-Net architecture for multi-task learning, the decoder is duplicated such that one decoder output is the predicted SoS map and the other is the predicted root image. Skip connections from the encoder layers are added to both SoS and image decoder layers. This single-input, multi-output (SIMO) network is depicted in Fig. 4.

C. Multi-Input, Multi-Output Network

While image reconstruction assuming a constant SoS in the soil results in poor quality root images, the variations in these images for different constant SoS values encode

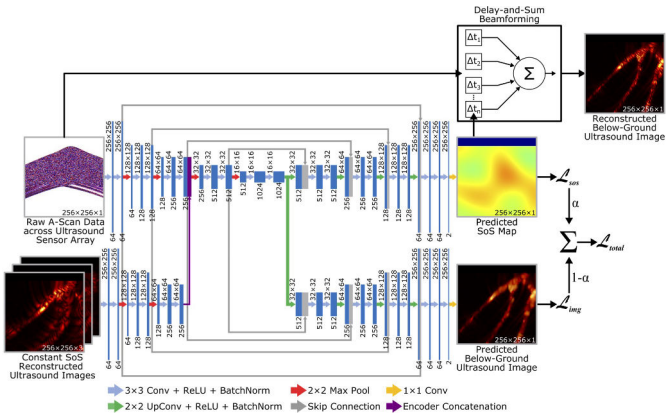

Fig. 5. MIMO U-Net architecture using both A-Scan data and assumed constant SoS root images for simultaneous SoS map and root image outputs.

information about both the true SoS profile and root structure. To provide this additional information to the network, we use a multi-input, multi-output (MIMO) network which includes an auxiliary encoder branch (Fig. 5). This additional branch takes in a three-channel image, where each channel is a reconstructed image assuming a different constant SoS (500, 650, and 800 m/s). The auxiliary encoder branch uses 2 downsampling steps before being concatenated with the main encoder branch. The skip connections are only taken from the main encoder branch.

D. Training

The 10,000 simulation results were divided into train, validation, and test datasets with 9,000, 500, and 500 samples respectively. The networks were all trained using the Adam optimizer [21] and mean square error (MSE) loss function over 100 epochs. Training was run on a system with an Intel Xeon E5 processor and an Nvidia Tesla V100 GPU with 16 GB VRAM. For the SIMO and MIMO networks, the total loss was the equally-weighted average of the SoS and image MSE losses. After each training epoch, the validation set was used to evaluate the degree of overfitting in the network. The final evaluation of the models was run on the test set using the weights from the epoch with the lowest validation loss.

Authorized licensed use limited to: Stanford University. Downloaded on March 27,2022 at 16:52:20 UTC from IEEE Xplore. Restrictions apply.

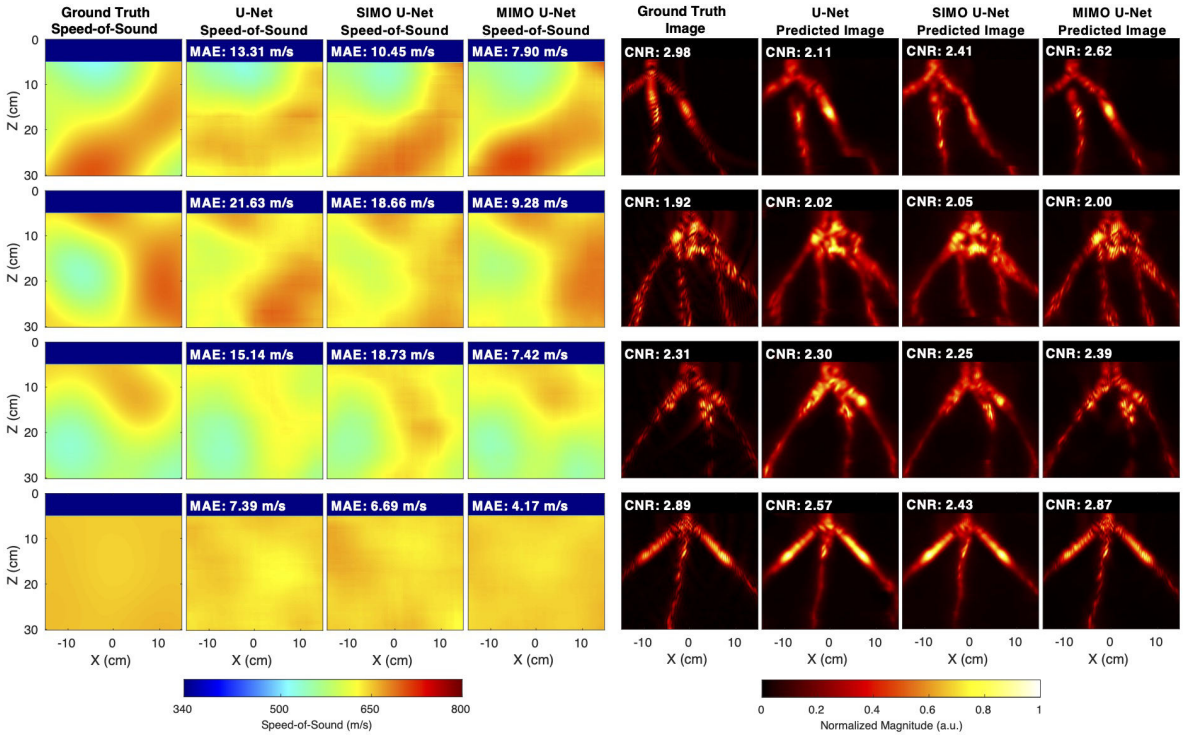

Fig. 6. Qualitative performance comparison of SoS and root image prediction across network architectures.

V. Results

To evaluate the performance of the above network architectures, we qualitatively comment on example predictions as well as quantitatively benchmark key metrics across the entire test set. Fig. 6 displays example predictions from the three network architectures. In comparison to the baseline U-Net architecture, it is clear that SoS prediction is enhanced by the SIMO U-Net. Whereas the baseline U-Nets individually predict SoS and root image from the TA measurements, exploiting the co-dependence of the two desired outputs through a shared encoder branch provides the network with additional context and reduces the ill-posed nature of SoS prediction. That being said, performance on root image prediction was largely unimproved using the SIMO network.

We hypothesized that while the co-dependent outputs assist the network in learning a better feature extraction from the TA measurement data, it is possible that this resource sharing between the two tasks ultimately was a bottleneck for further performance enhancement. This hypothesis led to a third architecture: the MIMO U-Net. As seen in Fig. 6, the MIMO U-Net significantly improves performance both on SoS and root image prediction – even when there are large spatial variations in the SoS distribution.

To quantify the relative performance of these three architectures, we analyze the mean absolute error (MAE) of the predicted SoS distributions and the contrast-to-noise ratio (CNR) of the predicted root images:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|SoS_{pred} - SoS_{truth}|, \tag{1}$$

$$CNR = \frac{\mu_{root} - \mu_{background}}{\sigma_{background}}, \tag{2}$$

where SoS_{pred} and SoS_{truth} are the predicted and ground-truth SoS maps, N is the number of pixels in the SoS map, μ_{root} is the mean value of the root pixels in the images, and $\mu_{background}$ and $\sigma_{background}$ are the mean and standard deviation of the background pixels in the images. Root versus background pixels were differentiated by using the root binary masks (Fig. 2(c)); the background is considered all pixels in an image that are not within the root mask.

The MAE and CNR of each model is summarized in Table I as mean $\pm$ standard deviation. For all metrics contained within, the MIMO U-Net model performs the best. In fact, the MIMO U-Net even predicts root images with higher average CNR than the ground-truth reconstructed images (2.40 ± 0.54); this is a result of the artifact suppression in the predicted images that is most easily noted in the second row of Fig. 6. Column (4) notes what percentage of predicted root images in the test set have higher CNR (CNR_{pred}) than the CNR of the image reconstructed (CNR_{const}) if we assume the soil has a constant SoS of 650 m/s. This metric demonstrates the value of our proposed approach for reconstructing higher quality root images from the TA measurements. Column (5) notes what percentage of predicted root images in the test set have higher CNR than the CNR of the image reconstructed (CNR_{recon})

Authorized licensed use limited to: Stanford University. Downloaded on March 27,2022 at 16:52:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON OF PERFORMANCE METRICS FOR THE U-NET, SIMO, AND MIMO ARCHITECTURES

	(1) Average MAE (m/s)	(2) Average CNR_{pred}	(3) Average CNR_{recon}	(4) $\text{CNR}_{pred} > \text{CNR}_{const}$	(5) $\text{CNR}_{pred} > \text{CNR}_{recon}$
U-Net	15.02 ± 6.00	2.30 ± 0.43	2.13 ± 0.43	97%	73%
SIMO	14.68 ± 5.93	2.39 ± 0.60	2.16 ± 0.48	98%	78%
MIMO	13.94 ± 5.76	2.43 ± 0.43	2.15 ± 0.40	99%	92%

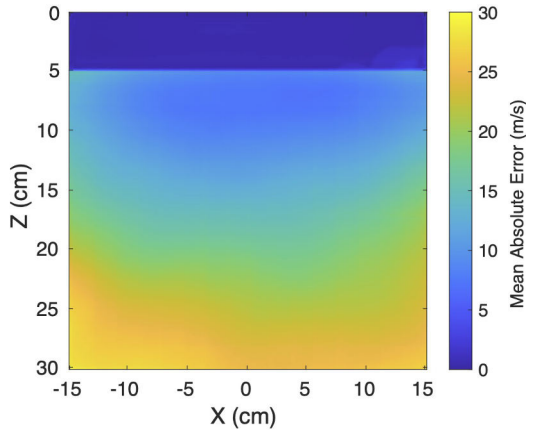

Fig. 7. Averaged spatial MAE across predicted SoS maps by the MIMO U-Net on the test set.

if we use the predicted SoS as input to the delay-and-sum algorithm. This metric demonstrates the value of the multi-task learning for simultaneous image reconstruction and SoS mapping rather than simply SoS mapping with subsequent conventional image reconstruction techniques.

VI. CONCLUSION

We have developed two MTL networks based on the U-Net architecture for predicting below-ground SoS maps and root images using TA imaging. Comparisons of the different networks showed that the MIMO U-Net performed consistently well for both tasks. Fundamentally, the more acoustic paths that intersect a region of the soil, the more information about the SoS at that location that is embedded in the TA measurements. While the MIMO network achieved reasonable accuracy, Fig. 7 shows that most of the SoS errors were concentrated towards the bottom of the imaging window corresponding to a region with the fewest acoustic paths due to the sparsity of root structures at depth. Since the relationship between soil moisture content and SoS is soil-type dependent, soil characterization and calibration could be performed for converting SoS to soil moisture in the field. Future work includes training the network on a variety of root types and experimentally validating network performance in the field for non-invasive, high-throughput soil moisture mapping and root biomass sensing applications.

REFERENCES

[1] H. F. Downie *et al.*, "Challenges and opportunities for quantifying roots and rhizosphere interactions through imaging and image analysis," *Plant, cell & environment*, vol. 38, no. 7, pp. 1213–1232, 2015.

[2] A. P. Wasson *et al.*, "Beyond digging: noninvasive root and rhizosphere phenotyping," *Trends in plant science*, vol. 25, no. 1, pp. 119–120, 2020.

[3] K. Paustian *et al.*, "Assessment of potential greenhouse gas mitigation from changes to crop root mass and architecture," Booz Allen Hamiltion Inc., McLean, VA (United States), Tech. Rep., 2016.

[4] A. Singhvi *et al.*, "Non-contact thermoacoustic sensing and characterization of plant root traits," in *2019 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2019, pp. 1992–1995.

[5] ——, "A microwave-induced thermoacoustic imaging system with non-contact ultrasound detection," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 2019.

[6] A. Fitzpatrick, A. Singhvi, and A. Arbabian, "Spatial reconstruction of soil moisture content using non-contact thermoacoustic imaging," in *2020 IEEE Sensors*. IEEE, 2020, pp. 1–4.

[7] W. Brutsaert and J. N. Luthin, "The velocity of sound in soils near the surface as a function of the moisture content," *Journal of Geophysical Research*, vol. 69, no. 4, pp. 643–652, 1964.

[8] F. Adamo *et al.*, "An acoustic method for soil moisture measurement," *IEEE transactions on instrumentation and measurement*, vol. 53, no. 4, pp. 891–898, 2004.

[9] J. Jose *et al.*, "Speed-of-sound compensated photoacoustic tomography for accurate imaging," *Medical physics*, vol. 39, no. 12, pp. 7262–7271, 2012.

[10] R. J. G. van Sloun, R. Cohen, and Y. C. Eldar, "Deep learning in ultrasound imaging," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 11–29, 2020.

[11] R. Almajalid *et al.*, "Development of a deep-learning-based method for breast ultrasound image segmentation," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1103–1108.

[12] T. Tong *et al.*, "Domain Transform Network for Photoacoustic Tomography from Limited-view and Sparsely Sampled Data," *Photoacoustics*, vol. 19, no. May, p. 100190, 2020.

[13] M. Feigin, D. Freedman, and B. W. Anthony, "A deep learning framework for single-sided sound speed inversion in medical ultrasound," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 1142–1151, 2020.

[14] F. K. Jush *et al.*, "Dnn-based speed-of-sound reconstruction for automated breast ultrasound," in *2020 IEEE International Ultrasonics Symposium (IUS)*, 2020, pp. 1–7.

[15] A. A. Nair *et al.*, "Deep Learning to Obtain Simultaneous Image and Segmentation Outputs from a Single Input of Raw Ultrasound Channel Data," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2493–2509, 2020.

[16] B. E. Treeby and B. T. Cox, "k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of biomedical optics*, vol. 15, no. 2, p. 021314, 2010.

[17] "Quantitative plant," https://www.quantitative-plant.org/dataset.

[18] J. A. Atkinson *et al.*, "Combining semi-automated image analysis techniques with machine learning algorithms to accelerate large-scale genetic studies," *GigaScience*, vol. 6, no. 10, p. gix084, 2017.

[19] A. Singhvi, A. Fitzpatrick, and A. Arbabian, "Resolution enhanced non-contact thermoacoustic imaging using coded pulse excitation," in *2020 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2020, pp. 1–4.

[20] B. Ma *et al.*, "Wide bandwidth and low driving voltage vented cmuts for airborne applications," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 66, no. 11, pp. 1777–1785, 2019.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

Authorized licensed use limited to: Stanford University. Downloaded on March 27,2022 at 16:52:20 UTC from IEEE Xplore. Restrictions apply.