RTS: Towards Underground Root Tuber Sensing via RF Sensor Networks

TAO WANG, YANG ZHAO*, ZHIBIN HUANG, and JIE LIU, Harbin Institute of Technology (Shenzhen), China and State Key Laboratory of Smart Farm Technologies and System, China
YANG ZHONG, Chinese Academy of Agricultural Sciences, Shenzhen, China

Underground root tuber sensing (RTS) is important for monitoring crop phenotypic traits in crop breeding and other smart agriculture applications. This paper proposes a novel RTS framework with a radio frequency (RF) sensor network and deep learning models, demonstrating the "see-through soil" capability of RF sensor networks in underground RTS. We build upon an RF tomography network system and propose a novel data-driven RTS model, TD-RTS, that uses transformer and diffusion neural networks for imaging cross-sections of potato root tubers. Furthermore, we propose a biomass sensing model by combining the transformer network in TD-RTS with a multilayer perceptron (MLP) to estimate the biomass of underground tubers. To achieve accurate sensing, we use both the frequency and spatial diversities of the networked sensing system in RTS, and use fade-level to facilitate the selection of RF channels in order to reduce the data processing overhead. We perform extensive experiments, demonstrating the efficacy of the RTS framework.

CCS Concepts: • **Computer systems organization → Sensor networks**.

Additional Key Words and Phrases: Cyber-physical systems, Radio frequency tomography, Image reconstruction.

ACM Reference Format:
Tao Wang, Yang Zhao, Zhibin Huang, Jie Liu, and Yang Zhong. 2026. RTS: Towards Underground Root Tuber Sensing via RF Sensor Networks. 1, 1 (February 2026), 31 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

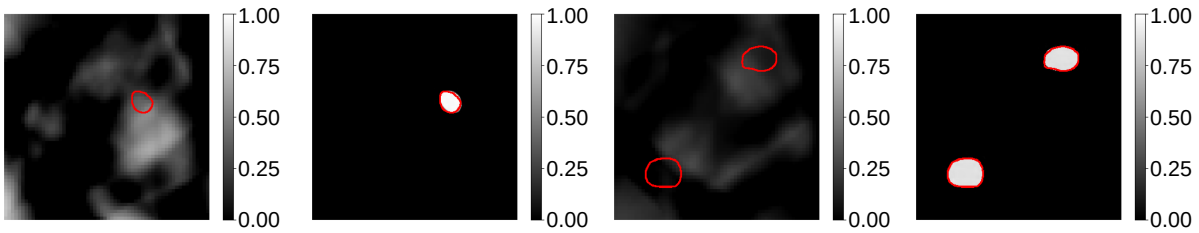
One significant factor contributing to food insecurity is the potential shortfall in crop yields to meet the demands of the growing population. Many research institutes are developing techniques required for future food production and food security without destroying our planet [14, 75]. Plant phenotyping is such an enabling technique that will contribute to making sustainable agriculture available, and crop biomass is an important phenotypic trait in crop breeding [56]. While various sensors and methods have been proposed for crop above-ground biomass (AGB) sensing, sensing techniques for below-ground biomass (BGB), defined as the biomass of live roots excluding fine roots less than 2 mm in diameter [55], remain largely understudied [8].

While cameras and computer vision techniques have been used to determine the size and shape of leaves, stems and the overall above-ground architecture of a single plant crop in greenhouse environments [14, 36], RF sensors provide a potentially cost-effective way of capturing the shape, size and location information of underground

*Yang Zhao is the corresponding author.

Authors' Contact Information: Tao Wang, hitszwangtao@gmail.com; Yang Zhao, yang.zhao@hit.edu.cn; Zhibin Huang, fingerh12138@gmail.com; Jie Liu, jieliu@hit.edu.cn, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong, China and State Key Laboratory of Smart Farm Technologies and System, China; Yang Zhong, zhongyang@caas.cn, Chinese Academy of Agricultural Sciences, Shenzhen, Shenzhen, Guangdong, China.

ACM XXXX-XXXX/2026/2-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) from a physical model. (b) from our DNN model. (c) from a physical model. (d) from our DNN model.

Fig. 1. Imaging results for (a) a physical RF tomography model [10] in a single-tuber scenario, (b) our DNN model in a single-tuber scenario, (c) a physical RF tomography model [10] in a double-tuber scenario, and (d) our DNN model in a double-tuber scenario. Red circles indicate the ground truth of the 2D cross-section areas of the potato root tubers.

roots, *e.g.*, potato tubers, due to their see-through soil capability. For example, [43] uses a ground penetrating radar (GPR) to detect and reconstruct images of below-ground roots. It proposes a deep neural network (DNN) to locate the root branches in each 2D GPR sensing slice, and then reconstructs root structures from multiple scans. However, the sensing range of a GPR sensor is constrained by the field of view of its antennas, and the applicability is hindered by its high cost. In this paper, we propose an alternative approach, a novel root tuber sensing (RTS) framework, which uses RF measurements from low-cost RF sensor networks for underground root sensing of potted plants in greenhouse environments.

The idea of proposing an RTS framework is inspired by RF tomography, a "see-through" sensing technique, which uses changes in received signal strength (RSS) measurements to detect and track objects even through walls [72, 83]. The goal of RF tomography is to determine an attenuation image, quantifying the influence caused by physical objects within the sensing area. While RF tomography was first proposed to detect and track human beings, different varieties of RF tomography methods were later proposed to retrieve properties of different objects, such as the inner structure of pillars [49], rice moisture levels [4], etc. However, when we directly use RF tomography for underground RTS, we find the following changes and issues. First, most previous model-based RF tomography methods focus on the localization and tracking of targets, whereas root tuber sensing requires extracting the shape and size information. Moreover, since RF tomography is essentially an ill-posed inverse problem, the higher the resolution of the reconstructed image is, the more serious the ill-posedness of the inverse problem would be. Thus, it is challenging for traditional physical model-based algorithms to obtain fine-grained imaging results. For example, when we apply the attenuation-based RF tomography algorithm [10], the evaluation metric, structural similarity index (SSIM), only reaches 0.45 (an example image shown in Fig. 1a), which is far below the result from our data-driven RTS framework. Additionally, RTS requires the capability of detecting multiple tubers underground. While physical model-based RF tomography methods can detect multiple targets [44], the imaging quality is further degraded because the effects of multiple targets on RF measurements are not linear combinations of those of individual targets. As shown in Fig. 1c, two tubers can be vaguely seen from the reconstructed RF tomography image, and the SSIM score calculated with ground truth is 0.36, even lower than the single-tuber case.

Second, although data-driven methods have achieved state-of-the-art (SOTA) performance in wireless sensing and show great potential for sensing underground tubers [46, 65], their performance degrades in dynamic environments, where environmental changes introduce varying multi-path effects on wireless signals. For example, human activities create time-varying multi-path effects [13] on wireless signals. As shown in Fig. 2, human activities induce significant short-term fluctuations in RSS measurements, degrading DNN model performance.

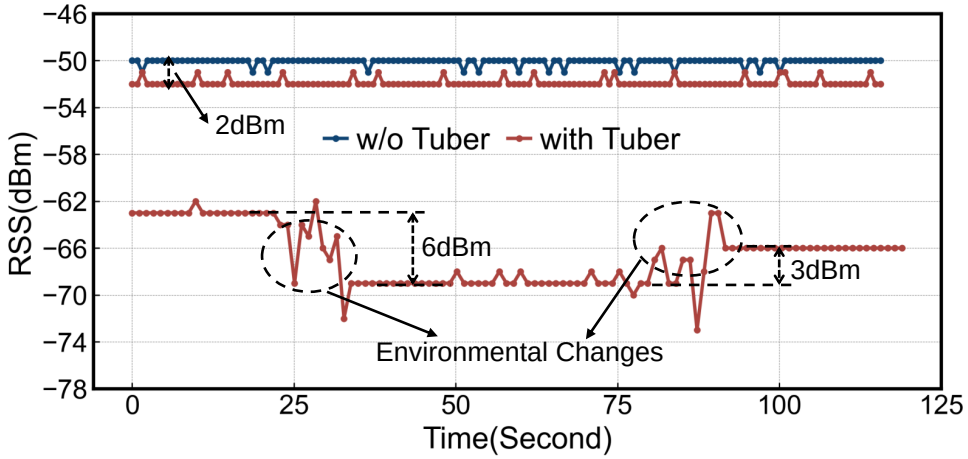


Fig. 2. RSS measurements in different environments: In a static environment, a 2 dB difference in RSS measurements is shown between scenarios with and without a root tuber. In a dynamic environment, RSS variations of 3 dB and 6 dB are observed due to environmental changes.

In addition, environmental layout changes also lead to RSS variations [21]. Fig. 2 shows a 2 dB difference in RSS measurements between scenarios with and without a tuber. However, it also shows variations of 3 dB and 6 dB in RSS measurements due to environmental layout changes, which can produce false root tuber imaging results.

Third, while multi-frequency sensor data enhance the accuracy of sensing approaches by providing additional information, they also increase the computational burden of the method and the overhead of the sensing system. For example, the study in [65] uses RSS measurements from multiple frequency channels, and increasing the number of channels from 8 to 16 not only doubles the data collection time and power consumption of sensor nodes, but also increases the deep learning neural network (DNN) model parameters and computation resource requirements of monitoring devices. Thus, it is important to find a trade-off between sensing accuracy and adequate channel measurements, especially for long-term underground tuber monitoring on resource-constrained embedded computing devices.

To solve the issues mentioned above, we propose an RTS framework using an RF sensor network and DNN models. First, we build upon our data acquisition system with a ZigBee sensor network [66], and perform extensive experiments in various scenarios, *e.g.*, different tuber counts and different environmental conditions, to build an underground potato root tuber sensing dataset. The dataset includes 58 potato tubers of varying sizes and shapes. Both the frequency and spatial diversities of the networked sensing system can be used to investigate how different root tubers affect RF measurements under various positions and environmental conditions.

Second, with the tuber sensing dataset, we propose a two-stage DNN model called TD-RTS to locate underground root tubers and reconstruct their maximum cross-sections using RF measurements from the RF sensor network. Transformer neural networks have shown greater performance in various fields, such as computer vision [22, 54] and wireless sensing [53]. By using the self-attention mechanism, they capture the dependencies between different frequency channels [84], allowing the model to learn complex relationships and extract discriminative features from RSS measurements. Accordingly, TD-RTS combines transformer and convolution networks for initial imaging in the first stage. In the second stage, a latent diffusion network [74] is used to remove noise that persists in initial images. Diffusion networks have achieved SOTA performance in various computer vision tasks, such as image synthesis, restoration, and denoising [24, 34, 41]. In TD-RTS, the diffusion network

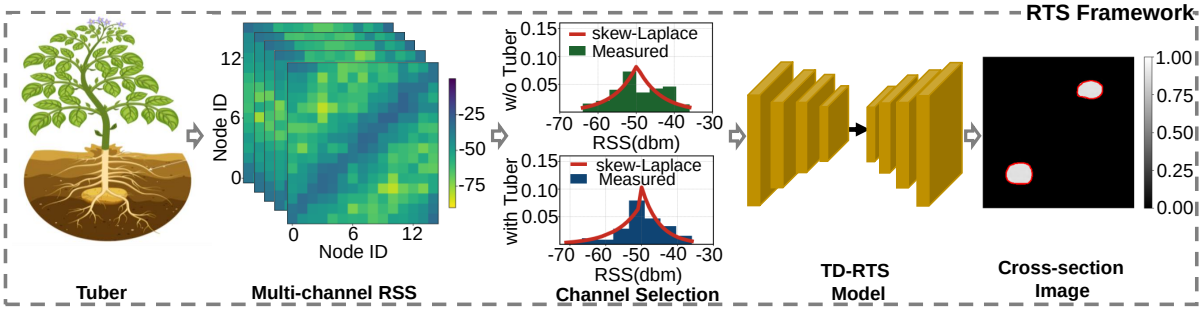


Fig. 3. Overview of the root tuber sensing (RTS) framework. An RF sensor network monitors underground tubers, with multi-channel RSS measurements collected and input into the sensing model to reconstruct a tuber cross-section image. To reduce the overhead of the sensing system, a channel selection method is proposed to select channels that are more sensitive for underground tubers.

encodes 2D images into 1D vectors to perform denoising, reducing memory and computation requirements while maintaining high performance. Furthermore, we propose a biomass sensing model by combining the transformer network in TD-RTS with an MLP to estimate the biomass of underground tubers. In addition, we propose a statistical method and a one-shot fine-tuning method to ensure robust imaging in a dynamic environment. The statistical method detects environmental changes [37, 69, 80], while the fine-tuning method [76] updates the neural networks online to maintain robust imaging.

Finally, to reduce the overhead of the sensing system, we propose a channel selection method based on the fade-level metric of wireless channels [73]. Based on our observations, RF measurements from links operating on deep fade channels are more sensitive to interference from regions outside the sensing area, whereas those from links on anti-fade channels are better fitted for underground sensing [30]. Accordingly, we use the fade level of each frequency channel to select anti-fade channels for underground tuber sensing, thereby reducing the number of channels used in image reconstruction and biomass estimation. Our experimental results demonstrate the efficacy of this channel selection method. To the best of our knowledge, this RTS framework is the first RF sensor network-based framework towards underground tuber imaging and biomass estimation.

In summary, this paper makes the following contributions.

- We propose RTS, an underground root tuber sensing framework with a low-cost RF sensor network. Both the frequency and space diversities of the RF tomography network are used in the framework, and an underground potato root sensing dataset is built by our extensive experiments.
- We propose DNN-based models for imaging root tuber cross-sections and estimating root tuber biomass. The imaging model uses a two-stage neural network to generate fine-grained images of underground potato tubers, along with an environmental change detection method and a one-shot fine-tuning method to ensure robust imaging in a dynamic environment.
- We propose a fade-level-based channel selection method to reduce the number of required channels, thereby improving the efficiency of the networked sensing system, while maintaining high accuracy.
- We perform extensive real-world experiments and evaluate our method using over 900,000 network link measurements. The proposed DNN models demonstrate superior imaging quality and biomass estimation accuracy compared to SOTA baselines.

2 Problem Statement and Overview

2.1 Problem Statement

Considering a two-dimensional sensing area surrounded by S RF sensor nodes, an RF tomography network is formed with $M = S(S-1)$ RF links, where the nodes operate on C frequency channels. The RSS measured by the receiving node of link i at time t on frequency channel c can be described as [29]:

$$y_{i,c}[t] = P_c - L_{i,c} - H_{i,c}[t] + F_{i,c}[t] - G_{i,c}[t], \tag{1}$$

where P_c is the transmit power, $L_{i,c}$ is the larger scale path loss, $F_{i,c}$ is the fading gain, $G_{i,c}$ is the measurement noise, and $H_{i,c}$ is the shadowing loss caused by objects blocking the signal propagation path. Note that the transmit power P_c is constant for all links operating on the same frequency channel, and the larger scale path loss $L_{i,c}$ remains unchanged over time. Therefore, we use a single subscript index for P_c and two subscript indices for $L_{i,c}$, respectively. By considering all links and channels, we construct the RSS data matrix $\mathbf{Y}$, where each column corresponds to an RF link and each row represents a frequency channel.

We use an image vector $\mathbf{r} = [r_0, \cdots, r_{N-1}]^T$ to represent the sensing area, where N denotes the pixel number of the image vector, and r_n is a measure of the current presence of the target, *i.e.*, root tuber, in pixel n. Since the presence of root tubers in the sensing area attenuates and reflects RF signals, based on [83] and [71], the relationship between the image vector $\mathbf{r}$ and the corresponding RSS data can be modeled as:

$$\mathbf{Y} = \mathcal{H}(\mathbf{r}) + \mathbf{b}, \tag{2}$$

where $\mathcal{H}$ is an observation function and $\mathbf{b}$ is the model error and measurement noise.

For root tuber cross-section imaging, we aim to find the inverse function $\mathcal{H}^{-1}$ of $\mathcal{H}$, that is, to use RSS measurements $\mathbf{Y}$ as input to generate the image vector $\mathbf{r}$. Previous works [29, 71, 83] model $\mathcal{H}$ as a linear function and compute $\mathcal{H}^{-1}$ by solving an inverse problem. As shown in Fig. 1a and Fig. 1c, the physical model-based method [83] provides a coarse location estimate for underground tubers but struggles to capture fine-grained information of tuber dimension and shape. Moreover, as the number of tubers increases, the imaging quality degrades significantly. In this paper, we propose to use a DNN model to learn $\mathcal{H}^{-1}$, the mapping between RSS data and tuber cross-section images. Specifically, we aim to train a DNN model $\mathcal{F} : \mathbf{Y} \rightarrow \mathbf{r}$ to estimate the image vector $\mathbf{r}$:

$$\hat{\mathbf{r}} = \mathcal{F}(\mathbf{Y}; \Theta), \tag{3}$$

where $\hat{\mathbf{r}}$ is the estimate of $\mathbf{r}$, and Θ represents the set of neural network parameters, which are iteratively optimized using various loss functions, as detailed in Section 3. After training, the optimized network $\mathcal{F}$ serves as $\mathcal{H}^{-1}$ to reconstruct cross-section images from RSS data.

2.2 Framework Overview

Fig. 3 provides an overview of the RTS framework, which includes the following modules. First, a data acquisition testbed is built to collect sensing data and ground truth. The testbed uses an RF sensor network to capture multi-channel RSS measurements from underground tubers. Additionally, the testbed uses a novel sensing toolkit, including "plug-and-play" containers and a rotating platform, to enhance the diversity of dimensions and locations of the underground tubers. Using the testbed, we perform extensive data collection experiments under various conditions, including the single-tuber case, the double-tuber case, and dynamic environments, thereby building a comprehensive dataset. Further details of the testbed, experiments, and dataset are provided in Section 4.

Second, a channel selection method is proposed to reduce the overhead of the sensing system. This method uses the fade-level metric [73] to classify frequency channels into anti-fade and deep-fade categories. Fig. 3 presents a sample of the RSS value histograms from an anti-fade channel, along with the fitted results using the skew-Laplace function [73], under conditions with and without tubers. The histograms and fitted results show

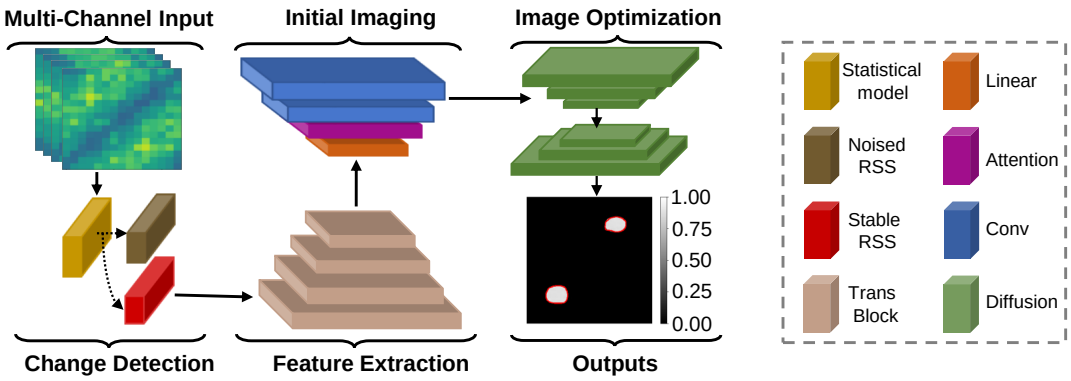


Fig. 4. Network architecture of the TD-RTS model. "Conv" and "Linear" denote the convolutional and linear layers, respectively. "Trans Block" represents the transformer block.

noticeable differences under different conditions, indicating that the anti-fade channel provides high sensing quality for detecting underground tubers. Thus, we use the fade-level-based method to select anti-fade channels for sensing tubers, reducing data collection time and the computational burden of the sensing model. Further details of the method are provided in Section 3.2, and additional analysis and results are presented in Section 5.7.

Third, a novel DNN model called TD-RTS is proposed to reconstruct cross-section images of underground tubers using multi-frequency RSS data. The model consists of multiple components: an environmental change detection component to identify changes in dynamic environments and update the pre-trained model online, a feature extraction component to learn high-dimensional features from RSS data, an initial imaging component to generate cross-section images of underground tubers, and an image optimization component to reduce residual noise in the imaging results. The architecture of TD-RTS is shown in Fig. 4 and more details are provided in Section 3.1. Additionally, to address the occasional blurriness of underground tuber edges in reconstructed images, TD-RTS uses the canny algorithm [11] as a post-processing step to detect edges and define cross-section regions of tubers. The region bounded by the detected edges is defined as the cross-section area of a tuber, with pixel values set to 1, while pixels outside this region are assigned a value of 0. This post-processing step facilitates the use of various evaluation metrics, which are discussed in Section 5.1.1.

Finally, missing data due to wireless interference is a common issue in networked sensing. To address this issue, we use the most recent packets to perform data imputation in our data preprocessing module. For instance, when a sensor node fails to receive packets from other nodes, we impute the missing RSS values using the latest data on the same frequency channel. If all packets in a particular frequency channel are lost, we use the most recent values within the same channel for imputation.

3 Algorithms

In this paper, we propose two novel algorithms: the TD-RTS model for accurate underground root tuber imaging and a channel selection method to reduce the overhead of the RF networked sensing system. More details are discussed below.

3.1 TD-RTS model

3.1.1 The Feature Extraction Component. We propose to use a transformer network [63] to model the relationships between different frequency channels and extract discriminative features from multi-frequency RSS data. Before

being input into the transformer network, the RSS data vector $\mathbf{Y}_c$ on frequency channel c is mapped using a linear projection, and the result is summed with a learnable channel embedding vector, as described by:

$$\mathbf{F}_c = \mathcal{L}(\mathbf{Y}_c) + \mathbf{E}, \tag{4}$$

where $\mathcal{L}$ represents the linear projection and $\mathbf{E}$ represents the learnable channel embedding vector, which is shared across all frequency channels. Additionally, to generate the global feature for multi-frequency RSS data, we introduce an additional learnable vector $\mathbf{Y_0}$, which has the same dimension as $\mathbf{Y_c}$. The output $\mathbf{F}_0$, derived from the linear projection and summation with the channel embedding, is concatenated with outputs from other channels to form a feature matrix $\mathbf{F} = [\mathbf{F}_0, \mathbf{F}_1, \cdots, \mathbf{F}_C]$.

Subsequently, $\mathbf{F}$ is fed into the transformer network. As shown in Fig. 4, this network consists of four transformer blocks, each containing a self-attention module and a feed-forward module. The self-attention module is a key component within the transformer block, capturing relationships between different frequency channels and generating refined attention features for multi-frequency RSS data. Specifically, three matrices $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are derived from the input matrix $\mathbf{F}$ through linear transformations [63]:

$$(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{F} \cdot (\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v), \tag{5}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are learnable weight matrices. $\mathbf{Q}$ and $\mathbf{K}$ are used to compute the normalized attention weight matrix $\mathbf{A} = \mathcal{S}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})$, where $\mathcal{S}$ represents the softmax function and d represents the column dimension of $\mathbf{Q}$ and $\mathbf{K}$. The matrix $\mathbf{A}$ is used to multiply $\mathbf{V}$, followed by the application of a layer normalization function and a residual connection to produce the attention feature matrix.

The attention feature matrix is fed into the feed-forward module, followed by layer normalization and a residual connection to generate the output matrix, which is then input into the next transformer block. Finally, the first row vector of the output matrix of the last transformer block is used as the global feature and processed through a linear layer to generate the output of the transformer network.

To further bridge the sensing-to-agronomy gap, the transformer network is also used to directly estimate the biomass of underground root tubers. Specifically, the features generated by the transformer network are fed into an additional multilayer perceptron (MLP) for biomass estimation. To further evaluate its performance, we perform extensive evaluations, with results presented in Section 5.6.

3.1.2 The Initial Imaging Component. To achieve initial imaging of underground tubers, we propose a novel neural network that incorporates attention and convolution layers. The attention layer adaptively emphasizes tuber-related features while attenuating irrelevant ones. Specifically, a high-dimensional vector from the feature extraction component is first fed into a linear layer to adjust its dimension. The output is then reshaped into a two-dimensional feature map, with its width and height scaled to one-ninth of the dimensions of the target image. Subsequently, an attention layer, implemented by a learnable weight matrix, adaptively adjusts the feature map. After that, the feature map is upsampled, increasing both its length and width by a factor of three. A 3×3 convolution layer is applied to smooth the interpolated result and maintain consistent feature map dimensions. Finally, the feature map is upsampled to the target size, and two 1×1 convolution layers are used to refine the output. In this paper, bilinear interpolation [33] is used for all upsampling operations. To enhance the imaging results, this network is jointly optimized with the network in the feature extraction component using a mean squared error (MSE) loss function, which is defined as [58]:

$$L_{mse} = \frac{1}{U} \sum_{u=1}^{u=U} ||\mathbf{r}_u - \mathcal{F}_{ini}(\mathcal{F}_{ext}(\mathbf{Y}_u))||^2, \tag{6}$$

where $\mathcal{F}_{ini}$ denotes the network used in the initial imaging component, $\mathcal{F}_{ext}$ represents the network used in the feature extraction component, and U denotes the number of training samples in a batch.

3.1.3 The Image Optimization Component. To further optimize the initial imaging result, we propose an image optimization component implemented using a latent diffusion network. This diffusion network compresses the initial imaging result into a feature vector and uses this vector to perform the diffusion process, thereby enhancing imaging quality while reducing model parameters and iterations.

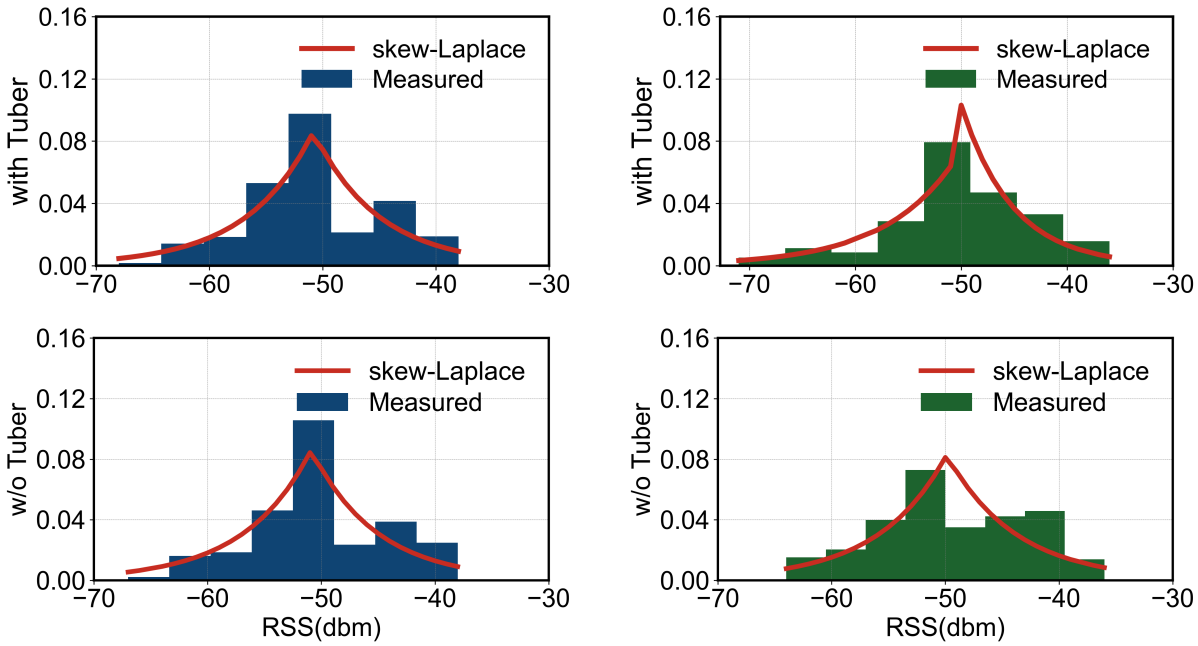
Specifically, the diffusion network first uses an encoder-decoder module to compress the image from the previous component into a one-dimensional feature vector, which is then decoded to produce a high-quality result. The structure of the encoder-decoder module follows that proposed by [74], including a prior representation learning network (PRN) and an Unet-shaped transformer network (UTN). PRN takes the concatenation of an initial image $\hat{\mathbf{r}}$ and the corresponding ground truth $\mathbf{r}$ as input to generate the prior representation $\mathbf{z} = PRN(\hat{\mathbf{r}}, \mathbf{r})$. The prior representation $\mathbf{z}$ is input into UTN along with the image sample $\hat{\mathbf{r}}$ as a dynamic modulation parameter to denoise the feature map, ensuring the decoding of a high-quality result. In this paper, PRN is implemented using a ViT network [18]. Both PRN and UTN are first optimized using an MSE loss function.

Subsequently, the diffusion process is applied to the prior representation obtained from PRN to generate $\mathbf{z}$ without relying on the ground truth $\mathbf{r}$. In the forward diffusion process, the prior representation from PRN is progressively corrupted by Gaussian noise, resulting in a noised result $\mathbf{z}_j$ at time step j. In the reverse diffusion process, a denoising network ϵ, consisting of linear layers, estimates the noise in $\mathbf{z}_j$. The inputs of ϵ are $\mathbf{z}_j$, j, and $\mathbf{o}$, where $\mathbf{o}$ is a condition vector used to control the reverse diffusion process. To generate $\mathbf{o}$, we define another prior representation learning network, PRN_2, which takes only the image $\hat{\mathbf{r}}$ as input. The estimated noise from ϵ is used to obtain $\mathbf{z}_{j-1}$ and start the next iteration. After J iterations, the estimated result $\hat{\mathbf{z}}$ is obtained and used as the prior representation, which is fed into UTN along with $\hat{\mathbf{r}}$ to produce the denoised image. We jointly train PRN_2, ϵ, and UTN using denoising losses from both ϵ and UTN. The mean absolute error (MAE) [70] and MSE are used as their respective loss functions.

In the inference phase, we begin by extracting a conditional vector $\mathbf{o}$ using PRN_2 from a test image sample. Then, we randomly sample a Gaussian noise $\mathbf{z}_J$, and the denoising network utilizes $\mathbf{z}_J$ and $\mathbf{o}$ to estimate $\hat{\mathbf{z}}$ after J iterations. The $\hat{\mathbf{z}}$ is used as the prior representation to input UTN along with the test image sample to generate a high-quality result.

3.1.4 The Environmental Change Detection Component. Before feeding RSS data into the DNN model, we use a statistical method to detect dynamic changes in the environment. When changes are detected, the corresponding RSS data with significant variations are discarded, and stable data from the new environmental condition are used to update the pre-trained model. Specifically, we observe that RSS values of most links in the RF sensor network remain relatively stable under static environmental conditions. However, environmental changes, such as human activities or alterations in the layout, may cause significantly short-term variations in RSS values, as shown in Fig. 2. Thus, we propose a statistical method that uses the standard deviation of RSS data within a sliding window to detect dynamic changes in the environment. For each link on each frequency channel, we first calculate the standard deviation of RSS data within the sliding window. Subsequently, we rank these deviation values in descending order and calculate the average of the top-k standard deviations. When the average exceeds a threshold, it indicates the occurrence of environmental changes at that time. The corresponding RSS data are discarded, and the system continues monitoring for environmental changes in the subsequent period. Once no changes are detected, stable RSS data from the new environmental condition are used to fine-tune the pre-trained model to achieve robust imaging.

Specifically, we propose a one-shot fine-tuning method [76] to adjust the network parameters after detecting the environmental changes. In practice, the growth rate of root tubers is much slower than the short-term dynamic changes induced by human activities and alterations in the environmental layout. This indicates that the dimensions of tubers remain consistent before and after these dynamic changes, and allows us to use the most recent stable RSS data to fine-tune the pre-trained model without the need to relabel the ground truth.



(a) Deep-fade channel.

(b) Anti-fade channel.

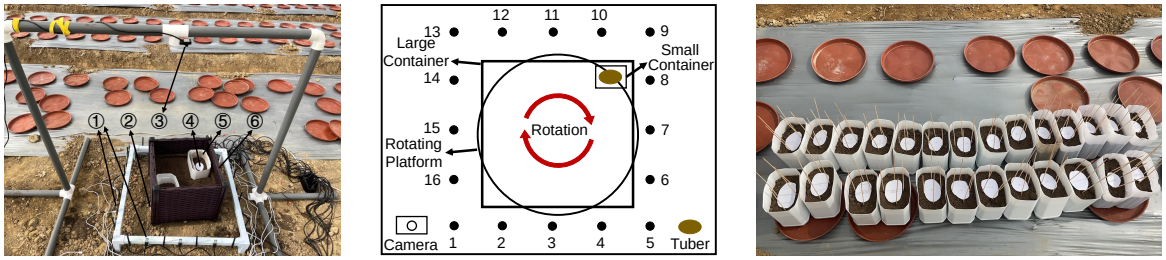
Fig. 5. Histograms of RSS values on deep-fade and anti-fade channels under conditions with and without tubers, along with fitted results using skew-Laplace [73]. The histograms and fitted results are similar for the deep-fade channel (a), while noticeable differences are observed for the anti-fade channel (b).

Specifically, when the statistical method detects environmental changes, the neural network parameters of the feature extraction and initial imaging components are fine-tuned using the most recent stable RSS data from a single tuber. In the fine-tuning process, the MSE loss function is used to optimize the neural networks, enabling them to adapt to new environments. Given that the fine-tuning process introduces additional computational cost, we perform further evaluations using different fine-tuning epochs to analyze the trade-off between fine-tuning time and sensing performance. The results are presented in Section 5.5.

3.2 Fade-level-based Channel Selection Method

To further reduce the overhead of a multi-channel RF sensor network, we propose a fade-level-based channel selection method, in which frequency channels are classified from anti-fade to deep-fade based on their fade levels [73]. Links on deep-fade channels are sensitive to interference from the environment, while links on anti-fade channels are better fitted for underground tuber sensing [30].

Histograms and fitted results using a skew-Laplace function [73] for RSS values from a deep-fade channel and an anti-fade channel, with and without tubers, are shown in Fig. 5. We see that the histograms and fitted results with and without tubers are similar on the deep-fade channel, while noticeable differences are observed on the anti-fade channel. This indicates that the anti-fade channel is more sensitive to underground tubers and provides higher sensing quality. Thus, we use the fade level of each channel to select anti-fade channels to reduce the number of channels.



(a) Testbed (b) Experimental layout (c) Small containers with various tubers.

Fig. 6. The used testbed comprises a ZigBee network with 16 nodes and a through-soil sensing toolkit featuring "plug-and-play" functionality and data augmentation capabilities. ① TI CC2531 nodes, ② Rotating platform, ③ RGB camera, ④ Marker with the same dimension as the tuber. ⑤ Small container with the tuber. ⑥ Larger container. Additional small containers containing potato tubers with various dimensions are shown in (c).

Specifically, we first collect a duration of RSS data under the non-tuber condition and calculate the fade level of each link on each channel, as formulated in [73]:

$$\eta_{i,c} = \overline{y}_{i,c} - \min_{c \in C} \quad \overline{y}_{i,c}, \tag{7}$$

where i and c represent the network link and the frequency channel indices, respectively. $\overline{y}_{i,c}$ denotes the mean RSS value of the link i on channel c during the data collection period. $\eta_{i,c}$ denotes the fade level of link i on channel c, with $\eta_{i,c} \geq 0$ for each link on each channel. Subsequently, we calculate the average fade level of the links on each channel to represent the fade level of the corresponding channel and then sort the channels in descending order. Finally, we select the top-k frequency channels for data collection from underground tubers, reducing the overhead of the RF sensor network while maintaining relatively high sensing performance. The evaluation results are discussed in Section 5.7.

Based on our investigation, although the channel selection method reduces sensing overhead, the selected channels may vary with changes in soil properties and moisture. This suggests that channel reselection should be performed as soil conditions change. To address this issue, we propose a scheme that periodically reselects channels. Specifically, the channels used in the current period are recorded. If the selected channels in the next period differ from the current ones, they are automatically updated, and the model is retrained using data from new channels. In practice, the time required for channel reselection and model retraining is much shorter than the growth period of underground tubers. This ensures that the shapes and positions of tubers remain unchanged during reselection and retraining. Consequently, the currently obtained imaging results and the data from the new channels can serve as ground truth and training data, respectively, for retraining the model without requiring additional data relabeling. Note that, we have performed experiments to investigate the stability of the selected channels, and more results are discussed in Section 5.7.

4 Experiments and Dataset

Using a multi-channel ZigBee RF sensor network, we perform extensive measurement campaigns and create an extended version of our previously published dataset [66], referred to as wireless potato sensing (WPS) 2.0. Further details are discussed below.

Table 1. Comparison of hardware cost, energy consumption, and sample size among our ZigBee sensor network, a GPR device [59], and a CT scanner [47].

	Cost (USD)	Power (W)	Battery	Mobility	Sample Size (KB)
GPR device [59]	39, 950	13.3	Yes	Yes	477.27
CT scanner [47]	150, 000	80×10^3	No	No	512.00
ZigBee network	10	2.82×10^{-3}	Yes	Yes	15.00

4.1 Data acquisition testbed

In this paper, we use the testbed proposed in [66], which consists of a ZigBee sensor network with 16 TI CC2531 sensor nodes, a through-soil sensing toolkit, a rotating platform, and an RGB camera, as shown in Fig. 6.

In addition to the ZigBee sensor network, other sensing solutions have been proposed for underground RTS in recent years. To further demonstrate the advantages of the ZigBee network, we compare it with a GPR system [59] and a computed tomography (CT) system [47] in terms of hardware cost, energy consumption, and sample size. As shown in Table 1, we first illustrate the hardware cost of these sensing systems. The GPR system uses a customized device manufactured by IDS GeoRadar, which generally costs more than standard commercial units, typically priced at USD 39,950. Moreover, the SIEMENS CT scanner used in [47] is considerably more expensive, with a used unit costing approximately USD 150,000. In contrast, the ZigBee sensor network used in our framework is low-cost, with each sensor node costing approximately USD 10. The low-cost advantage of the ZigBee network allows additional sensor nodes to be deployed with minimal increase in overall cost, enabling larger-scale data collection and improving adaptability to diverse sensing conditions. Second, we compare these sensing systems in terms of power consumption and battery support, which serve as indicators of energy efficiency. As shown in Table 1, each sensor node of the ZigBee network consumes 2.82 mW, substantially lower than 13.3 W and 80 kW consumed by the GPR and CT systems, respectively. The low-power design enables battery operation, enhancing mobility and providing a longer runtime than the GPR system under equivalent battery conditions. Third, we compare the sample size of different sensing solutions. Using the ZigBee sensor network with 16 nodes, each RTS network sample size is only 15 KB, which is significantly smaller than the corresponding sample sizes of the GPR and CT systems. The smaller sample size further accelerates data transmission and reduces storage requirements. These results demonstrate the superiority of our framework and highlight its practicality for real-world deployment. Note that we have tested using a subset of the 16 wireless nodes in our evaluation, as discussed in Section 5.8.

4.2 Experiments

We use the testbed with different configurations to perform measurement campaigns in various environments and conditions, and build an extended version of our previously published dataset [66], referred to as WPS 2.0. Table 2 lists all experiments in the measurement campaigns.

In Exp. 1, we perform experiments in a single-tuber case. The ZigBee nodes are deployed on a rack measuring 72 cm × 72 cm. We collect RSS data from 26 potato tubers, with lengths and widths shown in Table 2. The potato tubers weigh between 66.4 g and 209.8 g and are buried in a plastic container measuring 40 cm × 40 cm × 40 cm, with depths ranging from 11 cm to 13.5 cm. We rotate the platform 32 times for each tuber, positioning it at different angles and locations within the sensing area. In total, we collect RSS data with 832 unique tuber-position annotations, each corresponding to 40 seconds of RSS link measurements. The evaluation results for the single-tuber case are discussed in Section 5.3.

Table 2. Three sets of experiments to build WPS 2.0: imaging potato tubers in a single-tuber case, imaging potato tubers in a double-tuber case, imaging potato tubers in an environment with changes. A total of 922,384 RF sensor network measurements and 873 ground truth annotations are collected in these experiments.

Experiment	Tuber Dimensions (LxWxT) (cm)	RF network area (cm)	No. of tubers	No. of positions	No. of RF Measurements	Description
Exp. 1	L:5.3-9.5 W:4.5-6.5 T:4.0-6.4	72×72	26	32	822,608	Imaging of potato tubers in a single-tuber case.
Exp. 2	L:9.0-14.0 W:6.0-9.0 T:N/A	60×60	6	2	46,016	Imaging of potato tubers in a double-tuber case.
Exp. 3	L:2.0-10.0 W:1.0-8.0 T:1.5-6.3	72×72	26	1	53,760	Imaging of potato tubers in dynamic environments.

In Exp. 2, we perform experiments in a double-tuber case, with ZigBee nodes deployed in a 60 cm × 60 cm area. We collect RSS data from 6 tubers, with their lengths and widths provided in Table 2. During data collection, we randomly select two tubers and bury them at two predefined locations within the sensing area. In total, we use 15 tuber pairs and collect RSS data for 4 minutes per pair. The evaluation results for the double-tuber case are discussed in Section 5.4.

In Exp. 3, we perform experiments in dynamic environments. The ZigBee network is deployed in an indoor room, with the sensing area set to 72 cm by 72 cm. RSS data are collected from 26 potato tubers, with their dimensions presented in Table 2. These tubers are placed in a predefined position within the sensing area, with depths ranging from 10.7 cm to 15.5 cm. First, we collect RSS data from 26 potato tubers in the initial environment. Second, we randomly select five tubers with varying dimensions to collect RSS data in dynamic environments, enabling the evaluation of the sensing model. Specifically, during data collection, we continuously change the layout of the environment by moving furniture around the sensing area. Concurrently, human activities such as walking and replacing small containers are performed. We use $E_1 \sim E_4$ to represent four continuously changing environments, respectively: E_1 represents the initial environment, E_2 represents the environment after moving a paper box, E_3 represents the environment after placing a chair around the sensing area, and E_4 represents the environment after swapping the positions of the chair and the paper box. Overall, we construct three datasets from 5 potato tubers in three dynamic environments, along with an additional dataset from 26 potato tubers collected in the initial environment. Each dataset contains one minute of RSS data for each tuber. The evaluation results under dynamic environmental conditions are presented in Section 5.5.

Compared with the previously published dataset [66], WPS 2.0 includes data from greenhouse environments and multi-tuber scenarios, serving as an extension of the original dataset. Moreover, we collect RSS data from observation experiments to analyze the stability of the channel selection method, and from outdoor environments to evaluate the feasibility of our framework for outdoor deployments, with details provided in Sections 5.7 and 5.9, respectively. Note that our dataset is publicly available at http://zenodo.org/records/17477040.

4.3 Dataset Description

Our dataset contains over 900,000 ZigBee network measurements and over 800 ground truth annotations. We describe them in detail next.

4.3.1 Ground Truth Annotation. In contrast to the annotation method presented in [66], we propose a more accurate method for generating ground truth, as follows.

Table 3. Parameters used in the diffusion network.

Parameter description	Default Value
The learning rate of the diffusion neural network.	$1e^{-4}$
The input dimension of a training or testing sample.	16×240
The output dimension of a reconstructed image.	360×360
The number of layers used in UTN in the diffusion neural network.	8
The number of transformer blocks used in UTN in the diffusion neural network.	16
The number of attention heads used in UTN in the diffusion neural network.	32
The number of layers used in ViT (as PRN and PRN_2 in the diffusion neural network).	4
The number of attention heads in ViT (as PRN and PRN_2 in the diffusion neural network).	8
The patch size in ViT (as PRN and PRN_2 in the diffusion neural network).	18

First, the potato tuber is placed horizontally in a smaller container with soil, ensuring that its maximum cross-section remains parallel to the ground. One end of four sticks is inserted into the soil and fixed around the tuber, with the other end exposed above the soil surface, as shown in Fig. 6c. A marker with the same dimension as the potato tuber is placed on the soil surface within the region surrounded by the sticks, indicating the position and dimension of the underground potato tuber. When the potato tuber and container are rotated on the platform, a camera at a fixed location captures RGB images. An image segmentation algorithm [32] is then used to segment the pixels of the marker in the RGB image. Second, we establish a two-dimensional coordinate system for the sensing area, and the pixels corresponding to the marker in the RGB image are converted into coordinates within this system, representing the coordinates of the tuber cross-section in the sensing area. Third, we construct a ground truth image, where each pixel corresponds to a region in the two-dimensional coordinate system. Pixels corresponding to the coordinates of the tuber cross-section are assigned a value of 1, while all other pixels are assigned a value of 0.

4.3.2 ZigBee data. As shown in Table 2, we collect RSS data in three experiments. In Exp. 1, we record RSS data from 832 tuber-position pairs, generating 822,608 measurements with different frequency channels. In Exp. 2, we focus on a double-tuber case and collect a total of 46,016 ZigBee network measurements. In Exp. 3, the dataset can be divided into four subsets. The first subset contains RSS data from 26 fixed-position potato tubers in the initial environment, with a total of 34,016 network measurements. The second, third, and fourth datasets focus on environmental changes, which commonly occur in practical monitoring scenarios. RSS data are collected from 5 fixed-position potato tubers, resulting in a total of 19,744 network measurements across these datasets.

In all experiments, we use 58 potato tubers with varying shapes and dimensions, ensuring the diversity of potato tubers in the dataset. A total of 922,384 network measurements are collected.

5 Evaluation

To evaluate our TD-RTS model, we perform extensive experiments and use various metrics to assess imaging quality and biomass estimation accuracy for underground tubers. Furthermore, we compare its performance with that of different baseline models. More details are provided below.

5.1 Metrics and Model Parameters

5.1.1 Evaluation Metrics. We assess the imaging performance of TD-RTS using four metrics: structural similarity index (SSIM) [67], intersection over union (IoU) [20], equivalent diameter error (EDE) [7], and relative pixel difference (RPD). SSIM, ranging from 0 to 1, is a popular metric to quantify imaging quality, with higher values

indicating better reconstruction. IoU quantifies the similarity between the reconstructed cross-section image and its ground truth, accounting for both position and shape accuracy. It ranges from 0 to 1, with a higher value indicating greater similarity. EDE calculates the diameter of a circle whose area equals the absolute difference between the reconstructed and ground truth cross-section areas. In addition, we define RPD as the ratio of the pixel count difference between the reconstructed and ground truth cross-sections to the total pixel count of the ground truth cross-section. Both EDE and RPD quantify the accuracy of tuber cross-section reconstruction, with smaller values indicating better performance. Using the post-processing step described in Section 2.2, we can obtain the pixel counts for both the reconstructed and the ground truth cross-sections through simple summation. The cross-section area is subsequently obtained by multiplying the number of pixels by the real-world area represented by each pixel.

To assess biomass estimation performance, we use mean absolute error (MAE) [70], mean absolute percentage error (MAPE) [6], and root mean squared error (RMSE) [12] as evaluation metrics. First, MAE quantifies the average absolute difference between the estimated and true values, where lower MAE values indicate higher estimation accuracy. Second, MAPE expresses the absolute difference as a proportion of the true value, providing a more intuitive measure of estimation accuracy. Lower MAPE values correspond to higher accuracy. Third, RMSE computes the square root of the mean squared differences between the estimated and true values. It is widely used in estimation tasks because of its sensitivity to outliers [26]. Smaller RMSE values indicate better performance, with zero representing perfect estimation.

5.1.2 Model Parameters. To extract discriminative features for imaging, we use a transformer network with four transformer blocks, each containing a self-attention module with four attention heads and a feed-forward module composed of two linear layers. The output dimensions of the linear layers in the feed-forward module are set to 1024 and 256, respectively. The linear projection of the RSS data for each frequency channel is implemented using a linear layer with an output dimension of 256. The learnable channel embedding is implemented using a weight matrix with a dimension of 1×256. A linear layer with an output dimension of 1024 is used to generate the output of the transformer network. To further enhance imaging quality, a latent diffusion network is used, with its parameters listed in Table 3. To detect environmental changes, RSS data from 400 links on all frequency channels are used to calculate the average standard deviation. GELU is used as the activation function for the transformer network, while ReLU is used for the diffusion network and the network of the initial imaging component.

To perform biomass estimation, we employ a transformer network with eight blocks, each containing a self-attention module with a single attention head. The feed-forward module in each transformer block consists of two linear layers, each with an output dimension of 256. The output of the transformer network has a dimension of 256 and serves as the input to a two-layer MLP for biomass estimation, with hidden and output dimensions of 64 and 1, respectively.

5.2 Baselines

For comparison, we choose two SOTA data-driven imaging models as baselines [27, 46]. First, CNN-UNet [27] proposes a two-stage convolution neural network (CNN) for image reconstruction from RF data. The first stage uses a multilayer convolution network to generate initial images from measured data, while the second stage, relying on a UNet network [85], acts as a post-processing module to enhance the quality of the reconstructed image. Second, LSTM-UNet [46] captures relationships across frequency channels using an LSTM network [25] and generates high-dimensional features, which are fed into a UNet network to produce imaging results.

In addition, we select two SOTA biomass estimation models as baselines for comparison. The CNN-GRU model [64] combines a convolution neural network with a gated recurrent unit (GRU) network [15] for feature extraction, followed by fully connected layers for estimation. The CNN-Transformer model [19] first uses a

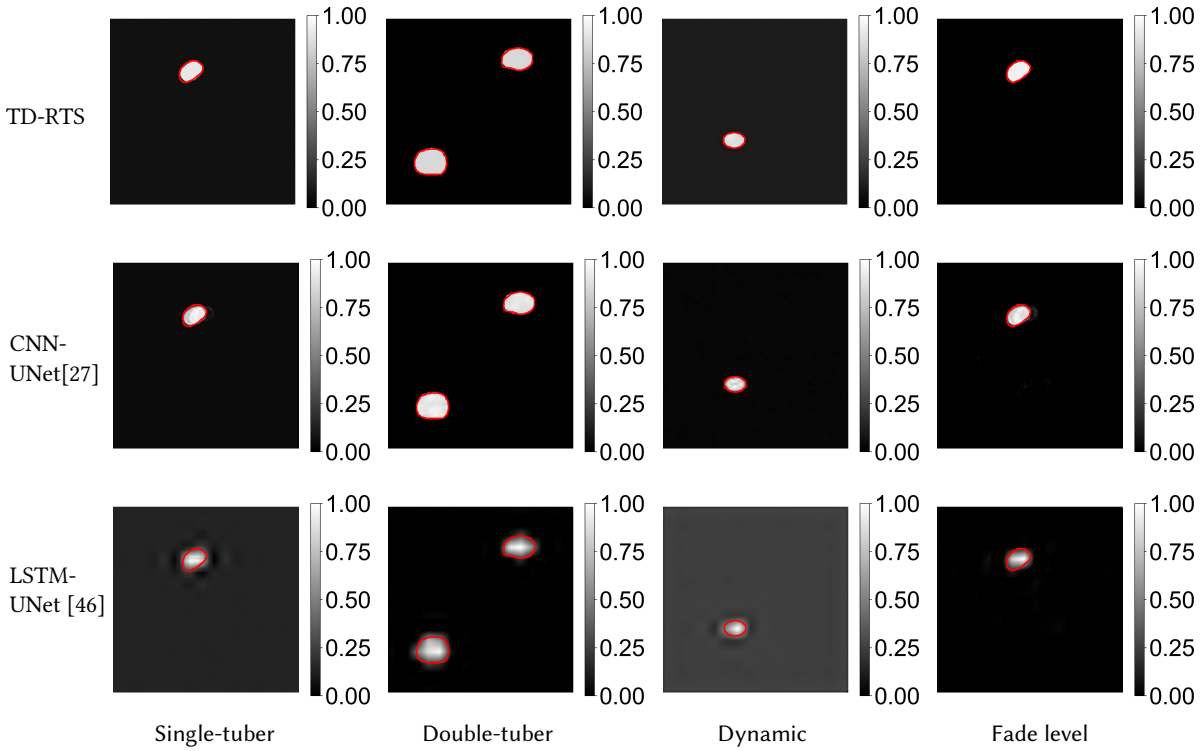


Fig. 7. Visualization results for different models and sensing scenarios. Each row corresponds to a method, while each column corresponds to a sensing case. "Dynamic" refers to using different models for imaging underground tubers in dynamic environments. "Fade level" indicates the use of the fade-level-based channel selection method to select three frequency channels for reconstructing images of underground tubers in the single-tuber scenario. Red circles indicate the ground truth of the 2D cross-section areas of the potato tubers.

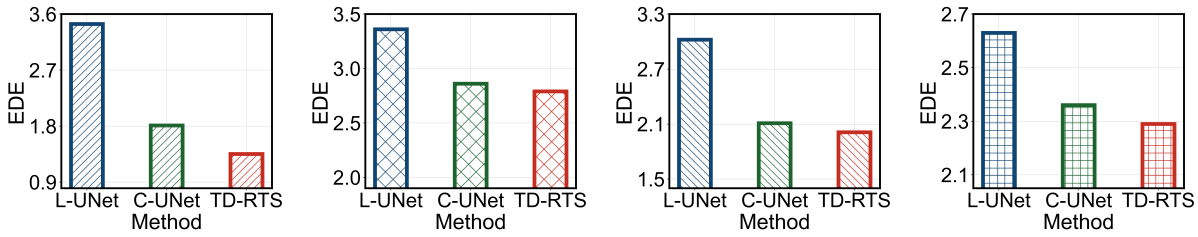
Table 4. Performance of TD-RTS in the single-tuber and double-tuber cases. We mark the best and second-best results using bold and underlined text, respectively.

Method \ Case	Single-tuber case				Double-tuber case			
	SSIM ↑	RPD ↓	IoU ↑	EDE(cm) ↓	SSIM ↑	RPD ↓	IoU ↑	EDE(cm) ↓
CNN-UNet [27]	**0.99**	<u>0.13</u>	<u>0.82</u>	<u>2.48</u>	**0.98**	**0.06**	<u>0.93</u>	<u>3.47</u>
LSTM-UNet [46]	0.95	0.23	0.73	3.34	0.80	0.11	0.83	3.97
TD-RTS	**0.99**	**0.08**	**0.86**	**2.01**	**0.99**	**0.05**	**0.94**	**3.05**

convolution neural network to extract local features from input data, which are then fed into a transformer network for global feature extraction and estimation.

Table 5. Leave-k-out performance of TD-RTS in the single-tuber case. The variable "k" represents the number of test tubers. The best and second-best results are indicated in bold and underlined text, respectively.

k value / Method	k=1			k=2			k=3			k=4		
	SSIM ↑	RPD ↓	IoU ↑	SSIM ↑	RPD ↓	IoU ↑	SSIM ↑	RPD ↓	IoU ↑	SSIM ↑	RPD ↓	IoU ↑
CNN-UNet [27]	<u>0.98</u>	<u>0.05</u>	<u>0.89</u>	<u>0.98</u>	<u>0.21</u>	<u>0.82</u>	<u>0.98</u>	<u>0.10</u>	<u>0.84</u>	<u>0.98</u>	<u>0.10</u>	<u>0.84</u>
LSTM-UNet [46]	0.93	0.19	0.76	0.93	0.29	0.73	0.93	0.21	0.73	0.92	0.13	0.78
Ours	**0.99**	**0.03**	**0.90**	**0.99**	**0.20**	**0.82**	**0.99**	**0.09**	**0.88**	**0.99**	**0.10**	**0.87**



(a) k=1 (b) k=2 (c) k=3 (d) k=4

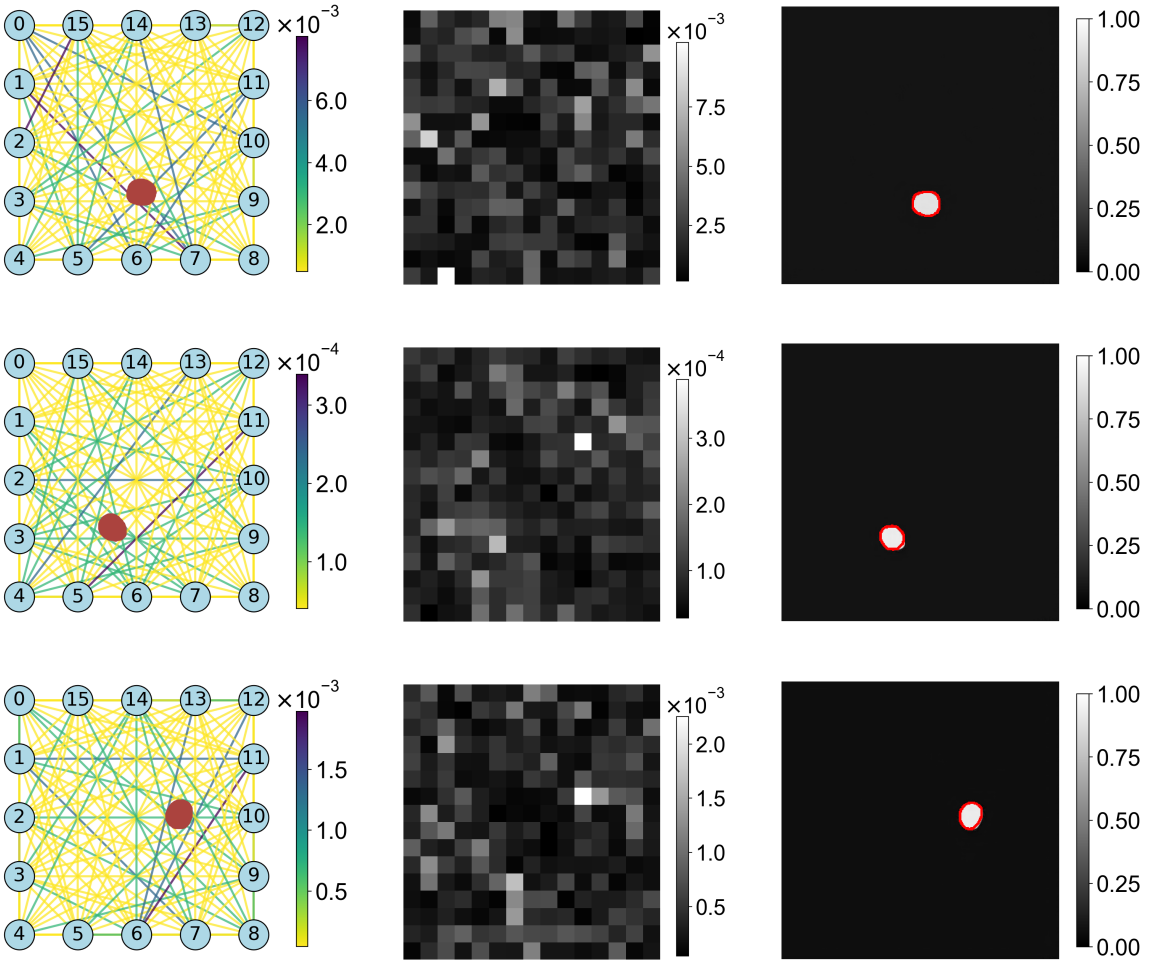
Fig. 8. Average EDE results in the single-tuber case for leave-k-out cross-validation with varying k values. The variable "k" represents the number of test tubers. "C-UNet" and "L-UNet" represent the CNN-UNet model [27] and the LSTM-UNet model [46], respectively.

Table 6. Computational cost of the TD-RTS model on different computing platforms, evaluated on a single RTS network sample and averaged over five runs.

	Parameters (MB)	FLOPs (G)	Power (W)	Energy (J)	Inference Time (S)
4090 GPU	71.16	25.53	79.17	37.35	0.47
10875H CPU	71.16	25.53	26.25	28.80	1.10
Jetson Nano	71.16	25.53	3.17	6.07	1.91

5.3 Performance in Single-tuber Case

First, we evaluate the performance of TD-RTS in the single-tuber case, where RSS data are collected from 26 potato tubers with varying dimensions and positions. We split tuber-position pairs in a 9:1 ratio for training and testing. As shown in the first column of Fig. 7, the visualization result of TD-RTS outperforms those of baseline models in both imaging quality and detection accuracy. As reported in Table 4, TD-RTS achieves an average SSIM value of 0.99, outperforming LSTM-UNet [46], which achieves 0.95. Moreover, TD-RTS achieves an average IoU value of 0.86, surpassing the 0.82 and 0.73 obtained by the baseline models. In addition, TD-RTS achieves average RPD and EDE values of 0.08 and 2.01, respectively, both lower than those of CNN-UNet [27] and LSTM-UNet [46]. Compared to LSTM-UNet [46], TD-RTS employs a two-stage neural network and uses a diffusion network as the post-processing component to enhance imaging quality and improve detection accuracy



(a) Sensor network (b) Saliency map (c) Estimation result

Fig. 9. Visualizations showing how the model maps RSS measurements to tuber locations and shapes. (a) The sensor network, in which each link weight represents the saliency of its RSS measurement, indicates that RSS measurements from links passing through or surrounding root tubers play a particularly critical role in imaging. (b) The saliency map, which quantifies the relevance of RSS measurements from all links to the estimation results, indicates that the sensor network produces distinct patterns for tubers located at different positions. (c) The reconstructed 2D cross-section images generated by TD-RTS, with red circles indicating the ground truth of the tuber cross-section areas.

for tubers of varying sizes and positions. Although CNN-UNet [27] also adopts a two-stage network, the diffusion network provides superior image refinement compared to the convolution network, further improving imaging quality and accuracy.

Second, we perform a leave-k-out evaluation using RSS data from 26 tubers, with k set to 1, 2, 3, and 4. As shown in Table 5, TD-RTS consistently achieves SSIM values above 0.98 across various configurations, demonstrating its

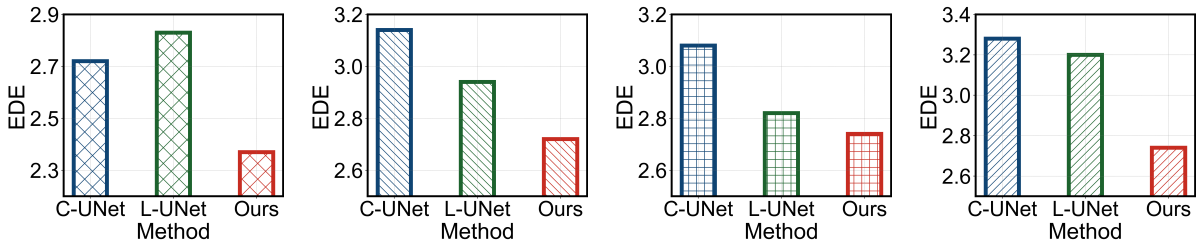
ability to generate high-quality cross-section images. In addition, TD-RTS outperforms the baseline models in terms of RPD and IoU, achieving average values of 0.11 and 0.87, respectively, across different configurations. As illustrated in Fig. 8, we also compare the EDE values obtained by different models. Our model exhibits lower EDE values in each sensing scenario compared to baseline models. These results not only demonstrate improvements over baseline models but also highlight the capability of TD-RTS for accurate imaging.

Third, to assess the feasibility of TD-RTS for real-time, low-power deployment, we evaluate its model size, floating-point operations (FLOPs), inference time, and energy consumption on three computing platforms: an NVIDIA RTX 4090 GPU, an Intel 10875H CPU, and an NVIDIA Jetson Nano edge device. Specifically, an RTS network sample comprising RSS measurements is used for evaluation, and the results are averaged over five runs. As shown in Table 6, inference on the RTX 4090 GPU is the fastest, requiring 0.47 seconds per sample and demonstrating the feasibility for real-time deployment. Inference on Jetson Nano is the slowest, taking 1.91 seconds per sample. However, Jetson Nano exhibits significantly lower power and energy consumption than the RTX 4090 GPU, demonstrating the feasibility for low-power deployment. In practice, although the inference time on Jetson Nano exceeds 1.9 seconds per sample, it remains significantly faster than the growth rate of underground tubers, confirming the practicality of TD-RTS for in-field deployment. In addition, the inference time and energy consumption on the Intel 10875H CPU fall between those of the RTX 4090 GPU and the Jetson Nano device, highlighting the versatility of TD-RTS across deployment platforms.

Fourth, to further enhance the interpretability of TD-RTS, we use saliency maps [62] to reveal how the model maps RSS measurements to tuber locations and shapes. In addition, since saliency maps demonstrate the contributions of RSS measurements from different wireless links to the reconstructed results, we use these contributions as link weights to visualize the mesh sensor network, highlighting the links that most influence the reconstruction. Specifically, an RTS network sample, comprising RSS measurements from distinct wireless links, is first fed into TD-RTS to generate a reconstructed image. The canny algorithm [11] is then used to detect edges in the reconstructed result, and the region enclosed by these edges is defined as the region of interest (ROI), corresponding to the cross-section area of the underground tuber. The estimated pixel values within ROI are used to compute the saliency of each RTS sample element. Finally, the saliency results of different RTS sample elements are assigned as link weights to visualize the mesh sensor network, highlighting the links most relevant to ROI. Fig. 9 shows three randomly selected imaging results, along with their corresponding saliency maps and mesh sensor networks. As shown in Fig. 9, when underground tubers are located at different positions, different RTS sample elements contribute most to estimating their locations and shapes. Moreover, using saliency results as link weights reveals that links passing through or surrounding tubers have the greatest impact on estimation. In summary, these results indicate that links are strongly affected by the locations and shapes of underground tubers, producing distinct RTS patterns captured by TD-RTS. The model then maps these patterns to tuber cross-section images.

5.4 Performance in Double-tuber Case

We evaluate the performance of TD-RTS in a double-tuber case, where RSS data are collected from tubers of varying dimensions placed at two fixed positions. We use a ratio of 9:1 to split the dataset for training and testing. As shown in the second column of Fig. 7, TD-RTS exhibits higher imaging quality and detection accuracy compared to the LSTM-UNet model [46]. Although both models capture relationships between frequency channels for imaging, the transformer network in TD-RTS is more effective than the LSTM network, and the diffusion network successfully removes residual noise, generating a more accurate result. Moreover, as shown in Table 4, TD-RTS outperforms the baseline models. For example, TD-RTS achieves an average IoU value of 0.94, surpassing the 0.93 and 0.83 obtained by baseline models. In addition, TD-RTS achieves an average EDE value of 3.05, reflecting improvements of 13.77% and 23.17% compared to CNN-UNet [27] and LSTM-UNet [46]. Note that we



(a) Environment 1 (b) Environment 2 (c) Environment 3 (d) Environment 4

Fig. 10. Average EDE values for various methods under different environmental conditions. "C-UNet" and "L-UNet" represent the CNN-UNet model [27] and the LSTM-UNet model [46], respectively.

Table 7. Performance of DNN models in a dynamic environment. $E_1 \sim E_4$ denote different environmental conditions. We mark the best and second-best results using bold and underlined text, respectively.

Test	Method	Leave-1-Out			Leave-2-Out			Leave-3-Out			Leave-4-Out		
		SSIM ↑	RPD ↓	IoU ↑	SSIM ↑	RPD ↓	IoU ↑	SSIM ↑	RPD ↓	IoU ↑	SSIM ↑	RPD ↓	IoU ↑
$E_1 \rightarrow E_1$	CNN-UNet [27]	0.98	0.15	0.86	0.97	0.18	0.84	0.98	0.18	0.84	0.97	0.20	0.80
	LSTM-UNet [46]	0.85	0.13	0.77	0.85	0.18	0.77	0.83	0.19	0.73	0.84	0.17	0.78
	TD-RTS	**0.99**	**0.11**	**0.90**	**0.99**	**0.15**	**0.87**	**0.99**	**0.15**	**0.87**	**0.99**	**0.14**	**0.88**
$E_1 \rightarrow E_2$	CNN-UNet [27]	0.97	0.21	0.80	0.95	0.27	0.76	**0.96**	0.22	0.79	0.95	0.26	0.71
	LSTM-UNet [46]	0.86	0.13	0.77	0.85	0.21	0.75	0.82	**0.19**	0.74	0.83	**0.21**	0.77
	TD-RTS	**0.99**	**0.12**	**0.88**	**0.98**	**0.19**	**0.84**	**0.96**	0.21	**0.83**	**0.98**	0.25	**0.80**
$E_2 \rightarrow E_3$	CNN-UNet [27]	0.95	0.25	0.78	0.92	0.25	0.76	0.93	0.20	0.80	0.88	0.22	0.75
	LSTM-UNet [46]	0.85	**0.11**	0.77	0.85	0.20	0.76	0.82	**0.18**	0.74	0.82	**0.16**	0.78
	TD-RTS	**0.98**	0.12	**0.88**	**0.98**	**0.19**	**0.84**	**0.98**	0.19	**0.84**	**0.98**	0.28	**0.79**
$E_3 \rightarrow E_4$	CNN-UNet [27]	0.95	0.19	0.80	0.92	0.25	0.75	0.92	0.23	0.76	0.87	0.29	0.69
	LSTM-UNet [46]	0.85	0.18	0.74	0.84	0.27	0.70	0.82	0.22	0.71	0.83	**0.18**	0.75
	TD-RTS	**0.98**	**0.12**	**0.88**	**0.98**	**0.19**	**0.84**	**0.96**	**0.20**	**0.84**	**0.97**	0.28	**0.79**

perform experiments in the double-tuber case to show the efficacy of TD-RTS for multi-tuber sensing scenarios, and we leave more evaluations of multi-target sensing as future work.

5.5 Performance in Dynamic Environments

To assess the robustness of our TD-RTS model, we continuously modify the environmental layout by randomly moving furniture three times during data collection. After detecting changes and removing noisy values, stable RSS data are recorded from four environmental conditions, denoted as E_1 to E_4. We use RSS data from E_1 to build a pre-trained model, which is then fine-tuned and tested on $E_2 \sim E_4$. During fine-tuning and testing, we select one tuber to fine-tune the pre-trained model and use the remaining four tubers for leave-k-out evaluations, where k ranges from 1 to 4. The third column of Fig. 7 shows visualization results generated by TD-RTS and the baseline models. Compared to LSTM-UNet [46], TD-RTS achieves higher detection accuracy and imaging

Table 8. Imaging performance and fine-tuning time across different fine-tuning epochs and leave-k-out scenarios. The best and second-best results are highlighted in bold and underlined text, respectively.

Epochs	Leave-1-Out			Leave-2-Out			Leave-3-Out			Leave-4-Out		
	RPD↓	IoU↑	Time(S)↓	RPD↓	IoU↑	Time(S)↓	RPD↓	IoU↑	Time(S)↓	RPD↓	IoU↑	Time(S)↓
5	0.18	0.85	**3.19**	0.22	0.82	**3.66**	0.23	0.81	**3.80**	**0.26**	0.79	**3.85**
25	0.12	0.88	15.91	**0.18**	0.84	18.15	**0.20**	**0.83**	18.96	0.27	0.79	18.94
50	0.12	0.88	30.11	**0.18**	0.84	36.22	**0.20**	**0.83**	36.97	0.27	**0.80**	37.34
75	**0.11**	0.88	44.65	**0.18**	**0.85**	54.91	**0.20**	**0.83**	55.52	0.27	**0.80**	55.46
100	**0.11**	0.88	59.88	**0.18**	**0.85**	72.08	**0.20**	**0.83**	73.66	0.27	**0.80**	73.60
150	**0.11**	**0.89**	91.52	**0.18**	0.84	107.95	**0.20**	**0.83**	110.75	0.27	0.79	112.80

quality. Although TD-RTS and CNN-UNet [27] exhibit similar imaging quality, TD-RTS achieves higher detection accuracy after fine-tuning.

Table 7 presents the SSIM, RPD, and IoU values of TD-RTS compared with the baseline models. TD-RTS achieves average RPD values of 0.14, 0.19, 0.20, and 0.20 for four environmental conditions, which are lower than those reported by CNN-UNet [27]. Meanwhile, TD-RTS achieves an average IoU of 0.85 across four environments and four evaluation ratios, outperforming the values of 0.74 and 0.79 reported by LSTM-UNet [46] and CNN-UNet [27], respectively. In addition, Fig. 10 illustrates the average EDE values across four evaluation ratios for each environment. TD-RTS exhibits lower EDE values for most environments compared to the baseline models. These evaluation results demonstrate the efficacy of TD-RTS in achieving robust imaging under environmental changes. Compared to baseline models, TD-RTS uses a one-shot fine-tuning method to update model parameters in dynamic environments, ensuring that it can automatically adapt to new environmental conditions.

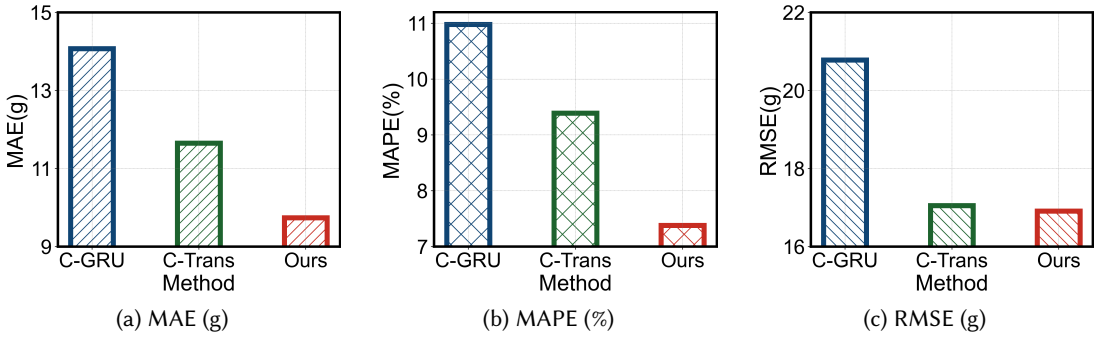
Given that the fine-tuning process introduces additional computational cost, the one-shot fine-tuning method updates only a subset of TD-RTS parameters for 50 epochs, requiring 35.81 seconds on average on an NVIDIA RTX 4090 GPU. To further evaluate the trade-off between fine-tuning time and sensing performance, we perform additional evaluations using different fine-tuning epochs, with results shown in Table 8. As expected, fine-tuning time increases with the number of epochs. Across four scenarios, fine-tuning takes an average of 3.67 seconds for 5 epochs and 35.81 seconds for 50 epochs. However, fine-tuning for 50 epochs achieves better imaging performance, and the total time remains below 40 seconds, which is much shorter than the growth period of underground root tubers. In addition, fine-tuning for 50 epochs achieves performance comparable to that obtained with a larger number of epochs, yielding an average IoU value of 0.84 across four evaluation scenarios. In practice, too few epochs may prevent the model from learning adequately, resulting in underfitting and poor adaptation to new environmental conditions. Conversely, too many epochs can lead to overfitting, reducing model stability and increasing fine-tuning time. Our setting not only reduces fine-tuning time but also maintains effective imaging.

5.6 Performance on Biomass Estimation

In this study, we combine the proposed transformer network with an MLP network for biomass estimation. First, we perform a leave-k-out evaluation using RSS data from 26 tubers, where k is set to 2 and 3. As shown in Table 9, the proposed model achieves average MAE values of 10.04 and 12.21 for the two configurations, respectively, both lower than those achieved by the baseline models. Moreover, it achieves superior MAPE values in each configuration, reaching 8.70% and 9.75%, respectively. In addition, although RMSE values of all models increase with the number of test tubers, the proposed model consistently achieves the best performance across configurations. These results not only demonstrate the efficacy of the proposed model in biomass estimation

Table 9. Leave-k-out performance for biomass estimation. The variable "k" represents the number of test tubers. The best and second-best results are indicated in bold and underlined text, respectively.

Method \ Leave-k-Out	Leave-2-Out			Leave-3-Out		
	MAE (g) ↓	MAPE(%) ↓	RMSE (g) ↓	MAE (g) ↓	MAPE(%) ↓	RMSE (g) ↓
CNN-GRU [64]	15.87	10.74	18.59	17.54	13.42	20.95
CNN-Transformer [19]	<u>10.55</u>	<u>9.32</u>	<u>12.77</u>	<u>12.96</u>	<u>11.30</u>	<u>15.28</u>
Ours	**10.04**	**8.70**	**12.04**	**12.21**	**9.75**	**14.60**



(a) MAE (g) (b) MAPE (%) (c) RMSE (g)

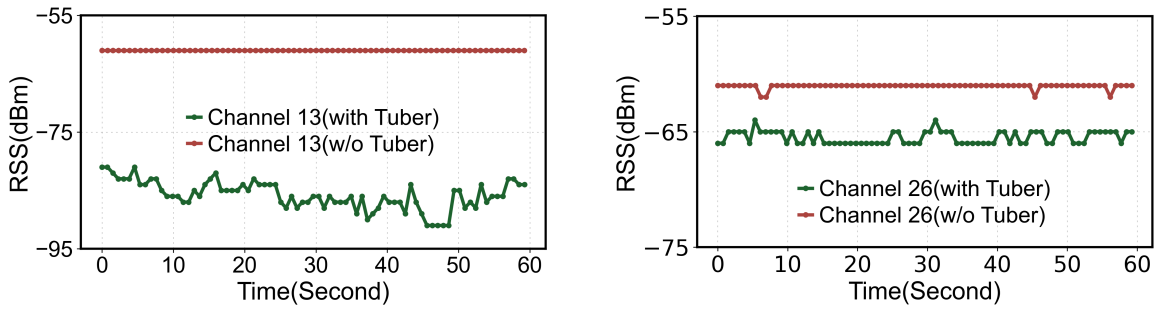
Fig. 11. MAE, MAPE, and RMSE performance in estimating biomass of underground root tubers placed at different positions. "C-GRU" and "C-Trans" denote the CNN-GRU [64] model and the CNN-Transformer [19] model, respectively.

but also highlight the generalizability of the proposed transformer network. Compared with the CNN-GRU model [64], the proposed model employs a transformer network to extract discriminative features from RSS data, which leverages the attention mechanism to adaptively capture information relevant to biomass. Although the CNN-Transformer [19] model also employs a transformer network for feature extraction, it only uses the transformer network to derive global representations from local features obtained by the convolution neural network, neglecting the relationships among different frequency channels. In contrast, the proposed model applies the attention mechanism to explicitly capture relationships across channels, providing more comprehensive representations for biomass estimation.

Furthermore, we evaluate the estimation performance for underground tubers at random positions. Specifically, RSS data are collected from 26 tubers placed at 32 positions, and the tuber-position pairs are randomly split in a 9:1 ratio for training and testing. As shown in Fig. 11, the proposed model outperforms the baseline models across all evaluation metrics. For example, the proposed model achieves average MAE and RMSE values of 9.74 and 16.91, respectively, outperforming the baseline models. These results further verify the efficacy of the proposed model for accurate biomass estimation. In addition, existing models use separate networks for imaging and biomass estimation, while our model employs the same transformer network architecture for both tasks, further demonstrating its efficacy and generalizability.

5.7 Performance on Channel Selection

In this study, we propose a fade-level-based channel selection method to reduce the overhead of the sensing system. First, we perform experiments to analyze the feasibility of selecting anti-fade channels for imaging



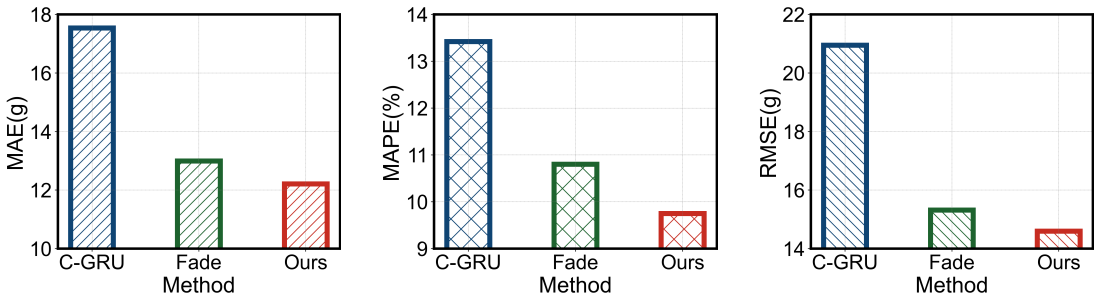
(a) Deep-fade channel.

(b) Anti-fade channel.

Fig. 12. RSS values from the same link on different channels with different fade levels. The variation of RSS on a deep-fade channel (a) is higher than that on an anti-fade channel (b), and the fade level of channel 26 is higher than that of channel 13. The link measurements on deep-fade channels have low sensing quality and will not be selected in RTS.

Table 10. Performance of using different frequency channels. We mark the best and second-best results using bold and underlined text, respectively.

Method \ No. channels	One frequency channel				Three frequency channels				All frequency channels			
	SSIM ↑	RPD ↓	IoU ↑	EDE ↓	SSIM ↑	RPD ↓	IoU ↑	EDE ↓	SSIM ↑	RPD ↓	IoU ↑	EDE ↓
CNN-UNet [27]	<u>0.98</u>	<u>0.13</u>	<u>0.82</u>	<u>2.47</u>	<u>0.98</u>	<u>0.15</u>	<u>0.81</u>	<u>2.63</u>	**0.99**	<u>0.13</u>	<u>0.82</u>	<u>2.48</u>
LSTM-UNet [46]	0.85	0.37	0.59	4.09	0.86	0.33	0.62	3.83	0.95	0.23	0.73	3.34
TD-RTS	**0.99**	**0.11**	**0.84**	**2.27**	**0.99**	**0.10**	**0.85**	**2.14**	**0.99**	**0.08**	**0.86**	**2.01**



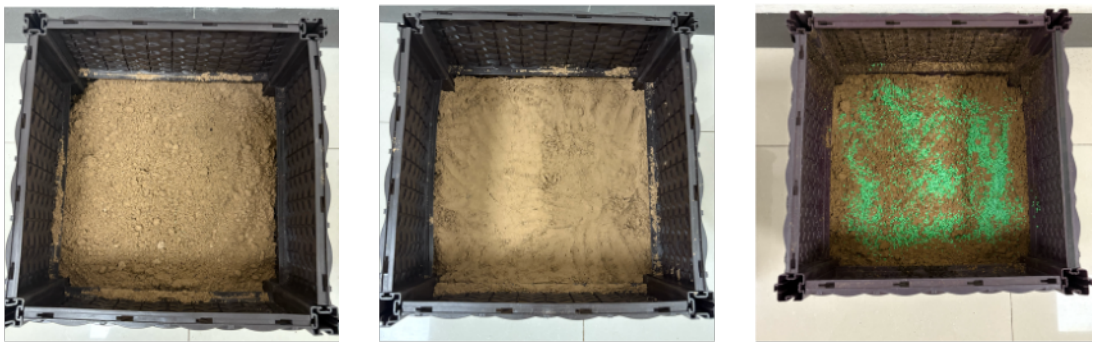
(a) MAE (g)

(b) MAPE (%)

(c) RMSE (g)

Fig. 13. MAE, MAPE, and RMSE performance in biomass estimation using different numbers of frequency channels. "Fade" refers to using the three channels selected by our fade-level method for biomass estimation, while "Ours" refers to using all 16 channels. "C-GRU" denotes the CNN-GRU [64] model, which also uses 16 frequency channels for biomass estimation.

underground tubers. As shown in Fig. 12, the RSS values of the same link on the anti-fade channel are more stable



(a) Loose soil (b) Compacted soil (c) Fertilized soil

Fig. 14. Different conditions of soils used to observe the stability of channel selection under different soil properties.

than those on the deep-fade channel. This indicates that RSS variations are higher on the deep-fade channel, resulting in lower sensing quality.

Subsequently, we compare the imaging performance of using different numbers of frequency channels. Specifically, RSS data from a single channel and three channels selected by our method are used to perform underground tuber imaging. As shown in Table 10, the performance of TD-RTS improves as the number of channels increases. However, the performance using 3 frequency channels is comparable to that achieved with 16 frequency channels. For example, the SSIM value obtained using 3 channels is the same as that obtained using 16 channels. The difference in IoU values between using 3 channels and 16 channels is only 0.01. As shown in the last column of Fig. 7, the visualization result using 3 channels closely resembles that obtained using all 16 channels. These results demonstrate the efficacy of our channel selection method. Moreover, the time required to collect a sample using 16 channels is more than five times longer than that for a sample collected using 3 channels. This indicates that the proposed channel selection method can significantly reduce data collection time and storage requirements. We also compare the performance of TD-RTS with baseline models across different frequency channels. TD-RTS outperforms the baseline models across all scenarios, demonstrating its robustness and generalizability under channel-limited conditions.

Furthermore, we evaluate the biomass estimation performance using three channels selected by our fade-level-based method. RSS data from 26 tubers are used, with 24 tubers for training and 2 tubers for testing. As shown in Fig. 13, using 16 channels achieves the best performance across all evaluation metrics. However, the performance of using 3 channels is comparable to that obtained with 16 channels. For example, the MAPE value obtained using 3 channels differs from that using 16 channels by only 1.05%, demonstrating the efficacy of our channel selection method for biomass estimation. Furthermore, compared with CNN-GRU [64] using RSS data from 16 channels, the proposed model using only three channels achieves superior performance across all evaluation metrics, further demonstrating the efficacy of both the channel selection method and the model design.

In addition, we perform a series of observation experiments to analyze the stability of the selected channels. First, we prepare different conditions of soils for experiments: relatively loose soil, more compacted soil, and soil with altered mineral composition, as shown in Fig. 14. For each soil condition, we collect RSS data for approximately one hour and define three sampling points for applying our channel selection method. Table 11 presents the channels selected by our method under three soil conditions. The selected results remain consistent within each soil condition but differ across soil conditions. Furthermore, we perform two long-term experiments to investigate how the selected channels vary with soil moisture. In the first experiment, we collect RSS data

Table 11. Selected channels using our fade-level-based method at three sampling points under three soil conditions. Frequency channel indices range from 11 to 26, and the channels with the top three fade-level values are selected. In each column, the channels are arranged from top to bottom in descending order of fade-level values.

	Loose soil			Compacted soil			Fertilized soil		
Sampling	First	Second	Third	First	Second	Third	First	Second	Third
First Channel	24	24	24	23	23	23	25	25	25
Second Channel	26	26	26	24	11	11	24	24	24
Third Channel	25	25	25	11	24	24	23	23	23

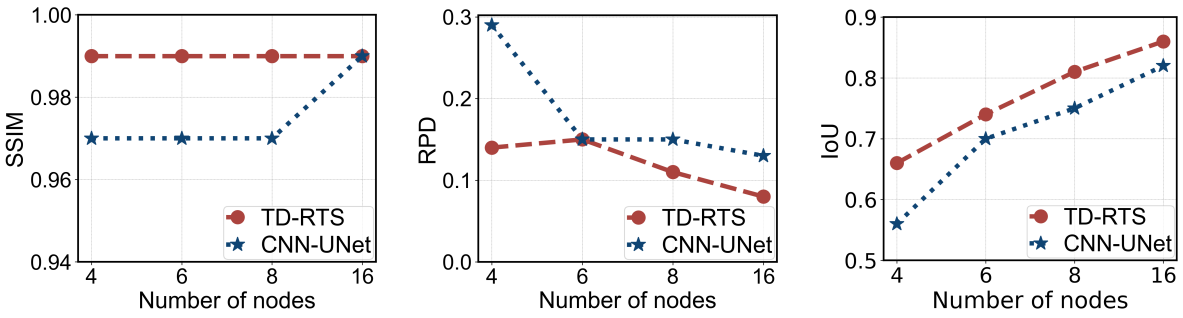
Table 12. Selected channels using our fade-level-based method at different soil moisture levels. Frequency channel indices range from 11 to 26, and the channels with the top three fade-level values are selected. In each column, the channels are arranged from top to bottom in descending order of fade-level values.

	Soil moisture decreasing					Soil moisture increasing				
Moisture	3.2%	1.5%	1.1%	0.5%	0.1%	6.7%	8.5%	11.9%	16.2%	17.5%
First Channel	14	14	14	20	20	24	14	25	23	23
Second Channel	16	16	21	14	14	23	24	24	24	22
Third Channel	15	19	20	21	21	25	23	14	22	24

for approximately 32 hours, during which the soil moisture gradually decreases. In the second experiment, we perform irrigation and collect data for approximately 14 hours. Four irrigation events are performed, and channel selection is performed after each event. Table 12 shows the channels selected under varying soil moisture levels, which change as the soil moisture decreases or increases. These results demonstrate that the selected channels are affected by soil properties and moisture levels, suggesting that the channel selection method should be reapplied when soil conditions change. To address this challenge, we propose a scheme that periodically reselects channels. In practice, channel reselection takes approximately 10 seconds, which is much shorter than the growth period of underground root tubers (75-110 days) [48]. Furthermore, we measure the model training time on an RTX 4090 GPU using a dataset of 26 potato tubers at 32 positions, which requires approximately 1.5 hours and is also far shorter than the growth period of underground tubers. These durations show the feasibility of periodic channel reselection and model retraining.

5.8 Performance on Number of Sensor Nodes

In this section, we evaluate the performance of TD-RTS in a single-tuber scenario with varying numbers of sensor nodes, and the results are presented in Fig. 15. As shown in Fig 15a, we compare the SSIM values obtained from TD-RTS and CNN-UNet [27]. The SSIM values of CNN-UNet [27] increase with the number of sensor nodes. However, TD-RTS consistently achieves higher SSIM values across all scenarios. As shown in Fig. 15b, TD-RTS consistently achieves lower RPD values than CNN-UNet [27] across all scenarios with varying numbers of sensor nodes. As shown in Fig. 15c, the IoU values of both TD-RTS and CNN-UNet [27] increase with the number of sensor nodes. However, TD-RTS consistently outperforms CNN-UNet [27] in each scenario. These results suggest that increasing the number of sensor nodes captures more semantic information about tubers, enhancing imaging quality and detection accuracy. Note that we also compare TD-RTS to LSTM-UNet [46], but we observe that the



(a) SSIM (b) RPD (c) IoU

Fig. 15. SSIM, RPD, and IoU performance of using different numbers of nodes.

(a) Outdoor environment and potato tubers. (b) Testbed deployment.

Fig. 16. Outdoor environment, potato tubers, and testbed deployment for RTS. Two ridges are constructed to bury potato tubers of different shapes and sizes, and the testbed is deployed and gradually moved to collect RSS data for each tuber.

performance of LSTM-UNet [46] significantly degrades as the number of sensor nodes decreases. For example, when using six sensor nodes, the average IoU value obtained by LSTM-UNet [46] is only 0.07, while TD-RTS achieves an average IoU value of 0.74 in the same case. These results demonstrate the robustness of TD-RTS for image reconstruction under information-limited conditions.

5.9 Discussion and Future Work

Extend to 3D imaging and root tuber sensing of other crop types. In this paper, we focus on 2D cross-section imaging of potato tubers to investigate the feasibility of networked sensing of underground root tubers. Our evaluation shows that a low-power RF tomography network deployed above ground can capture the footprints of potato root tubers below the soil surface. However, this is just the first step towards a general-purpose underground root sensing system. To extend this work in the future, we can deploy a multi-layer RF tomography network underground to reconstruct 3D images of root tubers by combining 2D images from each layer. In addition, we

focus on sensing of potato root tubers in this work, since it is the first step towards root tuber sensing using RF sensor networks. In the future, we plan to broaden our scope to a variety of root tubers and roots with more diverse sizes and shapes.

Towards outdoor deployment. In this study, the proposed framework is suitable for deployment in indoor environments and can be applied to sensing tasks such as greenhouse crop breeding. To further explore its feasibility in outdoor environments, we construct an outdoor underground tuber planting scenario and perform preliminary experiments, as shown in Fig. 16. Specifically, we build two ridges, and 18 potato tubers are evenly buried in them. RSS data from 15 tubers are used to train the deep learning model and data from the remaining 3 tubers are used for evaluation. The evaluation results indicate that the sensing performance in the outdoor environment is lower than that in the indoor environment. For example, the average RPD value in the outdoor environment is 0.18, while it is 0.08 in the indoor environment. In practice, sensing performance is often degraded by environmental noise [3], including other plants in the sensing area and weather conditions that can alter soil properties, such as wind, rain, and sunlight. To further enhance outdoor sensing performance, we first plan to replace the omni-directional antennas with directional ones for sensor nodes, thereby improving the quality of radio link observations [68]. Second, we plan to collect RSS data from a larger number of underground tubers for training, as increasing the diversity of training samples can further improve the generalizability of data-driven models [78]. Third, we plan to explore transfer learning methods [50] to enable a model trained in indoor environments to adapt to outdoor environments. Finally, we can increase the number of sensor nodes to further improve the performance of RTS.

Investigation on WiFi CSI. In this work, we focus on root tuber sensing using RSS measurements from a network. Since WiFi channel state information (CSI) data contain not only signal strength information but also phase information on the wireless channel, WiFi CSI data can also be used to train our image reconstruction DNN models. However, the phase measurements are more sensitive to motion in a dynamic environment, soil moisture level, and other environmental factors than the RSS measurements. DNN models trained by insufficient CSI data will suffer from the overfitting problem. Thus, we focus on the investigation of the sensing capability of RSS measurements in this work. We leave the investigation of WiFi systems for underground RTS as future work.

6 Related Work

RF-based Sensing. Despite the absence of existing work using RF tomography for sensing underground biomass, some previous RF-based works leveraging radio signals have been proposed for human sensing [39, 81], object localizing [52, 60], fruit ripeness detecting [2], and underground root sensing [1, 42, 43]. For example, [43] and [1] utilize the ground penetrating radar (GPR) to detect and reconstruct the underground plant roots. In [1], a novel processing procedure is proposed, encompassing noise removal, soil dielectric constant calculation, and wave migrations. Although these GPR-based methods demonstrate satisfactory performance in underground root sensing, their broader applicability is hindered by cost constraints, high power requirements, and substantial physical dimensions of the GPR.

Recent research endeavors aim to investigate the utilization of cost-effective RF-based devices for sensing [28, 31, 35, 38, 51, 65, 79]. For instance, a noteworthy contribution is presented by [4], where RF tomography is leveraged to assess the moisture levels in stored rice. In this research, the authors introduce an innovative approach that combines RF tomography with regression-based machine learning to offer a non-invasive, contactless method for obtaining a 3D volumetric distribution of moisture content within stored rice grains. Another noteworthy contribution is found in [65], which presents an RF-based solution for underground tuber sensing in a single-tuber scenario, employing a ZigBee RF sensor network. In contrast to [65], we first extend the sensing scenarios to include both single-tuber and double-tuber scenarios, while also enhancing the diversity of potato tubers. Then, we consider the interference caused by environmental changes to wireless signals, a common issue in practical

scenarios. Furthermore, we propose a novel sensing model that provides higher imaging quality and estimation accuracy. Additionally, we consider the overhead of the RF sensor network and propose a novel channel selection method to reduce the number of required channels, thereby decreasing network overhead while maintaining accurate estimation.

Domain Adaptation Learning. Domain adaptation learning is a long-standing field aimed at bridging the gaps between different domains. This method has been applied across various learning areas, including computer vision [23], natural language processing [82], and signal processing [5, 40]. A common approach to domain adaptation within deep neural networks is fine-tuning a model pre-trained on the source domain using data from the target domain. For example, [40] designs an innovative domain adaptation approach to mitigate the challenges faced by millimeter-wave radio-based gesture recognition in heterogeneous environments. This approach enables practical gesture recognition by leveraging the pre-learned model with minimal target samples for fine-tuning. Experimental results demonstrate that achieving comparable accuracy can be accomplished by retraining with as few as eight samples per gesture. Meanwhile, [5] introduces a novel method for fine-tuning the pre-trained model. This approach first adjusts the data distribution in the source domain to match that of the target domain, and then uses a smaller set of target domain data to further fine-tune the model. In this paper, the novelty of our proposed domain adaptation method lies in its ability to achieve robust imaging using RSS data from a single tuber in dynamic environments.

Diffusion Model. Recent years have seen diffusion models achieve remarkable success in computer vision tasks [9, 16, 57, 77]. For instance, [17] demonstrates the use of diffusion models for unconditional image synthesis, achieving image sample quality that surpasses existing state-of-the-art generative methods. Additionally, [77] enhances diffusion-based image synthesis by incorporating context prediction. Beyond image synthesis, diffusion models have been successfully applied to other computer vision tasks. SR3 [61] leverages a diffusion model for image super-resolution, outperforming GAN-based approaches. RePaint [45] introduces a diffusion model for image inpainting, with an improved denoising strategy through resampling iterations within the model. Traditional diffusion models operate directly on image pixels, requiring many iterations, substantial computational resources, and large model parameters to achieve high-quality predictions. In contrast, we adopt a latent diffusion model [57, 74] that uses an encoder-decoder network to compress the original image into a compact feature representation, which is then fed into the diffusion process. This approach significantly reduces the model size and the number of iterations needed to generate high-precision results.

7 Conclusion

This paper proposes a novel underground RTS framework, which leverages a low-cost RF sensor network and deep neural network models to reconstruct cross-section images and estimate the biomass of underground tubers. Furthermore, the adoption of a simple yet effective domain adaptation method significantly improves the robustness performance of the framework in dynamic environments. In addition, we construct a comprehensive dataset that is more realistic than the previous work and provides more accurate data annotations. The evaluation results demonstrate the efficacy of the RTS framework across various sensing cases.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62350710797.

References

[1] Abderrahmane Aboudourib, Mohammed Serhir, and Dominique Lesselier. 2021. A Processing Framework for Tree-Root Reconstruction Using Ground-Penetrating Radar Under Heterogeneous Soil Conditions. *IEEE Trans. Geosci. Remote. Sens.* 59, 1 (2021), 208–219.

[2] Sayed Saad Afzal, Atsutse Kludze, Subhajit Karmakar, Ranveer Chandra, and Yasaman Ghasempour. 2023. AgriTera: Accurate Non-Invasive Fruit Ripeness Sensing via Sub-Terahertz Wireless Signals. In *Proceedings of the 29th Annual International Conference on Mobile*

Computing and Networking. Association for Computing Machinery, New York, NY, USA, Article 61, 15 pages.
[3] Cesare Alippi, Maurizio Bocca, Giacomo Boracchi, Neal Patwari, and Manuel Roveri. 2016. RTI Goes Wild: Radio Tomographic Imaging for Outdoor People Detection and Localization. *IEEE Transactions on Mobile Computing* 15, 10 (2016), 2585–2598.
[4] Abd Alazeez Almaleeh, Ammar Zakaria, Latifah Munirah Kamarudin, Mohd Hafiz Fazalul Rahiman, David Lorater Ndzi, and Ismahadi Ismail. 2022. Inline 3D Volumetric Measurement of Moisture Content in Rice Using Regression-Based ML of RF Tomographic Imaging. *Sensors* 22, 1 (2022), 405.
[5] Maira Alvi, Rachel Cardell-Oliver, and Tim French. 2022. Utilizing autoencoders to improve transfer learning when sensor data is sparse. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. Association for Computing Machinery, New York, NY, USA, 500–503.
[6] J Scott Armstrong and Fred Collopy. 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting* 8, 1 (1992), 69–80.
[7] Feiyan Bai, Minqiang Fan, Hongli Yang, and Lianping Dong. 2021. Image segmentation method for coal particle size distribution analysis. *Particuology* 56 (2021), 163–170.
[8] Rinku Basak and Khan A Wahid. 2022. An in situ electrical impedance tomography sensor system for biomass estimation of tap roots. *Plants* 11, 13 (2022), 1713.
[9] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. 2023. Person Image Synthesis via Denoising Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, 5968–5976.
[10] Maurizio Bocca, Ossi Kaltiokallio, and Neal Patwari. 2013. Radio Tomographic Imaging for Ambient Assisted Living. In *Evaluating AAL Systems Through Competitive Benchmarking*, Stefano Chessa and Stefan Knauth (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 108–130.
[11] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.
[12] Samprit Chatterjee and Ali S Hadi. 2015. *Regression analysis by example*. John Wiley & Sons.
[13] Weiyan Chen, Hongliu Yang, Xiaoyang Bi, Rong Zheng, Fusang Zhang, Peng Bao, Zhaoxin Chang, Xujun Ma, and Daqing Zhang. 2023. Environment-aware Multi-person Tracking in Indoor Environments with MmWave Radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–29.
[14] Sruti Das Choudhury, Vincent Stoerger, Ashok Samal, James C Schnable, Zhikai Liang, and Jin-Gang Yu. 2016. Automated vegetative stage phenotyping analysis of maize plants using visible light images. In *KDD workshop on data science for food, energy and water, San Francisco, California, USA*.
[15] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[16] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
[17] Prafulla Dhariwal and Alex Nichol. 2024. Diffusion models beat GANs on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 672, 15 pages.
[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[19] Jiangli Du, Yue Zhang, Pengxin Wang, Kevin Tansey, Junming Liu, and Shuyu Zhang. 2025. Enhancing winter wheat yield estimation with a CNN-transformer hybrid framework utilizing multiple remotely sensed parameters. *IEEE Transactions on Geoscience and Remote Sensing* (2025).
[20] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88 (2010), 303–338.
[21] Xiuwen Fu, Giancarlo Fortino, Pasquale Pace, Gianluca Aloi, and Wenfeng Li. 2020. Environment-fusion multipath routing protocol for wireless sensor networks. *Information Fusion* 53 (2020), 4–19.
[22] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. MIST : Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 14773–14783.
[23] Xinyuan Gao, Yuhang He, Songlin Dong, Jie Cheng, Xing Wei, and Yihong Gong. 2023. DKT: Diverse Knowledge Transfer Transformer for Class Incremental Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 24236–24245.
[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
[25] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).

[26] Timothy O Hodson. 2022. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions* 2022 (2022), 1–10.

[27] Qiushi Huang, Guanghui Liang, Chao Tan, and Feng Dong. 2024. Res2-UNet++: A deep learning image post-processing method for electrical resistance tomography. *Measurement Science and Technology* 35, 10 (2024), 105403.

[28] Colleen Josephson, Manikanta Kotaru, Keith Winstein, Sachin Katti, and Ranveer Chandra. 2021. Low-cost In-ground Soil Moisture Sensing with Radar Backscatter Tags. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies* (Virtual Event, Australia) *(COMPASS '21)*. Association for Computing Machinery, New York, NY, USA, 299–311.

[29] Ossi Kaltiokallio, Maurizio Bocca, and Neal Patwari. 2012. Enhancing the accuracy of radio tomographic imaging using channel diversity. In *2012 IEEE 9th international conference on mobile ad-hoc and sensor systems (MASS 2012)*. IEEE, 254–262.

[30] Ossi Kaltiokallio, Maurizio Bocca, and Neal Patwari. 2013. A fade level-based spatial model for radio tomographic imaging. *IEEE Transactions on Mobile Computing* 13, 6 (2013), 1159–1172.

[31] Usman Mahmood Khan and Muhammad Shahzad. 2022. Estimating soil moisture using RF signals. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking* (Sydney, NSW, Australia) *(MobiCom '22)*. Association for Computing Machinery, New York, NY, USA, 242–254.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[33] Earl J Kirkland and Earl J Kirkland. 2010. Bilinear interpolation. *Advanced computing in electron microscopy* (2010), 261–263.

[34] Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. 2023. SinDDM: a single image denoising diffusion model. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 738, 11 pages.

[35] Branislav Kusy, David Abbott, Christian Richter, Cong Huynh, Mikhail Afanasyev, Wen Hu, Michael Brünig, Diethelm Ostry, and Raja Jurdak. 2014. Radio diversity for reliable communication in sensor networks. *ACM Transactions on Sensor Networks (TOSN)* 10, 2 (2014), 1–29.

[36] Narawitch Lertngim. 2023. *Improving non-invasive biomass estimation by combining different types of image-based phenotypic traits: a study in pea under drought stress.* Master's thesis. Wageningen University & Research.

[37] Jimeng Li, Xuefeng Chen, and Zhengjia He. 2013. Adaptive stochastic resonance method for impact signal detection based on sliding window. *Mechanical Systems and Signal Processing* 36, 2 (2013), 240–255.

[38] Zhuqi Li, Yaxiong Xie, Longfei Shangguan, R. Ivan Zelaya, Jeremy Gummeson, Wenjun Hu, and Kyle Jamieson. 2019. Towards programming the radio environment with large arrays of inexpensive antennas. In *Proceedings of the 16th USENIX Conference on Networked Systems Design and Implementation* (Boston, MA, USA) *(NSDI'19)*. USENIX Association, USA, 285–299.

[39] Peng Liao, Xuyu Wang, Lingling An, Shiwen Mao, Tianya Zhao, and Chao Yang. 2024. TFSemantic: A Time–Frequency Semantic GAN Framework for Imbalanced Classification Using Radio Signals. *ACM Trans. Sen. Netw.* 20, 4 (2024).

[40] Haipeng Liu, Kening Cui, Kaiyuan Hu, Yuheng Wang, Anfu Zhou, Liang Liu, and Huadong Ma. 2022. MTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–28.

[41] Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. 2024. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2773–2783.

[42] Qixin Liu, Xihong Cui, Xinbo Liu, Jin Chen, Xuehong Chen, and Xin Cao. 2017. Detection of Root Orientation Using Ground-Penetrating Radar. *IEEE Transactions on Geoscience and Remote Sensing* PP (10 2017), 1–12. https://doi.org/10.1109/TGRS.2017.2737003

[43] Yawen Lu and Guoyu Lu. 2022. 3D modeling beneath ground: Plant root detection and reconstruction based on ground-penetrating radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 68–77.

[44] Yiwei Lu, Yongtao Ma, Wanru Ning, and Haibo Zhao. 2023. Multi-Target Localization for RTI Based on Constructing Feature Region Combination. *IEEE Sensors Journal* (2023).

[45] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022*. IEEE, 11451–11461.

[46] Ben Martin, Colin Gilmore, and Ian Jeffrey. 2024. A Long Short-Term Memory Approach to Incorporating Multi-Frequency Data into Deep-Learning-Based Microwave Imaging. *IEEE Transactions on Antennas and Propagation* (2024).

[47] Jean-Pascal Matteau, Paul Celicourt, Elnaz Shahriarina, Philipe Letellier, Thiago Gumiere, and Silvio J Gumiere. 2022. Relationship between irrigation thresholds and potato tuber depth in sandy soil. *Frontiers in Soil Science* 2 (2022), 898618.

[48] Muhammad Waqar Nasir and Zoltan Toth. 2022. Effect of drought stress on potato production: A review. *Agronomy* 12, 3 (2022), 635.

[49] Tadahiro Negishi, Gianluca Gennarelli, Francesco Soldovieri, Yangqing Liu, and Danilo Erricolo. 2020. Radio frequency tomography for nondestructive testing of pillars. *IEEE Transactions on Geoscience and Remote Sensing* 58, 6 (2020), 3916–3926.

[50] Cong T. Nguyen, Nguyen Van Huynh, Nam H. Chu, Yuris Mulya Saputra, Dinh Thai Hoang, Diep N. Nguyen, Quoc-Viet Pham, Dusit Niyato, Eryk Dutkiewicz, and Won-Joo Hwang. 2021. Transfer Learning for Future Wireless Networks: A Comprehensive Survey. arXiv:2102.07572

[51] Jingyi Ning, Lei Xie, Chuyu Wang, Yanling Bu, Fu Xiao, Baoliu Ye, and Sanglu Lu. 2022. Revolving Scanning on Tagged Objects: 3D Structure Detection of Logistics Packages via RFID Systems. *ACM Trans. Sen. Netw.* 18, 2 (2022).
[52] Leonardo L. de Oliveira, Gabriel H. Eisenkraemer, Everton A. Carara, João B. Martins, and Jose Monteiro. 2023. Mobile Localization Techniques for Wireless Sensor Networks: Survey and Recommendations. *ACM Trans. Sen. Netw.* 19, 2 (2023).
[53] Gaurav Prasad, Aditya Gupta, Avnish Aryan, and Sudhir Kumar. 2024. A Vision Transformer Based Indoor Localization Using CSI Signals in IoT Networks. In *International Conference on Advanced Information Networking and Applications*. Springer, 73–83.
[54] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems* 34 (2021), 12116–12128.
[55] NH Ravindranath and Madelene Ostwald. 2008. Methods for below-ground biomass. *Carbon Inventory Methods Handbook for Greenhouse Gas Inventory, Carbon Mitigation and Roundwood Production Projects* (2008), 149–156.
[56] Thomas Roitsch, Llorenç Cabrera-Bosquet, Antoine Fournier, Kioumars Ghamkhar, José Jiménez-Berni, Francisco Pinto, and Eric S Ober. 2019. New sensors and data-driven approaches—A path to next generation phenomics. *Plant Science* 282 (2019), 2–10.
[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[58] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).
[59] Henry Ruiz-Guzman, Tyler Adams, Afolabi Agbona, Matthew Wolfe, Mark Everett, Jean-Francois Chamberland, and Dirk B Hays. 2025. Thresholding and continuous wavelet transform (CWT) analysis of Ground Penetrating Radar (GPR) data for estimation of potato biomass. *Computers and Electronics in Agriculture* 232 (2025), 110114.
[60] Dinesh Kumar Sah, Tu N. Nguyen, Manjusha Kandulna, Korhan Cengiz, and Tarachand Amgoth. 2022. 3D Localization and Error Minimization in Underwater Sensor Networks. *ACM Trans. Sen. Netw.* 18, 3 (2022).
[61] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence* 45, 4 (2022), 4713–4726.
[62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
[63] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
[64] Jie Wang, Pengxin Wang, Huiren Tian, Kevin Tansey, Junming Liu, and Wenting Quan. 2023. A deep learning framework combining CNN and GRU for improving wheat yield estimates using time series remotely sensed multi-variables. *Computers and Electronics in Agriculture* 206 (2023), 107705.
[65] Tao Wang, Yang Zhao, Jie Liu, and Yujie Zhuang. 2024. Demo Abstract: Underground Potato Root Tuber Sensing via a Wireless Network. In *2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 251–252.
[66] Tao Wang, Yang Zhao, Jinghua Wang, Zhibin Huang, Jie Liu, and Qiaorong Wei. 2025. See-through Soil: Underground Root Tuber Sensing with RF Sensor Networks. *IEEE Transactions on Geoscience and Remote Sensing* (2025).
[67] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
[68] Bo Wei, Ambuj Varshney, Neal Patwari, Wen Hu, Thiemo Voigt, and Chun Tung Chou. 2015. dRTI: directional radio tomographic imaging. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks* (Seattle, Washington) *(IPSN '15)*. Association for Computing Machinery, New York, NY, USA, 166–177.
[69] Graham V Weinberg. 2017. An invariant sliding window detection process. *IEEE Signal Processing Letters* 24, 7 (2017), 1093–1097.
[70] Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 1 (2005), 79–82.
[71] Joey Wilson and Neal Patwari. 2010. Radio Tomographic Imaging with Wireless Networks. *IEEE Trans. Mob. Comput.* 9, 5 (2010), 621–632.
[72] Joey Wilson and Neal Patwari. 2010. See-through walls: Motion tracking using variance-based radio tomography networks. *IEEE Transactions on Mobile Computing* 10, 5 (2010), 612–621.
[73] Joey Wilson and Neal Patwari. 2012. A fade-level skew-laplace signal strength model for device-free localization with wireless networks. *IEEE Transactions on Mobile Computing* 11, 6 (2012), 947–958.
[74] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. DiffIR: Efficient Diffusion Model for Image Restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13095–13105.
[75] Bolai Xin, Ji Sun, Harm Bartholomeus, and Gert Kootstra. 2023. 3D data-augmentation methods for semantic segmentation of tomato plant parts. *Frontiers in Plant Science* 14 (2023), 1045545.
[76] Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. 2023. One-Shot Generative Domain Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7733–7742.
[77] Ling Yang, Jingwei Liu, Shenda Hong, Zhilong Zhang, Zhilin Huang, Zheming Cai, Wentao Zhang, and Bin Cui. 2024. Improving diffusion-based image synthesis with context prediction. *Advances in Neural Information Processing Systems* 36 (2024), 37636–37656.

[78] Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can Data Diversity Enhance Learning Generalization?. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4933–4945.

[79] R Ivan Zelaya, William Sussman, Jeremy Gummeson, Kyle Jamieson, and Wenjun Hu. 2021. LAVA: fine-grained 3D indoor wireless coverage for small IoT devices. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. Association for Computing Machinery, New York, NY, USA, 123–136.

[80] Liangwei Zhang, Jing Lin, and Ramin Karim. 2016. Sliding window-based fault detection from high-dimensional data streams. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 2 (2016), 289–303.

[81] Yi Zhang, Zheng Yang, Guidong Zhang, Chenshu Wu, and Li Zhang. 2021. XGest: Enabling Cross-Label Gesture Recognition with RF Signals. *ACM Trans. Sen. Netw.* 17, 4 (2021).

[82] Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. Transfer Learning from Semantic Role Labeling to Event Argument Extraction with Template-based Slot Querying. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2627–2647.

[83] Yang Zhao and Neal Patwari. 2014. Robust estimators for variance-based device-free localization and tracking. *IEEE Transactions on Mobile Computing* 14, 10 (2014), 2116–2129.

[84] Tao Zhou, Xiangping Liu, Zuowei Xiang, Haitong Zhang, Bo Ai, Liu Liu, and Xiaorong Jing. 2024. Transformer Network Based Channel Prediction for CSI Feedback Enhancement in AI-Native Air Interface. *IEEE Transactions on Wireless Communications* 23, 9 (2024), 11154–11167.

[85] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 3–11.