See-through Soil: Underground Root Tuber Sensing with RF Sensor Networks

Tao Wang, *IEEE Student Member,* Yang Zhao, *IEEE Senior Member,* Jinghua Wang, *IEEE Member,* Zhibin Huang, *IEEE Student Member,* Jie Liu, *IEEE Fellow,* and Qiaorong Wei

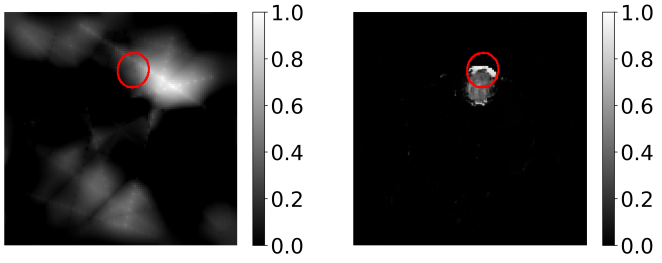
***Abstract*—This paper proposes a data-driven root tuber sensing (RTS) framework that uses the received signal strength (RSS) data from a radio frequency (RF) sensor network to reconstruct cross-section images of root tubers in soils. We perform extensive experiments with our data acquisition system in various environments to build a wireless potato sensing (WPS) dataset. We propose to integrate multi-branch convolutional neural networks with a diffusion neural network to enable fine-grained image reconstruction of root tubers. To deal with the multi-path effects on radio channels, we propose two domain adaptation methods: one-shot fine-tuning to update the neural network model online, and disentangled representation learning (DRL) to transfer a pre-trained model to unseen environments. Experimental results from over 1.7 million RF network measurements show the efficacy of the proposed methods across different environments. Our data and pre-trained models are publicly available on IEEE DataPort.**

***Index Terms*—Underground sensing, Radio frequency sensor network, Convolutional neural networks, Domain adaptation.**

I. Introduction

As the development of remote sensing techniques, various sensors and methods have been developed for monitoring plant above-ground phenotypic traits, e.g., leaf area index [1], in crop breeding [2], crop yield prediction [3] and other smart agricultural applications [4]. While there is also a pressing need to monitor underground phenotypic traits, such as below-ground biomass, underground root sensing remains an important research topic largely understudied, especially for root vegetables and crops bearing starchy tuberous roots, e.g., potato (Solanum tuberosum) [5].

For non-invasive RTS, computed tomography (CT) with X-ray scanning has been used to obtain root tuber images non-destructively in laboratory-scale environments [6]. However, CT machines have low mobility and are expensive for widespread use in smart agricultural applications. Ground penetrating radar (GPR) can also detect underground targets using RF signals, and previous studies have used GPR together with signal processing, machine learning and deep neural networks techniques to reconstruct 2D and 3D images of underground roots [7], [8], [9]. However, radar sensors experience distance loss inversely proportional to the fourth power of the distance ($1/D^4$), and higher frequency bands generally lead to lower penetration capability. As wireless devices are becoming ubiquitous nowadays, there is a research gap in investigating RF networks in underground RTS, especially considering the advantage of path loss in wireless communication devices over radar devices.



(a) Imaging from a physical model. (b) Imaging from a DNN model.

Fig. 1: Imaging results for an underground tuber: (a) from a physical RF tomography model [10], (b) from a DNN model [11]. Red circles indicate the ground truth of the 2D cross-section areas of the potato root tubers.

In fact, RF sensor networks have been widely used in various non-invasive sensing applications. For example, various RF tomography methods have been proposed to detect, locate people and even recognize their activities in a non-cooperative way [12], [10], [13]. Recent studies have shown that deep neural network (DNN) models achieve state-of-the-art (SOTA) performance in these RF sensing applications. For example, [11] uses a two-stage CNN model on the received signal strength (RSS) data from an RF sensor network to achieve radio tomographic imaging of objects in their simulation. [14] applies a lightweight neural network to WiFi channel state information (CSI) data for fruit ripeness sensing. However, we find the following challenges, when applying existing methods to networked RF sensing of underground root tubers.

First, previous physical model-based RF tomography methods focus on detection, localization and tracking of point targets, instead of area targets. To reconstruct images with higher resolution, the RF tomography inverse problem becomes more ill-posed. For example, when we apply the attenuation-based RF tomography algorithm [12] to underground sensing of root tubers, the reconstructed image shows high attenuation values near the correct tuber location (an example shown in Fig. 1a), but the structural similarity index (SSIM) with the 2D cross-section ground-truth is only 0.52. As shown in Fig. 1b, the state-of-the-art data-driven methods, e.g., CNN-based methods outperform physical model-based methods. However, deep neural networks (DNNs) require a large amount of training data, and to the best of our knowledge, no dataset is publicly available for networked RF sensing of underground root tubers. Indeed, it is laborious and time-consuming to annotate

the ground truth of root tubers with various dimensions and shapes. In this paper, we aim to fill in the gap by building an RF network-based RTS dataset and designing a data-driven RTS framework, as shown in Fig. 2.

Second, radio signals are sensitive to multi-path effects, and even small environmental changes can cause significant degradation of DNN models. As shown in Fig. 3b, a minor environmental change, e.g., moving a chair, can cause a 4 dBm variation in RSS data, which exceeds the 3 dBm difference observed between scenarios with and without the root tuber. In addition, a DNN model trained in an environment does not generalize well to a new one due to different multi-path effects. As shown in Fig. 3c, RSS values from the same RF link for the same tuber in different environments exhibit significant differences, degrading the performance of the DNN models. Imaging samples are shown in Fig. 5c and Fig. 5e, with details discussed in Section V-C2 and Section V-C3, respectively.

Finally, tuber dimensions, such as the cross-section area of root tubers, are important phenotypic traits that require high estimation accuracy. The estimated error of root tubers needs to be sufficiently small to meet the requirements of tuber genomics and phenomics research. However, even given a large amount of training data, existing deep neural network models, such as two-stage CNN models still fail to provide fine-grained image results, as shown in Fig. 1b.

To tackle these challenges, we propose a novel RTS framework with the following hardware and algorithm modules. First, we choose potato tuber as the object of study and design a data acquisition system called Spin, which includes an RF sensor network, as well as a rotating platform that enables data augmentation [15] for "through-soil sensing" in the data collection stage. We perform extensive measurement campaigns using the Spin testbed in various environments to build a wireless potato root tuber sensing (WPS) dataset, which contains over 30 hours of RSS measurements and over one thousand ground truth annotations for 42 potato tubers with different dimensions and shapes at different environments.

Second, to achieve fine-grained imaging from noisy RSS data, we propose a novel DNN model, MC-Diffusion, which is composed of multi-branch convolutional networks and a latent diffusion network. Since the attention mechanism can effectively extract useful information from noisy data [16], MC-Diffusion integrates it into convolutional neural networks to generate an initial image of the root tuber cross-section. Due to the efficacy of diffusion networks in image denoising [17], [18], we also design a novel latent diffusion network [19] to generate the final fine-grained image by eliminating the residual noise on the initial image, without adding many model parameters and increasing too much training time as traditional diffusion models.

Third, we propose two domain adaptation methods to address the multi-path effects caused by environmental changes and crossing environments, respectively. To robustly reconstruct images under environmental changes, we combine the MC-Diffusion model with a one-shot fine-tuning method [20], which updates the model using the most recent RSS data from a single tuber, enabling adaptation to a dynamic environment. For cross-environment imaging, we propose a DRL method [21] to extract environment-independent features for root tuber imaging, enabling our data-driven model trained in previous environments generalizable to unseen environments.

In summary, this paper makes the following contributions.

- We propose an RTS framework to reconstruct cross-section images for underground root tubers using RSS data from an RF sensor network. We build a wireless potato sensing (WPS) dataset with over 1.7 million network measurements collected in various environments.
- We propose a novel DNN-based model to achieve high-quality imaging of underground tubers. An attention mechanism, integrated into convolutional networks, adaptively adjusts the feature map, while a diffusion network generates a high-quality image by removing noise.
- We propose two domain adaptation methods to address the multi-path effects caused by environmental changes and transitions between different environments, respectively. A one-shot fine-tuning method is proposed to update the neural network online, enabling adaptation to environmental changes. A DRL method is proposed to extract environment-independent features, facilitating the transfer of our DNN model to unseen environments.
- We perform extensive real-world experiments, and experimental results from over 1.7 million RF network measurements show that our RTS model outperforms SOTA baselines in imaging quality and accuracy.

II. Problem Statement and Overview

A. Problem Statement

Given an RF sensor network consisting of S sensor nodes, there are $M = S(S-1)$ RF links, each operating on G frequency channels. We use $y_{g,l}[n]$ to denote the RSS time series from link l on channel g at time n, which can be described as [22], [10]:

$$y_{g,l}[n] = A_g - L_{g,l} - H_{g,l}[n] + F_{g,l}[n] - V_{g,l}[n], \quad (1)$$

where A_g is the transmit power, $L_{g,l}$ is the larger scale path loss, $F_{g,l}[n]$ is the fading gain, $V_{g,l}[n]$ is the measurement noise, and $H_{g,l}[n]$ is the shadowing loss caused by objects blocking the signal propagation path. Note that the transmit power A_g is constant for all links operating on the same frequency channel, and the larger scale path loss $L_{g,l}$ remains unchanged over time. Thus, we use a single subscript index for A_g and two subscript indices for $L_{g,l}$, respectively. Including all links and channels, we obtain the RSS data matrix $\mathbf{Y}^r$ with root tubers in soils, where each row corresponds to an RF link and each column represents a frequency channel.

To mitigate interference from environmental factors and soil conditions during underground tuber sensing, we collect RSS calibration data $\mathbf{Y}^c$ with no target, i.e., root tubers, present in the sensing area. Then, we obtain the corrected RSS data matrix $\mathbf{Y}$ for imaging underground tubers by subtracting $\mathbf{Y}^c$ from $\mathbf{Y}^r$:

$$\mathbf{Y} = \mathbf{Y}^r - \mathbf{Y}^c. \quad (2)$$

We let image vector $\mathbf{r} = [r_0, \cdots, r_{P-1}]^T$ represent the 2D cross-section image of the root tuber, where r_p represents

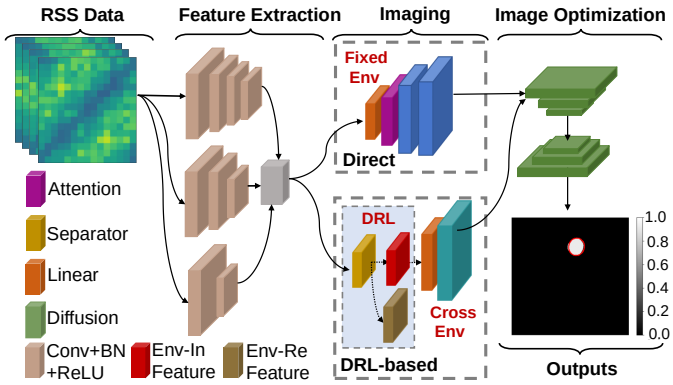


Fig. 2: Overview of the RTS framework. The RSS data from an RF sensor network is fed into an MC-Diffusion neural network, which is used to reconstruct tuber cross-section images. "Conv" and "BN" denote the convolutional and batch normalization layers, respectively. "Fixed Env" and "Cross Env" denote the model components used to reconstruct root tuber images in a particular environment and across different environments, respectively. "Env-In Feature" and "Env-Re Feature" respectively denote the environment-independent and environment-related features used in DRL.

the presence of the root tuber at pixel p, and P denotes the total pixel number. Based on RF tomographic imaging formulations [12], [10], the relationship between the image vector $\mathbf{r}$ and the corrected RSS data can be modeled as:

$$\mathbf{Y} = \mathcal{H}(\mathbf{r}) + \mathbf{b}, \tag{3}$$

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

For root tuber imaging, we aim to find the inverse function $\mathcal{H}^{-1}$ of $\mathcal{H}$, which estimates the image vector $\mathbf{r}$ from RSS data. Previous works in [12], [10], [22] model $\mathcal{H}$ as a linear function and compute $\mathcal{H}^{-1}$ by solving an inverse problem of $\mathcal{H}$. As shown in Fig. 1a, the attenuation-based RF tomography method [12] can provide a coarse location estimate for the underground tuber but struggles in capturing fine-grained tuber dimension and shape information. Thus, in this paper, we take advantage of the SOTA DNN models to learn $\mathcal{H}^{-1}$, the mapping between RSS data and tuber cross-section images. That is, we aim to train the DNN model $\mathcal{F} : \mathbf{Y} \rightarrow \mathbf{r}$ to estimate the image vector $\mathbf{r}$ from the RSS data matrix $\mathbf{Y}$:

$$\hat{\mathbf{r}} = \mathcal{F}(\mathbf{Y}; \Theta), \tag{4}$$

where $\hat{\mathbf{r}}$ is the estimate of $\mathbf{r}$ and Θ is the set of neural network parameters. During training, the neural network parameters are iteratively optimized using different loss functions, as discussed in Section III. During inference, the well-trained network $\mathcal{F}$ serves as $\mathcal{H}^{-1}$ to generate cross-section images from corrected RSS data.

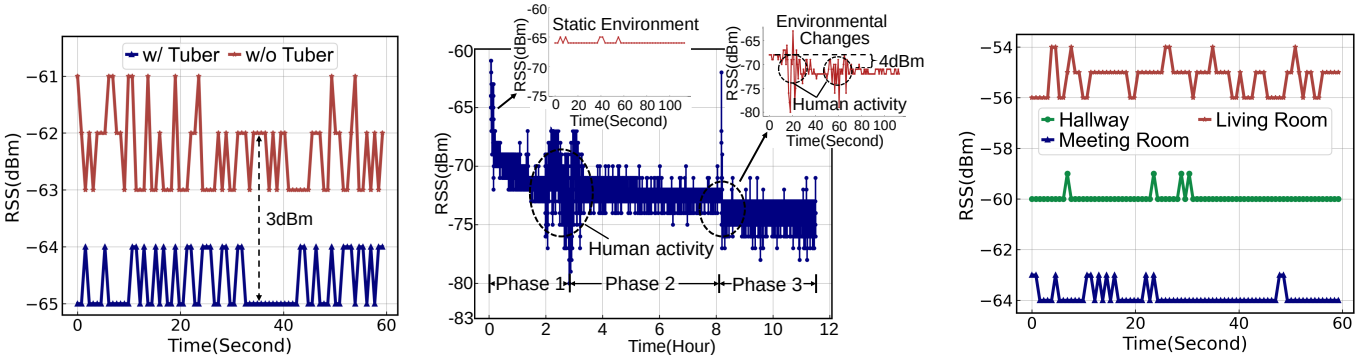
B. Framework Overview

The overview of the RTS framework is shown in Fig. 2, in which an RF sensor network-based testbed called Spin is used for data acquisition, and a novel DNN model called MC-Diffusion is used to reconstruct 2D cross-section images from RSS data. The Spin testbed is described in Section IV-A, and the MC-Diffusion model is described in Section III. We describe the RTS scenarios that our framework covers next.

In this paper, we define the following three RTS use-case scenarios: Case 1 - sensing in a static environment, Case 2 - sensing in an environment with environmental changes, and Case 3 - sensing across different environments. RSS time series data from these different scenarios are shown in Fig. 3. First, for RTS in a static environment (Case 1), the RSS time series $y_{g,l}[n]$ from a particular link l on a particular channel g can be very different for measurements $\mathbf{Y}^r$ with root tubers in soils and calibration measurements $\mathbf{Y}^c$ without tubers in soils. As shown in Fig. 3a, 3 dBm difference is observed due to the attenuation effect of a root tuber. Although RSS data from different links and channels have different variations due to the presence of a root tuber, "fingerprints" of all RSS data from an RF sensor network can be used to reconstruct root tuber cross-section images.

Second, RSS data is not only sensitive to the presence of root tubers, but also affected by the motion of other objects in the environment [10]. As shown in Fig. 3b, human activities occur at the beginning of phase 2, which causes short-term variations in RSS data. In addition, after the RSS data reaches a stable condition at the end of phase 2, a chair is moved at the beginning of phase 3, which causes an additional 4 dBm RSS variation. While the high variation RSS data caused by human motion can be detected and corresponding periods can be removed by using previous noise reduction methods [10], the RSS changes caused by other environmental conditions also need to be considered by the RTS framework. Thus, Case 2 refers to RTS with short-term variations in a dynamic environment, and a one-shot fine-tuning method is proposed to update the neural network online, which is discussed in detail in Section III-C.

Finally, we also aim to apply our DNN model trained by data collected in one environment to different environments, i.e., RTS across different environments (Case 3). From Fig. 3c, we see that for the same root tuber, RSS data from the same link on the same channel have significantly different values for three different environments. In addition, different soil moisture conditions can be seen as different environments, as RSS data from the same link also have significantly different values under different soil moisture conditions. As shown in Fig. 3b, an irrigation event occurs at the beginning of Phase 1, which causes the RSS data to decrease from -65 dBm to -73 dBm. While soil moisture and environmental changes can be partially compensated by subtracting calibration data $\mathbf{Y}^c$ from $\mathbf{Y}^r$, and using the one-shot fine-tuning method, we find that other domain adaptation methods are needed to deal with the cross-environment issues. Thus, we have designed different model components for fixed environment cases (Case 1 and Case 2) and cross-environment case (Case 3), as shown in Fig. 2. The DRL component designed for Case 3 will be discussed in detail in Section III-D. Detailed flowcharts for three use-case scenarios are provided in the supplementary material.



(a) RSS values in a static environment. (b) RSS values in a dynamic environment. (c) RSS values in different environments.

Fig. 3: RSS time series $y_{g,l}$ from a particular link l on a particular channel g in different sensing scenarios: (a) In a static environment (Case 1), 3 dBm difference is observed with and without root tubers. (b) In a dynamic environment, RSS variations of at least 4 dBm are observed with a person walking and moving a chair at the start of phase 3 (Case 2), and soil moisture level is changed after an irrigation event at the beginning of phase 1 (Case 3). (c) In different static environments, RSS values from the same RF link for the same tuber also show significant differences (Case 3).

III. MC-Diffusion Model

Our proposed MC-Diffusion model follows a two-stage imaging framework [11], [23], [24]. In the first stage, a feature extraction component is combined with a direct imaging component for initial imaging in the fixed environment, and with a DRL-based imaging component for initial imaging across environments. In the second stage, an image optimization component takes the initial reconstructions as input to further enhance imaging quality. Additionally, we propose a one-shot fine-tuning method to update the pre-trained model online, enabling adaptation to a dynamically changing environment. Details are provided below.

A. The Feature Extraction Component

In this study, we design a multi-branch convolutional neural network to extract discriminative features from RSS data, with each branch incorporating a pyramid structure. Since RF signals are collected from multiple links and frequency channels, the branches are configured with varying convolutional kernel sizes and depths to achieve significantly different receptive fields. This design allows each branch to focus on a scale-specific pattern in RF signals, enhancing the diversity of features and contributing to more comprehensive representations for imaging. As shown in Fig. 2, this network consists of three branches, each of which consists of multiple sequential convolutional blocks. Each block includes a convolutional layer for feature extraction, a BatchNorm layer to handle covariate shift and aid model convergence, and a ReLU activation layer to introduce nonlinearity to the neurons. The multi-branch CNN incrementally constructs representations from low to high levels. Subsequently, we sum features from the three convolutional branches and flatten the results to 1D vectors. These vectors are then fed into different components depending on different sensing scenarios, as discussed next.

B. The Direct Imaging Component

In a fixed environment, MC-Diffusion applies the direct imaging component to reconstruct initial images using high-dimensional features from the feature extraction component. To this end, we propose a novel neural network incorporating attention and convolution layers.

Specifically, the network first uses a linear layer to adjust the dimension of a vector from the feature extraction component. The output is reshaped into a two-dimensional feature map, with the width and height being one-ninth of the target image's dimensions. Then, an attention layer, implemented with a learnable weight matrix, is used to adaptively adjust the feature map, emphasizing target-related features while attenuating those unrelated to the target, i.e., root tuber. Subsequently, the network interpolates the feature map to triple both its length and width, and uses a 3×3 convolution layer to smooth the result. Finally, the network interpolates the feature map to the target size and refines the result using two 1×1 convolution layers. In this paper, bilinear interpolation is used for all interpolation operations. To improve the imaging quality in a fixed environment, this network is jointly optimized with the multi-branch CNN using a mean squared error (MSE) loss function, which is defined as [25]:

$$L_{mse} = \frac{1}{M} \sum_{k=1}^{k=M} ||\mathbf{r}_k - \mathcal{F}_{dir}(\mathcal{F}_{cnn}(\mathbf{Y}_k))||^2, \quad (5)$$

where $\mathcal{F}_{dir}$ denotes the network used in the direct imaging component, $\mathcal{F}_{cnn}$ represents the multi-branch CNN, and M denotes the number of training samples.

C. The One-shot Fine-tuning Component

In a dynamic environment, changes such as the relocation of environmental objects cause variations in RSS data, degrading the performance of the DNN model. To address this issue, we take advantage of the fact that a root tuber grows much slower than the human activity-induced or other environment-induced dynamic changes, and propose to use the one-shot

fine-tuning method to adjust the parameters of the pre-trained model, improving its generalization ability. Specifically, we first construct a well-trained MC-Diffusion model in a fixed environment, which includes the feature extraction component, the direct imaging component, and the image optimization component (discussed in Section III-E). Upon detecting an environmental change, the neural network parameters of the feature extraction and direct imaging components are fine-tuned using the most recent RSS data, while the parameters of the other components remain fixed. The MSE loss function is used to optimize neural networks during fine-tuning.

D. The DRL-based Imaging Component

Although corrected RSS data have been obtained for underground tuber imaging, they remain insufficient for accurate imaging across varying environments and soil conditions due to environmental multipath interference and differing attenuation caused by soil moisture variability [26]. In this study, we design a novel neural network based on DRL to enable robust imaging under varying multipath effects and signal attenuation caused by soil moisture variability. This component takes RF data collected under different environments as input and explicitly separates environment-invariant features, facilitating robust imaging across diverse environments. Additionally, by treating each soil condition as a distinct environment, the network can learn soil-independent features for imaging, ensuring robust performance and demonstrating its generalization ability across various sensing scenarios.

Specifically, a data pair $(\mathbf{Y}_u, \mathbf{Y}_q)$ is selected with the same ground truth, i.e., the same root tuber at the same location in the RF sensing area, from two environments u, q. This data pair is fed into the multi-branch CNN to generate the individual feature vectors, $\mathbf{O}_u$ and $\mathbf{O}_q$. Then, the DRL network uses a feature separator to divide features $\mathbf{O}_u$ and $\mathbf{O}_q$ into environment-related and environment-independent parts. The separator employs an attention mechanism that adaptively predicts the weight to split the original vector into two parts. The separating process can be expressed as:

$$\begin{aligned} \mathbf{O}_u^{ei} &= \mathcal{A}(\mathbf{O}_u) \times \mathbf{O}_u, \quad \mathbf{O}_u^{er} = (1 - \mathcal{A}(\mathbf{O}_u)) \times \mathbf{O}_u, \\ \mathbf{O}_q^{ei} &= \mathcal{A}(\mathbf{O}_q) \times \mathbf{O}_q, \quad \mathbf{O}_q^{er} = (1 - \mathcal{A}(\mathbf{O}_q)) \times \mathbf{O}_q, \end{aligned} \tag{6}$$

where $\mathbf{O}_u^{ei}$ and $\mathbf{O}_q^{ei}$ represent the environment-independent feature vectors, while $\mathbf{O}_u^{er}$ and $\mathbf{O}_q^{er}$ represent the environment-related feature vectors. The attention mechanism $\mathcal{A}$ is implemented by linear layers, and the outputs $\mathbf{O}_u^{ei}$ and $\mathbf{O}_q^{ei}$ are fed into additional linear layers to adjust their dimensions, followed by interpolation and reshaping for tuber imaging.

To ensure feature disentanglement, the DRL network uses a domain classifier that takes vectors generated by the feature separator as its inputs and aims to correctly distinguish the environment-independent vectors ($\mathbf{O}_u^{ei}$ and $\mathbf{O}_q^{ei}$) from the environment-related vectors ($\mathbf{O}_u^{er}$ and $\mathbf{O}_q^{er}$). The domain classifier is implemented through linear layers, producing a probability distribution corresponding to the environment-independent and environment-related classes. Additionally, since $\mathbf{O}_u^{ei}$ and $\mathbf{O}_q^{ei}$ correspond to the same tuber, the semantic information encoded in $\mathbf{O}_u^{ei}$ and $\mathbf{O}_q^{ei}$ should exhibit consistency or maximal similarity. Intuited by this, the DRL network performs the similarity measurement between $\mathbf{O}_u^{ei}$ and $\mathbf{O}_q^{ei}$, formulated as [25]:

$$L_{dis} = ||\mathbf{O}_u^{ei} - \mathbf{O}_q^{ei}||, \tag{7}$$

where $||\cdot||$ represents the mean absolute error (MAE) function. The similarity measurement result is used as a loss to optimize the neural network. In summary, three losses are produced by this network: an MSE loss from imaging, a cross-entropy loss from the domain classifier, and an MAE loss from the similarity measurement. The total loss is formulated as:

$$L_{drl} = L_{mse} + L_{dis} + L_{cro}, \tag{8}$$

where L_{cro} denotes the cross-entropy loss from the domain classifier. Finally, the DRL network is jointly optimized with the multi-branch CNN using the total loss L_{drl}.

During the inference phase, RSS data collected from a new environment or a new soil condition undergo feature extraction by the multi-branch CNN and feature separation by the DRL separator, generating environment-independent features for accurate cross-environment imaging.

E. The Image Optimization Component

Diffusion networks have demonstrated superior stability and performance compared to generative adversarial network-based (GAN-based) and autoencoder-based methods, as GANs often suffer from unstable training [27], and autoencoders may lose fine reconstruction details [28]. However, traditional diffusion models rely on high-parameter denoising networks and require numerous sampling steps to generate high-quality images, resulting in substantial computational burden and long inference time [29]. To address these challenges, we propose a novel latent diffusion network that first encodes 2D images into 1D vectors, and then performs the diffusion process on these representations. In contrast to existing works [30], [23] that perform diffusion directly in the image space, our network operates in a low-dimensional latent space, enabling the use of a lightweight denoising network and fewer sampling steps, thereby enhancing both architectural efficiency and sampling efficiency. After training the components for initial imaging, the latent diffusion network is trained from scratch using the initial reconstructions.

Specifically, the diffusion network first uses an encoder-decoder module to encode a 2D image sample $\hat{\mathbf{r}}$ into a vector, which is then decoded to produce a denoised result. The encoder-decoder module is implemented by a UNet-shaped transformer network (UTN), comprising transformer blocks with multi-head attention and gated feed-forward layers. To ensure that UTN decodes a denoised result, the diffusion network uses a prior representation learning module (PRM), which takes the concatenation of the image sample $\hat{\mathbf{r}}$ and its corresponding ground truth $\mathbf{r}$ as input to generate a prior representation $\mathbf{z} = PRM(\hat{\mathbf{r}}, \mathbf{r})$. The prior representation $\mathbf{z}$ is fed into UTN along with the image sample $\hat{\mathbf{r}}$ as a dynamic modulation parameter to denoise the feature map generated by UTN, ensuring a high-quality decoding result. In this paper,

PRM is implemented using a ViT network [31], and we first perform joint optimization of PRM and UTN using MSE.

Subsequently, the network applies the diffusion process to the prior representation, aiming to generate $\mathbf{z}$ without relying on the ground truth $\mathbf{r}$. During the training phase, the prior representation $\mathbf{z}$ from PRM is corrupted by noise in the forward diffusion process. In the backward diffusion process, a denoising network ϵ_θ, composed of linear layers, estimates the noise in $\mathbf{z}_n$, where $\mathbf{z}_n$ is the noisy result at time step n. The estimated noise is then used to obtain $\mathbf{z}_{n-1}$ to start the next iteration. After N iterations, the estimated result $\hat{\mathbf{z}}$ is obtained, which serves as a prior representation to be fed into UTN along with the sample $\hat{\mathbf{r}}$ to reconstruct a denoised image. The inputs of ϵ_θ are $\mathbf{z}_n$, n, and $\mathbf{c}$, where $\mathbf{c}$ is a condition vector used to control the backward diffusion process. To generate $\mathbf{c}$, the diffusion network uses a condition learning module (CLM) implemented with a ViT network, which takes only the image sample $\hat{\mathbf{r}}$ as input. Finally, CLM, ϵ_θ, and UTN are jointly optimized using losses from both the diffusion and image reconstruction processes, with MAE and MSE serving as their respective loss functions.

During the inference phase, a conditional vector $\mathbf{c}$ is first extracted from a test image sample using CLM. The denoising network ϵ_θ then uses a randomly sampled Gaussian noise $\mathbf{z}_N$ and $\mathbf{c}$ to estimate $\hat{\mathbf{z}}$ after N iterations. Finally, $\hat{\mathbf{z}}$ is used as the prior representation and input into UTN, along with the test image sample, to generate a high-quality result.Note that in this study, the denoising network consists of only 5 linear layers, significantly reducing the number of parameters and inference time. The diffusion process requires only 10 iterations, further enhancing computational efficiency.

Since edges of underground tubers in reconstructed images are occasionally blurry, MC-Diffusion uses the Canny algorithm [32] as a post-processing step to detect edges and define cross-section areas of tubers, further optimizing the reconstructed images. The region bounded by the detected edge is defined as the cross-section area of a tuber, with pixel values set to 1, while pixel values of other areas in the reconstructed image are assigned a value of 0. This post-processing step also facilitates the use of various evaluation metrics, which are discussed in Section V-A1.

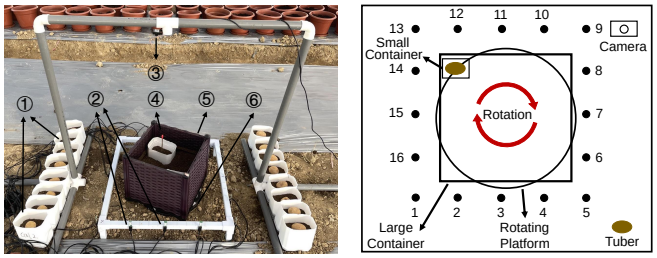
IV. Experiments, Dataset and Preprocessing

Using the Spin system, we conduct extensive measurement campaigns and build a novel wireless potato sensing (WPS) dataset for training and evaluation.

A. Data acquisition testbed

Our Spin testbed has the following hardware components.

1) RF sensor network with RSS measurements: The RF sensor network used in our experiments consists of 16 TI CC2531 (ZigBee) nodes deployed on a white rack, as shown in Fig. 4a. Each sensor node is programmed with a time-division multiple access (TDMA) communication protocol, and operates on one of the 16 frequency channels of the 2.4 GHz ISM band at a particular time [22]. In our experimental setup, a designated sink node receives all packets transmitted by RF network nodes. The sink node is connected to a laptop, on which the RSS measurements are stored and processed. Note that the RF sensor nodes used in this study are low cost, and our evaluations demonstrate that the cost increases by only USD 80, while the performance improves by more than 80%. Additional evaluation details are provided in the supplementary material.



(a) Spin testbed (b) Experimental layout

Fig. 4: The Spin data acquisition system: ① Small containers with different potato tubers, ② TI CC2531 nodes, ③ RGB camera, ④ Stick with colored marker, ⑤ Larger container, ⑥ Rotating platform.

2) Through-soil RTS toolkit: Collecting large amounts of RF sensing data is labor-intensive, especially for root tubers buried in soil. We build upon the data acquisition system developed in [33], which includes a rotating platform, and "plug-and-play" containers to facilitate data collection. First, we use two types of containers to avoid frequent soil digging and tuber burying. As shown in Fig. 4a, the large container contains soil with predefined locations to insert one or more small containers. Potato tubers with different dimensions are buried in different small containers with soil. Second, by placing the "nesting containers" on a rotating platform and rotating the platform with various predefined angles, root tubers can be located at various positions and orientations inside the sensing area. To automatically generate the ground truth for potato tubers, we use a stick with a colored marker, as shown in Fig. 4a, to indicate the position of the center of mass of a tuber within the sensing area. We use a camera and computer vision algorithms to capture the marker and generate tuber ground truth, which is discussed in more detail in Section IV-C1.

B. Experiments

Table I lists all experiments in the measurement campaigns. In Exp.1, we perform experiments in a static environment (Case 1 as described in Section II-B). The RF nodes are deployed on a rack of the size of 120 cm $\times$ 120 cm. We collect RSS data from 10 potato tubers, which are buried in a small container (container 1). The depth of potato tubers ranges from 11 cm to 13.5 cm. We randomly select 3 positions within a larger container (container 2) measuring 86 cm in length, 63 cm in width, and 48 cm in height, to place container 1, which are then rotated through 28 different angles on the rotating platform. In total, we collected RSS data with 840 different

TABLE I: Four sets of experiments in the WPS dataset.

Experiment	Tuber Dimensions (LxWxT) (cm)	RF sensor network area (cm)	No. of tubers	No. of positions	No. of RSS Measurements	Description
Exp.1	L:7.0-10.0 W:4.5-8.0 T:3.5-6.0	120×120	10	84	1,189,360	RTS at a static environment (Case 1).
Exp.2-1	L:2.0-10.0 W:1.0-8.0 T:1.5-6.3	72×72	6	4	178,016	RTS at a dynamic environment (Case 2).
Exp.2-2			26	1	69,312	
Exp.3-1	L:7.5-10.5 W:5.4-7.0 T:4.5-6.3	72×72	10	4	56,160	RTS across a hallway, living room and meeting room (Case 3).
Exp.3-2			10	4	55,744	
Exp.3-3			10	4	56,896	
Exp.4-1	L:5.1-12.8 W:4.8-7.2 T:N/A	60×60	6	2	43,536	RTS across two soil moisture levels (7.1% and 11.2%) and a new environment (Case 3).
Exp.4-2			6	2	43,856	
Exp.4-3			6	2	23,696	

tuber-position annotations, with each annotation matched to one-minute RSS measurements.

In Exp.2, we perform experiments in a dynamic environment with various environmental changes (Case 2). The sensing area of the RF sensor network is set to 72 cm by 72 cm for imaging 26 potato tubers, as shown in Exp.2-2 of Table I. Specifically, we perform two sets of experiments to investigate the effects of 1) human activity and 2) relocation of environmental objects. First, we collect RSS data from 6 tubers, each of which is placed in 4 different positions and recorded for 4 minutes at each position. During data collection, a person performs various activities within distances of 0.5 m to 2 m from the sensing area. Second, we collect RSS data from 26 tubers placed at a predefined position within the sensing area, while varying the environmental layout during data collection. We define the environment before and after altering the layout as E_1 and E_2, where E_1 is the initial environment and E_2 represents the environment after moving a chair and a paper box around the sensing area. Overall, we collect two sets of data with 26 potato tubers placed at depths ranging from 10.7 cm to 15.5 cm, with one-minute RSS data recorded for each tuber.

In Exp.3 and Exp.4, we perform experiments across different environments (Case 3). For Exp.3, we deploy our system in three environments: an empty hallway, an office meeting room, and a compact living room with furniture. For each environment, we collect RSS data from the RF sensor network for 10 potato tubers at 4 different locations, with one-minute RSS data recorded for each tuber at each position. Although we use the same RF sensor network with the same dimension and configuration for these experiments, different multi-path effects in these environments will cause domain shifts in collected RSS data. Thus, as discussed in Section III-D, RSS data from two environments are fed into the DRL network of MC-Diffusion to train for learning environment-independent features, while data from a third environment are used for evaluation. For Exp.4, we perform experiments using soil with different moisture levels. We use two types of soil with moisture levels of 7.1% and 11.2%, and conduct data collection in a hallway. As shown in Fig. 3b, RSS variations due to irrigation are time-varying. To account for the impact of different soil moisture conditions, we capture RSS data once they have stabilized. Specifically, we collect RSS data from six potato tubers at two positions using the RF sensor network, with each tuber recording four minutes of data at each position. To train and evaluate our DNN model, we also collect RSS measurements in a meeting room using the same tubers and soil with a moisture level of 7.1%, where we record two minutes of data for each tuber at each position.

C. Dataset Description

Our WPS dataset includes over 1.7 million RF network measurements and over one thousand ground truth annotations.

1) Ground Truth Annotation: We follow the procedures below to annotate the ground truth of the potato tuber cross-section image. First, the potato tuber is placed horizontally in the smaller container, ensuring its maximum cross-section is parallel to the ground. Then, one end of a stick is inserted at the center of mass of the potato tuber, with the other end marked with a colored marker and exposed above the soil surface, as shown in Fig. 4a. When the potato tuber and the container are rotated with the platform, a camera at a fixed location captures RGB images, and an object detection algorithm [34] is used to detect the position of the marker, i.e., the center of mass of the potato tuber, in the RGB image. Second, we construct a two-dimensional coordinate system for the sensing area and convert the position of the marker in the RGB image to the coordinate within the system, representing the center of mass of the tuber in the sensing area. Third, we use an elliptical function to approximately calculate the coordinates occupied by the potato tuber within the coordinate system. The calculation relies on the pre-measured dimension of the tuber, the detected coordinate of the center of mass of the tuber, and the recorded rotation angle. Fourth, we construct a ground truth image with each pixel corresponding to an individual region in the two-dimensional coordinate system. The ground truth image has values of 1 at pixels corresponding to the ellipse, and pixel values of 0 elsewhere. Note that, to speed up the annotation of the ground truth, we use the elliptical function to approximate the tuber cross-section area. We can use more advanced image segmentation algorithms to capture the ground truth of root tuber and we discuss it in Section V-E.

2) RSS data: In Exp.1, we collect RSS data from 840 tuber-position pairs, generating 1,189,360 measurements with different frequency channels. In Exp.2, the dataset can be

TABLE II: Parameters used in the MC-Diffusion model.

Parameter description	Default Value
Input dimension of the MC-Diffusion model.	$16 \times 16 \times 15$
Output dimension of the MC-Diffusion model.	360×360
The number of layers of the UTN encoder.	4
The number of layers of the UTN decoder.	4
The number of transformer blocks in UTN.	16
The number of attention heads in UTN	32
The number of attention heads in ViT	8
The number of diffusion sampling steps	10
The number of layers in ViT	1
Patch size used in ViT	30×30
Learning rate for initial imaging.	$5e^{-4}$
Learning rate for the diffusion network.	$1e^{-4}$

TABLE III: Performance of MC-Diffusion for Case 1. We mark the best and second-best results using bold and underlined text, respectively.

Method	Evaluation Metrics			
	SSIM↑	RPD↓	IoU↑	EDE(cm)↓
CD-EIT [30]	**0.99**	<u>0.09</u>	<u>0.81</u>	<u>2.70</u>
Two-CNN [11]	0.88	0.55	0.28	6.70
CNN-LSTM [35]	0.82	0.22	0.57	4.04
Trans-CNN [36]	0.96	0.16	0.78	3.45
Swin-Trans [37]	0.86	0.20	0.80	3.92
Linear-CD [23]	0.97	0.20	0.80	3.88
MC-GAN [24]	0.95	0.46	0.36	6.01
MC-VAE [38]	<u>0.98</u>	0.20	0.58	3.99
MC-Diffusion	**0.99**	**0.08**	**0.88**	**2.40**

categorized into two subsets. The first subset is collected during the introduction of human activity interference, with a total of 178,016 network measurements recorded. The second subset focuses on indoor deployment changes, frequently occurring during long-term monitoring periods. We gather RSS data from 26 fixed-position potato tubers, yielding a total of 69,312 measurements in this subset. In Exp.3, we collect 56,160, 55,744, and 56,896 measurements from various environments for 10 potato tubers and 4 positions, respectively. In Exp.4, we collect 43,536 and 43,856 measurements from 6 tubers placed in soil with moisture levels of 7.1% and 11.2%, respectively. To train and evaluate the MC-Diffusion model, we also collect 23,696 measurements in a new environment with soil moisture at 7.1%. In all experiments, we collect 1,716,576 network measurements from over 40 potato tubers with various shapes and dimensions. We make our WPS dataset publicly available at https://ieee-dataport.org/documents/underground-root-tuber-sensing-wireless-networks.

Note that to train and evaluate our RTS models, the RSS time series are preprocessed with data imputation and environmental change detection modules. We follow the method proposed in [33] to detect environmental changes. For data imputation, we use the nearest packets for imputing the missing data. To verify the efficacy of this imputation method, we compare its imaging quality against other methods, such as using a fixed value and the mean value. The results are discussed in the supplementary material.

V. Evaluation

We evaluate our framework through extensive experiments and use various metrics to assess imaging quality and detection accuracy for underground tubers.

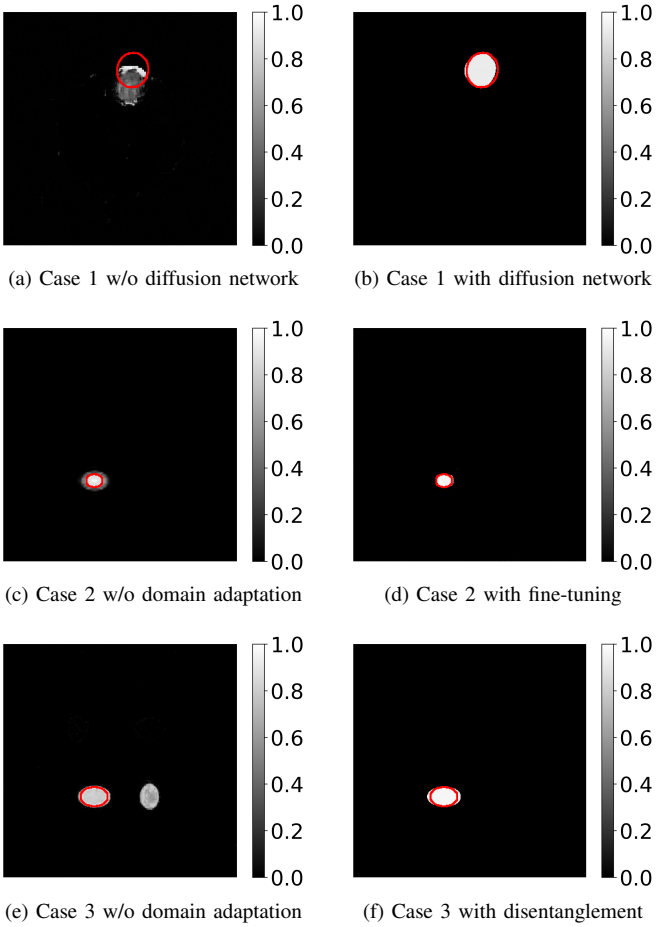
A. Metrics and Model Parameters

1) Evaluation metrics: Numerous evaluation metrics have been proposed to assess the quality of reconstructed images [39]. To comprehensively evaluate our framework, we use the following four metrics to assess the imaging quality and accuracy of MC-Diffusion: structural similarity index (SSIM) [40], intersection over union (IoU) [41], equivalent diameter error (EDE) [42] and relative pixel difference (RPD). First, SSIM is a popular metric used to quantify the imaging quality of an image reconstruction algorithm, ranging from 0 to 1, with higher values indicating better quality. Second, IoU is a widely used metric to quantify the similarity between the detection result and its ground truth, accounting for both position and shape accuracy. It ranges from 0 to 1, with a higher value indicating greater similarity. Third, we use two additional metrics to quantify the accuracy of tuber cross-section estimation. The EDE metric calculates the diameter of a circle whose area is equal to the absolute area difference between the estimated cross-section and the ground truth cross-section. We define the RPD metric as the ratio of the difference in pixel count between the estimated and ground truth cross-sections to the total pixel count of the ground truth cross-section. Smaller EDE and RPD values denote higher estimation accuracy. Note that, through the post-processing module described in Section III-E, the pixel counts for both the estimated and ground truth cross-sections can be obtained via simple summation. Furthermore, the cross-section area can be calculated by multiplying the pixel count by the real-world size represented by each pixel.

2) Model Parameters: In the multi-branch CNN, the first branch employs convolution kernels of sizes 3×3, 5×5, 5×5, and 6×6, with corresponding output channels of 128, 256, 512 and 1024, respectively. The second branch uses convolution kernels of sizes 9×9, 7×7, and 2×2, achieving 256, 512, and 1024 output channels. The third branch features convolution kernels of sizes 16×16 and 1×1, producing output channels of 512 and 1024. In the DRL network, the feature separator consists of two linear layers with output dimensions of 512 and 1, employing softmax as the activation function. The domain classifier comprises two linear layers with output dimensions of 128 and 2, using LeakyRelu as the activation function. Additionally, the DRL network includes two linear layers for imaging with output dimensions of 6400 and 14400, respectively, and uses LeakyRelu as the activation function. Other parameters are listed in Table II.

B. Baselines

For comparison, we select six state-of-the-art data-driven image reconstruction models as baselines, which can be



(a) Case 1 w/o diffusion network (b) Case 1 with diffusion network

(c) Case 2 w/o domain adaptation (d) Case 2 with fine-tuning

(e) Case 3 w/o domain adaptation (f) Case 3 with disentanglement

Fig. 5: Potato tuber imaging results from our model (right) and the Two-CNN model [11] (left) for three cases. The red circle indicates the cross-section ground truth of a potato tuber.

categorized into two groups: end-to-end models [35], [36], [37], [30] and two-stage models [11], [23]. First, for the end-to-end models, the CNN-LSTM model [35] incorporates convolutional networks into a long short-term memory model to directly reconstruct images from RSS time-series data. The Trans-CNN model [36] proposes an end-to-end model that uses a transformer network to extract features from RSS data, followed by convolutional networks for image reconstruction. The Swin-Trans model [37] uses a novel transformer network for feature extraction and image reconstruction. The CD-EIT model [30] proposes a diffusion network that takes RSS data as conditions to control the diffusion process in the image space, and then reconstructs the target images. Second, for the two-stage models, the Two-CNN model [11] develops a two-stage convolutional neural network for image reconstruction from RSS data. The first stage generates initial images directly from RSS data, while the second stage is used to further enhance image quality. The Linear-CD model [23] employs a two-stage network, in which the first stage uses linear layers for initial imaging, and the second stage applies a diffusion network operating in the image space to enhance image quality. Additionally, we employ RF-Diffusion [43], a generative AI

TABLE IV: Leave-k-out evaluation for Case 1. The variable "k" represents the number of test tubers.

k value / Method	k=1		k=2		k=3		k=4	
	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑
CD-EIT [30]	0.16	0.86	**0.14**	0.85	**0.17**	**0.83**	**0.14**	**0.85**
Two-CNN [11]	0.07	0.93	0.22	0.80	0.18	0.80	0.24	0.76
Linear-CD [23]	0.11	0.88	0.29	0.77	0.27	0.80	0.24	0.81
Trans-CNN [36]	0.24	0.60	0.16	0.78	0.32	0.49	0.30	0.49
Swin-Trans [37]	0.22	0.69	0.27	0.69	0.30	0.69	0.27	0.69
CNN-LSTM [35]	0.13	0.53	0.30	0.64	0.22	0.63	0.21	0.58
MC-Diffusion	**0.04**	**0.94**	0.15	**0.87**	0.18	**0.83**	0.16	0.83

model that has demonstrated superior performance compared to alternatives such as neural radiance fields [44], to synthesize RF data to improve DNN training. RF-Diffusion is integrated into both our model and the baselines for comparison, with the results provided in the supplementary material.

C. Evaluation for Various Use-cases

1) Case 1 Evaluation: We first evaluate the performance of MC-Diffusion in a static environment (Case 1), where RSS data are collected at 840 positions for 10 root tubers. All tuber-position pairs are split into a 9:1 ratio for training and evaluation. As shown in the first row of Fig. 5, our method exhibits higher imaging quality and estimation accuracy compared to the baseline model [11]. Although both models use convolutional neural networks for feature extraction and initial imaging, the diffusion network of our model effectively removes residual noise and generates more accurate results. As reported in Table III, our model achieves SSIM and IoU values of 0.99 and 0.88, respectively, outperforming all baseline models. Additionally, it achieves RPD and EDE values of 0.08 and 2.40, respectively, which are also superior to those of the baselines. These results demonstrate the efficacy of our model design in achieving high-quality and accurate imaging, compared to both two-stage and end-to-end models. Specifically, compared to end-to-end models [35], [36], [37], [30], our model first uses convolutional networks for feature extraction, which have shown superior performance over transformer-based models on datasets of limited sizes [31]. In addition, incorporating the diffusion network further reduces noise in the reconstructions, enabling the model to generate high-quality images. Compared to two-stage models [11], [23], our model not only uses a diffusion network for image refinement but also incorporates tailored components for initial imaging. These designs enable our model to outperform the baseline models.

To demonstrate the efficacy of the latent diffusion network in our model, we perform additional evaluations and compare it with other image refinement models, including GAN-based [24] and autoencoder-based [38] methods. We use the same neural networks for initial imaging, followed by different models for image refinement. To simplify notation, we refer to the model using GAN for image refinement as "MC-GAN" and the model using an autoencoder as "MC-VAE". Table III presents the evaluation results of various

TABLE V: Performance of MC-Diffusion for Case 2 (dynamic environments). E_1 and E_2 denote different conditions in the same environment. The best and second-best results are highlighted using bold and underlined text, respectively.

Method \ Test	$E_1 \rightarrow E_2$				$E_2 \rightarrow E_1$			
	Group 1		Group 2		Group 1		Group 2	
	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑
CD-EIT [30]	<u>0.13</u>	<u>0.82</u>	<u>0.16</u>	<u>0.80</u>	0.26	0.69	<u>0.21</u>	0.74
Two-CNN [11]	0.21	0.81	0.79	0.58	0.17	<u>0.86</u>	0.48	0.70
Linear-CD [23]	0.30	0.68	0.85	0.52	0.13	0.81	0.61	0.62
Trans-CNN [36]	0.36	0.75	0.98	0.54	0.78	0.63	0.67	0.49
Swin-Trans [37]	0.92	0.57	0.90	0.54	0.91	0.51	0.87	0.39
CNN-LSTM [35]	0.49	0.75	0.71	0.60	<u>0.12</u>	**0.87**	0.38	<u>0.75</u>
MC-Diffusion	**0.07**	**0.92**	**0.15**	**0.86**	**0.09**	<u>0.86</u>	**0.09**	**0.91**

TABLE VI: Performance of MC-Diffusion for Case 3 (cross environments). E_h, E_m, and E_l represent three different environments: hallway, meeting room, and living room.

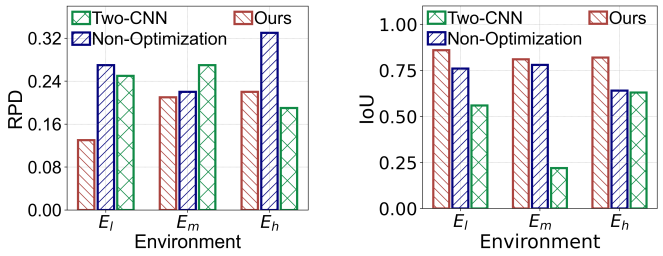
Method \ Test	$E_{h,m} \rightarrow E_l$		$E_{h,l} \rightarrow E_m$		$E_{m,l} \rightarrow E_h$	
	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑
CD-EIT [30]	0.20	0.70	0.24	0.78	0.23	<u>0.79</u>
Two-CNN [11]	0.25	0.56	0.27	0.22	**0.19**	0.63
Linear-CD [23]	0.26	0.71	0.33	0.70	0.31	0.59
Trans-CNN [36]	<u>0.17</u>	0.62	**0.17**	0.63	<u>0.22</u>	**0.82**
Swin-Trans [37]	0.35	0.44	0.29	0.38	0.28	0.63
CNN-LSTM [35]	0.33	0.23	0.30	0.13	0.27	0.34
MC-GAN [24]	**0.13**	<u>0.83</u>	0.20	<u>0.80</u>	**0.19**	0.76
MC-VAE [38]	<u>0.17</u>	0.73	<u>0.19</u>	0.68	0.30	0.77
MC-Diffusion	**0.13**	**0.86**	0.21	**0.81**	<u>0.22</u>	**0.82**

refinement methods in a static environment. First, the average RPD and EDE values of our model are 0.08 and 2.40, respectively, both lower than those of MC-GAN [24] and MC-VAE [38]. Second, our model achieves an average IoU value of 0.88, outperforming both MC-GAN [24] and MC-VAE [38]. These results indicate that the stable and progressive denoising process of the diffusion network enables it to capture more details of the sizes and shapes of underground tubers, thereby enhancing imaging quality and accuracy in the fixed-environment case. To further assess the efficiency of our diffusion network, we have performed comparisons with other lightweight diffusion networks, and the results are provided in the supplementary material. In addition, we have performed a parameter sensitivity experiment on the number of diffusion steps N, and the results demonstrate that increasing N beyond 10 yields diminishing returns in model accuracy. More details are provided in the supplementary material.

Furthermore, we perform leave-k-out evaluation. We use RSS data from 10 tubers, and the training-to-testing ratios for tubers are configured as 9:1, 8:2, 7:3, and 6:4. According to the results in Table IV, our model outperforms the baseline models in terms of RPD and IoU, achieving average values of 0.13 and 0.87, respectively, across different configurations. These results not only demonstrate improvements over the baseline models but also highlight the robustness of our model for accurate imaging across different conditions.

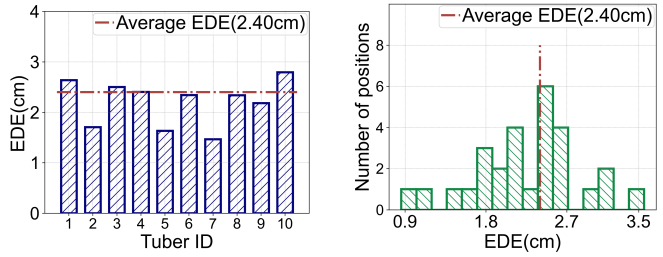
2) Case 2 Evaluation: As described in the experimental section, environmental changes can lead to noticeable variations in RSS measurements. To verify the robustness of our MC-Diffusion model, we randomly modify the environmental layout during data collection, thereby generating variations in the RSS data and creating different environmental conditions E_1 and E_2. We use data from one condition to build the pre-trained model, which is then fine-tuned and evaluated in the other condition. During testing, we select two groups: group 1 contains 5 tubers with an average size larger than that of group 2, which also consists of 5 tubers. The second row of Fig. 5 presents visualizations obtained using our method and the baseline. Although both methods have similar imaging quality, our method demonstrates higher estimated accuracy after fine-tuning. Table V compares the RPD and IoU values of our model under dynamic changes (Case 2) with those of the baseline models. Our model achieves average RPD values of 0.08 and 0.12 for two groups of tubers under different environmental conditions, which are lower than those reported by the baseline models. Moreover, it achieves an average IoU value of 0.89 for two groups, outperforming the baseline models. On the one hand, the limited data prevent transformer-based models [36], [37] from fully leveraging their strengths, resulting in suboptimal performance. On the other hand, RSS variations caused by dynamic changes lead to instability in the diffusion model [30], which directly uses RSS data as conditioning input. These variations in RSS data also affect the first-stage modules of two-stage networks [11], [23], resulting in performance degradation. In our model, we propose a one-shot fine-tuning method to update model parameters online, enabling automatic adaptation to new environmental conditions and mitigating the impact of data variations.

3) Case 3 Evaluation: To verify the ability of the MC-Diffusion model for cross-environment imaging, we collect RSS data from three different environments: a hallway, a meeting room, and a living room. We denote these environments as E_h, E_m, and E_l, respectively. As shown in the third row of Fig. 5, our method presents more accurate imaging compared to the baseline method, which exhibits a false positive blob adjacent to the tuber ground truth. We quantitatively evaluate our model using the RPD and IoU metrics for cross-environment imaging. As shown in Table VI, our model achieves average RPD and IoU values of 0.19 and 0.83, respectively, outperforming all baseline models. The performance of baseline models degrades when applied to new environments, as variations in multipath interference lead to significant changes in RSS data. This not only reduces imaging accuracy but also results in unstable performance. To address this issue, our model incorporates a DRL component that extracts environment-invariant features, so as to mitigate overfitting to specific environments and enable more robust and accurate imaging. We further compare different image refinement methods for cross-environment imaging. As shown in Table VII, our model achieves the best IoU performance,



(a) RPD (b) IoU

Fig. 6: Results of an ablation study: RPD and IoU metrics for the MC-Diffusion, Two-CNN, and Non-Optimization models. E_l,E_m and E_h denote the living room, the meeting room, and the hallway environments, respectively.



(a) EDE of each test tuber used in our experiment.

(b) EDE histogram for all test positions.

Fig. 7: EDE value for each test tuber and EDE histogram for all test positions.

with values of 0.86, 0.81, and 0.82 in the three target environments. All results outperform those of MC-GAN and MC-VAE, demonstrating the efficacy of the diffusion network in improving imaging quality and accuracy across environments.

Additionally, we perform an ablation study by removing the image optimization component from the MC-Diffusion model to evaluate cross-environment imaging. The remaining model is referred to as "Non-Optimization". Fig. 6 illustrates the cross-environment performance of three models: MC-Diffusion, Non-Optimization, and Two-CNN [11]. According to Fig. 6a, MC-Diffusion achieves RPD values of 0.13 and 0.22 in two target environments: E_l and E_h, each of which is lower than that of Non-Optimization (0.27 and 0.33). From Fig. 6b, it can be observed that MC-Diffusion achieves the highest IoU values in three environments. These results demonstrate the effectiveness of the diffusion network in enhancing the accuracy of estimating the shapes and sizes of tubers. Moreover, Non-Optimization achieves IoU values of 0.76, 0.78, and 0.64 in three environments, outperforming Two-CNN in each scene. This demonstrates the effectiveness of DRL in accurately estimating cross-sections and reducing interference from diverse environments.

Furthermore, we evaluate MC-Diffusion using RSS data collected under different soil conditions. As described in the experiment section, we perform experiments in a hallway under two soil moisture levels: 7.1% and 11.2%. Each soil condition is treated as an environment. We also collect data in a meeting room with soil at a moisture level of 7.1% to train and evaluate MC-Diffusion. We denote E_{h_1} and E_{h_2} as two soil conditions with moisture levels of 7.1% and 11.2%, respectively, and E_m represents the meeting room. Table VII presents the RPD and IoU values of MC-Diffusion and baseline models for imaging under different soil moisture levels. MC-Diffusion achieves an average IoU value of 0.79 under two soil moisture levels (E_{h_1} and E_{h_2}), outperforming all baseline models. The performance of baseline models degrades under new soil conditions, as varying moisture levels lead to different levels of RF signal attenuation, resulting in significant changes in RSS data. To address this problem, MC-Diffusion treats each soil condition as a distinct environment and uses the DRL component to extract soil-independent features from RSS data, enabling robust and accurate imaging across different soil conditions. We further compare various image refinement methods under different soil moisture levels. As shown in Table VIII, MC-Diffusion achieves IoU values of 0.82 and 0.75 under two soil conditions, respectively, outperforming MC-GAN [24] and MC-VAE [38]. In addition, the average RPD value of MC-Diffusion across two soil conditions is 0.17, which is lower than the 0.41 and 0.20 reported by MC-GAN and MC-VAE, respectively. These results not only demonstrate the efficacy of the latent diffusion network in enhancing imaging quality and accuracy across different soil conditions, but also highlight its generalization ability across diverse sensing scenarios.

TABLE VII: Performance of MC-Diffusion for Case 3. E_{h_1} and E_{h_2} represent two soil conditions with moisture levels of 7.1% and 11.2%, respectively, with data collected in a hallway. E_m represents a meeting room with the soil having a moisture level of 7.1%.

Method \ Test	$E_{h_1,m} \rightarrow E_{h_2}$		$E_{h_1,h_2} \rightarrow E_m$		$E_{h_2,m} \rightarrow E_{h_1}$	
	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑
CD-EIT [30]	0.15	0.80	0.22	0.56	0.28	0.68
Two-CNN [11]	**0.10**	0.79	**0.12**	0.42	**0.16**	0.47
Linear-CD [23]	0.70	0.35	0.48	0.49	0.87	0.20
Trans-CNN [36]	0.28	0.41	0.17	0.69	0.28	0.39
Swin-Trans [37]	0.48	0.63	0.23	0.54	0.34	0.59
CNN-LSTM [35]	0.24	0.50	0.21	0.40	0.28	0.38
MC-GAN [24]	0.36	0.69	0.20	0.75	0.46	0.54
MC-VAE [38]	0.22	0.73	0.19	0.76	0.17	0.74
MC-Diffusion	0.15	**0.82**	0.18	**0.80**	0.18	**0.75**

D. Ablation Study

In this study, we propose a novel data-driven model comprising distinct components to reconstruct cross-section images of underground tubers, enabling robust imaging across varying environments and soil conditions. To evaluate the contribution of each component, we perform ablation studies by individually removing the feature extraction, DRL-based imaging, and image optimization components. The resulting variants are referred to as "Non-Feature", "Non-DRL", and "Non-Optimization", respectively. We first evaluate the contribution of each component in achieving cross-environment

TABLE VIII: Performance of MC-Diffusion in the ablation study. E_h, E_m, and E_l represent three different environments: hallway, meeting room, and living room.

Method \ Test	$E_{h,m} \rightarrow E_l$		$E_{h,l} \rightarrow E_m$		$E_{m,l} \rightarrow E_h$	
	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑
Non-Feature	0.27	0.58	0.23	0.76	0.29	0.54
Non-DRL	0.15	0.82	**0.17**	0.78	0.23	0.66
Non-Optimization	0.27	0.76	0.22	0.78	0.33	0.64
MC-Diffusion	**0.13**	**0.86**	0.21	**0.81**	**0.22**	**0.82**

TABLE IX: Performance of MC-Diffusion in the ablation study. E_{h_1} and E_{h_2} represent two soil conditions with moisture levels of 7.1% and 11.2%, respectively, with data collected in a hallway. E_m represents a meeting room with the soil having a moisture level of 7.1%.

Method \ Test	$E_{h_1,m} \rightarrow E_{h_2}$		$E_{h_1,h_2} \rightarrow E_m$		$E_{h_2,m} \rightarrow E_{h_1}$	
	RPD↓	IoU↑	RPD↓	IoU↑	RPD↓	IoU↑
Non-Feature	**0.13**	0.78	0.20	0.71	0.29	0.69
Non-DRL	0.16	0.80	0.34	0.64	0.21	0.50
Non-Optimization	**0.13**	0.82	0.20	0.72	0.25	0.72
MC-Diffusion	0.15	**0.82**	**0.18**	**0.80**	**0.18**	**0.75**

imaging. As shown in Table VIII, Non-DRL achieves an average IoU value of 0.75 across all target environments, which is lower than 0.83 achieved by the complete model. This demonstrates the efficacy of the DRL-based component in enabling high-quality imaging across environments. Meanwhile, the complete model achieves an average RPD value of 0.18, outperforming 0.26 from Non-Feature and 0.27 from Non-Optimization. These results verify the efficacy of both the feature extraction and image optimization components in enhancing imaging accuracy. Table IX further presents the RPD and IoU values obtained from ablation studies in imaging across different soil moisture levels. The complete model achieves an average RPD value of 0.17 across all target scenarios, which is lower than the values reported by Non-Feature (0.21), Non-DRL (0.24), and Non-Optimization (0.19). Moreover, the complete model achieves IoU values of 0.82 and 0.75 for two soil moisture levels, respectively, demonstrating consistent improvements over the Non-Feature and Non-DRL models. These results further highlight the efficacy of the components incorporated in our model.

E. Discussion and Future Work

To show the feasibility of a data-driven approach in wireless RTS, we investigate the 2D cross-section image reconstruction of a single potato tuber, which is placed underground at a fixed depth with its maximum cross-section almost parallel to the soil surface. We use an elliptical function to approximate the truth shape of the root tuber to automate our cross-section ground truth annotation. While this approximation can introduce errors, we significantly speed up the building of a large RF sensing dataset. We have performed an initial study using a segmentation method to generate the ground truth, and have provided the experimental details and evaluation results in the supplementary material. Building upon our RTS framework, more advanced image segmentation and other algorithms can be used in future work to improve ground truth annotation accuracy. In addition, we can use multiple layers of networked nodes to perform 3D root tuber imaging, given more data collected from various tuber orientations and depths in the soil. We also leave it in future.

This paper focuses on 2D cross-section imaging of potato tubers to investigate the feasibility of using an RF sensor network for underground RTS. For cross-species RTS, we have performed an initial transfer learning experiment by training MC-Diffusion on potato root tuber data and testing it on carrot taproot data, with evaluation results provided in the supplementary material. The results demonstrate the feasibility of our framework for cross-species sensing, but more investigations are needed in future.

To achieve high-quality imaging, the MC-Diffusion model uses a latent diffusion network for image optimization and takes RSS data as input. In case 1, the average equivalent diameter error (EDE) of our model improves by 40.59% compared to the baseline method [35], but still achieves an EDE value of 2.40 cm. To further analyze this error, we compute the EDE value for each tuber and position, respectively. The results are shown in Fig. 7. First, Fig. 7a illustrates the average EDE for each potato tuber across all positions. We find that the EDE values of most tubers are below the average of 2.40 cm, whereas tuber 1 and tuber 10 exhibit noticeably higher values of 2.64 cm and 2.80 cm, respectively. In our experiment, tuber 1 has the largest cross-section area (72 cm^2), while tuber 10 has the smallest cross-section area (33.75 cm^2). This indicates that our model is effective for most tubers, but its performance decreases for those with areas not fully represented in the dataset. Second, we compute the average EDE of all tubers at each test position and categorize these values into 16 bins. As shown in Fig. 7b, the EDE values at most positions are below the average value, confirming the effectiveness of our model in these positions. However, there are still several positions where the model's error exceeds the average, indicating the potential for improvement. In future work, the diversity of potato tubers in the current WPS dataset can be further increased.

VI. Related Work

RF sensing. Recently RF sensing techniques have been proposed for various smart agriculture, food and forestry applications. For example, WiFi CSI data and Sub-Terahertz wireless signals are used for fruit ripeness sensing [14], [45]. RF tomographic imaging and machine learning algorithms extract moisture content in rice from signal strength data from a wireless network [46]. Ground penetrating radar (GPR) sensors are used to detect and retrieve images of underground tree roots [47], [48]. For underground tuber sensing, the study [49] uses acoustic signals to monitor the growth condition of sweet potatoes. While they show that sweet potatoes of different sizes have different patterns on an acoustic spectrum from 3.2 KHz to 20KHz, this research is in the preliminary stages.

Domain Adaptation. Domain adaptation learning is a long-stand field that enables to reduction of the gaps across various

domains. This ability has also been applied to many other learning areas, such as computer vision [50], nature language processing [51] and signal processing [52], [53]. The typical way of conducting domain adaptation with deep neural networks is to fine-tune a model pre-trained on the source domain using data from the target domain. For instance, To mitigate the challenges faced by millimeter-wave radio-based gesture recognition in heterogeneous environments, [53] designs an innovative domain adaptation approach. This approach allows for practical gesture recognition using pre-learned experiences with minimal target samples for fine-tuning. Based on the experimental findings, it has been demonstrated that achieving the same level of accuracy only requires retraining with as few as 8 samples per gesture. Meanwhile, [52] employs a novel method to fine-tune the pre-trained model. This process involves initially adjusting the data distribution in the source domain to align it with that of the target domain, followed by the utilization of a smaller amount of data from the target domain to further fine-tune the pre-trained model. In this paper, the novelty of our proposed domain adaptation learning strategy lies in its ability to address challenges using only one-shot samples in dynamic environments or without samples collected from new environments.

Diffusion Model. In recent years, diffusion models have demonstrated impressive performance in computer vision [54], [27], [55], [56]. Diffusion models adopt a parameterized Markov chain and generate a more accurate target distribution than other generative models [19]. For example, [57] employs the diffusion model for unconditional image synthesis, demonstrating image sample quality that surpasses the current state-of-the-art generative models. [56] further improves diffusion-based image synthesis with context prediction. In addition to image synthesis, diffusion models are widely used in other computer vision tasks. For instance, SR3 [58] introduces a diffusion model to image super-resolution, performing better than SoTA GAN-based methods. RePaint [59] applies the diffusion model to the image inpainting task, designing an improved denoising strategy by resampling iterations within the model. Traditional diffusion models operate directly on image pixels, necessitating numerous iterations, computational resources, and model parameters to achieve accurate predictions. In this paper, we adopt a latent diffusion model [19], [54] that initially trains an encoder-decoder network to compress the original image into a compact feature, which is then fed into the diffusion process. The reduced size of the feature map decreases both the model size and the number of iterations required for the diffusion network.

VII. Conclusion

This paper proposes a novel underground RTS framework that leverages RSS data from an RF sensor network and deep neural networks to reconstruct fine-grained cross-section images of root tubers. To enable the data-driven RTS, we first build a comprehensive dataset, and then combine the MC-Diffusion model with the one-shot fine-tuning method to make RTS robust in a dynamic environment. Furthermore, we propose and implement a novel DRL network to enable transferring the image reconstruction model across environments. Evaluation results show the efficacy of our RTS framework in various conditions and environments.

References

[1] Z. Zhang, W. Jin, R. Dou, Z. Cai, H. Wei, T. Wu, S. Yang, M. Tan, Z. Li, C. Wang *et al.*, "Improved estimation of leaf area index by reducing leaf chlorophyll content and saturation effects based on red-edge bands," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[2] P. Castro-Valdecantos, O. E. Apolo-Apolo, M. Pérez-Ruiz, and G. Egea, "Leaf area index estimations by deep learning models using rgb images and data fusion in maize," *Precision agriculture*, vol. 23, no. 6, pp. 1949–1966, 2022.

[3] E. Cheng, F. Wang, D. Peng, B. Zhang, B. Zhao, W. Zhang, J. Hu, Z. Lou, S. Yang, H. Zhang *et al.*, "A gt-lstm spatio-temporal approach for winter wheat yield prediction: From the field scale to county scale," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

[4] T. Roitsch, L. Cabrera-Bosquet, A. Fournier, K. Ghamkhar, J. Jiménez-Berni, F. Pinto, and E. S. Ober, "New sensors and data-driven approaches—a path to next generation phenomics," *Plant Science*, vol. 282, pp. 2–10, 2019.

[5] Y. Lin, S. Li, S. Duan, Y. Ye, B. Li, G. Li, D. Lyv, L. Jin, C. Bian, and J. Liu, "Methodological evolution of potato yield prediction: a comprehensive review," *Frontiers in Plant Science*, vol. 14, p. 1214006, 2023.

[6] J. K. Van Harsselaar, J. Claußen, J. Lübeck, N. Wörlein, N. Uhlmann, U. Sonnewald, and S. Gerth, "X-ray ct phenotyping reveals bi-phasic growth phases of potato tubers exposed to combined abiotic stress," *Frontiers in Plant Science*, vol. 12, p. 613108, 2021.

[7] W. Luo, Y. H. Lee, T. Hao, M. L. M. Yusof, and A. C. Yucel, "Automatic dual-polarized ground penetrating radar for enhanced 3d tree roots system architecture reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[8] S. Li, X. Cui, L. Guo, L. Zhang, X. Chen, and X. Cao, "Enhanced automatic root recognition and localization in gpr images through a yolov4-based deep learning approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[9] A. Aboudourib, M. Serhir, and D. Lesselier, "A processing framework for tree-root reconstruction using ground-penetrating radar under heterogeneous soil conditions," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 208–219, 2020.

[10] Y. Zhao and N. Patwari, "Robust estimators for variance-based device-free localization and tracking," *IEEE Transactions on Mobile Computing*, vol. 14, no. 10, pp. 2116–2129, 2014.

[11] H. Wu, X. Ma, C.-H. H. Yang, and S. Liu, "Convolutional neural network based radio tomographic imaging," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2020, pp. 1–6.

[12] J. Wilson and N. Patwari, "Radio tomographic imaging with wireless networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 5, pp. 621–632, 2010.

[13] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6258–6267, 2016.

[14] Y. Liu, L. Jiang, L. Kong, Q. Xiang, X. Liu, and G. Chen, "Wi-fruit: See through fruits with smart devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–29, 2021.

[15] L. Dunlap, A. Umino, H. Zhang, J. Yang, J. E. Gonzalez, and T. Darrell, "Diversify your vision datasets with automatic diffusion-based augmentation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[16] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided cnn for image denoising," *Neural Networks*, vol. 124, pp. 117–129, 2020.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[19] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.

[20] C. Yang, Y. Shen, Z. Zhang, Y. Xu, J. Zhu, Z. Wu, and B. Zhou, "One-shot generative domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7733–7742.

[21] S. Lee, S. Cho, and S. Im, "Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 252–15 261.

[22] O. Kaltiokallio, M. Bocca, and N. Patwari, "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, 2012, pp. 254–262.

[23] A. Denker, Željko Kereta, I. Singh, T. Freudenberg, T. Kluth, P. Maass, and S. Arridge, "Data-driven approaches for electrical impedance tomography image segmentation from partial boundary data," *Applied Mathematics for Modern Challenges*, vol. 2, no. 2, pp. 119–139, 2024.

[24] Z. Chen, C. Chen, C. Shao, C. Cai, X. Song, Y. Xiang, R. Liu, and Q. Xuan, "Mitnet: Gan enhanced magnetic induction tomography based on complex cnn," *IEEE Sensors Journal*, 2024.

[25] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[26] M. C. Vuran and I. F. Akyildiz, "Channel model and analysis for wireless underground sensor networks in soil medium," *Physical communication*, vol. 3, no. 4, pp. 245–254, 2010.

[27] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[28] K. Pandey, A. Mukherjee, P. Rai, and A. Kumar, "Vaes meet diffusion models: Efficient and high-fidelity generation," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[29] Y. Zhao, Y. Xu, Z. Xiao, H. Jia, and T. Hou, "Mobilediffusion: Instant text-to-image generation on mobile devices," in *European Conference on Computer Vision*. Springer, 2024, pp. 225–242.

[30] S. Shi, R. Kang, and P. Liatsis, "A conditional diffusion model for electrical impedance tomography image reconstruction," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–16, 2025.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[33] Y. Zhao, T. Wang, and S. Elhadi, "Data-driven rf tomography via cross-modal sensing and continual learning," in *2025 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2025, pp. 1–6.

[34] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[35] H. Wu, X. Ma, C. H. Yang, and S. Liu, "Attention based bidirectional convolutional LSTM for high-resolution radio tomographic imaging," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 68, no. 4, pp. 1482–1486, 2021.

[36] H. Wu, C. Cheng, T. Peng, H. Zhou, and T. Chen, "Combining transformer with a latent variable model for radio tomography based robust device-free localization," *Computer Communications*, vol. 231, p. 108022, 2025.

[37] N. Fan, Z. Tian, A. Dubey, S. Deshmukh, R. Murch, and Q. Chen, "Multitarget device-free localization via cross-domain wi-fi rss training data and attentional prior fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 91–99.

[38] H. Wang, L. Wang, Z. Wang, L. Ma, and Y. Luo, "Ssc-vae: Structured sparse coding based variational autoencoder for detail preserved image reconstruction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7665–7673.

[39] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[41] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[42] Z. Luo, H. Liu, D. Li, and K. Tian, "Analysis and compensation of equivalent diameter error of articulated arm coordinate measuring machine," *Measurement and Control*, vol. 51, no. 1-2, pp. 16–26, 2018.

[43] G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. X. Han, "Rf-diffusion: Radio signal generation via time-frequency diffusion," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 77–92.

[44] X. Zhao, Z. An, Q. Pan, and L. Yang, "Nerf2: Neural radio-frequency radiance fields," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023, pp. 1–15.

[45] S. S. Afzal, A. Kludze, S. Karmakar, R. Chandra, and Y. Ghasempour, "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*, 2023, pp. 1–15.

[46] A. A. Almaleeh, A. Zakaria, L. M. Kamarudin, M. H. F. Rahiman, D. L. Ndzi, and I. Ismail, "Inline 3d volumetric measurement of moisture content in rice using regression-based ml of rf tomographic imaging," *Sensors*, vol. 22, no. 1, p. 405, 2022.

[47] Y. Lu and G. Lu, "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*, 2022, pp. 68–77.

[48] A. Aboudourib, M. Serhir, and D. Lesselier, "A processing framework for tree-root reconstruction using ground-penetrating radar under heterogeneous soil conditions," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 1, pp. 208–219, 2021.

[49] J. Iwase, Y. Sato, D. Comparini, E. Masi, S. Mancuso, and T. Kawano, "Non-invasive acoustic sensing of tuberous roots of sweet potato (ipomoea batatas) growing belowground," *Advances in Horticultural Science*, vol. 29, no. 4, pp. 176–180, 2015.

[50] X. Gao, Y. He, S. Dong, J. Cheng, X. Wei, and Y. Gong, "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, 2023, pp. 24 236–24 245.

[51] Z. Zhang, E. Strubell, and E. Hovy, "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*, 2022, pp. 2627–2647.

[52] M. Alvi, R. Cardell-Oliver, and T. French, "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, BuildSys 2022,*. ACM, 2022, pp. 500–503.

[53] H. Liu, K. Cui, K. Hu, Y. Wang, A. Zhou, L. Liu, and H. Ma, "Mtranssee: Enabling environment-independent mmwave sensing based gesture recognition via transfer learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–28, 2022.

[54] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[55] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, "Person image synthesis via denoising diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5968–5976.

[56] L. Yang, J. Liu, S. Hong, Z. Zhang, Z. Huang, Z. Cai, W. Zhang, and B. Cui, "Improving diffusion-based image synthesis with context prediction," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[57] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021, pp. 8780–8794.

[58] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[59] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022*. IEEE, 2022, pp. 11 451–11 461.

Tao Wang (Student Member, IEEE) received the B.S. degree (2019) in computer science and technology from Jilin University and the MS degree (2021) in computer science and technology from Harbin Institute of Technology, respectively. He is currently pursuing the PhD degree at International Research Institute for Artificial Intelligence, Harbin Institute of Technology, Shenzhen. His current research interests include wireless sensing and artificial intelligence. He is a student member of the IEEE.

Yang Zhao (Senior Member, IEEE) received the B.S. degree (2003) in electrical engineering from Shandong University, the MS degree (2006) in electrical engineering from the Beijing University of Aeronautics and Astronautics, and the PhD degree (2012) in electrical and computer engineering from the University of Utah. He was a lead research engineer at GE Global Research between 2013 and 2021. He is currently a research professor at Harbin Institute of Technology, Shenzhen. His research interests include wireless sensing, edge computing, and autonomous intelligent system. He is a senior member of the IEEE.

Jinghua Wang (Member, IEEE) received his B.S. degree from Shandong University, MS degree from Harbin Institute of Technology, and PhD degree from The Hong Kong Polytechnic University. From 2014 to 2016, he was a research fellow with Nanyang Technological University, Singapore. From 2017 to 2022, he was a Research Assistant Professor with Shenzhen University. He is currently an Associate Professor with School of Computer Science and Software Engineering, Harbin Institute of Technology (Shenzhen), China. His current research interests include computer vision, multimodal learning and machine learning. He is a member of the IEEE.

Zhibin Huang (Student Member, IEEE) received the B.S degree(2024) in computer science and technology from Harbin Institute of Technology, Shenzhen. He is currently pursuing the MS degree at International Research Institute for Artificial Intelligence, Harbin Institute of Technology, Shenzhen. His current research interests include computer vision and artificial intelligence. He is a student member of the IEEE.

Jie Liu (Fellow, IEEE) is a Chair Professor at Harbin Institute of Technology Shenzhen (HIT Shenzhen), China and the Dean of its AI Research Institute. Before joining HIT, he spent 18 years at Xerox PARC and Microsoft. He was a Principal Research Manager at Microsoft Research, Redmond and a partner of the company. His research interests are Cyber-Physical Systems, AI for IoT, and energy-efficient computing. He received IEEE TCCPS Distinguished Leadership Award and 6 Best Paper Awards from top conferences. He is an IEEE Fellow and an ACM Distinguished Scientist, and founding Chair of ACM SIGBED China.

Qiaorong Wei received the B.S. degree (2001) in Agronomy from Northeast Agricultural University(NEAU), the M.S. degree(2005) in Crop Genetics and Breeding from NEAU. She is currently pursuing a PhD degree at NEAU. She has been serving on the faculty of the College of Agronomy at NEAU since 2006. She is Associate Director of the Crop Modeling Research Center at the State Key Laboratory of Smart Farm Technologies and Systems. Her research focuses on cultivation physiology and smart production of potatoes.