Multitarget Device-Free Localization via Cross-Domain Wi-Fi RSS Training Data and Attentional Prior Fusion

Na Fan*, Zeyue Tian*, Amartansh Dubey, Samruddhi Deshmukh, Ross Murch, Qifeng Chen

The Hong Kong University of Science and Technology
{nfanaa, ztianad, adubey, ssdeshmukh}@connect.ust.hk, {eermurch, cqf}@ust.hk

Abstract

Device-free localization (DFL) using easily-obtained Wi-Fi received signal strength (RSS) has wide real-world applications for not requiring people to carry trackable devices. However, accurate multitarget DFL remains challenging due to the unknown number of targets, multipath interference (MPI), especially between nearby targets, and limited real-world data. In this study, we pioneeringly propose a transformer-based learning method with Wi-Fi RSS as input, and an attentional prior fusion module, to simultaneously locate an unknown number of people at random positions. To overcome the multitarget data collection challenges, we contribute a large-scale cross-domain real-simulation-augmentation training dataset with one and two real-world nearby non-person objects at limited positions and up to five simulated and augmented randomly distributed targets. Experimental results demonstrate our method's improved accuracy, generalization ability, and robustness with fewer Wi-Fi nodes than previous methods.

Introduction

Device-free localization (DFL) is a passive non-cooperative localization method that eliminates the need for targets carrying trackable devices, ideal for applications like intelligent buildings, intrusion detection, emergency rescue, and healthcare monitoring. In these scenarios, multitarget DFL systems are usually required to simultaneously detect and accurately locate multiple targets for practical use.

To deploy such a system, several factors must be considered, such as cost-effectiveness, commercial viability, and adaptability to environmental changes. Free from issues like lighting conditions, occlusion, and privacy concerns, Wi-Fi signals that can even go through walls provide an alternative to cameras. Some recent works utilize Wi-Fi phase information (Adib et al. 2015), but the requirements for expensive duplexers and probably new spectrum allocations for required bandwidth may make it commercially unviable. Instead, we focus on cost-effective Wi-Fi received signal strength (RSS) systems as shown in Fig. 1a that are easily accessed from existing wireless communication systems. While numerous Wi-Fi nodes around the domain of interest (DOI) pose a challenge for deployment, researchers are exploring ways to maintain performance with fewer nodes for commercial use. (Xu et al. 2012; Bocca et al. 2014; Xu et al. 2016; Wang et al. 2017; Ma et al. 2023). The deployment can also be simplified using drones carrying nodes (Karanam and Mostofi 2017) or intelligent reflective surfaces integrated into building materials (Wu et al. 2021).

Despite the extensive research for single-target scenarios (Wilson and Patwari 2010; Shit et al. 2019), multitarget localization in an RSS-based DFL system is still challenging, especially in cluttered indoor environments (Nannuru et al. 2013; Xu et al. 2013; Sabek, oustafa Youssef, and Vasilakos 2015; Shit et al. 2019). Firstly, the association between the RSS measurements and each target is difficult to draw, particularly with an unknown number of targets and a lack of shape information (Thouin, Nannuru, and Coates 2011; Xu et al. 2012; Wilson and Patwari 2012; Nannuru et al. 2013; Wang et al. 2020). Secondly, the simultaneous presence of multiple targets makes the RSS changes more unpredictable, degrading algorithms that utilize line-of-sight information (Bocca et al. 2014; Wang et al. 2017). Thirdly, MPI between multiple targets results in unpredictable noise (Zhang et al. 2022), which becomes more serious when some targets are nearby but hard to simulate due to complex physics. Additionally, the difficulty of collecting real-world data for multiple targets at different positions increase exponentially as the number of targets grows, making it hard to obtain training dataset for data-driven methods. Furthermore, the metrics used in single-target systems, such as MSE, need to be revised for evaluating localization accuracy in multitarget systems due to uncertainties in associating targets and potential variations in the number of estimated targets compared to the ground truth.

This paper aims to develop a learning-based algorithm that improves multitarget localization performance in a Wi-Fi RSS-based DFL system. We approach this task as an inverse imaging problem, utilizing measured RSS as input and generating an estimated probability map for the presence of each target in the DOI. Our end-to-end approach avoids information loss that may occur with methods through subsequent processing pipelines (Ma et al. 2020a,b; Fu et al. 2022). With the probability map representation, we do not require the number of targets known in advance.

*These authors contributed equally.

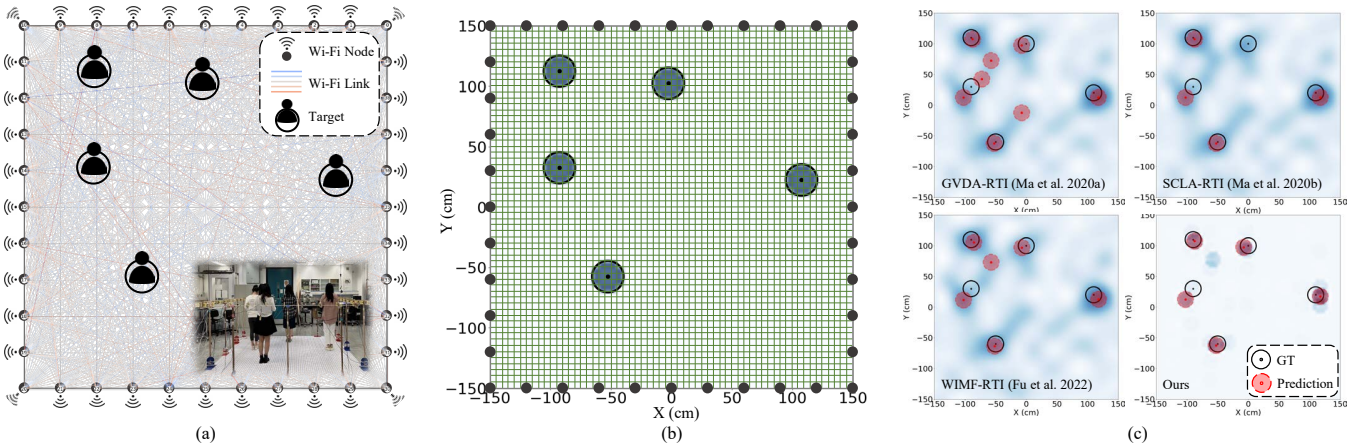


Figure 1: (a) A real-world test scenario with five people randomly distributed in the domain of interest (DOI) surrounded by Wi-Fi transceiver nodes. (b) The 3×3 m^2 DOI is discretized into 60×60 squares for ground truth construction. (c) Qualitative comparison of our method with baseline methods (Ma et al. 2020a,b; Fu et al. 2022) for this test scenario.

Our learning-based method pioneers adopting the Swin-Transformer (Liu et al. 2021) into this area as a network backbone to process measured RSS and estimate an initial probability map. To enhance generalization by leveraging prior localization information, we introduce a prior fusion module (PFM), utilizing the convolutional block attention modules (CBAM) (Woo et al. 2018). During training, we constrain losses between the ground truth and the initial and final fused estimations. To overcome multitarget data collection challenges, we propose a mixed real-simulation-augmentation training dataset incorporating manageable high-quality real-world nearby objects, and simulated and augmented random multitarget positions to improve generalization. Experimental results demonstrate the superior performance of our approach in accurate multitarget localization, effective generalization, and adaptation to environmental changes. Moreover, our method maintains comparable performance with fewer Wi-Fi nodes, highlighting its robustness and potential for energy-efficient deployment. Our contributions can be summarized as follows:

- We pioneeringly propose an end-to-end transformer-based learning method with an attentional prior fusion module within an Wi-Fi RSS-based DFL system to accurately localize an unknown number of multiple targets, outperforming previous methods.
- Our mixed real-simulation-augmentation training dataset addresses the challenges associated with collecting multitarget data. To our knowledge, this is the first large-scale multitarget Wi-Fi RSS dataset containing real-world data.
- Our method demonstrates generalization ability by accurately localizing multiple people at random positions, even without real people, more than two or randomly distributed real-world objects in the training data. Furthermore, it maintains competitive performance with fewer Wi-Fi nodes, indicating robustness and energy efficiency.

Related Work

Fingerprint-Based Multitarget RSS-DFL

Fingerprint-based methods utilize RSS as fingerprints and involve offline training and online testing. During training, a database is built using fingerprints at all possible positions. During testing, the fingerprint in the database is matched to determine the position. For single-target localization, fingerprints can be exhaustively collected, but for multitarget localization, building the database becomes impractical (Sabek and Youssef 2012; Sabek, oustafa Youssef, and Vasilakos 2015). Some methods address this issue by training on data from a single person and testing on multiple people, achieving good results by using probabilistic classification models (Xu et al. 2012), sequential counting and parallel localization of each target (Xu et al. 2013), cross-calibration (Sabek and Youssef 2012), conditional random Markov field (Sabek, oustafa Youssef, and Vasilakos 2015), and dictionary learning (Li et al. 2017). However, these methods trained on data from a single person assume target sparsity, which limits their accuracy with nearby targets. Our approach, in contrast, faces data collection challenges by introducing a mixed training dataset, incorporating real-world nearby objects to eliminate sparsity assumptions, and simulated/augmented randomly distributed targets to minimize exhaustive data collection.

RTI-Based Multitarget RSS-DFL

Radio Tomography Imaging (RTI) (Wilson and Patwari 2010), which models the change of RSS with the target's presence as a linear system, has been developed for multitarget localization. With a fixed, known number of targets, an additive likelihood moment filter can track simulated targets (Thouin, Nannuru, and Coates 2011), and up to three real-world targets can be tracked by Nannuru et al. (Nannuru et al. 2013). To handle the unknown number of targets, Bocca et al. (Bocca et al. 2014) use a fade level weight to average RSS on multiple frequency channels and track up to

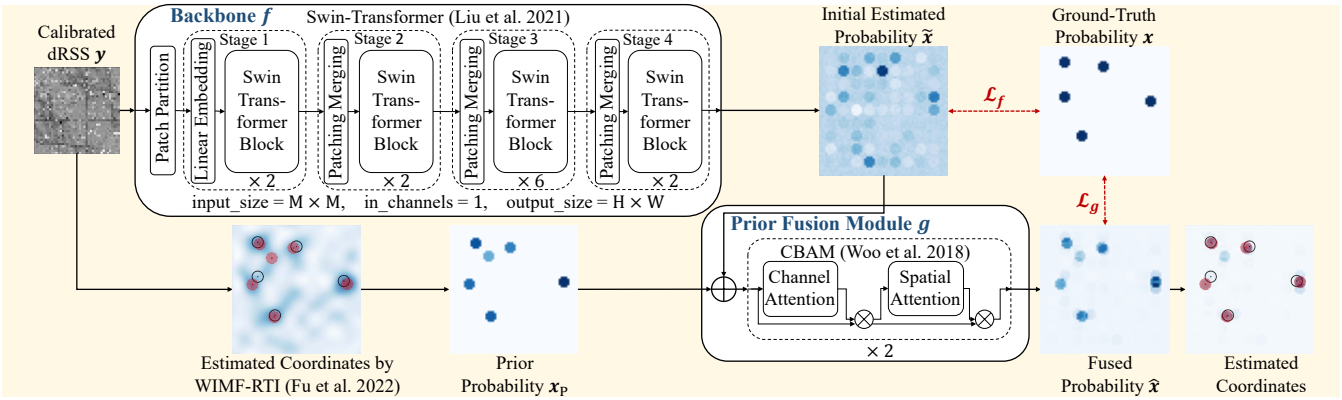


Figure 2: Our network architecture mainly consists of a Swin-Transformer-based (Liu et al. 2021) backbone and a doubled CBAM (Woo et al. 2018) attentional prior fusion module. The backbone takes the calibrated dRSS as input to estimate an initial probability map of the targets' presence. The prior fusion module further fuses a localization prior to the initial estimation to enhance robust generalization to an unknown number of targets at random, unseen positions. The estimated probability map is then post-processed to output coordinates for quantitative evaluation.

four targets with the Kalman filter; Wang et al. (Wang et al. 2016) propose an extended variational Bayesian Gaussian mixture model. Some recent works consider local maxima in the RTI image as candidates and use classifiers with extracted features: GVDA-RTI (Ma et al. 2020a) uses a gray value distribution analysis with a Bayes classifier, but the accuracy decreases with the increasing number of targets. The scanning circle link analysis (SCLA-RTI) (Ma et al. 2020b) improves GVDA-RTI with a random forest classifier but requires the distance among multiple targets to be at least 1 m. The weighted intersection multidimensional feature (WIMF-RTI) (Fu et al. 2022) is further developed to relocalize the final targets' positions. While reliable for identifying real targets from candidates, these methods that rely on local maxima in RTI images face limitations in restoring lost or distorted information during RTI imaging. These limitations become particularly evident when targets are nearby, leading to blurred areas and false high probabilities in the RTI image. Our learning-based method takes RSS as input to preserve as much information as possible to provide a high delineation of each target in the output probability map with an unknown number of targets.

Vision Transformer and Attentional Modules

Recent advances in Transformers have shown their potential to serve as superior alternatives to convolutional neural networks such as ResNet (He et al. 2016). The Vision Transformer (ViT) (Dosovitskiy et al. 2021) paves the concept of image tokenization and has achieved remarkable breakthroughs. Swin Transformer (Liu et al. 2021), an enhanced version of ViT, has become the state-of-the-art backbone in various vision tasks. Our work introduces the Swin Transformer as the backbone of our learning-based method, adapting it specifically to the Wi-Fi RSS-based DFL domain.

Attentional mechanisms have also been widely applied to improve feature representation through plug-and-play self-attentional blocks, such as squeeze-and-excitation (SE) (Hu, Shen, and Sun 2018), convolutional block attention module (CBAM) (Woo et al. 2018), and attentional feature fusion (AFF) (Dai et al. 2021). Inspired by these works, we introduce an attentional fusion module, utilizing a modified doubled CBAM block (Woo et al. 2018), to effectively fuse the localization priors with our initial estimation.

Method

Hardware System

Our DOI (Fig. 1a) is a 2D planar cross-section of a 3D space, surrounded by M Wi-Fi transceiver nodes evenly distributed at the boundary. The DOI's size is limited to 3×3 m^2, so signal fading is not a significant concern. Each node primarily consists of a SparkFun ESP32 Thing board (Espressif 2016) with an integrated 802.11 bgn Wi-Fi transceiver operating at 2.4 GHz. The nodes can transmit and receive signals from each other, allowing us to measure the RSS for each wireless link between any two nodes. They are lifted at the height of 1.2 m from the floor to minimize signal scattering due to the floor and objects outside the DOI.

For imaging, we discretize the DOI into $H \times W = 60 \times 60$ grids (Fig. 1b), each with a side length of 5 cm, which is smaller than the Wi-Fi wavelength of 12.5 cm, following the convention in inverse scattering (Deshmukh et al. 2022). People with larger radius as our localization targets therefore occupy several grids.

Problem Formulation

We formulate multitarget RSS-based DFL as a 2D inverse imaging problem: The presence of targets affects the RSS of wireless links passing through the DOI, allowing us to image a probability map of the targets' locations inversely. Given the input $\mathbf{y} \in \mathbb{R}^{M \times M}$ of calibrated RSS (dRSS) of all links formed by M nodes, we estimate each target's presence probability $\hat{\mathbf{x}} \in \mathbb{R}^{H \times W}$ at all grids.

Ground Truth Construction

Each target's presence can be modeled as a uniform circular area following the convention (Wilson and Patwari 2010). Knowing the position $\mathbf{p}_k \in \mathbb{R}^2$ within the DOI and radius $r_k \in \mathbb{R}^+$ of the k-th target, its ground-truth probability $\mathbf{x}_k \in \mathbb{R}^{H\times W}$ at each grid $\mathbf{q}$ can be constructed as

$$\mathbf{x}_k(\mathbf{q}) = \begin{cases} 1, & \text{if } \|\mathbf{q}-\mathbf{p}_k\|_2 \le r_k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Then assuming that no two targets overlap, the final ground-truth probability map $\mathbf{x}$ of simultaneously presented multiple targets can be represented as a sum of each target:

$$\mathbf{x} = \sum_k \mathbf{x}_k. \tag{2}$$

Learning-Based Model

Fig. 2 provides an overview of our learning-based approach. Our network architecture comprises two main components: a backbone network f_{θ_1} and a prior fusion module g_{θ_2}, each with learnable parameters θ_1 and θ_2, respectively.

Backbone. The backbone network f_{θ_1} is based on the Swin-Transformer architecture (Liu et al. 2021), which has been shown to be effective for a variety of vision tasks. Given a calibrated input dRSS $\mathbf{y} \in \mathbb{R}^{M\times M}$, we pass it through the backbone network to obtain a initial probability map estimation $\tilde{\mathbf{x}} \in \mathbb{R}^{H\times W}$, i.e.,

$$\tilde{\mathbf{x}} = f_{\theta_1}(\mathbf{y}). \tag{3}$$

We employ the loss function $\mathcal{L}_f$ to encourage the initial estimation to closely match the ground truth $\mathbf{x}$:

$$\mathcal{L}_f(\mathbf{x}, \tilde{\mathbf{x}}) = \mathcal{L}_f(\mathbf{x}, f_{\theta_1}(\mathbf{y})). \tag{4}$$

Prior fusion module. To improve the generalization ability and robustness of the network, we also incorporate localization priors. The prior $\mathbf{x}_\text{P} \in \mathbb{R}^{H\times W}$ can be an estimation result obtained from previous methods such as WIMF-RTI (Fu et al. 2022). We fuse the prior with the initial estimation using the prior fusion module g_{θ_2}, which is a doubled CBAM (Woo et al. 2018), to obtain the fused probability $\hat{\mathbf{x}} \in \mathbb{R}^{H\times W}$. Specifically,

$$\hat{\mathbf{x}} = g_{\theta_2}(\tilde{\mathbf{x}}, \mathbf{x}_\text{P}). \tag{5}$$

We employ another loss function $\mathcal{L}_g$ to guide the fused output consistent with the ground truth, i.e.,

$$\mathcal{L}_g(\mathbf{x}, \hat{\mathbf{x}}) = \mathcal{L}_g(\mathbf{x}, g_{\theta_2}(\tilde{\mathbf{x}}, \mathbf{x}_\text{P})) = \mathcal{L}_g(\mathbf{x}, g_{\theta_2}(f_{\theta_1}(\mathbf{y}), \mathbf{x}_\text{P})). \tag{6}$$

Loss. The total loss function is defined as

$$\mathcal{L} = \lambda_1\mathcal{L}_f + \lambda_2\mathcal{L}_g, \tag{7}$$

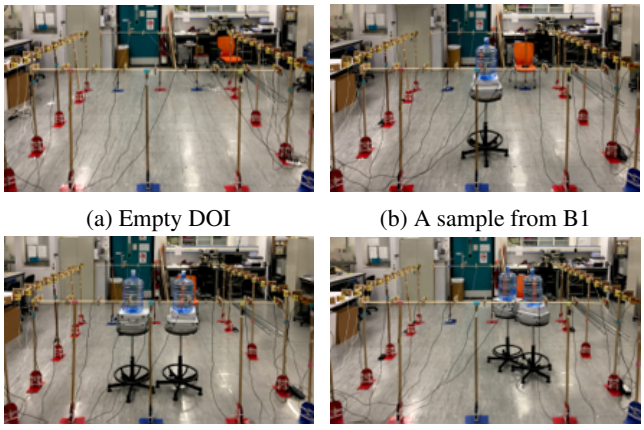
where λ_1 and λ_2 are hyperparameters that control the relative importance of the two loss terms. We adopt the mean squared error (MSE) loss for both $\mathcal{L}_f$ and $\mathcal{L}_g$, and set $\lambda_1 = \lambda_2 = 1$ in our experiments.

Optimization objective. We aim to find the optimal values of the learnable parameters θ_1 and θ_2 that minimize the total loss, i.e.,

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg\min_{\theta_1,\theta_2} \mathcal{L}. \tag{8}$$

	Collect.	Targets	Dist (cm)	Pos	Rds
Train	B1	1 bottle	-	81	150
	B2-60	2 bottles	60	126	75
	B2-42	2 bottles	42	128	150
	Sim	1-5 targets	> 60	10k	1
	Aug	2-5 bottles	> 30	120k	1
Test	P5-42	1-5 people	≥ 42	159	10

Table 1: Collections of real-world / simulated / augmented datasets. Distance (Dist) is measured between the centers of any two targets. Number of collection rounds (Rds) is an approximation, measured for each position combination (Pos).



(a) Empty DOI (b) A sample from B1

(c) A sample from B2-60 (d) A sample from B2-42

Figure 3: Real-world training data collection scenarios.

Dataset

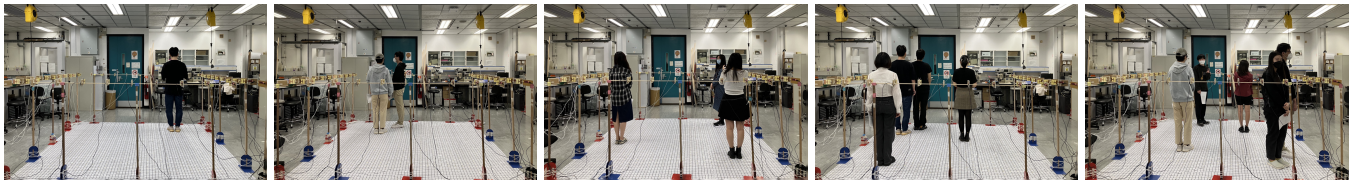
High-quality training data is crucial for achieving accuracy and robustness. However, large-scale multitarget Wi-Fi RSS datasets containing nearby targets are scarce due to the difficulty in simulating nearby targets and the significant effort required to collect real-world data for all possible target positions. We, therefore, propose a mixed dataset of real-world, simulated, and augmented multitarget RSS data (Table 1). The real-world data of single and two nearby water bottles focuses on the localization of nearby targets with complex MPI that is difficult to simulate. The simulated and augmented randomly distributed positions of non-nearby targets help to generalize to more targets at random positions.

Real-World Data Collection

One round of Wi-Fi signal collection. As nodes cannot transmit and receive simultaneously, they take turns transmitting while the rest receive. When all nodes have taken turns transmitting, we can form an $M \times M$ RSS measurement matrix $\mathbf{y}$, where each entry $\mathbf{y}_{i,j}$ represents the RSS received at node j from a signal transmitted by node i.

Calibration. As the RSS of the empty DOI can vary across different domains, we subtract the empty DOI's measurement from RSS to obtain the calibrated RSS (dRSS).

Real-world cross-domain training data. The real-world portion of our training data only consists of one and two



(a) 1 person (b) 2 people (c) 3 people (d) 4 people (e) 5 people

Figure 4: Test scenarios with an unknown number of 1-5 people randomly distributed at unseen position combinations.

nearby water bottles with similar scattering properties and radius (12.96 cm) to humans, as shown in Fig. 3b, 3c, 3d, and the first three rows of Table 1. Collection B1 consists of a single bottle at 81 positions with about 150 rounds per position. Collections B2-60 and B2-42 each involve two bottles, with a distance of about 60 cm and 42 cm from their cylinder centers, at 126 and 128 position combinations, respectively, and measured multiple rounds at each position.

Data Simulation

We simulate dRSS of randomly distributed targets ('Sim' in Table 1), adapting the simulation method (Deshmukh et al. 2022) to our setting. Due to the complexity of the MPI of nearby targets, a minimum center-to-center distance of 60 cm is kept between any two targets. We include 2000×5 position combinations, one round each, for 1-5 targets.

Data Augmentation

We augment multitarget dRSS from our real-world data to bridge the gap between simulated and real-world data ('Aug' in Table 1). We randomly select samples from B1, B2-60, and B2-42, compute their minimum dRSS value, and introduce noise to generate 2-5 targets at random positions.

Experimental Setup

Real-World Test Scenarios of Multiple People

We collected a test set of 14 people split into three groups after four months of training data collection, with changes to the environment layout. For each scenario, 1-5 people randomly stand in the DOI, with a minimum distance of about 42 cm between any two (Fig. 4). Finally, 159 scenarios, each with 10 rounds, are collected, with 31, 35, 33, 40, and 20 scenarios of 1-target to 5-target, respectively ('P5-42' in Table 1). We call the data of each round a sample.

Validation and Test Set Splitting

We randomly split the above data into non-overlapping validation and test scenarios, with approximately half the samples in each set. This ensures that all test samples are unseen in the training or validation sets. We denote the test set of 1-5 targets as $T_{1\text{-}5}$. Since our real training data is limited to 1-2 bottles, we further divide $T_{1\text{-}5}$ into $T_{1\text{-}2}$ and $T_{3\text{-}5}$ for 1-2 and 3-5 targets, respectively, to assess the performance of our proposed method across these different scenarios.

Using Fewer Wi-Fi Nodes

We reduce the number of utilized nodes M to 20 and 16, retaining the nodes' uniform distribution by setting the selected unused entries of the dRSS matrix to 0. As the priors of fewer nodes are not robust, we also take into account our models without the prior fusion module.

Training Details

Our experiments are performed on an Ubuntu 18.04 server with an NVIDIA GeForce RTX 2080 Ti. The network is trained end-to-end using a batch size of 256, a learning rate of 10^{-6}, and the Adam optimizer for 1000 epochs.

Post-Process: Probability Map to Coordinates

We post-process the probability map to obtain coordinates for quantitative evaluation. We pick the local maximum pixels as candidates and set a threshold of the number of targets to report according to the validation set performance.

Evaluation Metrics

Following previous works (Nannuru et al. 2013; Bocca et al. 2014; Sabek, oustafa Youssef, and Vasilakos 2015; Wang et al. 2016; Ma et al. 2020a,b; Fu et al. 2022), we adopt two widely used metrics for multitarget localization:

1) **Cardinality accuracy (CAcc).** (alias accuracy score, higher is better ↑) Let N be the total number of test samples and N_c be the number of test samples with the target number correctly estimated. We calculate the rate N_c/N of the number of targets identified correctly on the test set.

2) **Optimal subpattern assignment (OSPA).** (Schuhmacher, Vo, and Vo 2008) (lower is better ↓) Suppose $\mathcal{U} = \{u_1, \cdots, u_m\}$ and $\mathcal{V} = \{v_1, \cdots, v_n\}$ are the estimated and real target positions respectively. When $m \leq n$, we calculate

$$d_p^{(c)}(\mathcal{U}, \mathcal{V}) = \left(\frac{1}{n} \min_{\pi \in \Pi} \sum_{i=1}^{m} d^{(c)}\left(u_i, v_{\pi(i)}\right)^p + c^p(n-m) \right)^{1/p} \tag{9}$$

where Π is all possible permutations of $\mathcal{V}$, and $d^{(c)}(u, v) = \min\{d(u, v), c\}$. When $m > n$, we calculate $d_p^{(c)}(\mathcal{V}, \mathcal{U})$. We set $c = 40$ cm to penalize cardinality error and $p = 2$ to measure the Euclidean distance. OSPA is a metric dedicated to multitarget localization that finds the best permutation of the larger set to minimize its distance to the smaller set and penalizes the cardinality error. We calculate OSPA for each test sample and report the average on the test set.

		$T_{1\text{-}2}$				$T_{3\text{-}5}$				$T_{1\text{-}5}$			
Metric	M	GVDA	SCLA	WIMF	Ours	GVDA	SCLA	WIMF	Ours	GVDA	SCLA	WIMF	Ours
CAcc↑	16	0.209	0.009	0.116	**0.480**	0.040	0.000	0.020	**0.077**	0.111	0.003	0.057	**0.234**
	20	0.364	0.244	0.329	**0.516**	0.122	0.065	0.111	**0.131**	0.217	0.131	0.196	**0.260**
	40	0.676	0.427	0.711	**0.853**	0.435	0.290	0.557	**0.619**	0.529	0.343	0.617	**0.712**
OSPA↓	16	60.346	49.299	71.167	**42.423**	85.976	77.779	95.751	**64.933**	72.469	65.023	86.164	**55.674**
	20	42.043	49.251	49.509	**39.925**	65.102	62.209	67.529	**60.568**	56.110	56.873	60.151	**52.829**
	40	22.237	34.636	20.488	**14.545**	35.093	45.153	29.348	**25.545**	30.079	41.052	25.893	**21.301**

Table 2: Quantitative evaluation of our approach compared to GVDA-RTI (Ma et al. 2020a), SCLA-RTI (Ma et al. 2020b), and WIMF-RTI (Fu et al. 2022) (abbreviated as GVDA, SCLA, WIMF respectively in the table).

Experimental Results

Quantitative Evaluation

Table 2 demonstrates the performance of our method:

Accuracy. As shown in $M = 40$ rows, our method outperforms other methods on all test sets, with increases in CAcc values ranging from 11.13% to 19.97%, and decreases in OSPA values ranging from 12.96% to 29.01%, compared to the best baseline method.

With fewer nodes. Our method remains competitive with fewer Wi-Fi nodes after retraining the model with the unused entries of dRSS matrices set to 0 as input. (see Table 2 $M = 20$ and $M = 16$ rows). Despite the expected performance drop with fewer nodes, our method exhibits strong competitiveness in accuracy.

Performance differences on $T_{1\text{-}2}$, $T_{3\text{-}5}$, and $T_{1\text{-}5}$. Our model shows varying performance on different test sets, with the best performance on $T_{1\text{-}2}$, worst on $T_{3\text{-}5}$, and intermediate on $T_{1\text{-}5}$. This is probably due to the domain gap between the training and test sets. With the help of real-world training data of 1 and 2 nearby bottles, the model generalizes better to scenarios with 1 or 2 individuals. The domain gap is larger for scenarios with 3 to 5 individuals, as training data is obtained solely through simulation and augmentation.

Qualitative Evaluation

The samples shown in Fig. 5 provide more detailed insights.

Separate nearby targets. We show scenarios with two nearby targets in the second and third columns. The RTI (Wilson and Patwari 2010) creates a high-probability connected area around the targets while producing false high probabilities far from the targets, leading to inaccurate results for RTI-based methods. In contrast, our method produces a clear separation between the nearby targets.

Propose target beyond prior. In the last column, none of the baseline methods detect the target at the top right. Although our prior does not include it, our network predicts and shows it in the output probability map, indicating our method's generalization ability. Unfortunately, the target fails to pass post-processing and is not reported.

Ablation Study

We conduct experiments on $T_{1\text{-}5}$ using 40 nodes unless stated otherwise.

Training Set	OSPA↓ (w.o. / w. prior)		
	$T_{1\text{-}2}$	$T_{3\text{-}5}$	$T_{1\text{-}5}$
$s_{1\text{-}5}$	60.41 / 20.63	68.60 / 34.79	65.41 / 29.27
$r_{1\text{-}2}$+$s_{3\text{-}5}$	23.27 / 17.75	57.44 / 29.13	44.11 / 24.69
$r_{1\text{-}2}$+$a_{3\text{-}5}$	24.88 / 15.40	41.87 / 25.81	33.75 / 21.55
$r_{1\text{-}2}$+$s_{3\text{-}5}$+$a_{3\text{-}5}$	**22.45 / 14.54**	**41.25 / 25.55**	**33.62 / 21.30**

Table 3: Using different training sets without/with priors. 'r', 's', and 'a' denote real, simulated, and augmented data, respectively. The subscripts denote the target number range.

Arch.	CAcc↑ / OSPA↓		
	$T_{1\text{-}2}$	$T_{3\text{-}5}$	$T_{1\text{-}5}$
ResNet	0.827 / 15.597	0.591 / 27.536	0.672 / 23.183
ViT	0.831 / 15.790	0.617 / 25.713	0.690 / 22.429
Swin	**0.853 / 14.545**	**0.619 / 25.545**	**0.712 / 21.301**

Table 4: Using different network backbone architectures.

Mixed Cross-Domain Training Dataset

Table 3 shows the effectiveness of our training using mixed real-world, simulated, and augmented data (denoted by 'r','s','a'). The subscripts refer to the target number. For example, $r_{1\text{-}2}$ is the real data of 1 to 2 targets, using all 32,802 real samples. We control the number of samples in each training set to ensure a fair comparison. The first three rows in Table 3 use 20k simulated/augmented samples for each of the 3-5 target cases, while the last row uses 10k simulated and 10k augmented samples. Therefore, we have 100k training samples in $s_{1\text{-}5}$, and 92,802 samples in each of the other three training sets. We also avoid prior quality differences by using only initial estimations. The results show that using only simulation data for training leads to lower performance even with slightly more samples, justifying the need for real-world data. Using mixed real-simulation-augmentation datasets for training outperforms all other sets.

Network Backbone

We replaced the Swin-Transformer (Liu et al. 2021) with modified ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2021) while keeping the other structures unchanged. Table 4 shows that utilizing the Swin-Transformer as the network backbone is more effective than other alternatives.

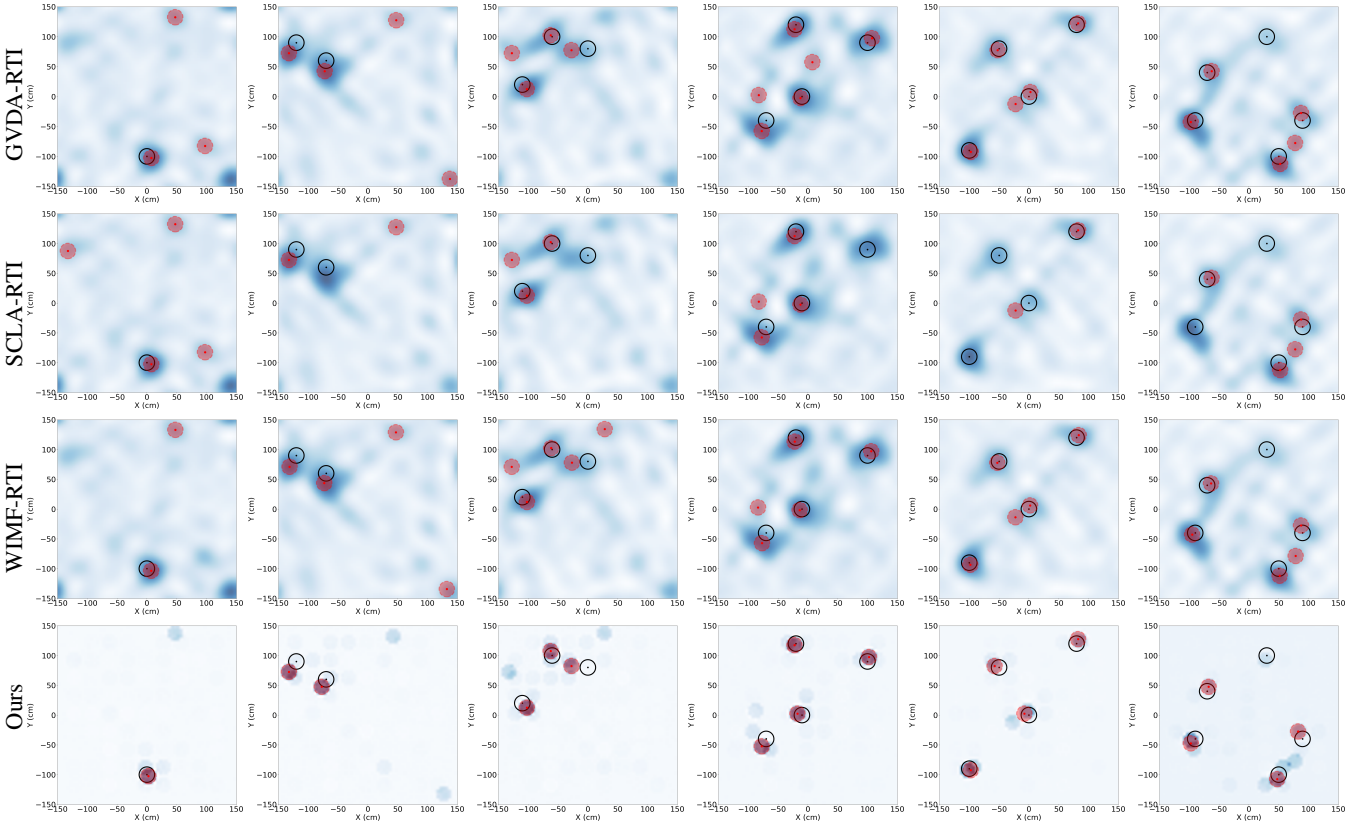


Figure 5: Qualitative comparison between our approach and GVDA-RTI (Ma et al. 2020a), SCLA-RTI (Ma et al. 2020b), and WIMF-RTI (Fu et al. 2022). The black circles denote ground truth, and the red circular areas denote estimations. The probability map background uses RTI results (Wilson and Patwari 2010) for the RTI-based methods and our network output for our method.

PFM	CAcc↑	OSPA↓	PFM	CAcc↑	OSPA↓
conv_{k1}	0.428	37.862	AFF	0.614	27.953
conv_{k3}	0.698	21.968	iAFF	0.684	23.864
conv_{k5}	0.697	22.690	CBAM_1	0.676	22.818
conv_{k7}	0.693	23.350	CBAM_2	**0.712**	**21.301**
SE_1	0.679	22.669	CBAM_3	0.651	24.952
SE_2	0.655	24.219	CBAM_4	0.653	24.886

Table 5: Using different prior fusion modules (PFM). conv_{ki} denotes a one-layer convolution of kernel size i. The other subscripts indicate the number of passes through the module.

Prior Fusion Module (PFM)

Table 5 shows our exploration of constructing the prior fusion module. Inspired by AFF and iAFF (Dai et al. 2021), we take the simple addition of initial estimation and prior as input, pass through a convolution layer, or the modified SE (Hu, Shen, and Sun 2018) or CBAM (Woo et al. 2018) multiple times. Finally, a doubled CBAM block is chosen.

Loss Constraint on Initial Estimation

While it may seem natural to directly constrain the loss between our fused probability map and the ground truth, Table 6 shows adding loss constraint $\mathcal{L}_f$ to encourage initial estimation closer to the GT enhances overall performance.

Loss	CAcc↑ / OSPA↓		
	$T_{1\text{-}2}$	$T_{3\text{-}5}$	$T_{1\text{-}5}$
$\mathcal{L}_g$	0.769 / 18.734	0.616 / 26.448	0.711 / 22.536
$\mathcal{L}_f + \mathcal{L}_g$	**0.853 / 14.545**	**0.619 / 25.545**	**0.712 / 21.301**

Table 6: Adding loss term $\mathcal{L}_f$ on the initial estimation.

Conclusion

This paper proposes an end-to-end transformer-based learning approach and contributes a large-scale mixed real-simulation-augmentation training dataset for accurate multi-target device-free localization using Wi-Fi RSS. Despite our real-world training data consisting of only up to two nearby bottles at limited position combinations, our transformer-based network architecture, along with an attentional prior fusion module and the simulated/augmented data, successfully localizes an unknown number of up to five people randomly distributed at previously unseen positions, with some may be nearby each other. Experimental results showcase our method's accuracy, generalization ability, and robustness with fewer nodes, outperforming the previous methods.

Ethical Statement

Our volunteers are informed of potential risks and provide consent for their images and Wi-Fi RSS measurements to be publicly available for research purposes.

If widely adopted, this technology could involve RSS data collection without consent. While RSS information is less directly linked to personal identification, collecting data in public spaces poses fewer risks. However, covert monitoring in private buildings may lead to potential societal harm.

Acknowledgments

This work was supported by the Hong Kong Grants Council Collaborative Research Fund (CRF) under Grant C6012-20G.

References

Adib, F.; Hsu, C.; Mao, H.; Katabi, D.; and Durand, F. 2015. Capturing the human figure through a wall. *ACM Trans. Graph.*, 34(6): 219:1–219:13.

Bocca, M.; Kaltiokallio, O.; Patwari, N.; and Venkatasubramanian, S. 2014. Multiple Target Tracking with RF Sensor Networks. *IEEE Trans. Mob. Comput.*, 13(8): 1787–1800.

Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; and Barnard, K. 2021. Attentional Feature Fusion. In *WACV*.

Deshmukh, S.; Dubey, A.; Ma, D.; Chen, Q.; and Murch, R. D. 2022. Physics assisted deep learning for indoor imaging using phaseless Wi-Fi measurements. *IEEE Transactions on Antennas and Propagation*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Espressif, S. 2016. ESP32 datasheet. https://www.sparkfun.com/products/13907. Accessed: 2024-01-16.

Fu, H.; Ma, Y.; Gong, X.; Zhang, X.; Wang, B.; Ning, W.; and Liang, X. 2022. Device-Free Multitarget Localization With Weighted Intersection Multidimensional Feature for Passive UHF RFID. *IEEE Sensors Journal*, 22(7): 7300–7310.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*.

Karanam, C. R.; and Mostofi, Y. 2017. 3D through-wall imaging with unmanned aerial vehicles using wifi. In *IPSN*, 131–142.

Li, X.; Ding, S.; Li, Z.; and Tan, B. 2017. Device-free localization via dictionary learning with difference of convex programming. *IEEE sensors journal*, 17(17): 5599–5608.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*.

Ma, D.; Zhang, Y.; Dubey, A.; Deshmukh, S.; Shen, S.; Zhang, Q.; and Murch, R. 2023. Millimeter-Wave 3-D Imaging Using Leaky-Wave Antennas and an Extended Rytov Approximation in a Frequency-Diverse MIMO System. *IEEE Transactions on Microwave Theory and Techniques*, 71(4): 1809–1825.

Ma, Y.; Wang, B.; Gao, X.; and Ning, W. 2020a. The Gray Analysis and Machine Learning for Device-Free Multitarget Localization in Passive UHF RFID Environments. *IEEE Trans. Ind. Informatics*, 16(2): 802–813.

Ma, Y.; Zhang, Y.; Wang, B.; and Ning, W. 2020b. SCLA-RTI: A novel device-free multi-target localization method based on link analysis in passive UHF RFID environment. *IEEE Sensors Journal*, 21(3): 3879–3887.

Nannuru, S.; Li, Y.; Zeng, Y.; Coates, M.; and Yang, B. 2013. Radio-Frequency Tomography for Passive Indoor Multitarget Tracking. *IEEE Trans. Mob. Comput.*, 12(12): 2322–2333.

Sabek, I.; oustafa Youssef; and Vasilakos, A. V. 2015. ACE: An Accurate and Efficient Multi-Entity Device-Free WLAN Localization System. *IEEE Trans. Mob. Comput.*, 14(2): 261–273.

Sabek, I.; and Youssef, M. 2012. Multi-entity device-free WLAN localization. In *GLOBECOM*, 2018–2023.

Schuhmacher, D.; Vo, B.; and Vo, B. 2008. A Consistent Metric for Performance Evaluation of Multi-Object Filters. *IEEE Trans. Signal Process.*, 56(8-1): 3447–3457.

Shit, R. C.; Sharma, S.; Puthal, D.; James, P.; Pradhan, B.; van Moorsel, A.; Zomaya, A. Y.; and Ranjan, R. 2019. Ubiquitous Localization (UbiLoc): A Survey and Taxonomy on Device Free Localization for Smart World. *IEEE Commun. Surv. Tutorials*, 21(4): 3532–3564.

Thouin, F.; Nannuru, S.; and Coates, M. 2011. Multi-target tracking for measurement models with additive contributions. In *FUSION*, 1–8.

Wang, J.; Fang, D.; Yang, Z.; Jiang, H.; Chen, X.; Xing, T.; and Cai, L. 2017. E-HIPA: An Energy-Efficient Framework for High-Precision Multi-Target-Adaptive Device-Free Localization. *IEEE Trans. Mob. Comput.*, 16(3): 716–729.

Wang, Q.; Yigitler, H.; Jäntti, R.; and Huang, X. 2016. Localizing Multiple Objects Using Radio Tomographic Imaging Technology. *IEEE Trans. Veh. Technol.*, 65(5): 3641–3656.

Wang, Z.; Qin, L.; Guo, X.; and Wang, G. 2020. Dual-Radio Tomographic Imaging With Shadowing-Measurement Awareness. *IEEE Trans. Instrum. Meas.*, 69(7): 4453–4464.

Wilson, J.; and Patwari, N. 2010. Radio Tomographic Imaging with Wireless Networks. *IEEE Trans. Mob. Comput.*, 9(5): 621–632.

Wilson, J.; and Patwari, N. 2012. A Fade-Level Skew-Laplace Signal Strength Model for Device-Free Localization with Wireless Networks. *IEEE Trans. Mob. Comput.*, 11(6): 947–958.

Woo, S.; Park, J.; Lee, J.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *ECCV*.

Wu, Q.; Zhang, S.; Zheng, B.; You, C.; and Zhang, R. 2021. Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial. *IEEE Trans. Commun.*, 69(5): 3313–3351.

Xu, C.; Firner, B.; Moore, R. S.; Zhang, Y.; Trappe, W.; Howard, R. E.; Zhang, F.; and An, N. 2013. SCPL: indoor device-free multi-subject counting and localization using radio signal strength. In *IPSN*, 79–90.

Xu, C.; Firner, B.; Zhang, Y.; and Howard, R. E. 2016. The Case for Efficient and Robust RF-Based Device-Free Localization. *IEEE Trans. Mob. Comput.*, 15(9): 2362–2375.

Xu, C.; Firner, B.; Zhang, Y.; Howard, R. E.; Li, J.; and Lin, X. 2012. Improving RF-based device-free passive localization in cluttered indoor environments through probabilistic classification methods. In *IPSN*, 209–220.

Zhang, X.; Ma, Y.; Gong, X.; Fu, H.; Wang, B.; Ning, W.; and Liang, X. 2022. A Training-Free Multipath Enhancement (TFME-RTI) Method for Device-Free Multi-Target Localization. *IEEE Sensors Journal*, 22(7): 7399–7410.