RF-Prox: Radio-based Proximity Estimation of Non-directly Connected Devices

Yuchong Gao, *Student Member, IEEE,* Guoxuan Chi*, *Member, IEEE,*
Zheng Yang, *Fellow, IEEE,* Shijie Cheng, *Student Member, IEEE*, Zhiqing Wei, *Member, IEEE*

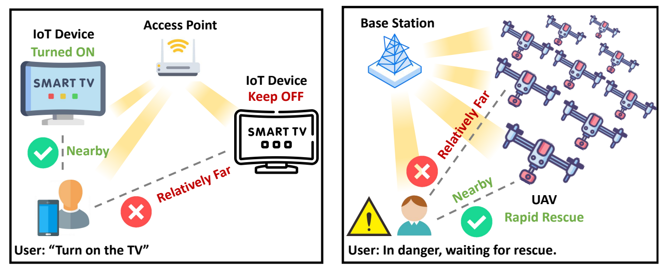
***Abstract*—Recent years have witnessed an increasing number of mobile devices, posing a more diversified demand for device localization solutions. Existing methods can locate connected devices but fail to address the spatial proximity between devices lacking direct communication links. This limitation impedes numerous emerging applications, such as implicit control of IoT device and proximity-based UAVs scheduling. In response to this technical challenge, we introduce *RF-Prox*, the pioneering system designed for the proximity estimation of non-directly connected devices. *RF-Prox* determines the proximity between devices by extracting and analyzing the spatio-temporal correlation between two signals. RF-Prox introduces a Multi-Resolution Spatio-Temporal Encoder (MRSTE) that extracts multi-scale features from complex-valued wireless signals, capturing both spatial and dynamic temporal characteristics. Additionally, the Proximity Metric Adaptation Network (PMAN) bridges the gap between high-dimensional signal characteristics and physical proximity. To enhance scalability, we leverage a transfer learning framework, significantly reducing the need for extensive data collection and retraining. Extensive experiments demonstrate RF-Prox's outstanding performance across Wi-Fi and cellular networks, achieving fine-tuned accuracy rates of 98.6% indoors and 91.3% outdoors. Even without fine-tuning, the pre-trained model achieves strong zero-shot performance, showcasing its exceptional performance in both proximity estimation accuracy and domain generalizability.**

***Index Terms*—Domain adaptation, proximity estimation, transfer learning, spatio-temporal encoder.**

I. Introduction

LOCATION awareness is a key enabler for a wide range of applications such as smart homes, augmented reality, and security monitoring [1]. With the increasing number of mobile devices, extensive research efforts have been devoted to wireless-based localization, which infers the devices' relative locations from ubiquitous radio signals.

Despite the advances in positioning technologies such as the Global Positioning System (GPS), which offers satisfactory outdoor positioning accuracy, its performance is significantly impaired under extreme weather conditions due to interference or obstruction of satellite signals with ground devices [2]. Moreover, rapid urbanization has led to an increased indoor time for individuals, escalating the demand for indoor positioning services across commercial, governmental, and communication sectors. This surge in demand places higher requirements on the coverage of positioning services [3], highlighting the limitations of GPS in indoor environments and emphasizing the need for integrated solutions suitable for both outdoor and indoor wireless sensing in scenarios where GPS is unavailable.



(a) Implicit control of IoT device (b) Proximity-based UAVs scheduling

Fig. 1. Illustration of two application scenarios of *RF-Prox*.

Current wireless localization methods are primarily designed for devices with direct communication links, such as wireless access points (AP) and user equipment (UE). However, these methods fall short in determining spatial relationships between non-directly connected devices (e.g., UE and IoT devices), a capability that is pivotal for a range of emerging applications including implicit control of IoT devices and proximity-based unmanned aerial vehicle (UAVs) scheduling [4], as illustrated in Fig. 1.

One straightforward approach involves estimating the location of each device independently, then deducing their relative proximity from these estimates. However, geometric-based approaches that rely on channel parameters like angle-of-arrival (AoA) [5], [6], time-of-flight (ToF) [7], [8], and their fusion [9], [10], are prone to significant errors in non-line-of-sight (NLoS) conditions. On the other hand, the fingerprint-based localization technique [11]–[13] is effective, where radio-assisted LiDAR SLAM [13] improves accuracy and speed significantly by integrating radio fingerprints with LiDAR for mapping. However, fingerprint-based methods require extensive labeled data collection and faces major generalization challenges across different domains. In summary, both traditional device localization methods exhibit significant limitations in terms of practicality and scalability, making them unsuitable for the task of proximity estimation.

Unlike traditional device localization methods, our approach focuses on proximity estimation from a high-dimensional feature perspective, rather than relying on the physical coordinates of the devices. Inspired by the principle of 'estimating by comparing', we observe that wireless devices nearby share similar signal propagation characteristics. By analyzing and comparing the spatio-temporal features encoded in the signals received from two devices, we can estimate their proximity.

Yuchong Gao, Guoxuan Chi, Zheng Yang and Shijie Cheng are with the School of Software and BNRist, Tsinghua University, Beijing 100084, China (e-mail: gaoyc01@gmail.com, chiguoxuan@gmail.com, hmilyyz@gmail.com, chengsj23@mails.tsinghua.edu.cn).

Zhiqing Wei is with Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: weizhiqing@bupt.edu.cn).

*Guoxuan Chi is the corresponding author.

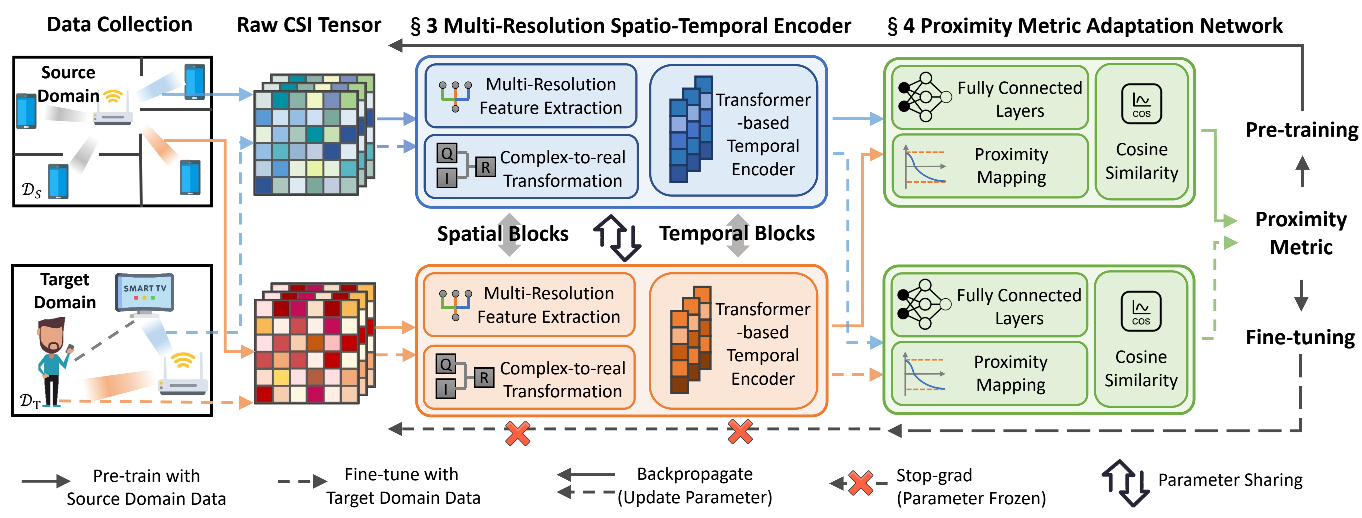


Fig. 2. An overview of the *RF-Prox*, where solid and dashed lines represent data collection from source domain and target domain, respectively, with blue and orange used to distinguish devices. The system workflow involves transforming CSI data into a proximity metric through the MRSTE and PMAN modules, training a pre-trained model in the source domain, and then applying a transfer learning framework to fine-tune *RF-Prox* in the target domain. Here, different domains represent varying scene setups, device placement configurations, and other environmental factors.

This approach eliminates the need for explicit estimation of devices' physical locations, thereby reducing errors associated with direct geometric parameter estimation in NLOS environments. Additionally, when RF-Prox is transferred to a new environment, devices with close physical proximity will still exhibit high proximity in high-dimensional feature space, making RF-Prox significantly superior to traditional localization-based methods in both practicality and scalability.

Nonetheless, actualizing this concept into a functional system presents formidable challenges. Firstly, the precise extraction of spatio-temporal features is complex, as conventional geometric parameters like Angle of Arrival (AoA) and Time of Flight (ToF) are plagued by significant inaccuracies in Non-Line-of-Sight (NLoS) conditions [10]. Secondly, devising a domain-adaptive proximity metric is essential, considering the variability in signal propagation due to differing scenario configurations and device placements, which directly affects proximity assessments.

To address these challenges, we introduce *RF-Prox*, the first proximity estimation system for wireless devices that are not directly connected, adaptable to a wide range of radio frequency signals. For accurate spatio-temporal feature extraction, we employ a data-driven strategy and create a complex-valued neural network module called Multi-Resolution Spatio-Temporal Encoder (MRSTE). This encoder excels at deriving multi-scale latent representations from the wireless signal and fusing them into a comprehensive feature vector that captures the spatio-temporal characteristics of the wireless channel. To establish a domain-adaptive proximity metric that enables *RF-Prox* to quickly adapt to new environments, we formulate a Proximity Metric Adaptation Network (PMAN), which compares the spatio-temporal features of two wireless channels to assess device proximity, incorporating domain adaptation techniques. In *RF-Prox*, we utilize a transfer learning approach [14], enabling model pre-training on source domain data followed by minimal fine-tuning with target domain data, significantly diminishing the need for extensive target domain data collection while maintaining broad generalization capabilities. The transfer strategy is based on the assumption that similar CSI distributions correspond to proximate distances within a scene. This assumption is consistent across all scenarios and is fundamental to the effectiveness of transfer learning in *RF-Prox*.

RF-Prox's efficacy is validated through extensive evaluations across over 9,000 domains, encompassing Wi-Fi-based indoor environments and cellular-based outdoor scenarios, with more than 1,000,000 data samples gathered. The evaluation results highlight that a fine-tuned *RF-Prox* achieves remarkable accuracy of 98.6% and 91.3% in identifying the most proximate device in indoor and outdoor scenarios, respectively. Impressively, even without fine-tuning, the pre-trained model demonstrates substantial zero-shot accuracy, reaching 92.2% and 88.9%, underscoring its exceptional performance in proximity estimation accuracy and domain adaptability. Additionally, *RF-Prox*'s superior spatial-awareness capability is evidenced through its application of a sorting metric, the Normalized Discounted Cumulative Gain (NDCG), in both scenarios.

We summarize our contributions as follows.

- We propose *RF-Prox*, the first proximity estimation system for non-directly connected wireless devices. *RF-Prox* shows the domain-adaptive capability and can be easily deployed in any target domain environment, making it a promising step towards integrated sensing and communication.
- Our proposed Multi-Resolution Spatio-Temporal Encoder is a pioneering attempt at applying complex-valued neural networks to wireless sensing. The multi-resolution design and transformer-based temporal processing model have unique advantages and can also be integrated into other types of wireless sensing applications.
- The transfer learning mechanism adopted by our system has been proven effective, providing a new approach to enhance the generalizability of data-driven wireless systems.
- We implement and evaluate *RF-Prox* on both Wi-Fi-based indoor environments and cellular-based outdoor scenarios, which showcases the practicality and effectiveness of

deploying *RF-Prox* in target domain scenarios. We made our codes, data and pre-trained models publicly available[1] to facilitate the research community.

Compared to our prior conference version [15], we have made significant enhancements including expanding the system to support full-scene proximity estimation across different wireless signal types, thereby improving universality, scaling, and facilitating Integrated Sensing and Communication (ISAC) in the 6G era. In our module design, we have seamlessly integrated transformer-based temporal analysis with CNN-based spatial feature extraction, offering improved adaptability to device mobility and bolstered support for analyzing temporal data in practical applications. Additionally, we've replaced the previous ordered MLP-based feature mapping module with an unordered cosine similarity module, and extend the transfer learning of simulation-to-reality to a more pervasive transfer task under source and target domains. To validate these enhancements, we have comprehensively revised all experiments, incorporated cellular-based outdoor scenarios, redesigned and retrained our models, ensuring they are fine-tuned for the new case. We've also conducted a full evaluation and optimization of these additions. To foster transparency and collaboration, we have made all related codes publicly available.

The rest of this paper is organized as follows. We begin with an overview of *RF-Prox* in Section II, followed by the detailed design of the MRSTE in Section III and PMAN in Section IV. Our implementation and evaluation of *RF-Prox* are shown in Section V, followed by the related work in Section VI, and the conclusion in Section VIII.

II. System Overview

In this section, we provide a high-level overview of how data flows through the designed neural networks and is ultimately transformed into the proximity metric. We also provide a detailed explanation of how the transfer learning framework enables the model to adapt from the source domain to the target domain.

Acted as a device proximity estimation system based on wireless signals, *RF-Prox* integrates two pivotal components: the *Multi-Resolution Spatio-Temporal Encoder (MRSTE)* and the *Proximity Metric Adaptation Network (PMAN)*. As depicted in Fig. 2, the workflow of *RF-Prox* initiates by capturing the Channel State Information (CSI) from wireless links between two distinct mobile devices. This CSI data is then processed through the MRSTE module to extract relevant features. The MRSTE module employs a complex-valued neural network, incorporating residual convolution blocks to derive multi-resolution latent representations from CSI's real and imaginary components. Following this, a complex-to-real transformation is performed, converting these complex representations into real-valued spatial features for subsequent analysis. The MRSTE concludes with a transformer-based temporal module, which compresses and extracts latent temporal information.

After processing through MRSTE, the domain-agnostic spatio-temporal features are amalgamated and fed into the PMAN module's fully connected layers for a comprehensive analysis and comparison. The proximity metric between two devices is then determined using cosine similarity, after transformation by an elaborately designed proximity mapping function.

RF-Prox adopts a transfer learning framework, initiating with a model pre-trained in a source domain, which can then be directly applied and fine-tuned within a target domain as necessary, where the source and target domains differ in environments and AP/BS deployments. Our transfer strategy relies on the assumption that similar CSI distributions correspond to proximate distances within a scene, which holds across all scenarios and is crucial for the effectiveness of transfer learning in *RF-Prox*. During the pre-training phase, a substantial dataset from the source domain $\mathcal{D}_{\mathrm{S}}$ is utilized to enhance the model's ability to generalize and extract domain-independent features. After pre-training, the model can be fine-tuned using only a minimal dataset from the target domain $\mathcal{D}_{\mathrm{T}}$, with MRSTE parameters kept frozen while only the parameters of PMAN are optimized. By leveraging the transfer learning mechanism, *RF-Prox* can be efficiently adapted to new environments, facilitating its practical deployment in varied scenarios.

III. Multi-Resolution Spatio-Temporal Encoder

In this section, we introduce the *Multi-Resolution Spatio-Temporal Encoder (MRSTE)*, designed to extract the domain-independent spatio-temporal information embedded in the CSI. As depicted in Fig. 3, MRSTE takes the complex-valued CSI tensor as input and transforms it into multi-resolution latent spaces via paralleled residual convolution blocks. Latent representations with different resolutions are then fused by channel concatenation. After passing through a fully connected layer, the fused complex-valued representation is converted to a real-valued spatial feature. The converted spatial features of a time-series are then sent to the Transformer block for temporal feature extraction, which could be further used for robust device proximity estimation.

Compared to geometric-based algorithms [6], [9], the MRSTE adopts a data-driven approach, analyzing signal statistical information in high-dimensional space. This approach significantly enhances the efficacy of the system in diverse settings, including both indoor and outdoor environments, particularly in scenarios afflicted by Non-Line-of-Sight (NLoS) conditions.

A. CSI Preliminary

Considering the phenomenon of multipath propagation, the wireless channel can be modeled as a function of frequency f and time t, expressed as

$$H(f,t) = \sum_{l=1}^{L} \alpha_l(t,f) e^{-j2\pi f \tau_l(t)}, \tag{1}$$

where L denotes the number of multipath components, $\alpha_l(t,f)$ embodies the complex attenuation factor, and $\tau_l(t)$ the propagation delay corresponding to the l-th path, respectively. CSI represents a discretized sampling of the channel response [16], with frequency domain samples positioned on specific OFDM subcarriers, time domain samples corresponding to each received packet, and spatial domain samples for each radio chain (*i.e.*, Tx-Rx pair), rendering CSI a complex-valued tensor

[1]Our project is available here.

$\mathbf{H} \in \mathbb{C}^{T\times S\times A}$, with T, S, and A indicating the number of time samples, subcarriers, and radio chains, respectively.

B. Complex-valued Network for CSI Processing

Previous studies have often utilized processed CSI data, such as the short-time Fourier transform and ToF-AoA spectrogram [17], [18], as inputs to classification network models for learning, or have divided the CSI into its real and imaginary components for independent processing within deep neural networks [19]. Conversely, our approach leverages the raw, unmodified CSI to mine richer spatio-temporal information. Therefore, we embrace the concept of the complex-valued neural network, incorporating novel elements such as complex-valued linear and convolutional layers into the MRSTE.

To start with, a linear transformation for a CSI matrix $\mathbf{H} = \mathbf{H}_r + j\mathbf{H}_i$ with complex-valued weight $\mathbf{W} = \mathbf{W}_r + j\mathbf{W}_i$ can be decomposed into several real-valued transformations:

$$\mathrm{Linear}(\mathbf{H};\mathbf{W}) = \begin{bmatrix}\Re(\mathbf{WH})\\ \Im(\mathbf{WH})\end{bmatrix} = \begin{bmatrix}\mathbf{W}_r & -\mathbf{W}_i\\ \mathbf{W}_r & \mathbf{W}_i\end{bmatrix}\begin{bmatrix}\mathbf{H}_r\\ \mathbf{H}_i\end{bmatrix}. \tag{2}$$

Similarly, given a complex kernel $\mathbf{C} = \mathbf{C}_r + j\mathbf{C}_i$, the convolution operation $\mathbf{C} * \mathbf{H}$ on the complex domain can also be equivalently written into the following form:

$$\mathrm{Conv}(\mathbf{H};\mathbf{C}) = \begin{bmatrix}\Re(\mathbf{C}*\mathbf{H})\\ \Im(\mathbf{C}*\mathbf{H})\end{bmatrix} = \begin{bmatrix}\mathbf{C}_r & -\mathbf{C}_i\\ \mathbf{C}_r & \mathbf{C}_i\end{bmatrix} * \begin{bmatrix}\mathbf{H}_r\\ \mathbf{H}_i\end{bmatrix}. \tag{3}$$

Research has affirmed [20] the feasibility of implementing dropout, batch normalization, and activation mechanisms directly within the complex domain by independently manipulating the real and imaginary components of the input. This approach ensures that each complex module within the MRSTE is a synthesis of operations conducted in the real domain, thereby preserving the differentiability across the entirety of the MRSTE module. Additionally, complex-valued neural networks differ from real-valued neural networks in terms of computational overhead. In complex-valued neural networks, the parameter count is twice that of a real-valued neural network since each complex parameter consists of a real part and an imaginary part. As shown in Eq. (2) and Eq. (3), complex multiplication is equivalent to four real multiplications and two real additions, resulting in a computational load four times greater than that of a real-valued neural network.

C. Multi-Resolution Feature Extraction

The core design of MRSTE is the principle of fusing CSI features across multiple resolutions, a strategy that has demonstrated its efficacy within the domain of computer vision [21]. The underlying logic behind this methodology is that variations in antenna spacing, when measuring the Angle-of-Arrival (AoA) from CSI, can introduce a balance between resolution and the operational range [22].

Fig. 3 showcases the MRSTE's structure, which includes four residual convolution blocks, each characterized by unique output channels, kernel dimensions, and stride lengths, thereby forging four concurrent processing pathways. Let us represent a residual block as $\mathrm{ResBlock}(\cdot)$ and define $\mathbf{C}_i$ as the parameter set corresponding to the i-th residual block. For an input

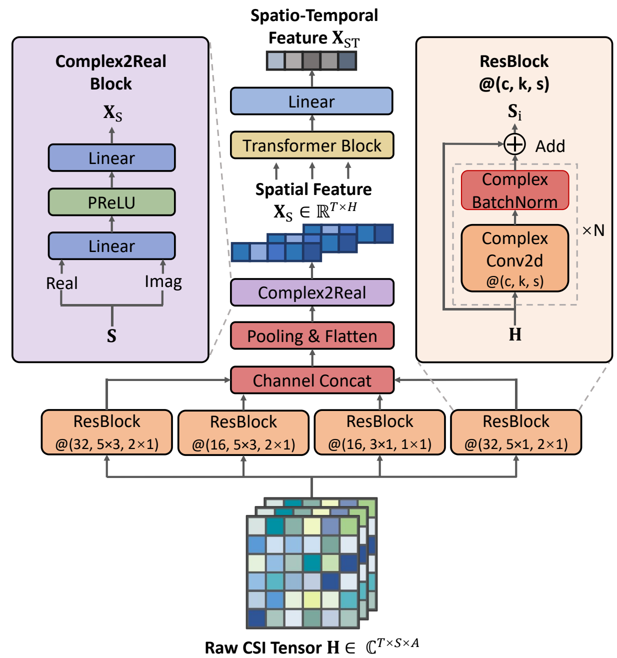


Fig. 3. Illustration of Multi-Resolution Spatio-Temporal Encoder.

CSI tensor $\mathbf{H}$, the output feature from the i-th block can be articulated as

$$\mathbf{S}_i = \mathrm{ResBlock}(\mathbf{H};\mathbf{C}_i), \quad i = 0,1,2,3, \tag{4}$$

where the residual block is basically a convolution with shortcut connection [23], which makes the model easier to train by solving the gradient disappearance problem during the training for better expressive ability.

$$\mathrm{ResBlock}(\mathbf{H};\mathbf{C}_i) = \mathrm{BatchNorm}(\mathrm{Conv}(\mathbf{H};\mathbf{C}_i)) + \mathbf{H}. \tag{5}$$

Features extracted from parallel residual blocks are then concatenated along the channel dimension and fuse into a latent representation $\mathbf{S} = \mathrm{Concat}(\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3)$. The concatenated $\mathbf{S}$ contains multi-level features of the CSI input, which greatly improves the receptive field of MRSTE and thus enhances the generalization performance of *RF-Prox*.

D. Complex-to-Real Transformation

After processing CSI with paralleled residual blocks, multi-level features can be extracted. In order to transform the complex-valued latent representation $\mathbf{S}$ to the real-valued spatial feature $\mathbf{X}_\mathrm{S} \in \mathbb{R}^{T\times H}$, where H is hidden dimension, we design a complex-to-real transformation module C2R, which applies two linear operations on the real and imaginary part:

$$\begin{aligned}\mathbf{X}_\mathrm{S} &= \mathrm{C2R}(\mathbf{S};\mathbf{W}^\mathrm{R},\mathbf{W}^\mathrm{I})\\ &= \mathrm{PReLU}(\mathrm{Linear}(\Re(\mathbf{S}),\mathbf{W}^\mathrm{R}) + \mathrm{Linear}(\Im(\mathbf{S}),\mathbf{W}^\mathrm{I})),\end{aligned} \tag{6}$$

where $\mathbf{W}^\mathrm{R}$ and $\mathbf{W}^\mathrm{I}$ are the real-valued linear weights. Note that for better expressive ability of the model, the PReLU activation function is leveraged to add non-linear factors to the features:

$$\mathbf{Y} = \mathrm{PReLU}(\mathbf{X}) = \begin{cases}\mathbf{X}, & \text{if } X \geq 0,\\ \theta\mathbf{X}, & \text{if } X < 0,\end{cases} \tag{7}$$

where θ is a learnable parameter.

E. Transformer-based Temporal Encoder

Through the complex-to-real transformation module C2R, a time-series of real-valued spatial features are extracted. In order to extract the temporal features carried by the device as it moves, we use the transformer encoder module with the attention mechanism [24]. We first embed the series of $\mathbf{X}_{\mathrm{S}}$ into a high-dimensional representation $\mathbf{X}_{\mathrm{E}} = \mathrm{FC}(\mathbf{X}_{\mathrm{S}})$ with fully connected layers $\mathrm{FC}(\cdot)$. To introduce temporal information, we use absolute positional encoding for $\mathbf{X}_{\mathrm{E}}$ to get embedded positional information $\mathbf{P}$, with each element as follows:

$$\begin{aligned}\mathbf{P}(pos, 2i) &= \sin\left(pos/10000^{2i/d}\right),\\ \mathbf{P}(pos, 2i+1) &= \cos\left(pos/10000^{2i/d}\right),\end{aligned} \tag{8}$$

where pos represents the sequence element's ordinal position, and i refers to the dimension index within the embedding space. Then, we formulate the transformer's encoded input as $\mathbf{X}_{\mathrm{PE}} = \mathbf{X}_{\mathrm{E}} + \mathbf{P}$.

In order to better learn the relationships between the elements inside the sequence, we use the attention mechanism to linearly transform the input $\mathbf{X}_{\mathrm{PE}}$ into queries, keys and values:

$$\begin{aligned}\mathbf{Q} &= \mathbf{X}_{\mathrm{PE}}\mathbf{W}_{\mathrm{Q}},\\ \mathbf{K} &= \mathbf{X}_{\mathrm{PE}}\mathbf{W}_{\mathrm{K}},\\ \mathbf{V} &= \mathbf{X}_{\mathrm{PE}}\mathbf{W}_{\mathrm{V}},\end{aligned} \tag{9}$$

where $\mathbf{W}_{\mathrm{Q}}$, $\mathbf{W}_{\mathrm{K}}$ and $\mathbf{W}_{\mathrm{V}}$ are the weight matrices used for linear transformation. Attention features $\mathbf{X}_{\mathrm{A}}$ are obtained as

$$\begin{aligned}\mathbf{X}_{\mathrm{A}} &= \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})\\ &= \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V},\end{aligned} \tag{10}$$

where d_k is the dimension of keys.

After residual connection and layer normalization, the spatio-temporal features $\mathbf{X}_{\mathrm{ST}}$ in the CSI can be finally extracted as

$$\mathbf{X}_{\mathrm{ST}} = \mathrm{LayerNorm}(\mathbf{X}_{\mathrm{A}} + \mathbf{X}_{\mathrm{PE}}), \tag{11}$$

which encapsulates the spatio-temporal relationship between the device and the access point. Thus, the spatio-temporal features corresponding to two different terminal devices can be further utilized to determine their proximity relationship, which will be detailed in Section IV.

IV. Proximity Metric Adaptation Network

A. Metric Network Design

In this section, we introduce the *Proximity Metric Adaptation Network (PMAN)*. As illustrated in Fig. 4, PMAN transforms the spatio-temporal features extracted from two wireless devices $\mathbf{X}_{\mathrm{ST}}$ to their proximity metric with domain adaptation capability. Serving as a proximity comparing network, PMAN determines device proximity at the feature level, leveraging the powerful nonlinear fitting capabilities of neural networks. Compared to traditional methods that rely on determining physical coordinates, PMAN is more effective in mitigating the adverse effects of non-line-of-sight (NLOS) conditions. PMAN is implemented based on the fully connected layers, cosine similarity and an elaborately designed proximity loss function $L_{\mathrm{P}}(\cdot)$. When being deployed in a new environment with different landscapes, building structures and device deployment, the PMAN learns the appropriate transformation and adapts to the new environment with only a small amount of fine-tuning data.

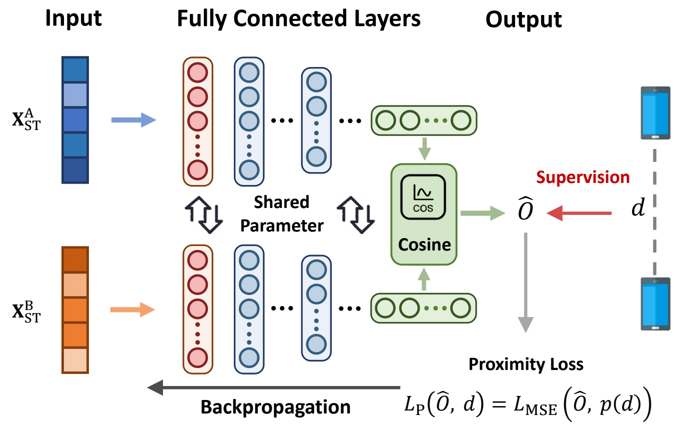


Fig. 4. An overview of the Proximity Metric Adaptation Network.

The PMAN takes the spatio-temporal features from two wireless devices as inputs, which are denoted as $\mathbf{X}_{\mathrm{ST}}^{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{ST}}^{\mathrm{B}}$ respectively. Two features first pass through the fully connected layers $\mathrm{FC}(\cdot)$ to get the latent representation for cosine similarity calculation $\hat{\mathbf{O}}$:

$$\hat{\mathbf{O}} = \frac{\mathrm{FC}(\mathbf{X}_{\mathrm{ST}}^{\mathrm{A}}) \cdot \mathrm{FC}(\mathbf{X}_{\mathrm{ST}}^{\mathrm{B}})}{\max\left(\left\|\mathrm{FC}(\mathbf{X}_{\mathrm{ST}}^{\mathrm{A}})\right\|_2 \cdot \left\|\mathrm{FC}(\mathbf{X}_{\mathrm{ST}}^{\mathrm{B}})\right\|_2, \epsilon\right)}, \tag{12}$$

where ϵ is a small value to avoid division by zero.

In environments where indoor or complex outdoor conditions prevail, the use of Euclidean distance for quantifying device proximity may not accurately reflect the perceived nearness due to potential obstructions such as walls and buildings. Consequently, our system adopts the Shortest Non-Blocking Distance (SNBD) as the ground truth label, which represents the minimum path length between wireless devices that does not intersect any physical barriers. Therefore, during the training phase, the model inputs CSI as the raw feature and uses SNBD as the ground truth label. The definition of SNBD assists in aligning the proximity metric more closely with user perception. To optimize model learning during the backpropagation phase, we introduce a novel proximity loss function, $L_{\mathrm{P}}(\cdot)$. This function is computed using a batch of predicted proximity outputs, $\hat{\mathbf{O}}$, and the corresponding SNBD values, $\mathbf{d}$.

$$L_{\mathrm{P}}(\hat{\mathbf{O}}, \mathbf{d}) = L_{\mathrm{MSE}}(\hat{\mathbf{O}}, p(\mathbf{d})) = \frac{1}{N}\|\hat{\mathbf{O}} - p(\mathbf{d})\|_2^2, \tag{13}$$

where $L_{\mathrm{MSE}}(\cdot)$ indicates the mean square error, N is the batch size and $p(\mathbf{d}) = \tanh(-\log(\alpha\mathbf{d}))$, where α is the elastic parameter to control the distribution and the steepness of the function $p(\cdot)$.

The proximity estimation mechanism within our system is designed to be more sensitive to devices in closer proximity than those further away. Our evaluations indicate that, within indoor settings, relevant proximity distances typically span from 0 to 4 meters. Outdoor scenarios, however, present a more variable range of interest. The model's focus is modulated by an elasticity parameter, α, which adjusts according to the environmental context. For instance, within indoor settings,

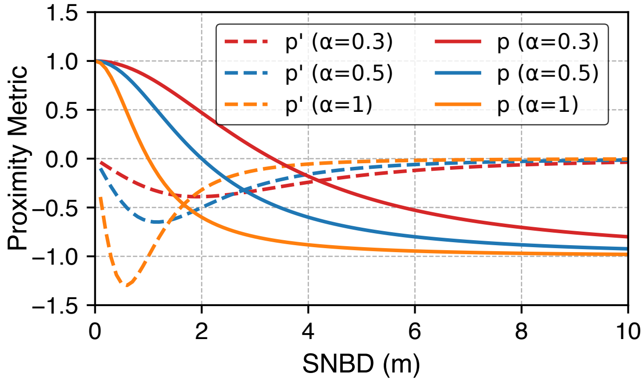


Fig. 5. Illustration of the proximity mapping function varied by α.

Fig. 5 demonstrates how different α values influence the mapping function, $p(\cdot)$, and its derivative, $p'(\cdot)$. An α value of 0.5 is chosen to ensure a steep gradient in the function between 0 and 4 meters, with a plateau beyond this range, thereby optimizing the model for indoor applications.

To sum up, the mapping function $p(\cdot)$ guides the model to pay more attention to the distance range of interest, helping PMAN converge quickly and perform better.

B. Transfer Learning

Our goal is to develop a full-scenario proximity detection system for various wireless signals. However, in some complex scenarios, data acquisition becomes very difficult and consumes a lot of manpower and resources. In order to solve this problem, we propose a pre-training and fine-tuning strategy to fill the performance gap from the source domain to the target domain, based on the assumption that similar CSI distributions correspond to proximate distances within a scene, which holds across all scenarios and is crucial for the effectiveness of transfer learning in *RF-Prox*. This strategy has been proven to greatly enhance the model's domain adaptation capability.

Pre-training in source domain. We build both indoor environments and outdoor scenarios on MATLAB for integrated sensing and communication based on the ray tracing model and collect labeled CSI data under thousands of different deployment cases. pre-training with a large amount of collected data from the source domain helps the MRSTE module learn domain generalized spatio-temporal features without overfitting specific structures. During the pre-training process, all the parameters in *RF-Prox* are jointly optimized. Suppose our pre-training model is $\mathrm{M_P}$. Denote the source domain dataset as $\mathcal{D}_\mathrm{S}$, and the set of parameter in MRSTE and PMAN as Θ_M and Θ_P respectively, the optimization (a.k.a backpropagation) process can be written as

$$\{\Theta_\mathrm{M}, \Theta_\mathrm{P}\} = \underset{\{\Theta_\mathrm{M},\Theta_\mathrm{P}\}}{\arg\min} \sum_{(\mathbf{H},\mathbf{d})\sim\mathcal{D}_\mathrm{S}} \frac{1}{|\mathcal{D}_\mathrm{S}|} L_\mathrm{P}(\mathrm{M_P}(\mathbf{H};\Theta_\mathrm{M},\Theta_\mathrm{P}),\mathbf{d}). \tag{14}$$

Fine-tuning for real-world application. The pre-trained model based on the source domain data has a certain generalization capability for domain transfer. To achieve better transfer performance from source domain to target domain, a small amount of target domain data is collected to fine-tune the pre-trained model. During the fine-tuning process, the parameters of the MRSTE are frozen and only the parameters of the PMAN are optimized. The optimization process can be written as

$$\Theta_\mathrm{P} = \underset{\Theta_\mathrm{P}}{\arg\min} \sum_{(\mathbf{H},\mathbf{d})\sim\mathcal{D}_\mathrm{T}} \frac{1}{|\mathcal{D}_\mathrm{T}|} L_\mathrm{P}(\mathrm{M_F}(\mathbf{H};\Theta_\mathrm{M},\Theta_\mathrm{P}),\mathbf{d}), \tag{15}$$

where $\mathrm{M_F}$ refers to the fine-tuning model, and $\mathcal{D}_\mathrm{T}$ indicates the target domain dataset.

Upon completion of the pre-training and fine-tuning processes, we get a pre-trained model $\mathrm{M_P}$ with high generalizability, and a fine-tuned model $\mathrm{M_P}$ which adapts to a specific target domain environment.

V. Evaluation

A. Experimental Methodology

1) Experimental Scenarios: To rigorously assess the performance of *RF-Prox*, we conducted comprehensive evaluations through two distinct case studies focused on proximity detection: one within indoor settings involving UEs and IoT devices utilizing Wi-Fi signals, and the other in outdoor UAV scenarios leveraging cellular signals.

For these studies, we constructed source domains employing the MATLAB Communication Toolbox and the Deep MIMO toolkit [25], where different domains represent varying scene setups, device placement configurations, and other environmental factors. Specifically, this construction involved setting more than 30 distinct environments, each featuring 300 varied access point/base station (AP/BS) deployment scenarios. Within each scenario, over 20 UEs/UAVs were maneuvered across various locations and orientations to gather a comprehensive dataset of temporal Channel State Information (CSI) for pre-training. For the evaluation within the target domain, we generated novel scenarios with differing AP/BS configurations, where 3-10 UEs/UAVs at varying locations were permitted to move, thereby enabling the collection of corresponding CSI data.

2) System Implementation: Within Wi-Fi-enabled indoor settings, *RF-Prox* is configured with one AP and multiple client devices operating at 5.6 GHz, where both the transmitter and receiver are outfitted with three antennas each, spaced at $\lambda/2$, thereby forming a 3×3 antenna array. Conversely, for cellular-based outdoor scenarios, *RF-Prox* encompasses one BS and multiple UAVs operating at 200 GHz, with the BS equipped with a 4×4 antenna array (spaced at $\lambda/2$) and UAVs equipped with a single antenna, culminating in a 1×16 array configuration. The target domain's evaluation was conducted under scenarios with a Signal-to-Noise Ratio (SNR) of 30 dB.

RF-Prox employs a hybrid programming approach, utilizing both MATLAB and Python to facilitate rapid and efficient processing. Specifically, MATLAB is utilized for the collection and preprocessing of CSI data, while a Python-based deep learning model is leveraged for real-time proximity estimation.

3) Comparative Methods: To thoroughly benchmark *RF-Prox*'s efficacy, we juxtaposed it against two state-of-the-art Wi-Fi-based localization methods: SpotFi [6] and mD-Track [9]. This comparison was executed by substituting our MRSTE module with each alternative, thereby highlighting the superiority of our module design.

In scenarios featuring $N+1$ UEs/UAVs, with one randomly selected as the reference, the **Top-1 Accuracy** metric reflects

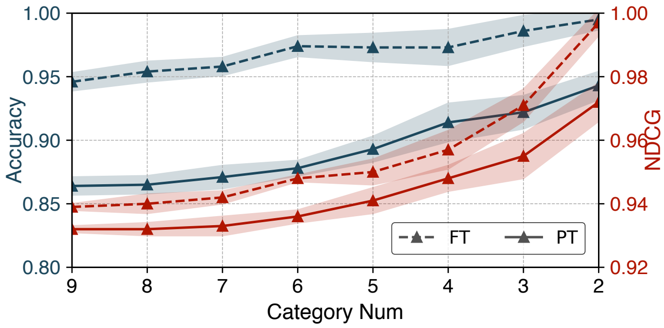


Fig. 6. Top-1 Accuracy and NDCG for Wi-Fi scenario.

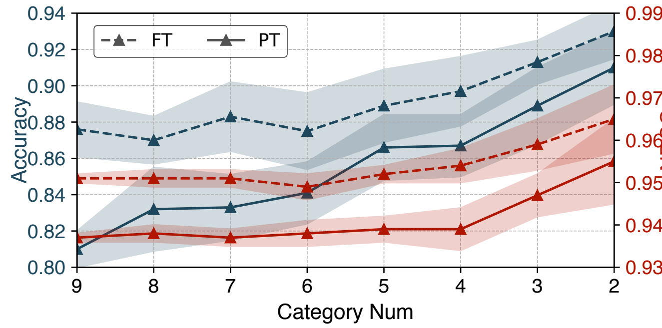


Fig. 7. Top-1 Accuracy and NDCG for cellular scenario.

the success rate of identifying the most proximate device amongst the remaining N. Additionally, the **NDCG** [26], a prevalent ranking metric within sorting problems ranging from 0 to 1, evaluates the accuracy of the proximity estimation results in a ranked order.

B. Overall Performance

In this section, we investigate the comprehensive efficacy of *RF-Prox* across both Wi-Fi-enabled indoor environments and cellular-based outdoor scenarios, denoted as W and C, respectively. The performance of the system is examined using both a pre-trained model (PT) and a fine-tuned model (FT).

1) Top-1 accuracy: As depicted in Fig. 6 and Fig. 7, the mean accuracy and its variability are illustrated by dots and shaded regions, respectively. The performance of *RF-Prox* in terms of Top-1 accuracy is represented by blue lines, with dashed lines highlighting enhancements post fine-tuning. Initially, accuracy for the pre-trained models (PT-W, PT-C) ranges between 86.4%/81.0% and 94.3%/91.0% for categories decreasing from nine to two, underscoring the system's robust generalizability. After applying domain transfer learning, the fine-tuned models (FT-W, FT-C) demonstrate improved accuracies ranging from 94.6%/87.6% to 99.5%/93.0%, indicating superior adaptation across varied categorical scenarios and effectively bridging the source-target discrepancy through transfer learning.

2) NDCG: The NDCG performance of *RF-Prox* is portrayed by red lines in Fig. 6 and Fig. 7, with dashed lines denoting enhancements following fine-tuning. To assess the system's distance-awareness capability, a reference point is chosen at random, and additional UEs are positioned linearly at uniform intervals. The pre-trained models (PT-W, PT-C) achieve NDCG scores ranging from 0.932/0.937 to 0.972/0.955 for categories decreasing from nine to two, highlighting the system's generalizability. After domain transfer learning, the fine-tuned models (FT-W, FT-C) attain NDCG scores between 0.939/0.951 and 0.997/0.965, showcasing the system's potent distance-awareness and the successful mitigation of the source-target gap via transfer learning.

3) System latency & model parameters: Utilizing the PyTorch Profiler, we assessed the computational demands, including the floating point operations (FLOPs), model parameters, and inference timing for each component, as documented in Table I. Remarkably, the total number of model parameters is lower than that of typically employed small-scale models (e.g., ResNet-18 [23]), suggesting significant potential for direct deployment on various edge-embedded devices for real-time inference. Additionally, the PMAN's notably smaller parameter count compared to the MRSTE emphasizes efficient fine-tuning and domain adaptation with minimal data volume.

4) Scalability analysis: Scalability refers to a model's ability to improve performance with increasing size, which is crucial for setting appropriate parameter quantities based on performance requirements. Benefiting from the complex-domain neural network integrating MRSTE and PMAN, RF-Prox effectively processes both the amplitude and phase information of complex-valued CSI signals, allowing it to better capture hidden spatio-temporal features. This design enables the model to leverage increased parameters when handling large datasets.

To verify the scalability of *RF-Prox*, we trained 9 models of different sizes, exploring different numbers of residual convolution blocks (2C, 4C, 6C) and transformer blocks (2T, 4T, 6T). As shown in Fig. 8 and Fig. 9, the average inference time across all model sizes is kept within 20 ms, enabling at least 50 operations per second, thereby meeting real-time requirements. When deploying smaller models, *RF-Prox* significantly reduces inference time while maintaining satisfactory performance. For instance, in the cellular-based scenario, the '2C2T' model, with approximately 100K parameters, achieves nearly 86% proximity estimation accuracy while reducing inference time to 6 ms, making it suitable for embedded systems. Conversely, scaling up model parameters and increasing model computation can enhance accuracy but inevitably increases inference time, which may be preferable in scenarios with lower real-time requirements.

5) Comparison with localization methods: To validate the robustness and expressiveness of the high-order spatio-temporal features extracted by MRSTE, we compared its performance to that achieved using multipath AoA and ToF data processed by SpotFi and mD-Track. As depicted in Fig. 10, the pre-trained *RF-Prox* model surpasses both mD-Track and SpotFi in accuracy and NDCG within Wi-Fi-based indoor and cellular-based outdoor scenarios, with respective gains of 9.0%/0.033 and 15.0%/0.057 for indoor, and 8.3%/0.036 and 14.1%/0.062 for outdoor scenarios. These improvements underscore the advanced capabilities of MRSTE in leveraging high-order spatio-temporal features for superior performance, particularly when deployed in NLOS environments or transferred to new settings. This highlights how RF-Prox not only achieves high proximity estimation accuracy but also outperforms traditional localization-based methods in terms of practicality and scalability.

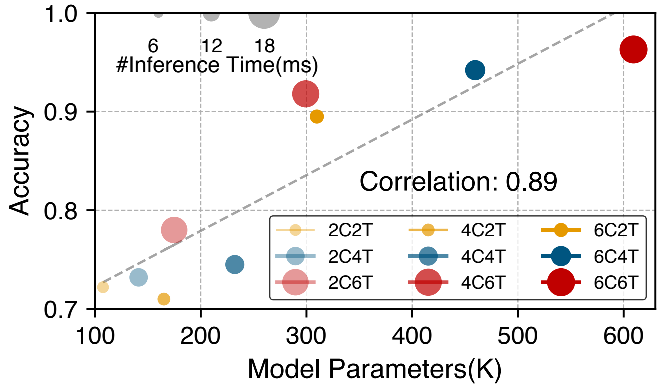


Fig. 8. Scalability analysis for Wi-Fi scenario.

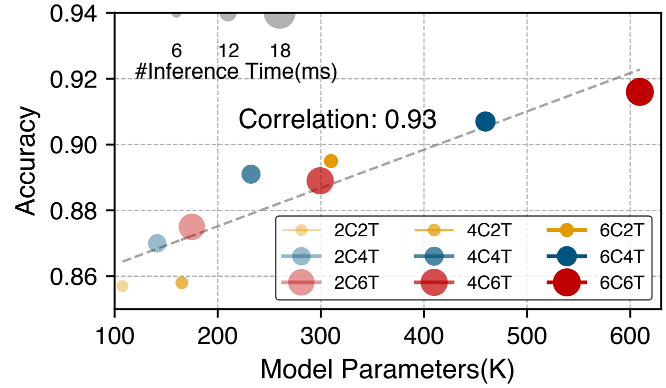


Fig. 9. Scalability analysis for cellular scenario.

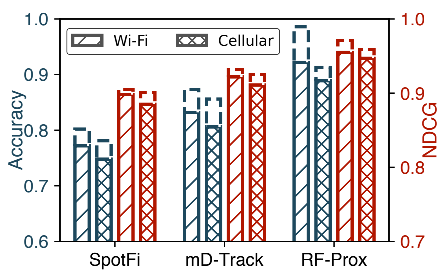


Fig. 10. Comparison with localization methods.

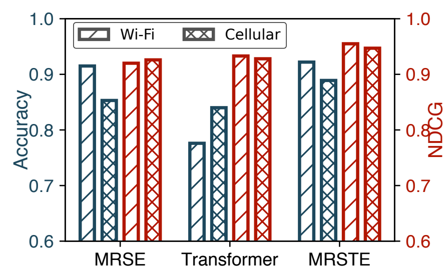


Fig. 11. Effectiveness of each component.

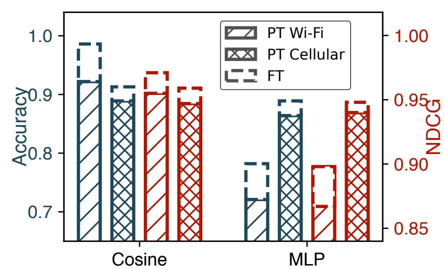


Fig. 12. Comparison of fusion methods.

We analyze that MRSTE can extract generalizable features across different environments due to the consistent relationship between similar CSI distributions and proximate distances within a scene. This principle allows the MRSTE module to maintain consistent feature extraction across diverse environments. It's important to note that the reported performances are averaged across various common category numbers, maintaining consistency in reporting metrics across the following sections.

TABLE I
SYSTEM LATENCY & NUMBER OF MODEL PARAMETERS.

Parameters	Multi-resolution Spatio-Temporal Encoder	Proximity Metric Adaptation Network	Overall
FLOPs (M)	26.12	0.2	26.32
Model Parameters (k)	280.73	18.58	299.31
Inference Time (ms)	13.67	0.13	13.8

C. Component Study

In this section, we undertake a component study to evaluate the significance of each module within *RF-Prox*.

1) MRSTE: The core of *RF-Prox* comprises two main components: a CNN-based multi-resolution spatial feature extraction module and a transformer-based temporal feature processing module. The spatial features are derived from varying numbers of antennas and subcarriers, while temporal features are extracted from device motion. As illustrated in Fig. 11, *RF-Prox* demonstrates superior performance over both MRSE and Transformer across all metrics in both indoor and outdoor scenarios. This underscores RF-Prox's efficacy in integrating the strengths of both components to enhance spatio-temporal feature extraction, thereby boosting overall performance.

2) PMAN: Within the PMAN module, we explore two fusion methods applied to pairs of Channel State Information (CSI): cosine similarity and CSI concatenation followed by Multi-Layer Perceptron (MLP) processing. As depicted in Fig. 12, the cosine similarity approach consistently outperforms the CSI concatenation approach in all evaluated metrics within both environments. This discrepancy is attributed to the fact that concatenating CSIs introduces superfluous sequential data, whereas CSI pairs are inherently unordered and independent. Consequently, the unordered nature of cosine similarity yields superior performance.

D. System Robustness

In this section, we evaluate the system's robustness across various antenna types, SNR levels, and dynamic environments to assess their impact on overall efficacy.

1) Antenna type: To demonstrate the superior domain adaptability of the PMAN module, especially when deploying the *RF-Prox* system in scenarios with significant device heterogeneity, we test the impact of different types of antennas with distinct radiation patterns on *RF-Prox*'s performance, where the isotropic, dipole, patch, and monopole antennas are selected for this evaluation. As shown in Fig. 13, different antenna configurations result in only minor performance degradation in zero-shot scenarios for both Wi-Fi-based indoor and cellular-based outdoor environments, where the performance can be effectively restored to similar levels through fine-tuning.

We analyze that different antenna types, due to their unique radiation patterns, affect the input system's CSI in multipath environments. However, the joint training of the PMAN and the MRSTE enables the MRSTE to implicitly learn the computation of spatio-temporal features in the CSI that are indicative of proximity. These extracted features are independent of antenna radiation characteristics, allowing *RF-Prox* to easily

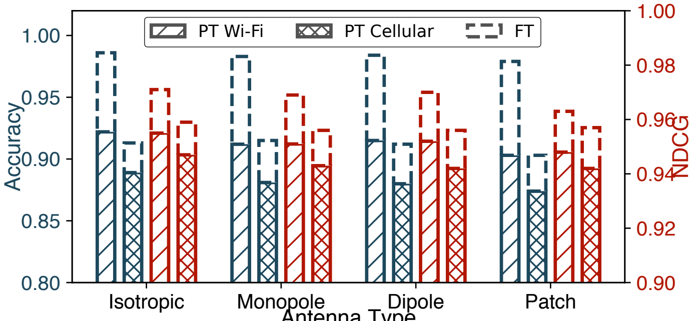


Fig. 13. Performance with different antenna types.

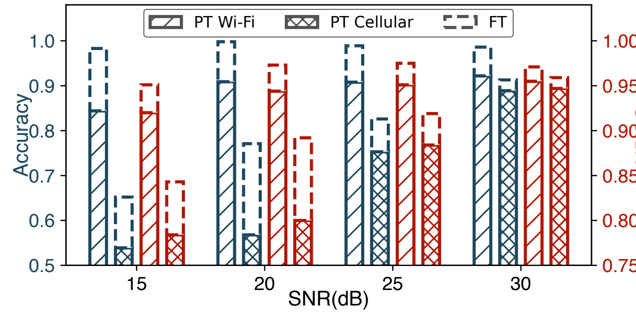
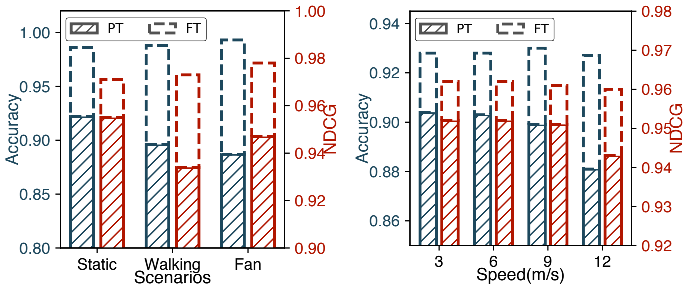
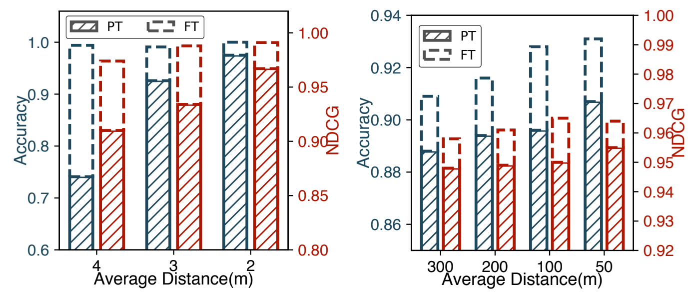


Fig. 14. Performance under different SNR.



(a) Wi-Fi (b) Cellular

Fig. 15. Performance in dynamic environments.



(a) Wi-Fi (b) Cellular

Fig. 16. Impact of distance between AP/BS and UEs/UAVs.

adapt to different antenna types, thereby addressing device heterogeneity issues (antenna radiation characteristic heterogeneity). Furthermore, the PMAN's fully connected layers can be fine-tuned to easily learn domain-specific parameters related to antenna radiation characteristic heterogeneity, thereby enhancing the model's adaptive capabilities.

2) SNR: To demonstrate the robust interference resistance of *RF-Prox*, we evaluate its performance under varying SNR conditions, which can reflect signal quality. The SNR settings are established at $15, 20, 25, 30$ dB, reflective of typical conditions for communication environments. As shown in Fig. 14, the system maintains robust performance across various SNR levels in Wi-Fi-based indoor environments. For instance, under challenging communication conditions with an SNR of 15 dB, the system achieves an accuracy of 84.4%/98.3% and an NDCG of 0.920/0.951 for the pre-trained and fine-tuned models, respectively. Conversely, the complex and dynamic nature of cellular-based outdoor environments exacerbates the degradation of performance at lower SNRs, especially when unmanned aerial vehicles (UAVs) are involved. revNotably, fine-tuning enhances the model resilience in lower SNR environments, evidencing the model's adaptability through domain transfer even under challenging conditions.

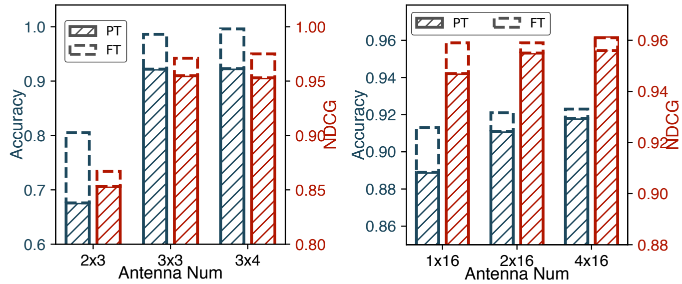
3) Dynamic environments: To further demonstrate the robustness of *RF-Prox*, we evaluate its performance under highly dynamic environments. As shown in Fig. 15, we introduce additional pedestrian movement and rotating fans as interference in Wi-Fi-based indoor environments. With pedestrian movement, the system achieves an accuracy of 89.6%/98.8% and an NDCG of 0.934/0.973 for the pre-trained and fine-tuned models. For rotating fans, the system achieves an accuracy of 88.7%/99.3% and an NDCG of 0.947/0.978 for the respective models. The results indicate a slight decrease in zero-shot performance, which is easily mitigated through fine-tuning.

For cellular-based outdoor environments, we test the impact of UAVs flying at different speeds. With a high-speed UAV moving at 12 m/s, the zero-shot accuracy and NDCG remain at 88.1% and 0.943, respectively, and can be fine-tuned to an outstanding 92.7% and 0.960. The results show a minor degradation in zero-shot performance with increasing UAV speed, which is also easily compensated by fine-tuning. These experiments demonstrate that the *RF-Prox* maintains strong robustness in highly dynamic environments, attributed to the MRSTE's feature extraction module, which incorporates a transformer-based temporal block to capture dynamic temporal features and ensure stable performance in highly dynamic settings.

E. Micro-benchmarks

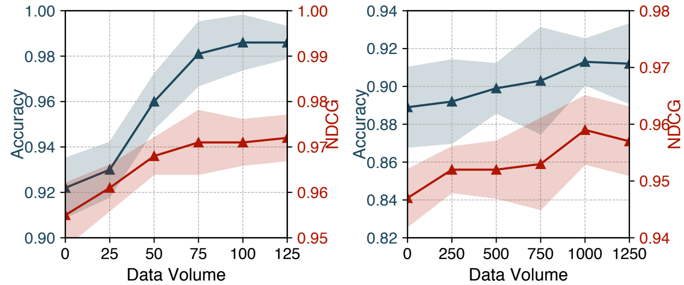
In this section, we conduct a robustness analysis focusing on the average distance between AP/BS and UEs/UAVs, antenna number, and the volume of data used for fine-tuning, to assess their impact on system efficacy.

1) Distance between AP/BS and UEs/UAVs: The indoor Wi-Fi coverage is approximately 20 meters [27], while the outdoor LTE coverage extends to about 500 meters [28]. Focusing on specific application scenarios, we set a distance of 4 meters for Wi-Fi and 300 meters for cellular base stations in our experiments. As shown in Fig. 16, as the distance between AP/BS and UEs/UAVs increases in both Wi-Fi-based indoor environments and cellular-based outdoor scenarios, zero-shot performance exhibits a decreasing trend. This is likely due to the introduction of more environmental interference and NLOS conditions before parsing the CSI. However, the fine-tuning strategy still improves the model performance to a satisfactory level, demonstrating *RF-Prox*'s excellent domain generalization capabilities that enable the model to adapt and perform effectively across various environments without the need for significant retraining.



(a) Wi-Fi (b) Cellular

Fig. 17. Impact of antenna number.



(a) Wi-Fi (b) Cellular

Fig. 18. Impact of fine-tune data volume.

2) Antenna Number: Antenna configurations are varied as 2×3, 3×3, and 3×4 arrays for indoor settings, and 1×16, 2×16, or 3×16 for outdoor scenarios. As depicted in Fig. 17, both accuracy and NDCG metrics exhibit an upward trend with the increase in the number of antennas in both scenarios. In Wi-Fi-based environments, a 3×3 array provides satisfactory results, achieving accuracy and NDCG of 92.2% / 0.955 with the pre-trained model, and 98.6% / 0.971 post fine-tuning. Similarly, for cellular-based environments, a 1×16 array also shows commendable performance, with accuracy and NDCG of 88.9% / 0.947 (pre-trained) and 91.3% / 0.959 (fine-tuned) respectively. Although additional antennas could potentially improve performance by providing more channel information for enhanced multipath resolution discrimination, the marginal gains diminish beyond a certain point. From a practical perspective, it is prudent to balance the benefits against the costs of using an excessive number of antennas.

3) Fine-tuning data volume: During the transfer learning process, the fine-tuning data volume affects the model's performance in the target domain. For fine-tuning, data volumes are set at $0, 25, 50, 75, 100, 125$ for indoor environments and $0, 250, 500, 750, 1000, 1250$ for outdoor scenarios, with zero data equivalent to employing the pre-trained model. As illustrated in Fig. 18, the system exhibits substantial performance improvements with minimal fine-tuning data. The performance plateau observed at 125 data points indoors and 1250 outdoors suggests scenario-specific data requirements for optimal performance, influenced by device mobility and environmental complexity. The findings affirm the PMAN's robust adaptability across various settings.

VI. Related Work

This section offers an insightful summary of the research landscape surrounding our work.

Wireless-based Localization Techniques. Cellular networks support a wide range of positioning methods. For outdoor scenarios, the Cell ID (CID) method [29] leverages the cellular network's awareness of the user equipment's (UE) serving cell to provide basic location insights, albeit with constrained accuracy. Observed Time Difference Of Arrival (OTDOA) [30], [31] employs multilateration, estimating positions through the Time of Arrival (ToA) from several base stations. This technique, reliant on base station infrastructure, exhibits efficacy predominantly in Line-Of-Sight (LOS) situations. Assisted Global Navigation Satellite System (AGNSS) [32] utilizes satellite signal measurements retrieved by systems such as Galileo (Europe) and GPS (US) with high accuracy (i.e. few meters), but AGNSS can be compromised by extreme weather conditions which disrupt satellite signal communication with ground devices [2].

Indoor localization solutions exploit various channel attributes, such as Angle of Arrival (AoA) [6], Time of Flight (ToF) [7], and their fusion [9], [33], [34], achieving excellent centimeter-level accuracy in Line-of-Sight (LOS) scenarios. However, these approaches are prone to significant errors in Non-Line-of-Sight (NLoS) settings. Wireless fingerprinting techniques [11]–[13] achieve finer accuracy by matching signal features against a pre-compiled database, where radio-assisted LiDAR SLAM [13] improves accuracy and speed significantly by integrating radio fingerprints with LiDAR for mapping. However, the adaptability of wireless fingerprinting techniques is limited.

In recent years, combining deep learning with domain adaptation [35]–[37] has been shown to significantly enhance a system's ability to adapt to varying environments. For instance, Fidora [36] proposes a Wi-Fi-based indoor localization system that leverages domain adaptation to localize different users with minimal labeled data. Similarly, DAFI [37] introduces a domain adaptation technique that addresses the challenge of fingerprint inconsistency caused by small environmental changes. However, these methods still face limitations in generalizing to highly dynamic or diverse environments, where substantial changes in indoor layouts or user behavior can lead to a degradation in accuracy. Additionally, several studies have focused on improving localization accuracy in Non-Line-of-Sight (NLOS) conditions [38], [39], but the improvements remain limited when compared to ideal line-of-sight (LOS) conditions.

In summary, traditional device localization methods exhibit significant limitations in terms of practicality and scalability, making them unsuitable for the task of proximity estimation. Prior research primarily focused on pinpointing the location of individual devices, whereas *RF-Prox* innovatively facilitates proximity assessments between two indirectly connected devices for the first time. This novel capability sets our approach apart from traditional localization methods and demonstrates its potential for applications such as proximity-based services in IoT, UAVs, and beyond.

Deep Learning in Wireless Sensing. Deep learning architectures have been extensively applied across a variety of wireless sensing tasks, including gesture [17], [40]–[44] and gait recognition [45]–[49], respiration monitoring [50]–[54], fall detection [55]–[60], tracking [9], [61]–[65], and depth estimation [66]. Acted as a pioneering work in the emerging task of proximity estimation for non-directly connected wireless devices, *RF-Prox* is pivotal for a range of applications, including implicit control of IoT devices and proximity-based unmanned aerial vehicle (UAV) scheduling [4]. Recent innovations have incorporated advanced deep learning concepts such as adversarial [67], meta-learning [19], and generative model [68]. By incorporating data augmentation based on generative models like RF-Diffusion, *RF-Prox* holds promise for further performance improvement. Contrary to the conventional reliance on time-frequency spectrograms as inputs [69], *RF-Prox* represents a pioneering approach by utilizing end-to-end complex-valued neural networks for wireless sensing applications, further enhanced by a transfer learning framework to excel in domain generalization.

VII. Discussion and Future Work

A. Application Potential

RF-Prox's ability to estimate proximity between non-connected devices opens promising applications in various industries. In smart cities [70], [71], RF-Prox could enhance public safety, pedestrian monitoring, and transportation flow by enabling communication and distance measurement between non-connected devices in real-time. In autonomous vehicles, proximity estimation between non-connected devices (e.g., vehicles and infrastructure) could support vehicle-to-infrastructure (V2I) systems [72], [73] for safer navigation, particularly in crowded or low-visibility environments. Additionally, industries like healthcare [74] and industrial automation [75], [76] could leverage RF-Prox to monitor device interactions in environments where direct connectivity may be challenging. For example, in healthcare, RF-Prox could support non-invasive patient monitoring by assessing the proximity between wearable devices, medical instruments, and staff devices.

Further adaptation and fine-tuning would be necessary to optimize RF-Prox for healthcare and industrial automation. Specific adjustments could include incorporating customized transfer learning to handle strict privacy constraints in healthcare and reducing latency for fast-response automation in industrial environments. Additionally, these applications could benefit from further refinement in handling dynamic environments and various proximity thresholds to suit specific industry needs.

B. Emerging Wireless Technologies

With the rapid advancement of wireless communication, there are significant opportunities to expand RF-Prox's capabilities by integrating it with emerging technologies such as 5G, 6G [77], and Wi-Fi 6/7. For instance, integrating with 6G networks would enhance RF-Prox's ability to handle high-density environments and high-frequency communications, crucial for smart cities and autonomous systems. Newer wireless technologies also introduce ultra-reliable low-latency communication (URLLC) [77], which could further reduce response times, making RF-Prox suitable for real-time proximity detection in more demanding applications.

Future research will focus on extending RF-Prox to support 6G standards and explore domain adaptation improvements to streamline deployment in varied environments without large-scale data collection. Additionally, enhancing domain adaptation efficiency, possibly through self-supervised learning techniques [78], could reduce the dependence on extensive labeled data during fine-tuning, making RF-Prox more efficient for applications across different domains.

VIII. Conclusion

This paper introduces *RF-Prox*, a novel system designed for the proximity estimation of non-directly connected devices, marking a significant innovation in this domain. Utilizing a sophisticated Multi-Resolution Spatio-Temporal Encoder (MRSTE), *RF-Prox* is capable of extracting domain-agnostic spatio-temporal features from wireless signals. These features are then processed through the Proximity Metric Adaptation Network (PMAN), which converts the extracted latent representations into a set of proximity metrics specifically tailored to the target domain. We implement and evaluate *RF-Prox* on both Wi-Fi-based indoor environments and cellular-based outdoor scenarios. Our results demonstrate that through the incorporation of a transfer learning mechanism, *RF-Prox* efficiently leverages extensive source domain data to learn generalized representations. Moreover, it exhibits remarkable adaptability to new target domains with minimal fine-tuning data. As the inaugural system of its kind, *RF-Prox* represents a pivotal breakthrough in the proximity estimation landscape for non-directly connected devices, offering substantial potential for future applications and research.

Acknowledgment

This work is supported by the NSFC under grant No. 62402276, 62372265, 62271081.

References

[1] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity wi-fi," in *Proceedings of the ACM MobiHoc*, 2017.
[2] "Chapter 23 - extreme ionospheric storms and their effects on gps systems," in *Extreme Events in Geospace*, N. Buzulukova, Ed., 2018.
[3] A. Bensky, *Wireless positioning technologies and applications*, 2016.
[4] Y. Gao, H. Feng, J. Chen, J. Li, and Z. Wei, "A scalable distributed control algorithm for bearing-only passive uav formation maintenance," *Sensors*, 2023.
[5] J. Xiong and K. Jamieson, "Arraytrack: a fine-grained indoor location system," in *Proceedings of the USENIX NSDI*, 2013.
[6] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015.
[7] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-level localization with a single wifi access point," in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, 2016.
[8] E. Soltanaghaei *et al.*, "Multipath triangulation: Decimeter-level wifi localization and orientation with a single unaided receiver," in *Proceedings of the ACM MobiSys*, 2018.
[9] Y. Xie, J. Xiong, M. Li, and K. Jamieson, "md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking," in *Proceedings of the ACM MobiCom*, 2019.
[10] G. Chi, Z. Yang, J. Xu, C. Wu, J. Zhang, J. Liang, and Y. Liu, "Wi-drone: wi-fi-based 6-dof tracking for indoor drone flight control," in *Proceedings of the ACM MobiSys*, 2022.
[11] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proceedings of the ACM MobiCom*, 2012.
[12] D. Li, J. Xu, Z. Yang, Y. Lu, Q. Zhang, and X. Zhang, "Train once, locate anytime for anyone: Adversarial learning based wireless localization," in *Proceedings of the IEEE INFOCOM*, 2021.
[13] R. Liu, B. P. L. Lau, K. Ismail, A. Chathuranga, C. Yuen, S. X. Yang, Y. L. Guan, S. Mao, and U.-X. Tan, "Exploiting radio fingerprints for simultaneous localization and mapping," *IEEE Pervasive Computing*, 2023.

[14] J. Shi, M. Sha, and X. Peng, "Adapting wireless mesh network configuration from simulation to reality via deep learning based domain adaptation," in *Proceedings of the USENIX NSDI*, 2021.

[15] Y. Gao, G. Chi, G. Zhang, and Z. Yang, "Wi-prox: Proximity estimation of non-directly connected devices via sim2real transfer learning," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023.

[16] Z. Yang, Z. Zhou, and Y. Liu, "From rssi to csi: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, 2013.

[17] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] R. Song, D. Zhang, Z. Wu, C. Yu, C. Xie, S. Yang, Y. Hu, and Y. Chen, "Rf-url: unsupervised representation learning for rf sensing," in *Proceedings of the ACM MobiCom*, 2022.

[19] S. Ding, Z. Chen, T. Zheng, and J. Luo, "Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition," in *Proceedings of the ACM SenSys*, 2020.

[20] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *International Conference on Learning Representations*, 2018.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[22] J. Wang, D. Vasisht, and D. Katabi, "Rf-idraw: Virtual touch screen in the air using rf signals," *ACM SIGCOMM Computer Communication Review*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[25] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. of Information Theory and Applications Workshop (ITA)*, 2019.

[26] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, "A theoretical analysis of ndcg type ranking measures," in *Conference on learning theory*, 2013.

[27] M. E. Berezin, F. Rousseau, and A. Duda, "Citywide mobile internet access using dense urban wifi coverage," 2012.

[28] R. S. Hassan, T. Rahman, and A. Abdulrahman, "Lte coverage network planning and comparison with different propagation models," *Telkomnika (Telecommunication Computing Electronics and Control)*, 2014.

[29] S. M. Razavi, F. Gunnarsson, H. Rydén, Å. Busin, X. Lin, X. Zhang, S. Dwivedi, I. Siomina, and R. Shreevastav, "Positioning in cellular networks: Past, present, future," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, 2018.

[30] M. Huang and W. Xu, "Enhanced lte toa/otdoa estimation with first arriving path detection," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 2013.

[31] H. Ryden, A. A. Zaidi, S. M. Razavi, F. Gunnarsson, and I. Siomina, "Enhanced time of arrival estimation and quantization for positioning in lte networks," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016.

[32] S. Madry *et al.*, *Global navigation satellite systems and their applications*, 2015.

[33] G. Zhang, G. Chi, Y. Zhang, X. Ding, and Z. Yang, "Push the limit of millimeter-wave radar localization," *ACM Trans. Sen. Netw.*, 2023.

[34] Z. Yang, Y. Zhang, G. Chi, and G. Zhang, "Hands-on wireless sensing with wi-fi: A tutorial," *arXiv preprint arXiv:2206.09532*, 2022.

[35] Y. Tian, J. Wang, and Z. Zhao, "Wi-fi fingerprint update for indoor localization via domain adaptation," in *2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS)*, 2021.

[36] X. Chen, H. Li, C. Zhou, X. Liu, D. Wu, and G. Dudek, "Fidora: Robust wifi-based indoor localization via unsupervised domain adaptation," *IEEE Internet of Things Journal*, 2022.

[37] H. Li, X. Chen, J. Wang, D. Wu, and X. Liu, "Dafi: Wifi-based device-free indoor localization via domain adaptation," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2022.

[38] P. Wang, T. Koike-Akino, and P. V. Orlik, "Fingerprinting-based indoor localization with commercial mmwave wifi: Nlos propagation," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020.

[39] W. Wang, Y. Zhang, and L. Tian, "Toa-based nlos error mitigation algorithm for 3d indoor localization," *China Communications*, 2020.

[40] N. Yu, W. Wang, A. X. Liu, and L. Kong, "Qgesture: Quantifying gesture distance and direction with wifi signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

[41] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: A ubiquitous wifi-based gesture recognition system," in *2015 IEEE conference on computer communications (INFOCOM)*, 2015.

[42] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the 19th annual international conference on Mobile computing & networking*, 2013.

[43] S. Tan and J. Yang, "Wifinger: Leveraging commodity wifi for fine-grained finger gesture recognition," in *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*, 2016.

[44] R. H. Venkatnarayan, G. Page, and M. Shahzad, "Multi-user gesture recognition using wifi," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018.

[45] Y. Zhang, Y. Zheng, G. Zhang, K. Qian, C. Qian, and Z. Yang, "Gaitsense: towards ubiquitous gait-based human identification with wi-fi," *ACM Transactions on Sensor Networks*, 2021.

[46] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wifi signals," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.

[47] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Wiwho: Wifi-based person identification in smart spaces," in *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2016.

[48] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015.

[49] C. Wu, F. Zhang, Y. Hu, and K. R. Liu, "Gaitway: Monitoring and recognizing gait speed through the walls," *IEEE Transactions on Mobile Computing*, 2020.

[50] H. Abdelnasser, K. A. Harras, and M. Youssef, "Ubibreathe: A ubiquitous non-invasive wifi-based breathing estimator," in *Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing*, 2015.

[51] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf wifi," in *Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing*, 2015.

[52] X. Wang, C. Yang, and S. Mao, "Tensorbeat: Tensor decomposition for monitoring multiperson breathing beats with commodity wifi," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2017.

[53] C. Wu, Z. Yang, Z. Zhou, X. Liu, Y. Liu, and J. Cao, "Non-invasive detection of moving and stationary human with wifi," *IEEE Journal on Selected Areas in Communications*, 2015.

[54] F. Zhang, C. Wu, B. Wang, M. Wu, D. Bugos, H. Zhang, and K. R. Liu, "Smars: Sleep monitoring via ambient radio signals," *IEEE Transactions on Mobile Computing*, 2019.

[55] Z. Yang, Y. Zhang, and Q. Zhang, "Rethinking fall detection with wi-fi," *IEEE Transactions on Mobile Computing*, 2022.

[56] S. Ji, Y. Xie, and M. Li, "Sifall: Practical online fall detection with rf sensing," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022.

[57] Y. Hu, F. Zhang, C. Wu, B. Wang, and K. R. Liu, "A wifi-based passive fall detection system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[58] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "Falldefi: Ubiquitous fall detection using commodity wi-fi devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

[59] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "Rf-based fall monitoring using convolutional neural networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

[60] G. Chi, G. Zhang, X. Ding, Q. Ma, Z. Yang, Z. Du, H. Xiao, and Z. Liu, "Xfall: Domain adaptive wi-fi-based fall detection with cross-modal supervision," *IEEE Journal on Selected Areas in Communications*, 2024.

[61] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person localization via rf body reflections," in *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, 2015.

[62] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2. 0: Passive human tracking with a single wi-fi link," in *Proceedings of the 16th annual international conference on mobile systems, applications, and services*, 2018.

[63] C. Wu, F. Zhang, Y. Fan, and K. R. Liu, "Rf-based inertial measurement," in *Proceedings of the ACM Special Interest Group on Data Communication*, 2019.

[64] K. Ismail, R. Liu, A. Athukorala, B. K. K. Ng, C. Yuen, and U.-X. Tan, "Wifi similarity-based odometry," *IEEE Transactions on Automation Science and Engineering*, 2023.

[65] J. Xu, G. Chi, Z. Yang, D. Li, Q. Zhang, Q. Ma, and X. Miao, "Followupar: enabling follow-up effects in mobile ar applications," in

Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, 2021.

[66] S. Cheng, Y. Gao, Z. Yang, G. Chi, and T. X. Han, “Wivid: Leveraging wi-fi and vision for depth estimation via multimodal diffusion,” in *2024 20th International Conference on Mobility, Sensing and Networking (MSN)*, 2024.

[67] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, “Towards environment independent device free human activity recognition,” in *Proceedings of the ACM MobiCom*, 2018.

[68] 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.

[69] Z. Yang, Y. Zhang, K. Qian, and C. Wu, “Slnet: A spectrogram learning neural network for deep wireless sensing,” in *Proceedings of the USENIX NSDI*, 2023.

[70] R. Rathnayake, M. W. P. Maduranga, V. Tilwari, and M. B. Dissanayake, “Rssi and machine learning-based indoor localization systems for smart cities,” *Eng*, 2023.

[71] J. S. Gracias, G. S. Parnell, E. Specking, E. A. Pohl, and R. Buchanan, “Smart cities—a structured literature review,” *Smart Cities*, 2023.

[72] D. Kanthavel, S. Sangeetha, and K. Keerthana, “An empirical study of vehicle to infrastructure communications-an intense learning of smart infrastructure for safety and mobility,” *International Journal of Intelligent Networks*, 2021.

[73] T. Sukuvaara and P. Nurmi, “Wireless traffic service platform for combined vehicle-to-vehicle and vehicle-to-infrastructure communications,” *IEEE Wireless Communications*, 2009.

[74] L. Lubna, H. Hameed, S. Ansari, A. Zahid, A. Sharif, H. T. Abbas, F. Alqahtani, N. Mufti, S. Ullah, M. A. Imran *et al.*, “Radio frequency sensing and its innovative applications in diverse sectors: A comprehensive study,” *Frontiers in Communications and Networks*, 2022.

[75] P. Neumann, “Communication in industrial automation—what is going on?” *Control Engineering Practice*, 2007.

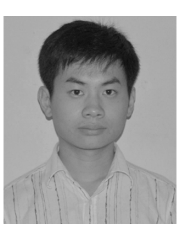
[76] C. Dey and S. K. Sen, *Industrial automation technologies*, 2020.

[77] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, “Toward 6g networks: Use cases and technologies,” *IEEE communications magazine*, 2020.

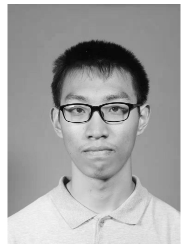
[78] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” *Technologies*, 2020.

Yuchong Gao (Student Member, IEEE) received his B.E. degree in School of Information and Communication Engineering from Beijing University of Posts and Telecommunications in 2024. He is currently working towards the Master degree in School of Software, Tsinghua University. His research interests include wireless sensing and mobile computing.

Guoxuan Chi (Member, IEEE) received his B.E. degree in School of Information and Communication Engineering from Beijing University of Posts and Telecommunications in 2019, and the Ph.D. degree in the School of Software, Tsinghua University in 2024. He is currently a research assistant in Tsinghua University. His research interests include wireless sensing and mobile computing.



Zheng Yang (Fellow, IEEE) received the B.E. degree in computer science from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He is currently an associate professor with Tsinghua University. His main research interests include Internet of Things and mobile computing. He is the PI of National Natural Science Fund for Excellent Young Scientist. He was the recipient of the State Natural Science Award (second class).



Shijie Cheng (Student Member, IEEE) received his B.E. degree in Department of Intelligence and Computing from Tianjin University in 2023. He is currently working towards the Master degree in School of Software, Tsinghua University. His research interest is wireless sensing.

Zhiqing Wei (Member, IEEE) received the B.E. and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2010 and 2015, respectively. He is an Associate Professor with BUPT. He has authored one book, three book chapters, and more than 50 papers. His research interests include the performance analysis and optimization of intelligent machine networks. Dr. Wei was granted the Exemplary Reviewer of IEEE Wireless Communications Letters in 2017 and the Best Paper Award of International Conference on Wireless Communications and Signal Processing (WCSP) in 2018. He was the Registration Co-Chair of IEEE/Chinese Institute of Communications (CIC) International Conference on Computer and Communications (ICCC) 2018 and the Publication Co-Chair of IEEE/CIC ICCC 2019 and IEEE/CIC ICCC 2020.