Wi-Prox: Proximity Estimation of Non-directly Connected Devices via Sim2Real Transfer Learning

Yuchong Gao[1]*, Guoxuan Chi[2]*, Guidong Zhang[2], Zheng Yang[2]†
*Co-primary author †Corresponding author
[1]Beijing University of Posts and Telecommunications, China
[2]School of Software and BNRist, Tsinghua University, China

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437884

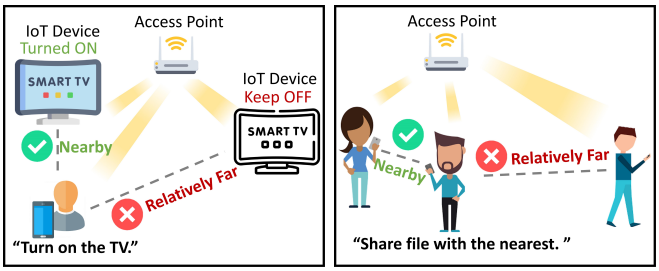
***Abstract*—Recent years have witnessed an increasing number of mobile devices, posing a more diversified demand for device localization solutions. While existing wireless localization solutions can obtain the relative locations of connected devices, they fall short in estimating the spatial relationships between devices that are not directly connected. To address this technical gap, we propose *Wi-Prox*, the first proximity estimation system for non-directly connected devices. *Wi-Prox* evaluates the spatial proximity of two devices by analyzing their received wireless signals. It integrates a novel multi-resolution spatial encoder that extracts multi-scale spatial features from complex-valued wireless signals, which are then analyzed and transformed into a domain-adaptive proximity metric. To enhance the generalizability of *Wi-Prox*, we adopt a simulation-to-reality transfer learning framework. *Wi-Prox* is pre-trained with a large amount of simulated data and then fine-tuned for real-world deployment, significantly reducing the need for real-world data collection. We implement *Wi-Prox* and evaluate its performance in both simulated and real environments. Our results indicate that a fine-tuned *Wi-Prox* achieves an average accuracy of 97.2% in selecting the most proximate device. Even without fine-tuning, a pre-trained *Wi-Prox* still manages an average accuracy of 93.8%, thereby demonstrating impressive performance in terms of both proximity estimation accuracy and domain generalizability.**

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.

Existing wireless localization solutions are designed for devices with a direct communication link, such as wireless access point (AP) and user equipment (UE). Unfortunately, there is no effective wireless-based solution to obtain spatial relationships between non-directly connected terminals (e.g., UE and IoT devices). And this capability, as shown in Fig. 1, is fundamental for many novel applications, such as implicit control of IoT devices and proximity-based user discovery.

One straightforward approach is to estimate the location of each device independently, and then infer their relative proximity from these location estimates. However, geometric-based solutions relying on the channel parameters, such as angle-of-arrival (AoA) [2], [3], time-of-flight (ToF) [4], [5], and their fusion [6], [7], yield significant localization errors in non-line-of-sight (NLoS) environments. Alternatively, the fingerprint-based localization method requires collecting a large amount of labeled data and suffers from severe generalization problems for cross-domain application [8], [9].



(a) Implicit control of IoT device (b) Proximity-based user discovery

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

Different from the traditional device localization approaches, we draw our inspiration from the concept of "estimating by comparing," based on the observation that wireless devices in close proximity exhibit similar signal propagation processes. By analyzing and comparing the spatial features implied in the received signal of two devices, their proximity can be inferred. However, translating this intuitive idea into a practical system poses significant challenges. First, accurate extraction of spatial features is difficult as traditional geometric features like AoA and ToF suffer from intolerable errors in non-line-of-sight (NLoS) conditions [7]. Second, formulating a domain-adaptive proximity metric is crucial because signal propagation characteristics can vary due to different indoor layouts and device deployments, leading to variations in the proximity metrics.

To overcome the above challenges, we propose *Wi-Prox*, the first proximity estimation system for wireless devices that are not directly connected. To extract accurate spatial features, we adopt a data-driven approach and design a complex-valued neural network module called Multi-Resolution Spatial Encoder (MRSE). The MRSE is capable of extracting multi-scale latent representations from the wireless signal and fusing them into a feature vector that represents the spatial characteristics of the wireless channel. To formulate a domain-adaptive proximity metric, we construct a Proximity Metric Adaptation Network (PMAN) that compares the spatial features of two wireless channels and evaluates the devices' proximity with domain adaptation capability. In *Wi-Prox*, we leverage a simulation-to-reality (Sim2Real) transfer learning mechanism [10], which allows us to pre-train the model using simulated data and then fine-tune it with a minimal amount of real-world data. This approach significantly reduces the need for real-world data collection, while ensuring its generalization capability. We implement *Wi-Prox* and evaluate

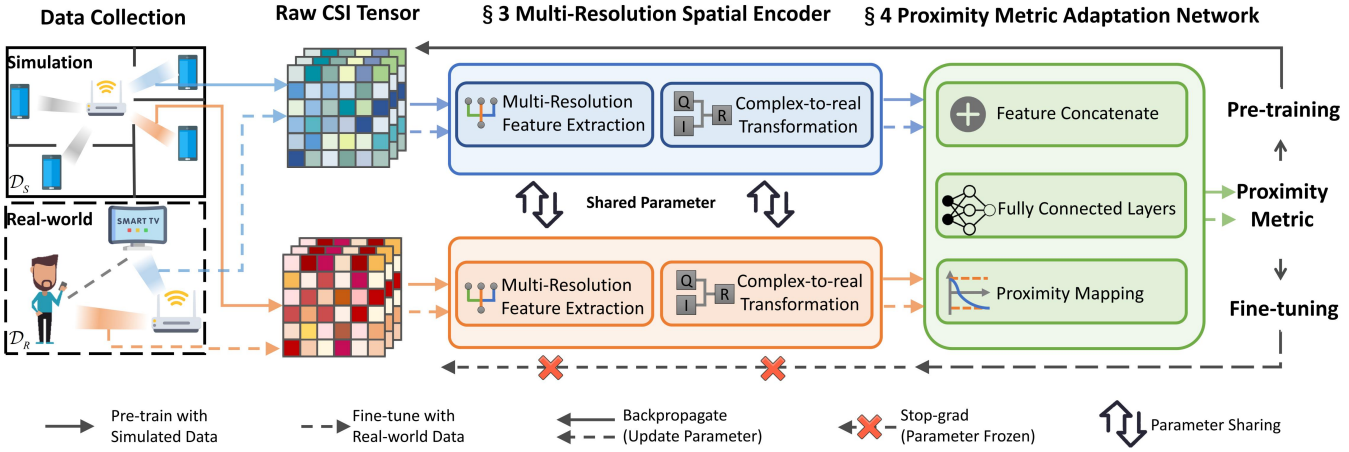


Fig. 2. An overview of the *Wi-Prox*, where solid and dashed lines represent data collection from simulated and real-world environments, respectively, with blue and orange used to distinguish devices.

its performance across more than 3,000 domains in both simulated environments and real-world scenarios. During the evaluation process, over 7,700,000 data samples are collected. The results show that a fine-tuned *Wi-Prox* achieves an average accuracy of 97.2% in selecting the most proximate device, and even a pre-trained model without fine-tuning achieves an average of 93.8% accuracy, demonstrating its impressive performance in both proximity estimation accuracy and domain generalizability.

We summarize our contributions as follows: 1) We propose *Wi-Prox*, the first proximity estimation system for non-directly connected wireless devices. *Wi-Prox* shows the domain-adaptive capability and can be easily deployed in any real-world environment, making a promising step towards integrated sensing and communication. 2) Our proposed Multi-Resolution Spatial Encoder is a pioneering attempt at applying complex-valued neural networks to wireless sensing. The multi-resolution design has its unique advantages and can also be integrated into other types of wireless sensing applications. 3) The simulation-to-reality transfer learning adopted by our system has been proven effective, providing a new approach to enhance the generalizability of data-driven wireless systems. 4) We implement and evaluate *Wi-Prox* on commercial hardware, which showcases the practicality and effectiveness of deploying *Wi-Prox* in real-world scenarios. Our work will be open-sourced after acceptance to facilitate the research community.

The rest of this paper is organized as follows. We first present the overview of *Wi-Prox* in Section II, followed by the detailed design of the MRSE in Section III and PMAN in Section IV. Our implementation and evaluation of *Wi-Prox* are shown in Section V, and the conclusion in Section VI.

II. System Overview

Wi-Prox is a device proximity estimation system based on the wireless signal, which consists of two key components: *Multi-Resolution Spatial Encoder (MRSE)* and *Proximity Metric Adaptation Network (PMAN)*. As illustrated in Fig. 2, the process of *Wi-Prox* begins by extracting the channel state information (CSI) of wireless links corresponding to two different mobile devices. The collected CSI tensors are then fed into the MRSE module for feature extraction. MRSE is a complex-valued neural network composed of residual convolution blocks designed to extract multi-resolution latent representations from both the real and imaginary parts of the CSI. Subsequently, a complex-to-real transformation is applied to convert the complex-valued representation to a real-valued spatial feature for further analysis. Once encoded with MRSE, the domain-independent spatial features are concatenated and subsequently passed through fully connected layers of the PMAN module for joint analysis and comparison. After being transformed by an elaborately designed proximity mapping function, the output proximity metric of two devices can be obtained.

Wi-Prox employs a simulation-to-reality transfer learning framework, whereby the model is initially pre-trained in a simulated environment. This pre-trained model can be directly deployed in real-world settings and fine-tuned as needed. During pre-training, a large amount of data generated from a simulated data domain $\mathcal{D}_S$ are leveraged to enhance the model's generalization capability in extracting domain-independent features. Following pre-training, the model is fine-tuned with only a small amount of real-world data from $\mathcal{D}_R$. By leveraging this approach, *Wi-Prox* can be easily adapted to new scenarios for practical application.

III. Multi-Resolution Spatial Encoder

In this section, we introduce the *Multi-Resolution Spatial Encoder (MRSE)*, designed to extract the domain-independent spatial information embedded in the CSI. As illustrated in Fig. 3, MRSE 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, which could be further used for robust device proximity estimation.

Compared with geometric-based algorithms [3], [6], our proposed MRSE leverage a data-driven approach, analyzing signal statistical information in high-dimensional space. This strategy enhances system performance in complex indoor scenarios, particularly under NLoS conditions.

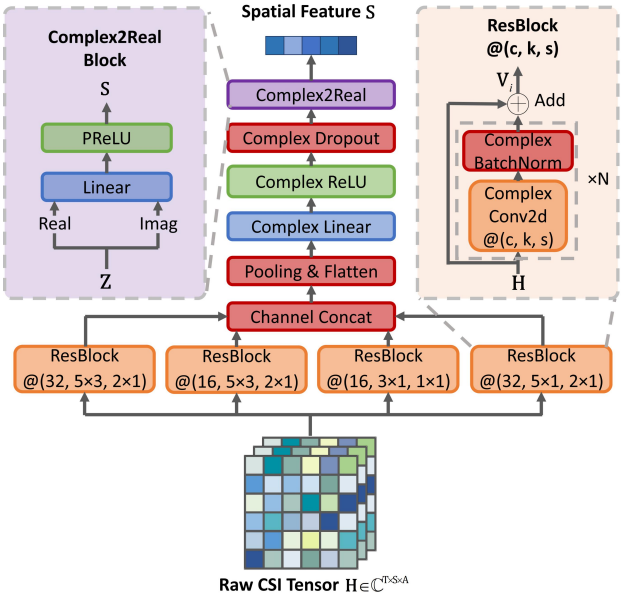


Fig. 3. Illustration of Multi-Resolution Spatial Encoder.

A. CSI Preliminary

Taking multipath propagation into account, the wireless channel can be formulated by frequency f and time t as:

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

where L represents the number of multipath components. $\alpha_l(t,f)$ and $\tau_k(t)$ are the complex attenuation factor and propagation delay for the l-th path respectively. CSI is a discretely sampled version of channel response [11]. In the frequency domain, CSI is sampled on certain OFDM subcarriers, in the time domain, CSI is measured for each received packet, while in the spatial domain, CSI can be measured on each radio chain (i.e., Tx-Rx pair). Therefore, CSI is generally considered as a complex-valued tensor $\mathbf{H} \in \mathbb{C}^{T\times S\times A}$, where T, S, A are the number of time samples, subcarriers and radio chains.

B. Complex-valued Network for CSI Processing

Previous researches usually take the pre-processing results of CSI, such as short-time Fourier transform and ToF-AoA spectrogram [12], [13]) as the input of a classification network model for learning, or split the original CSI into real and imaginary parts [14] for separate processing within a deep neural network. In contrast, the raw CSI as a whole can be used as input to extract richer spatial information. Therefore, we exploit the idea of the complex-valued neural network and integrate several innovative components including complex-valued linear layers and complex-valued convolutional layers into our proposed MRSE.

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 demonstrated [15] that dropout, batch normalization, and activation operations can be directly applied in the complex domain by individually manipulating the real and imaginary components of the input. In this way, each complex module in MRSE is a linear combination of real domain operations, which guarantees the differentiability of the entire MRSE module.

C. Multi-Resolution Feature Extraction

The core design of MRSE is to fuse CSI features from multiple scales, which has been proven effective in the field of computer vision [16]. One intuitive explanation for the rationale behind this design is that angle-of-arrival (AoA) measurement from CSI obtained using different antenna spacing can form a trade-off between resolution and range [17].

As illustrated in Fig. 3, MRSE comprises four residual convolution blocks, each with distinct output channels, kernel sizes, and strides, thus establishing four parallel paths. Denote the residual block as $\mathrm{ResBlock}(\cdot)$, and let $\mathbf{C}_i$ be the set of parameters of the i-th residual block. Given the input CSI tensor $\mathbf{H}$, the feature extracted from the i-th block can be written as:

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

where the residual block is basically a convolution with shortcut connection [18], 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 to a latent representation $\mathbf{Z} = \mathrm{Concat}(\mathbf{V}_0,\mathbf{V}_1,\mathbf{V}_2,\mathbf{V}_3)$. The concatenated $\mathbf{Z}$ contains multi-level features of the CSI input, which greatly improves the receptive field of MRSE and thus enhances the generalization performance of *Wi-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 $\mathbf{Z}$ to the real-valued spatial feature $\mathbf{S}$, we design a complex-to-real transformation module C2R, which applies two linear operations on the real and imaginary part:

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

where $\mathbf{W}^{\mathrm{R}}$ and $\mathbf{W}^{\mathrm{I}}$ are the real-valued linear weights.

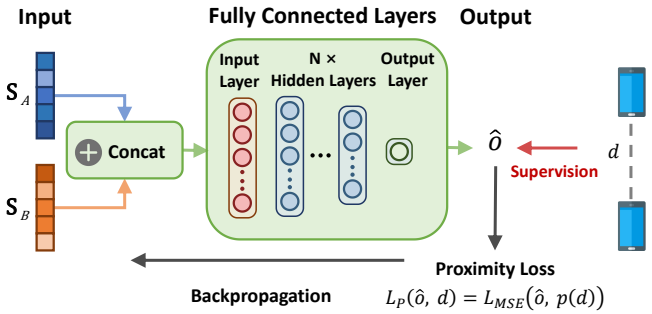


Fig. 4. An overview of the Proximity Metric Adaptation Network.

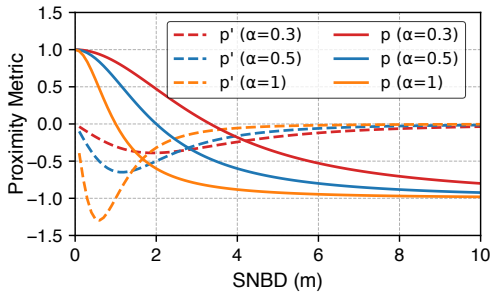


Fig. 5. Illustration of the proximity mapping function varied by α.

IV. Proximity Metric Adaptation Network

In this section, we introduce the *Proximity Metric Adaptation Network (PMAN)*. As illustrated in Fig. 4, PMAN transforms the spatial features extracted from two wireless devices $\mathbf{S}$ to their proximity metric with domain adaptation capability. PMAN is implemented based on the fully connected layers and an elaborately designed proximity loss function $L_P(\cdot)$. When being deployed in a new environment with different 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.

A. Metric Network Design

The PMAN takes the spatial features from two wireless devices as inputs, which are denoted as $\mathbf{S}_A$ and $\mathbf{S}_B$ respectively. Two features are first concatenated into $\mathbf{S}' = \mathrm{Concat}(\mathbf{S}_A, \mathbf{S}_B)$, and then put into the fully connected layers $\mathrm{FC}(\cdot)$ to get the proximity estimation $\hat{o} = \mathrm{FC}(\mathbf{S}'; \Theta_{\mathrm{FC}})$, where Θ_{FC} is the set of parameters of fully connected layers.

In typical indoor environments, using Euclidean distance to describe the proximity of two devices may not be a wise choice, since two devices with close Euclidean distance may be blocked by obstacles (e.g., the walls). This is not in accordance with the proximity of the user perception. As such, we define the ground truth label in our system as the Shortest Non-Blocking Distance (SNBD), signifying the minimal path length between wireless devices that does not traverse any obstacles. To enable the model to learn more effective parameters during the backpropagation process, we designed a novel proximity loss function $L_P(\cdot)$. Given a batch of proximity output $\hat{\mathbf{O}}$ and corresponding SNBD $\mathbf{d}$, the loss can calculated as:

$$L_P(\hat{\mathbf{O}}, \mathbf{d}) = L_{\mathrm{MSE}}(\hat{\mathbf{O}}, p(\mathbf{d})) = \frac{1}{N}\|\hat{\mathbf{O}} - p(\mathbf{d})\|_2^2, \quad (7)$$

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 steepness of the function $p(\cdot)$.

In proximity estimation, the system should exhibit greater sensitivity to nearby devices as opposed to distant ones, considering that device proximity distance typically ranges from 0 to 4 m based on our comprehensive evaluation. The elastic parameter α governs the model's distance of interest. Figure 5 illustrates the variation in the mapping function $p(\cdot)$ and its derivative $p'(\cdot)$ with different values of α. We select $\alpha = 0.5$ to ensure that $p(\cdot)$ exhibits a sharp gradient from 0 to 4 m and a flat gradient at longer distances, thus adapting the model to indoor environments.

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. Sim2Real Transfer Learning

Our goal is to develop a proximity estimator for real-world applications. However, collecting large amounts of CSI data in the real world is a laborious task, and simulated data, while easily accessible, may not maximize the real-world performance of our system. Therefore, we propose a pre-training and fine-tuning strategy from simulation to reality to enhance the model's domain adaptation capability.

Pre-training in simulated environments. We build a simulation environment for integrated sensing and communication based on the ray tracing model and collected labeled CSI data under tens of different indoor building structures. pre-training with a large amount of simulated data helps the MSRE module to learn domain generalized spatial features without overfitting specific indoor structures. During the pre-training process, all the parameters in *Wi-Prox* are jointly optimized. Suppose our pre-training model is M_P. Denote the simulated data domain as $\mathcal{D}_S$, and the set of parameter in MRSE and PMAN as Θ_{M} and Θ_{P} respectively, the optimization (a.k.a backpropagation) process can be written as:

$$\{\Theta_{\mathrm{M}}, \Theta_{\mathrm{P}}\} = \arg\min_{\{\Theta_{\mathrm{M}}, \Theta_{\mathrm{P}}\}} \sum_{(\mathbf{H}, d) \sim \mathcal{D}_S} \frac{1}{|\mathcal{D}_S|} L_P(\mathrm{M}_P(\mathbf{H}; \Theta_{\mathrm{M}}, \Theta_{\mathrm{P}}), d). \quad (8)$$

Fine-tuning for real-world application. The pre-trained model based on the simulation data has a certain generalization capability. In order to achieve better real-world performance, a small amount of real-world data are collected to fine-tune the pre-trained model. During the fine-tuning process, the parameters of the MRSE are frozen and only the parameters of the PMAN are optimized. the optimization process can be written as:

$$\Theta_{\mathrm{P}} = \arg\min_{\Theta_{\mathrm{P}}} \sum_{(\mathbf{H}, d) \sim \mathcal{D}_R} \frac{1}{|\mathcal{D}_R|} L_P(\mathrm{M}_F(\mathbf{H}; \Theta_{\mathrm{M}}, \Theta_{\mathrm{P}}), d), \quad (9)$$

where M_F refers to the fine-tuning model, and $\mathcal{D}_R$ indicates the real-world data domain.

Upon completion of the pre-training and fine-tuning processes, we get a pre-trained model M_P with high generalizability, and a fine-tuned model M_P which adapts to a specific real environment.

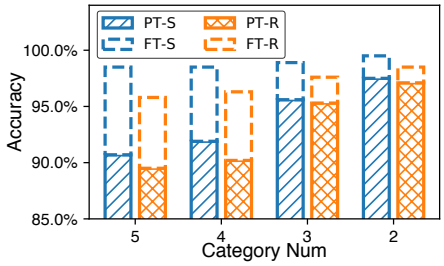


Fig. 6. Top-1 Accuracy for N Category

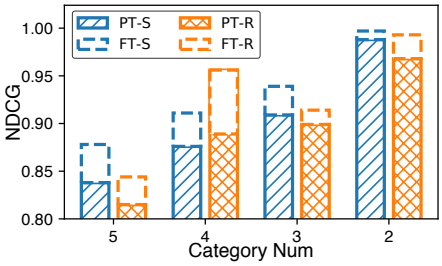


Fig. 7. NDCG for N-device sorting

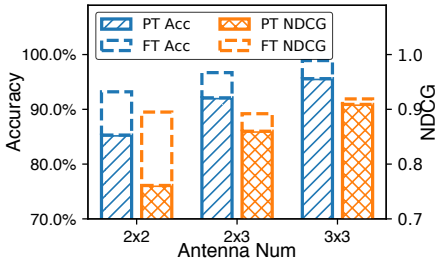


Fig. 8. Performance with different antenna number

V. Evaluation

A. Experimental Methodology

1) Experimental Scenarios: In order to fully evaluate the performance of *Wi-Prox*, we set two kinds of experimental environments: a simulation environment and a real-world environment. We build a simulation environment based on the MATLAB Communication Toolbox. During the simulation process, more than 30 different indoor environments are set, each with 100 different AP deployment scenarios. In each deployment case, 3-6 UEs are randomly placed in different locations to collect CSI data. For the real-world evaluation, the system is deployed in an office building with 3 different AP deployments. In each deployment case, 3-6 UEs are allowed to move within the given range and the corresponding CSI can be acquired in real time. The devices' ground-truth locations are obtained in real time through surveillance cameras.

2) System Implementation: *Wi-Prox* consists of 1 AP and multiple clients working at 5.6 GHz. Both the transmitter and the receiver are equipped with 3 antennas with a spacing of $\lambda/4$, forming into a maximum of 3×3 MIMO array. For real-world evaluation, we fully implement *Wi-Prox* with commercial Wi-Fi NICs AR9580. *Wi-Prox* utilizes a hybrid programming approach in MATLAB and Python to enable fast and efficient processing. Specifically, MATLAB is used for collecting CSI data from either the simulation environment or the real-world platform. The collected CSI data is instantly streamed to the server, where the deep learning model built with Python is deployed, making predictions in real time.

3) Comparative Methods: To extensively evaluate the performance of *Wi-Prox*, we implement two state-of-the-art Wi-Fi-based localization approaches for comparison. We compare our *Wi-Prox* with SpotFi [3] and mD-Track [6] by replacing our MRSE module with each of these approaches to demonstrate the superiority of our module design.

4) Evaluation Metrics: Suppose there are $N+1$ UEs in the experimental scenario, where we randomly choose one as the reference and estimate its proximity with the rest of N devices, **Top-1 Accuracy** indicates the success rate of selecting the most proximate device. **Normalized Discounted Cumulative Gain (NDCG)** [19] is a common ranking metric in sorting problems ranging from 0 to 1, which is used to evaluate the sorted proximity estimation result.

B. Overall Performance

In this section, we evaluate the overall performance of *Wi-Prox*. Denote the evaluation under a new virtual simulated environment and a new real-world environment as 'S' and 'R' and set the pre-trained model and fine-tuned model as 'PT' and 'FT' respectively.

1) Top-1 accuracy: Fig. 6 depicts the performance of the *Wi-Prox*, where the dashed box indicates the performance improvement after fine-tuning. As shown, the pre-trained model ('PT-R', 'PT-S') achieves accuracy from 89.5% / 90.7% to 97.1% / 97.5% with categories from 5 to 2, respectively, which shows good generalizability of the system. After fine-tuning, the average Top-1 accuracy exceeds 95.0% for all categories, demonstrating that the system is able to handle different numbers of categories in common scenarios with good adaptation capability, where the simulation-to-reality gap is well bridged by the sim2real transfer learning strategy.

2) NDCG: We set the reference point and other N UEs on a line. As shown in Fig. 7, the NDCG with the pre-trained model ('PT-R', 'PT-S') varies from 0.815 / 0.838 to 0.968 / 0.988 with categories from 5 to 2, while the fine-tuned model shows good performance with over 0.844 for all categories, manifesting its good spatial distance perception ability.

3) System latency & model parameters: We employ the PyTorch Profiler to count the number of floating point operations (FLOPs), model parameters, and inference time of each component. The data is shown in Table I. It's noted that the overall number of model parameters is smaller than commonly used small-scale models (e.g. ResNet-18 [18]), which shows that the model has great potential to be deployed directly on various edge-embedded devices and realize real-time inference. Besides, the extremely small number of parameters in the PMAN compared to the MRSE enables minimal data volume fine-tuning and domain adaptation.

TABLE I
System Latency & Number of Model Parameters

Parameters	Multi-resolution CSI Encoder	Adative Metric Network	Overall
FLOPs (k)	35220.4	3.726	35224.1
Model Parameters (k)	82.56	3.818	86.38
Inference Time (ms)	2.18	0.11	2.29

C. Micro-benchmarks

In this section, we set the category number to 3 and implement robustness analysis for antenna number, SNR, CSI-based feature extraction models (e.g. MRSE), and fine-tuning data volume to evaluate their impact on system performance.

1) Antenna Number: We configure the antenna size as a 2×2, 2×3, or 3×3 MIMO array. Fig. 8 indicates a slight increase in both accuracy and NDCG as the number of antennas increases. For a 3×3 MIMO array, the pre-trained

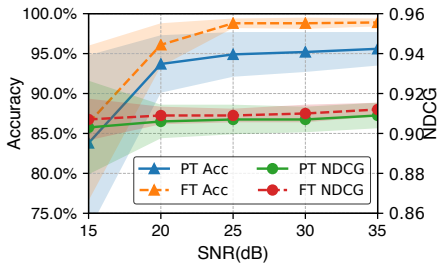


Fig. 9. Performance under different SNR

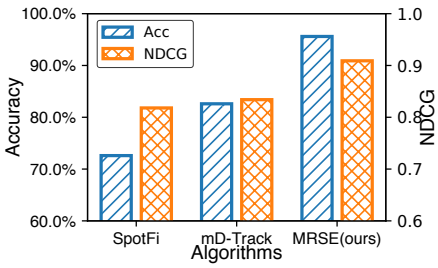


Fig. 10. Effectiveness of MRSE

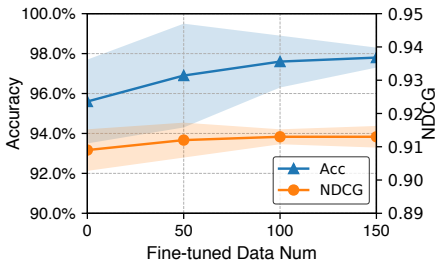


Fig. 11. The impact of fine-tuning data volume

model achieves accuracy and NDCG of 95.6% and 0.909, respectively, while the fine-tuned model reaches 98.9% and 0.919. We anticipate further improvement in performance with more antennas because more channel information provided indirectly enhances the capacity for multipath resolution discrimination, which benefits the extraction of potential spatial features.

2) SNR: We set SNR as $\{15, 20, 25, 30, 35\}$ dB, which is typical in indoor environments [20]. In Fig. 9, the mean and standard deviation are represented by dots and color blocks respectively. The results demonstrate that the system can perform well in different common indoor SNRs. For example, when SNR is 25, the accuracy reaches 94.9±2.8% / 98.8±0.6% and NDCG 0.907±0.007 / 0.907±0.003 under pre-trained and fine-tuned model respectively.

3) Effectiveness of MRSE: To verify the robustness and expressiveness of the high-order spatial features extracted by MRSE, we replaced that with the multipath AoA and ToF acquired by SpotFi and mD-Track to train models. As shown in Fig. 10, with the mere pre-trained model, MRSE outperforms mD-Track and SpotFi on both accuracy and NDCG, achieving improvements of 13.0% / 0.075 and 23.0% / 0.091, respectively.

4) Fine-tuning data volume: We set data volume as $\{0, 50, 100, 150\}$. It is worth noting that data amounts being 0 is equivalent to using the pre-trained model. As shown in Fig. 11, the system can achieve good performance given only a small amount of data to fine-tune. When fine-tuning data amounts reach 150, the model achieves 97.8±0.5 and 0.913±0.003 on both accuracy and NDCG, which demonstrates the good adaptive capability of the PMAN.

VI. Conclusion

This paper proposes the design and implementation of *Wi-Prox*, a proximity estimation system for non-directly connected devices. *Wi-Prox* utilizes a complex-valued Multi-Resolution Spatial Encoder (MRSE) to extract domain-independent spatial features and leverages the Proximity Metric Adaption Network (PMAN) to transform spatial features into proximity metrics. By employing Sim2Real transfer learning, *Wi-Prox* learns generalized representations in the simulated environment and adapts to specific real-world environments using only a minimal amount of fine-tuning data.

Acknowledgment

This work is supported in part by the NSFC under grant 61832010, 62372265.

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] J. Xiong and K. Jamieson, "Arraytrack: a fine-grained indoor location system," in *Proceedings of the USENIX NSDI*, 2013.

[3] 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, pp. 269–282.

[4] 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, pp. 165–178.

[5] E. Soltanaghaei *et al.*, "Multipath triangulation: Decimeter-level wifi localization and orientation with a single unaided receiver," in *Proceedings of the ACM MobiSys*, 2018.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Z. Yang, Z. Zhou, and Y. Liu, "From rssi to csi: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–32, 2013.

[12] 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*, vol. 44, no. 11, pp. 8671–8688, 2021.

[13] 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.

[14] 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.

[15] 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.

[16] 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*, vol. 31, no. 1, 2017.

[17] J. Wang, D. Vasisht, and D. Katabi, "Rf-idraw: Virtual touch screen in the air using rf signals," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 235–246, 2014.

[18] 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, pp. 770–778.

[19] 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*. PMLR, 2013, pp. 25–54.

[20] I. S. Association, "Part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications," *802.11-2016 - IEEE Standard for Information Technology*, 2016.