The Summary of Difference from the Conference Version in IEEE GLOBECOM 2023

Preliminary conference paper:

Yuchong Gao, Guoxuan Chi, Guidong Zhang, and Zheng Yang. "*Wi-Prox: Proximity Estimation of Non-Directly Connected Devices via Sim2Real Transfer Learning*", IEEE Global Communications Conference (GLOBECOM) 2023.

Main differences:

(1) We expand 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.

a) We have modified some expressions in Section 1 about background and motivation:

> *"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. 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, 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."*

b) We modify the System overview to expand origin simulation-to-reality transfer learning to a more general transfer learning between source and target domains and introduce temporal data into the system in Section 2:

> *" RF-Prox is a device proximity estimation system based on the wireless signal, which 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 the 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 a carefully 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. During the pre-training phase, a substantial dataset from the source domain $\mathcal{D}_{\mathrm{S}}$ is utilized to enhence the model's ability to generalize and extract domain-independent features. After pre-training, fine-tuning the model requires only a minimal dataset from the target domain $\mathcal{D}_{\mathrm{T}}$. This method enables RF-Prox to be efficiently adapted to new environments, facilitating its practical deployment in varied scenarios."*

c) We add a related work section in Section 6 to better articulate previous work in the wireless signal-based spatial perception series including localization and tracking, and to facilitate readers' understanding of the unique contributions of our work:

> *" This section offers an insightful summary of the research landscape surrounding our work.*
>
> ***Wireless-based Localization Techniques.*** *Cellular networks have support for a wide range of positioning methods. For outdoor scenarios, the Cell ID (CID) method 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) 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) 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. Indoor localization solutions exploit various channel attributes, such as Angle of Arrival (AoA), Time of Flight (ToF), and their fusion, aiming for meter-level accuracy. However, these approaches are prone to significant errors in Non-Line-of-Sight (NLoS) settings. Wireless fingerprinting techniques achieve finer accuracy by matching signal features against a pre-compiled database, but their adaptability is limited. 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.*
>
> ***Deep Learning in Wireless Sensing.*** *Deep learning architectures have been extensively applied across a variety of wireless sensing tasks, including gesture and gait recognition, respiration monitoring, fall detection, and tracking. Recent innovations have incorporated advanced deep learning concepts such as adversarial and meta-learning. Contrary to the conventional reliance on time-frequency spectrograms as inputs, 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."*

(2) 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.

a) We introduce the temporal data in Section 2 as described above.

b) We explain in detail how we integrated the transformer into our network in Section 3.5:

> *" 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. 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:*
>
> $$\mathbf{P}(pos, 2i) = \sin\left(pos/10000^{2i/d}\right),$$
> $$\mathbf{P}(pos, 2i+1) = \cos\left(pos/10000^{2i/d}\right),$$
>
> *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:*
>
> $$\mathbf{Q} = \mathbf{X}_{\mathrm{PE}}\mathbf{W}_{\mathrm{Q}},$$
> $$\mathbf{K} = \mathbf{X}_{\mathrm{PE}}\mathbf{W}_{\mathrm{K}},$$
> $$\mathbf{V} = \mathbf{X}_{\mathrm{PE}}\mathbf{W}_{\mathrm{V}},$$
>
> *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*
>
> $$\begin{aligned}\mathbf{X}_{\mathrm{A}} &= \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V},\end{aligned}$$
>
> *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*
>
> $$\mathbf{X}_{\mathrm{ST}} = \mathrm{LayerNorm}(\mathbf{X}_{\mathrm{A}} + \mathbf{X}_{\mathrm{PE}}),$$
>
> *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. "*

c) We modified the description of the MRSTE model in the contribution to emphasize the integration of the timing model in Section 1:

> *" 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. "*

(3) We have improved some of the modules to increase the performance of the models.

a) We replaced the previous ordered MLP-based feature mapping module with an unordered cosine similarity module in Section 5.3:

> *" 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.10, cosine similarity 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. "*

b) We extend the transfer learning of simulation-to-reality to a more pervasive transfer task under source and target domains in Section 4.2:

> *" 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, which greatly enhances 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 collected labeled CSI data under thousands of different deployment cases. Pre-training with a large amount of collected data from source domain helps the MRSTE module to 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*
>
> $$\{\Theta_\mathrm{M}, \Theta_\mathrm{P}\} = \underset{\{\Theta_\mathrm{M}, \Theta_\mathrm{P}\}}{\arg\min} \sum_{(\mathbf{H}, \mathbf{d}) \sim \mathcal{D}_\mathrm{S}} \frac{1}{|\mathcal{D}_\mathrm{S}|} L_\mathrm{P}(\mathrm{M_P}(\mathbf{H}; \Theta_\mathrm{M}, \Theta_\mathrm{P}), \mathbf{d}).$$
>
> ***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"*

"

$$\Theta_{\mathrm{P}} = \arg\min_{\Theta_{\mathrm{P}}} \sum_{(\mathbf{H},\mathbf{d})\sim\mathcal{D}_{\mathrm{T}}} \frac{1}{|\mathcal{D}_{\mathrm{T}}|} L_{\mathrm{P}}(\mathrm{M}_{\mathrm{F}}(\mathbf{H};\Theta_{\mathrm{M}},\Theta_{\mathrm{P}}),\mathbf{d}),$$

where 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 M_{P} *with high generalizability, and a fine-tuned model* M_{P} *which adapts to a specific target domain environment. "*

(4) To validate the above 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.

a) We tried different optimizers, learning rates, and learning rate decay strategies to tune the model for the best performance in Section 5.

b) We used the newly proposed MRSTE model to test the overall performance in 8 classification situations in the newly proposed cellular-based outdoor scenarios and the previously mentioned Wi-Fi-based indoor environments in Section 5.2:

"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. Post 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."

c) We tested the newly proposed MRSTE model's parameters such as parameter size and inference experiments to prove its ability to perform real-time inference in Section 5.2:

> *"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 below. Remarkably, the total number of model parameters is lower than that of typically employed small-scale models (e.g., ResNet-18), 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."*

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

d) We tested the performance improvement of the newly proposed model compared to the traditional state-of-the-art SpotFi and mD-Track in both indoor and outdoor scenarios in Section 5.2:

> *"4) 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. 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. 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."*

e) By conducting component studies on the newly proposed MRSTE and PMAN, we proved the excellent spatio-temporal feature extraction capability of the MRSTE module and the excellent performance of cosine similarity in PMAN in Section 5.3:

"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.9, RF-Prox demonstrates superior performance over both MRSE and Transformer across all metrics in both indoor and outdoor senarios. This underscores the efficacy of RF-Prox 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.10, cosine similarity 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."

f) we conduct a robustness analysis focusing on antenna number, SNR, and the volume of data used for fine-tuning, to assess their impact on system efficacy in Section 5.4:

"1) 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.11, 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.*

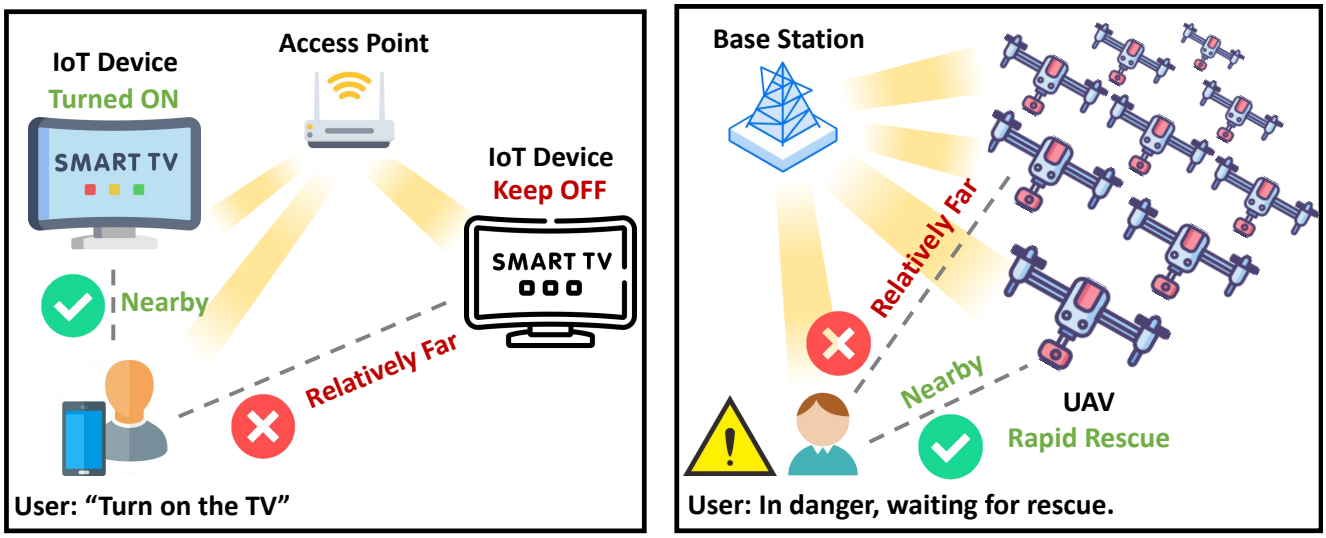
2) SNR: The SNR settings are established at $15, 20, 25, 30$ dB*, reflective of typical conditions in communication environments. As shown in Fig.13, the system maintains robust performance across various SNR levels in Wi-Fi-based indoor environments. For instance, at an SNR of 25* dB*, the system achieves an accuracy of 90.8±1.8% / 98.9±0.7% and an NDCG of 0.951±0.005 / 0.975±0.005 for the pre-trained and fine-tuned models, respectively. Conversely, performance in cellular-based outdoor environments degrades significantly at lower SNRs, attributed to the complex and dynamic nature of these settings, particularly when unmanned aerial vehicles (UAVs) are involved. Notably, fine-tuning enhances model resilience in lower SNR environments, evidencing the model's adaptability through domain transfer even under challenging conditions."*

3) Fine-tuning data volume: 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.12, 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.

g) To foster transparency and collaboration, we have made all related code publicly available[1].

REMARK: Please note that all of the figures and tables below are **NEW** and different from the original paper.

We believe that the above aspects significantly set this paper apart from the previous conference paper.



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

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

[1]Our project is available here.

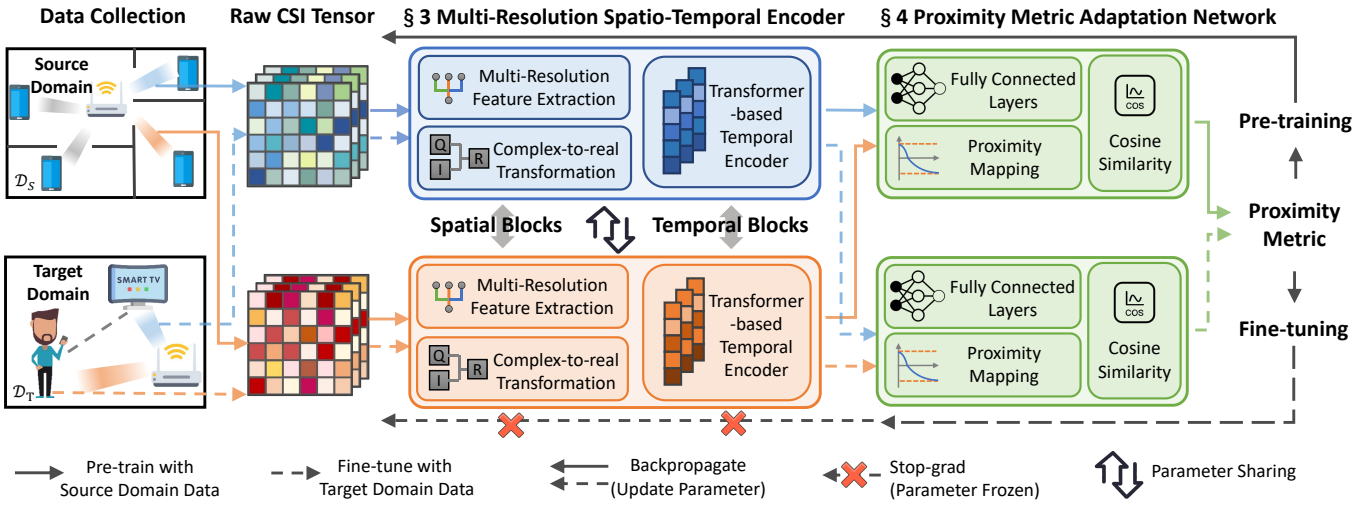


Fig. 2: An overview of the *RF-Prox*, where solid and dashed lines represent data collection from source domain and target domain, respectively, with blue and orange used to distinguish devices.

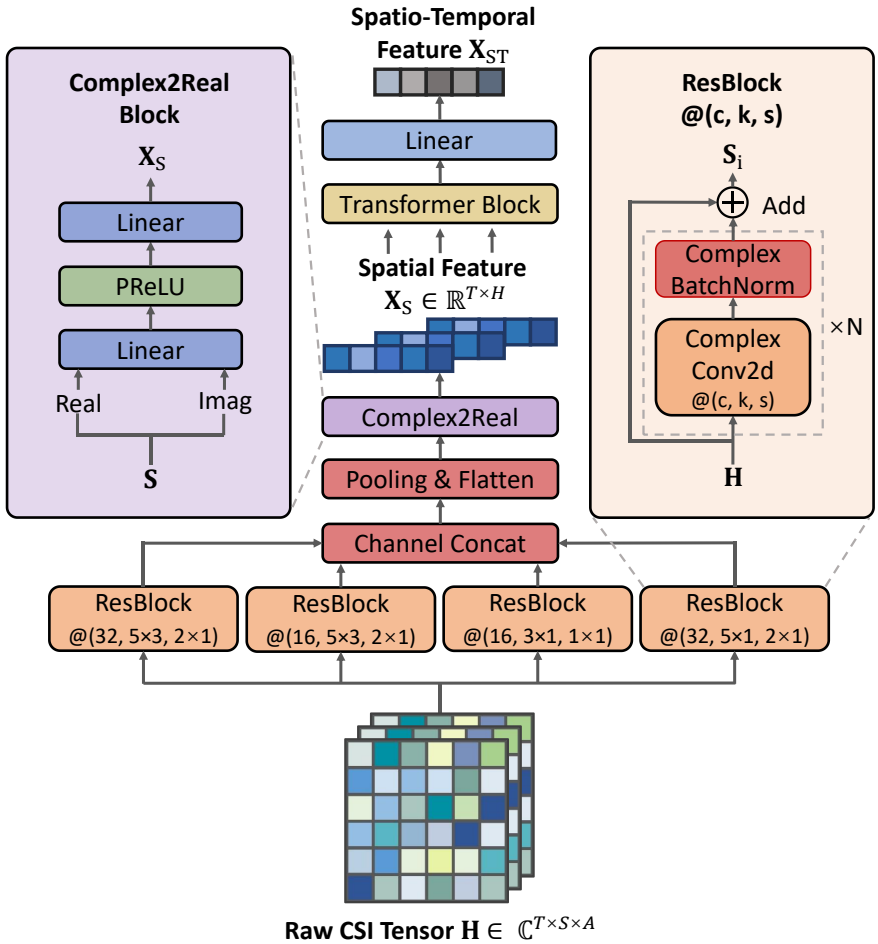


Fig. 3: Illustration of Multi-Resolution Spatio-Temporal Encoder.

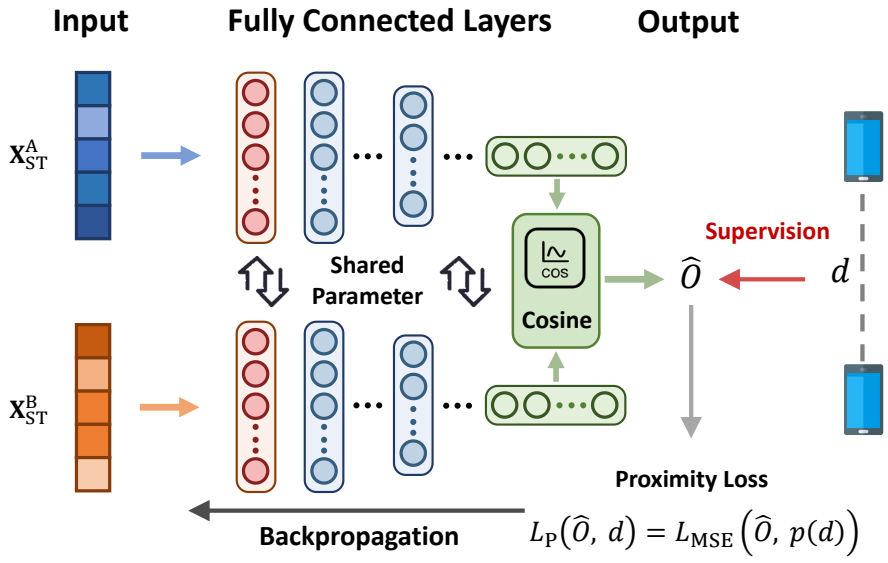


Fig. 4: An overview of the Proximity Metric Adaptation Network.

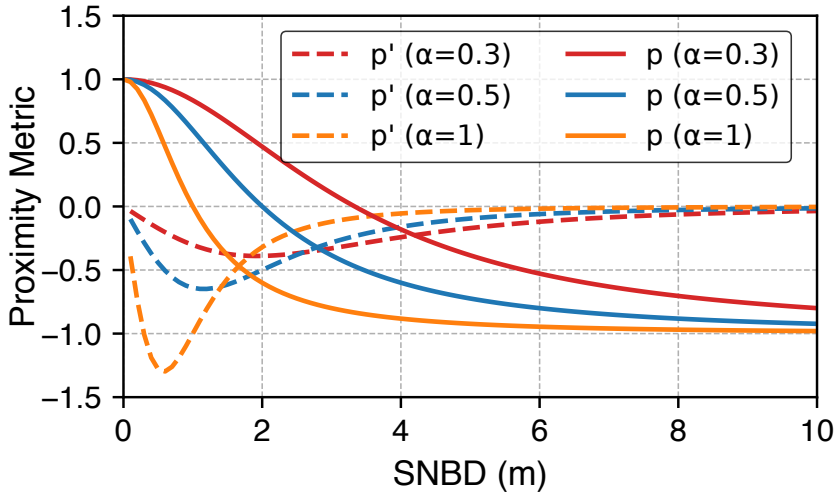


Fig. 5: Illustration of the proximity mapping function varied by α.

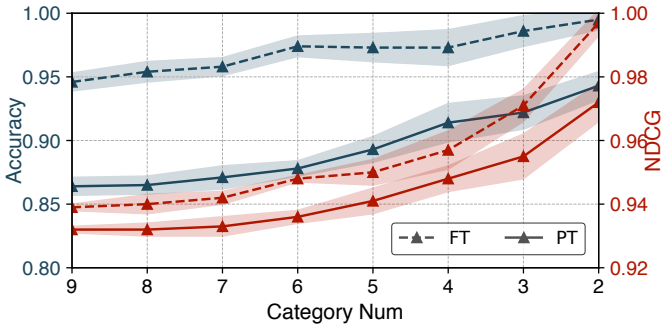


Fig. 6: Top-1 Accuracy and NDCG for Wi-Fi scenario

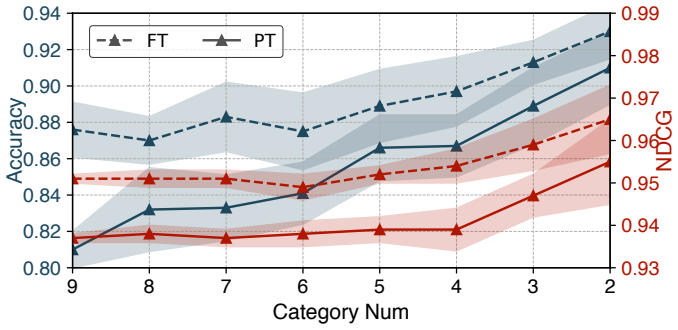


Fig. 7: Top-1 Accuracy and NDCG for cellular scenario

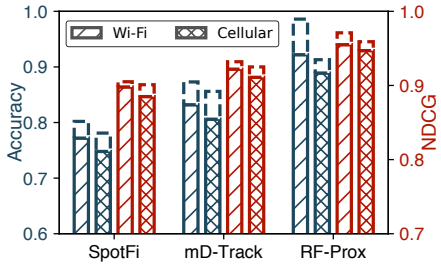


Fig. 8: Comparison with localization methods

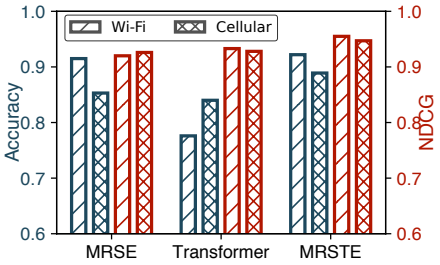


Fig. 9: Effectiveness of each component

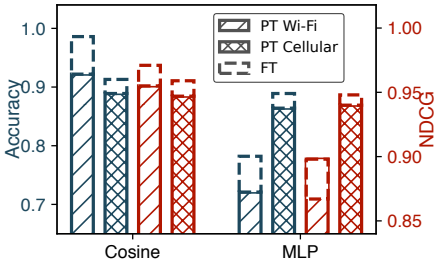
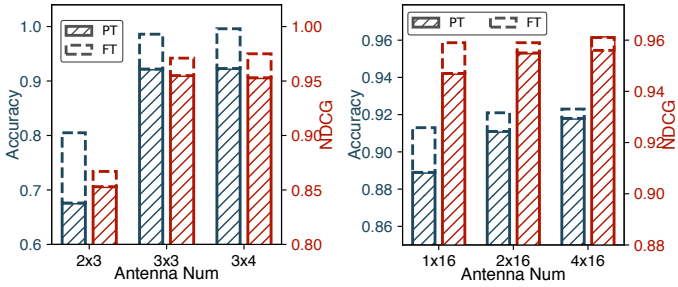
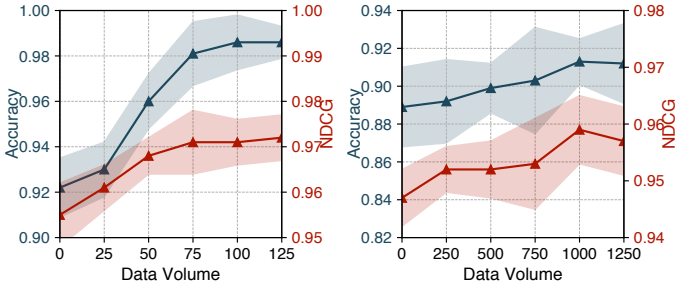


Fig. 10: Comparison of fusion methods



(a) Wi-Fi (b) Cellular

Fig. 11: Impact of antenna number



(a) Wi-Fi (b) Cellular

Fig. 12: Impact of fine-tune data volume

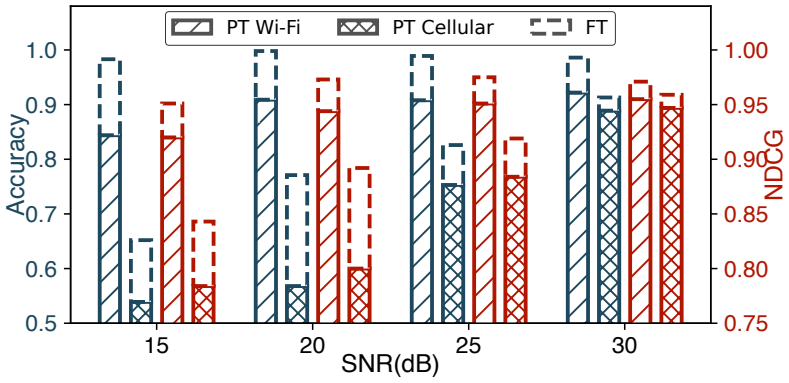


Fig. 13: Performance under different SNR