VIn-Loc: Towards Deep Learning-Based Environment Agnostic Localization

Rushabha Balaji[1*], **Aditya Sant**[2], **Akshay Malhotra**[2], **Keya Patani**[2], **Shahab Hamidi-Rad**[2]

[1]University of California Los Angeles
[2] InterDigital Communications, Inc

Abstract

User localization in wireless sensing and communication systems are essential for ensuring improved quality of service. To this end, the field of radio-frequency (RF) fingerprinting, extracting user location information from multiple transmit receive points (TRPs) using the wireless channel information between the user and different TRP, has been heavily studied and developed, both through model-based and deep neural network (DNN)-aided methods. However, conventional RF fingerprinting approaches face two major limitations *(i)* Dependence on fixed environment and TRP locations, and *(ii)* Relying on inputs from a fixed number of TRPs. Addressing these limitations, this work introduces the view independent localization (VIn-Loc) model - A transformer-based DNN framework that localizes users invariant to the number and location of the TRPs. This work presents the first step towards environment-agnostic user localization using both line-of-sight (LoS) and non-LoS channel information. This approach is rigorously validated on different statistical and ray tracing models, from wireless channels with only NLoS paths, to outdoor city blocks with both LOS and NLoS paths. Experimental results highlight the strength of the proposed model, VIn-Loc, over DNN-based RF fingerprinting.

1 Introduction

Today, the era of wireless IoT with high-speed interconnected devices and users (Fang et al. 2023) has brought in a need for high reliability communication and sensing systems. Accurate user localization in these systems is an important aspect and a crucial step to ensure high quality of service (QoS) between a user node and an interconnected network of transmit receive points (TRPs). As an example, beamforming in mmWave systems require accurate information about the user position for maintaining high data throughput (Ayach et al. 2014; Van Trees 2002). Building on traditional radar technologies, conventional algorithm-based methods for user localization include the time of arrival (ToA), and its variant the time difference of arrival (TDoA) (Shen, Zetik, and Thoma 2008), which require a minimum of 3 TRPs for localization. However, these classical methods require a line of sight (LoS) path between the user and each of the different TRPs, and fail to operate if even one of the LoS paths are blocked. Fig. 1 shows the difference between a LoS scenario, and a scenario where the LoS is blocked and only a non-line of sight (NLoS) path exists between the TRP and the user. Model-based approaches, extracting user location from NLoS wireless channels, require the use of more advanced methods and algorithms. To this end, deep learning based RF fingerprinting has gained immense research interest.

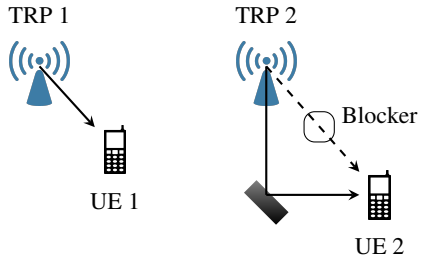

Figure 1: Two user equipments each connected to a different TRP. UE 1 has a LoS path to TRP 1 and UE 2 has its LoS path blocked and only has a non-LoS path to TRP 2.

Existing approach: DNN-aided RF fingerprinting

Contrary to the LoS channel using radar, for the NLoS wireless channel the mapping between the received signal at the user, and the user's location, is non-linear and intractable to compute. DNNs, with the ability for universal functional approximation (Hornik, Stinchcombe, and White 1989), especially in modeling highly nonlinear systems, present a promising solution to overcome the NLoS localization problem. Several deep learning methods have been developed to learn this mapping through data. For deep neural network (DNN)-aided radio-frequency (RF) fingerprinting, the model is trained to associate a wireless channel statistic with a particular location in the environment (Wang, Wang, and Mao 2018). The network is trained to encode the physical environment within the network parameters, in order to map the RF fingerprint or channel response to the user position. At inference, the model follows an approach analogous to a

*Corresponding author of this work rubalaji99@g.ucla.edu.
This work was completed by Rushabha Balaji as a part of the Summer Internship program at the InterDigital AI Lab in 2024.

look-up table, mapping the wireless channel input to a location associated with that statistic. Although this method is a very efficient way to learn an intractable relation between the RF input and user location, its applicability is limited to the given environment, as discussed next.

Limitations of DNN-aided RF fingerprinting: Existing approaches are specialized to a given environment. In particular, the DNN-aided methods assume the following

1. The layout of the physical environment is fixed; the location of the blockers and obstacles do not change
2. The number and position of the TRP nodes, relaying the channel information, remain fixed.

However, in dynamic settings—where either the environmental blockers frequently change and the environment itself shifts, or the TRP configurations are adjusted, the RF fingerprinting model may become ineffective. In such cases, it necessitates re-training with new data, often also requiring the redesign of the DNN architecture, to maintain accuracy. Though different networks have been proposed to circumvent this issue (Klus et al. 2024), these models still require re-training and data collection to operate successfully in the changed environment. Deep learning based methods beyond RF fingerprinting have been developed (Ayyalasomayajula et al. 2020), however the localization performance under changes to TRP configurations are not studied. This limits the applicability of these models in modern smart environments, like smart factories and warehouses, where the TRP nodes may be mobile and vary in number, and the physical environment can continuously change (Syberfeldt et al. 2016). These limitations are overcome through the contributions in this work, discussed below.

Our contribution: Beyond RF fingerprinting and towards environment agnostic localization

In this work, we propose a new methodology for user localization which discourages the model from memorizing the environment—the root cause for the lack of generalizability of RF fingerprinting methods. Herein, a different strategy is needed, one in which a model is trained to learn the functional mapping between the wireless channels collected from various TRPs and the UE location, independent of the specific environment in which the measurements were taken. This is a very challenging task since it requires the model to re-create the environment from the wireless channel measurements and extract the UE's location from it. As a first step in achieving this goal we propose a model which is view independent. Specifically, in a fixed environment, the model localizes the UE irrespective of the TRP locations. We posit that every environment-agnostic model should be view-independent, and in this work we propose an architecture that achieves this property. View-independent localization is addressed in (Wu et al. 2024), where the authors go beyond the wireless channel by incorporating multi-modal inputs from camera, audio, and mmWave radar to achieve localization in an indoor environment. However, this approach is not invariant to the number and location of the TRP nodes. Further, the availability of such multi-modal data is not always possible. To the best of the authors' knowledge this is the first work which addresses the following.

1. User localization using the limited data modality of wireless channel impulse response (either LoS or NLoS), and TRP locations
2. A universal model that exhibits the view-independent property for localization of a UE in a given environment irrespective of the number and position of TRPs

The rest of the paper is organized as follows: Section 2 introduces the localization problem and the wireless channel model. We introduce our deep learning architecture in Section 3. A brief overview of the datasets along with their results is presented in Section 4 and we conclude the paper in Section 5. We use lower case boldfont to represent vector $\mathbf{a}$ and upper case for matrix $\mathbf{A}$. The magnitude of a complex vector is denoted by $|.|$, and the convolutional operator is denoted by $\circledast$.

2 System Model

Consider the localization of a user equipment (UE) using N_{TRP} TRPs. We consider the downlink scenario where each TRP communicates with the UE. The received signals from all the TRPs are used by the UE to estimate its position. We assume that the UE is aware of the positions of all the TRPs, that is, the $\mathbf{p}_i = (x_i, y_i, z_i)$ coordinates of the i^{th} TRP is known at the UE. The UE has N_{u} receive antennas and each TRP is a single antenna transmitter. The received signal at the UE $\mathbf{y} \in \mathbb{C}^{N_{\text{u}} \times 1}$ is given as

$$\mathbf{y} = \sum_{i=1}^{N_{\text{TRP}}} \mathbf{h}_i \circledast x_i + \mathbf{n}. \tag{1}$$

where the transmit signal from the i^{th} TRP is denoted as x_i, $\mathbf{h}_i \in \mathbb{C}^{N_{\text{u}} \times D}$ is the wireless channel impulse response (CIR) between the i^{th} TRP and the UE, and D is the number of taps. Given this, the wireless channel model between the i^{th} TRP and the UE follows the continuous time tap-delay line model (Tse and Viswanath 2005) given by

$$\mathbf{h}_i(t) = \sum_{p=1}^{P^i} g_p^i \mathbf{a}(\theta_p^i)\delta(t - \tau_p^i), \tag{2}$$

where P^i is the number of paths between the i^{th} TRP and the UE, τ_p^i, g_p^i, θ_p^i are the delay, complex gain and the angle of arrival associated with the p^{th} path, respectively. The array manifold vectors is denoted by $\mathbf{a}(\theta) = [1, e^{j\pi\cos(\theta)}, e^{j2\pi\cos(\theta)}, \ldots, e^{j(N_t-1)\pi\cos(\theta)}]^T$. The noise vector $\mathbf{n}$ is sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$, with the noise variance σ^2 depending on the signal-to-noise ratio (SNR). The baseband sampled signal of (2) with a sampling rate of $2B$ Hz where B is the bandwidth of the signal is considered in (1) (Tse and Viswanath 2005). We assume that perfect CIR information is available at the UE, which can be obtained through the various channel estimation schemes developed to estimate $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M\}$ from the received signal (Tse and Viswanath 2005).

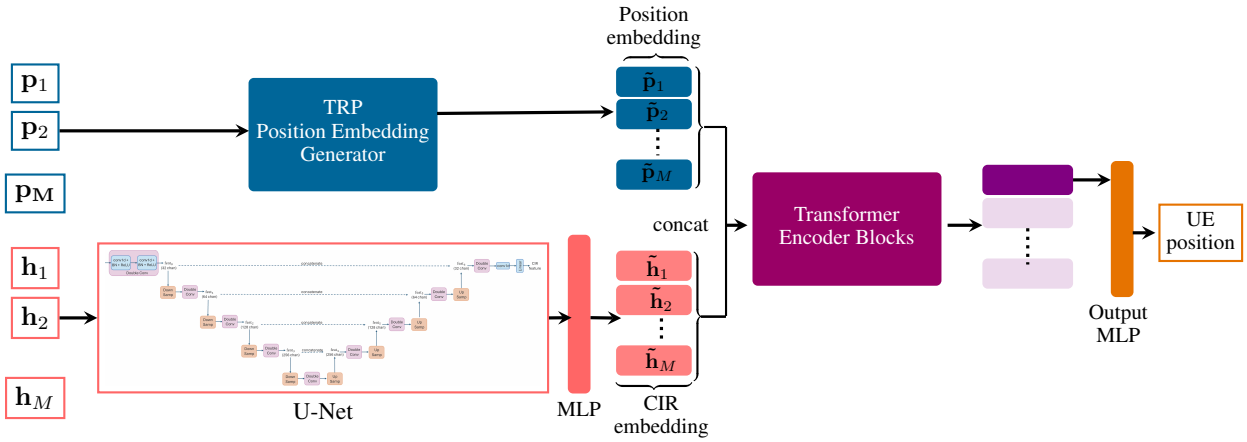

Figure 2: Network architecture of VIn-Loc which takes in the set of TRP locations $\{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$ and the corresponding wireless channel inputs $\{\mathbf{h}_1, \ldots, \mathbf{h}_M\}$ to output the UE's coordinates.

The UE location relative to each TRP is embedded in their respective channel estimates. In the LoS scenario, where a direct path exists between the UE and $M^{'}$ TRPs, with $3 \leq M^{'} \leq M$, the user's position can be triangulated by estimating $\{\tau_1^1, \ldots, \tau_1^{M'}\}$, and combining them with the known positions of the TRPs. This approach does not depend on the specific environment in which the TRPs and UE are located, making it environment-agnostic. However, accurately estimating the τ values from the baseband signal is challenging when the signal bandwidth is low, and the method fails to perform effectively in NLoS scenarios. In the next section, we aim to combine the powerful modeling capabilities of the deep learning model to localize users in the NLoS scenario with the view-independent property of the classical methods. We elucidate the details of our deep learning model which can localize users in a NLoS scenario under variable number and configuration of the TRPs. The following section describes the proposed approach in more detail.

3 VIn-Loc: View-Independent Localization

The deep learning model is trained to learn a mapping from the set of inputs $\{(\mathbf{h}_1, \mathbf{p}_1), (\mathbf{h}_2, \mathbf{p}_2), \ldots, (\mathbf{h}_M, \mathbf{p}_M)\}$ to the co-ordinates of the UE location $\mathbf{p}_\text{u} = (x_u, y_u, z_u)$. The architecture of our model is shown in Fig. 2. It can be broken down to into three main parts.

1. *Emebedding generation*: The position and wireless inputs are passed through their respective embedding generation blocks, each producing a set of N_TRP embeddings. Note that the wireless channel input to our model is in the complex field. We break each wireless input $\mathbf{h}_i$ into 3 feature channels, namely $\{\text{Real}(\mathbf{h}_i), \text{Imag}(\mathbf{h}_i), |\mathbf{h}_i|\}$ which is then fed to its embedding generation block.
2. *Embedding processor*: The positions embeddings of the TRP are combined with the wireless channel embeddings using a transformer encoder to capture the correlations that exist between the TRPs.
3. *Output regression*: The processed features from the transformer are passed through a set of multilayer perceptron (MLP) blocks to regress the output UE co-ordinates.

Additional details about each of the different blocks are given in the following sub-sections.

Part 1 - Embedding generation

In order to extract the UE location from the TRP information, the latter is converted into an appropriate embedding to be processed by the network. The embeddings are separately created for the TRP channel responses and the TRP locations, as explained below in detail.

UNet-Based CIR Embedding: We use the UNet architecture for generating the CIR embeddings (Ronneberger, Fischer, and Brox 2015). UNet is a convolutional neural network which allows us to mix features at different feature hierarchies. This property of a UNet allows us to extract wireless features at different timescales, from finer to coarser variations, hence creating a CIR embedding which is a richer representation of the input. Based on the number of antennaes at the UE we use two different convolutional filter sizes, for $N_\text{u} = 1$ we use a 1D filter and for $N_\text{u} > 1$ we use 2D filters. The output of the UNet is flattened and processed by an MLP with ReLU non-linearity to generate the embeddings. The above architecture is broadcasted across the N_TRP wireless channel inputs to obtain the set of CIR channel embedding $H \in \mathbb{R}^{M \times N_d}$, where N_d is the embedding dimension.

TRP Position Embedding: Inspired by the classical methods of ToA and TDoA, we posit that view-independent models should incorporate the position information of TRPs with the corresponding wireless channel inputs to accurately determine the UEs location. To generate the embeddings for the TRP positions we drew inspiration from previous works mainly involving neural radiance fields (Mildenhall et al. 2021). Based on empirically evaluating the performance of different types of position embeddings, we arrived at the following architecture. An embedding for each TRP coordinate

is generated by passing the co-ordinates through a 2 layer MLP with the ReLU non-linearity. A skip connection from the input to the ouput ensures that the input co-ordinates are concatenated with the output of the MLP to generate the final position embedding.

Part 2 - Embedding processing

The architecture obtains its invariance to the number of TRPs due to the emebedding processing block. The following section provides details on how this is achieved.

Transformer encoder: The encoder block of the transformer (Vaswani 2017) is used to process the correlations between the wireless channels from the different TRPs, whose values are conditioned on the position of the UE in the environment. In the traditional transformer model, the relative position information are encoded into the input tokens by using the position emebeddings. In our case, the position embedding is more interpretable since we use the global co-ordinates of the TRP to inform the CIR embedding inputs about their absolute position information. The position embedding and the wireless channel embedding are concatenated together to form the transformer tokens $F \in \mathbb{R}^{M \times 2N_d}$.

Location token to extract UE information: We append the input tokens F with an additional learnable token called as the `[LOC]` token. This special token represents our initial guess of the UE position in the joint embedding space, which is refined by the transformer through the contextual information presented in F. The `[LOC]` is very similar to the `[CLS]` token designed for the vision transformer (Dosovitskiy 2020) which captures the contextual information of image in terms of the class to which it belongs. Using this interpretation, we can think of each encoder block of the transformer as an iterative update to the position of the UE based on the TRP information in the high-dimensional embedding space.

Part 3 - Output regression

After N encoder blocks of the transformer, the output token corresponding to the `[LOC]` token represents the UE position in the joint embedding space. We use a series of MLP blocks with the ReLU non-linearity to transform this high dimensional position information into the three dimensional Cartesian co-ordinates of the position estimate of the UE in the given environment.

VIn-Loc: Network architecture and training

An overview of the different layers in the VIn-Loc architecture is given in Table 1, summarizing the DNN parameters shown in Fig. 2. The network is trained using the ℓ_2-norm loss between the predicted UE co-ordinates and the ground truth co-ordinates across 600 epochs. The AdamW (Loshchilov, Hutter et al. 2017) optimizer is used to train the network with weight decay of 0.1, momentum of 0.99, and a learning rate of 4×10^{-4}. We use a exponential decay for the learning rate with the decay factor set to 0.993 and the batch size is set to 16. Table 1 tabulates the detailed network sizes in each block. More details about the dataset are provided in Sec. 4.

Table 1: Network parameters for VIn-Loc

Network block	Layer	Parameters
Embedding generation	*Wireless channel embedding*	Input = $N_{\text{TRP}} \times 256 \times 3$
		UNet: Kernel size= $3(1) \times 3$ Pooling layers= 4 Output channels= 6
		MLP: Output dim = $N_{\text{TRP}} \times 256$
	TRP position embedding	Input = $N_{\text{TRP}} \times 3$
		Output = $N_{\text{TRP}} \times 256$
Embedding processing	*Transformer encoder*	Input = $(N_{\text{TRP}+1)} \times 512$
		Encoder blocks = 4 LayerNorm MLP=1024 neurons Heads = 4
		Output= $(N_{\text{TRP}} + 1) \times 512$
Output regression	*Output MLP*	Input = 512×1
		MLP: 4 layers Hidden layers= 256 neurons ReLU
		Output = 3×1

4 Results

This section presents the validation of the proposed VIn-Loc model across various scenarios, highlighting its performance and robustness in response to changes in the number of TRPs and their configurations, for both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) conditions. We begin by testing the model in a scenario currently being investigated by the 3rd generation partnership project (3GPP). Recently, 3GPP has initiated a study on AI/ML-based UE localization in environments with fixed TRP configurations, specifically under NLoS conditions, as part of a work item. RF fingerprinting methods are a promising solution to this problem, and our model is another strong candidate in this context. The following tests are used to benchmark and evaluate the performance of the VIn-Loc framework.

Static TRP network: We first benchmark our work against previously proposed RF fingerprinting models (3GPP 2023a) for the NLoS 3GPP channel scenario in the factory environment showcasing that our model can be used for the simpler case of fixed number and configuration of TRPs.

Variable number of TRPs: We then extend this to show that if the TRP configuration were changed then the previous RF fingerprinting methods fail but our model provides accurate localization.

The general structure of the channel, i.e., number of taps, is similar to the channel model given in eq. (2). Specifically, for each TRP the channel with $D = 256$ taps is considered. The following channel models have been used to benchmark the performance of the given networks, based on two different use cases of the localization problem.

1. Statistical channel models used by 3GPP
2. Ray traced channel models using Sionna

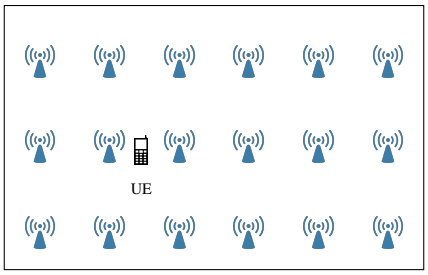

Figure 3: Illustration of the top-view of a 3GPP scenario consisting of 18 TRPs and the user located on the factory floor.

The specific structure of the channel models and the results for each are presented in the subsequent sub-sections.

Remark 1: Unless otherwise specified in the different experiments, it is assumed that the a single antenna system is used at the transmitted and receiver. The results for the multi-antennas receiver is given in Scenario 2 for the ray-traced channel model using Sionna.

3GPP channel models

The scenario of interest consists of a standard factory setting with 18 single antennae TRPs and a single antennae UE located on the factory floor (3GPP 2023b). Fig. 3 shows an example of the 3GPP scenario. It consists of only NLoS paths between each TRP and the UE. The NLoS positioning scenario presents a challenging case to localize the user, requiring the use of DNN-based localization frameworks. We benchmark our model against the ResNet-based model[1] proposed for RF fingerprinting(3GPP 2023a). The aim of this test is to showcase the performance of the view-agnostic VIn-Loc against the RF fingerprinting approach that will memorize the wireless environment.

Fixed TRP configuration The TRP configurations are fixed and the UE can be located anywhere at the ground level. The training set consists of 18K data samples and the test set consists of 2K samples. The results are shown in Table 2 under the fixed TRP configuration. We evaluate the models on three metrics namely, root mean square error (RMSE), mean absolute error (MAE) and 90^{th} percentile error. We see that our model has comparable performance to the ResNet based model. For a static TRP scenario, the VIn-Loc can also effectively memorize the environment, without being provided the information about the TRP positions. We present this version of the model that does not rely on the positions of the TRPs. The model is trained solely on the CIRs collected from different TRPs to predict the UEs coordinates. We see that due to the fixed configuration of the TRPs the location information does not affect localization performance of the VIn-Loc model. This is because our model can memorize the environment statistics and the positions of the TRPs, analogous to the ResNet-based RF fingerprinting approach. This validates the use of the VIn-Loc framework as an effective model for RF fingerprinting as well in an environment with static TRPs. However, the strength of this approach is portrayed in the variable TRP case, explained next.

Variable TRP configuration We test the view-independent property of the models in the 3GPP scenario, that is, of the 18 TRPs we subsample S TRPs and the wireless channel from each of the S TRPs to the UE is fed to VIn-Loc. Note that in the training phase, each training example is obtained from a different set of S TRPs. The subsampling method allows us to test the view-independent property of our model. This increases the size of our original dataset with static TRPs since each UE location is now associated with $\binom{M}{S}$ choices of the wireless channels. This increased diversity of the data makes the task of memorizing multiple such wireless environments, via conventional RF fingerprinting, extremely challenging. This is consistent with the performance results on the subsampled dataset in Table 2 for $S = 15$. We see that due to the many-to-one mapping present in the training dataset, RF fingeprinting methods which rely on a look-up table approach fail to localize the user. We also showcase the importance of the TRP positions to achieve the view-independent property by presenting results of VIn-Loc without the TRP positions. The reduced accuracy indicate that TRP position information is vital to achieve the view-independent property.

Ray-traced channels using Sionna

Sionna is a ray-tracing platform developed by NVIDIA which allows us to generate channels in new environments using the powerful computing capabilities of GPUs (Hoydis et al. 2022). We rendered a city block in Paris as our environment, with the top-view of the scenario shown in Fig. 6. We used Sionna to simulate the wireless channels by positioning the TRPs and UEs within the environment. We selected this particular environment because its dimensions of $280\text{m} \times 280\text{m}$ is large enough to effectively demonstrate the view-independent property of the model. The wireless system configuration of the UE and TRP are tabulated in Table 3. Given the size of the environment, further information had to be provided to maintain adequate localization accuracy, as explained below.

Use of time of arrival: To provide more information about the relative position of the UE with respect to the TRPs, we assumed additional global synchronization among the TRP nodes. This allows the model to capture the relative distances between them. Global synchronization (Lee et al. 2012) ensures a common time frame among all the TRPs, where the CIRs from the TRPs to the UE are shifted by the appropriate amount from a common reference point of $t = 0$.

Scenario 1: In this scenario the TRPs were restricted to being deployed on buildings that line the main street, and the UE was placed at street level. Fig. 4 shows an example of this simplified scenario where the shaded region shows the

[1]Model presented by InterDigital Communications, Inc. for the 3GPP standards

TRP Configuration	Metric	RF Fingerprinting (ResNet)	VIn-Loc	VIn-Loc (No TRP positions)
Fixed (18/18)	RMSE	0.81m	0.99m	0.98m
	MAE	0.76m	0.72m	0.73m
	90^{th} PE	1.99m	1.71m	1.62m
Variable (15/18)	RMSE	26.76m	0.95m	1.45m
	MAE	23.4m	0.82m	1.13m
	90^{th} PE	66.84m	1.68m	2.11m

Table 2: Performance of VIn-Loc compared to RF fingerprinting methods in the 3GPP scenario.

Parameters	Value
Center frequency	40 GHz
B	100 MHz
N_{u}	8

Table 3: Sionna configuration parameters

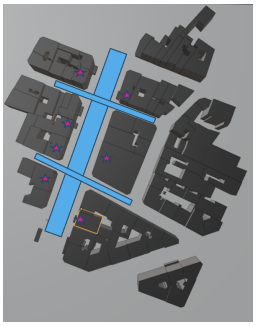

Figure 4: Scenario 1

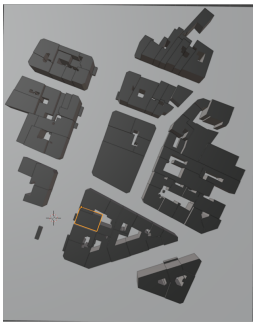

Figure 5: Scenario 2

Figure 6: Top-view of the environment loaded in Sionna to localize the users. The stars represent the position of the TRPs.

TRP Configuration	Metric	VIn-Loc
Scenario 1	RMSE	0.93m
	MAE	0.9m
	90^{th} PE	1.5m
Scenario 2	RMSE	7.3m
	MAE	6.1 m
	90^{th} PE	11m

Table 4: Performance of VIn-Loc on the Sionna dataset.

potential positions of the UE. We consider $N_{\text{TRP}} = 12$ TRPs, each positioned at random locations on top of the buildings, with their coordinates drawn from a uniform distribution to localize the UE along the street. Note that a maximum of N_{TRP} TRPs can have a path (LoS/NLoS) to the UE, however some TRPs can have no paths to the UE in such cases we have only $N^{'} < N_{\text{TRP}}$ wireless channel inputs. In such scenarios, the variable sequence modelling of a transformer helps us to localize the UE. The performance metrics of our model in scenario 1 is shown in Table 4. Due to the scale of the environment we conducted experiments in two scenarios to showcase the performance of VIN-Loc. Scenario 1 is a simpler case where only a portion of the environment is used to deploy TRPs and the UE. Scenario 2 is much more complex and the full environment is made available for deploying TRPs and UE. More information is provided next.

Scenario 2 In this scenario we consider the TRPs to be distributed uniformly above the city block and the UE is located on the street. The full scenario is shown in Fig. 5. Given the scale of the localization task, changes to the system configuration had to be made so that we achieve the required localization accuracy. First, to adequately cover the size of the city block we chose $N_{\text{TRP}} = 64$ single antenna TRPs. This number was chosen such that the UE has atleast a path (LoS/NLoS) to atleast 3 of the TRPs, irrespective of their positions in the environment. The single antenna UEs, which were assumed till now, were insufficient to meet the performance standards for localization. To provide additional information at the UE we considered a single input multiple output (SIMO) system with $N_{\text{u}} = 8$ uniform planar array with a spacing $\lambda_c/2$ where λ_c is the wavelength corresponding to the center frequency. This setup is summarized in Table 3. This allows the UE to exploit angular information along with the delay information for each path incident on the antennas. A 2D fast Fourier transform (FFT) is used to obtain a new representation of the channel in the angle of arrival (AoA) and time of flight (ToF) axes (Ayyalasomayajula et al. 2020). This representation is much richer as spatial correlation of the samples correspond to actual spatial information in the environment. Fig. 7 shows a two path wireless channel in the AoA-ToF plot. The results after introducing spatial information to VIn-Loc is shown in Table 4. Though there is a sharp increase in the error for scenario 2 compared to scenario 1, the scale of the problem has also substantially increased. The error tells us that irrespective of where we place the N_{TRP} TRPs in the environment, we get an average of 7 m error to localize the UE.

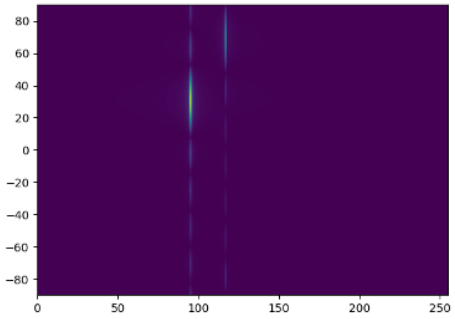

Figure 7: Angular domain of the channel

The performance results on the Sionna Scenario 2 goes on to illustrate the following additional takeaways for view-agnostic localization.

- Use of multi-antenna systems plays a crucial role to further improve localization performance in wireless channels with high variability in TRP positions. Angular domain representation of the can better extract the environment properties, proving unavoidable for larger and more complex wireless channels.
- Global synchronization plays a key role in better alignment of received CIR values from the different TRP nodes. In addition to the position information of the TRP nodes, this global synchronization plays a key role in inferring environment information to improve localization performance.

These results, to the best of the authors' knowledge, present the first application of the localization problem using variable number and configuration of the TRPs, in both indoor and outdoor environments. Further work in this area aims to better use these CIR parameters to fine tune the above results, and move in the direction of complete environment agnostic localization.

5 Conclusions

This work presents a novel model designed to address a challenge that extends beyond the current scope of RF fingerprinting-based methods by incorporating view-independence into the traditional localization problem. This property allows the model to be invariant to the number of TRPs as well as the configuration of the TRPs. We propose a transformer-based architecture called VIn-Loc that achieves the view-independence property by designing a special `[LOC]` token to captures the UE location information from the TRPs location as well as their CIR data. Experimental results highlight two key takeaways: first, the VIn-Loc framework can be utilized for RF finger-printing without any loss in performance, and second, it goes beyond RF fingerprinting to learn the mapping between the channel inputs and the UE position, becoming completely agnostic to the TRP configuration. This work presents a key step towards designing deep learning models that can achieve an environment-agnostic property, that is, learning a general mapping from wireless inputs to the UE location irrespective of the environment the wireless measurements were made.

References

3GPP. 2023a. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evaluation on AI/ML for positioning accuracy enhancement. Technical Report R1-2303450, 3GPP.

3GPP. 2023b. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz (Release 17). Technical Report TR 38.901 V17.1.0 (2023-12), 3GPP.

Ayach, O. E.; Rajagopal, S.; Abu-Surra, S.; Pi, Z.; and Heath, R. W. 2014. Spatially Sparse Precoding in Millimeter Wave MIMO Systems. *IEEE Transactions on Wireless Communications*, 13(3): 1499–1513.

Ayyalasomayajula, R.; Arun, A.; Wu, C.; Sharma, S.; Sethi, A. R.; Vasisht, D.; and Bharadia, D. 2020. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. New York, NY, USA: Association for Computing Machinery.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, X.; Feng, W.; Chen, Y.; Ge, N.; and Zhang, Y. 2023. Joint Communication and Sensing Toward 6G: Models and Potential of Using MIMO. *IEEE Internet of Things Journal*, 10(5): 4093–4116.

Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5): 359–366.

Hoydis, J.; Cammerer, S.; Aoudia, F. A.; Vem, A.; Binder, N.; Marcus, G.; and Keller, A. 2022. Sionna: An open-source library for next-generation physical layer research. *arXiv preprint arXiv:2203.11854*.

Klus, R.; Talvitie, J.; Domae, B.; Cabric, D.; and Valkama, M. 2024. Deep Hypernetwork-based Robust Localization in Millimeter-Wave Networks. In *2024 IEEE 35th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 1–7. IEEE.

Lee, D.; Seo, H.; Clerckx, B.; Hardouin, E.; Mazzarese, D.; Nagata, S.; and Sayana, K. 2012. Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges. *IEEE Communications Magazine*, 50(2): 148–155.

Loshchilov, I.; Hutter, F.; et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted*

intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, 234–241. Springer.

Shen, G.; Zetik, R.; and Thoma, R. S. 2008. Performance comparison of TOA and TDOA based location estimation algorithms in LOS environment. In *2008 5th Workshop on Positioning, Navigation and Communication*, 71–78.

Syberfeldt, A.; Ayani, M.; Holm, M.; Wang, L.; and Lindgren-Brewster, R. 2016. Localizing operators in the smart factory: A review of existing techniques and systems. In *2016 International Symposium on Flexible Automation (ISFA)*, 179–185.

Tse, D.; and Viswanath, P. 2005. *Fundamentals of wireless communication*. Cambridge university press.

Van Trees, H. L. 2002. *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, X.; Wang, X.; and Mao, S. 2018. RF Sensing in the Internet of Things: A General Deep Learning Framework. *IEEE Communications Magazine*, 56(9): 62–67.

Wu, J.; Wang, Z.; Ouyang, X.; Jeong, H. L.; Samplawski, C.; Kaplan, L.; Marlin, B.; and Srivastava, M. 2024. FlexLoc: Conditional Neural Networks for Zero-Shot Sensor Perspective Invariance in Object Localization with Distributed Multimodal Sensors. *arXiv preprint arXiv:2406.06796*.