Simplicity is Key: An Unsupervised Pretraining Approach for Sparse Radio Channels

Jonathan Ott [1] **Maximilian Stahlke** [1] **Tobias Feigl** [1] **Bjoern M. Eskofier** [2] **Christopher Mutschler** [1]

Abstract

We introduce the **Spa**rse pretrained **R**adio **Tran**sformer (SpaRTran), an unsupervised representation learning approach based on the concept of compressed sensing for radio channels. Our approach learns embeddings that focus on the physical properties of radio propagation, to create the optimal basis for fine-tuning on radio-based downstream tasks. SpaRTran uses a sparse gated autoencoder that induces a simplicity bias to the learned representations, resembling the sparse nature of radio propagation. For signal reconstruction, it learns a dictionary that holds atomic features, which increases flexibility across signal waveforms and spatiotemporal signal patterns.

Our experiments show that SpaRTran reduces errors by up to $85\,\%$ compared to state-of-the-art methods when fine-tuned on radio fingerprinting, a challenging downstream task. In addition, our method requires less pretraining effort and offers greater flexibility, as we train it solely on individual radio signals. SpaRTran serves as an excellent base model that can be fine-tuned for various radio-based downstream tasks, effectively reducing the cost for labeling. In addition, it is significantly more versatile than existing methods and demonstrates superior generalization.

1. Introduction

Wireless indoor localization is considered one of the most promising concepts for future position-aware applications. The key advantages include low deployment costs, the ability to cover large and complex areas with minimal infrastructure, and the simultaneous use of the system for communication purposes. Areas of application include health-care, industry, and emergency services (Laoudias et al., 2018). Conventional wireless localization methods, achieve centimeter-level accuracies utilizing signal properties such as the time-of-arrival (Gifford et al., 2022) or the direction-of-arrival (DOA) (Pang et al., 2020; Yen et al., 2022). These approaches operate under the assumption of line-of-sight (LoS) conditions, wherein the majority of the base-stations maintain direct links to the target. However, typical indoor deployments are characterized by significant signal blockage, non-line-of-sight (nLoS), due to objects and walls (O'Lone et al., 2022). Deep learning-based Fingerprinting (FP) approaches perform well in nLoS-dominated areas as they map channel state information (CSI) measurements to prerecorded position labels. CSI provides rich spatial information as it captures the reflection, scattering, and absorption of the signal, i.e., multipath components (MPCs), that are characterized by the environment (Niitsoo et al., 2019; Stahlke et al., 2022). While FP leverages the full potential of wireless localization, it incurs increased costs due to the complex and labor-intensive process of collecting position labels. Moreover, FP assumes that the environment is wide-sense static, i.e., the radio environment does not change considerably between training and inference phase. Hence, with every significant change in the environment updating the method is unavoidable involving the acquisition of labeled data (Stahlke et al., 2022; Widmaier et al., 2019).

Unsupervised learning has demonstrated significant improvements in domains such as natural-language processing (Devlin et al., 2019; Radford et al., 2018) and computer-vision (Grill et al., 2020; Caron et al., 2021; He et al., 2020; Chen et al., 2020), often requiring fewer labeled samples for fine-tuning. Existing methods are predominantly categorized as self-supervised learning (SSL). They generate supervisory signals directly from the input data, enabling models to learn meaningful representations without relying on labeled data. Self-supervision has also been applied to wireless positioning, achieving state-of-the-art accuracy with significantly less labeled data. Here, contrastive methods (Salihu et al., 2024; 2020) as well as generative methods (Ott et al., 2024) have been adapted to the radio domain. However, existing approaches still face three key challenges: First, it is essential for radio foundation models to be pretrained on single-channel measurements rather than full CSI data to ensure generalization across different system setups and environments. Second, the assumptions of prominent SSL methods may not align well with the characteristics of CSI. For instance, while it is reasonable in vision tasks to separate representations of different classes, the relationships between CSI measurements are more nu-

anced, as they vary gradually across space (Studer et al., 2018). Third, most methods operate in large, unconstrained solution spaces. In contrast, SpaRTran introduces inductive biases by embedding physical knowledge into both the model architecture and the training process, thereby improving training efficiency and the quality of the learned representations.

Rather than adapting existing SSL methods, we propose a novel, purely unsupervised method specifically tailored for CSI. Our method draws inspiration from the concept of sparse autoencoders (Lee et al., 2007) and compressive sensing (CS) (Donoho & Huo, 2001; Candes et al., 2006). The central premise of CS is that sufficiently sparse representations reduce ambiguity; in contrast, non-sparse representations typically contain numerous insignificant components, complicating both analysis and the signal recovery process (Donoho & Huo, 2001). We propose three contributions: First, we design a pretext task encouraging the model to represent signals using as few signal components as possible, corresponding to interactions with the radio environment. We hypothesize that this increases robustness against non-deterministic MPCs caused by diffuse signal scattering. Second, we propose a pretraining framework based on a transformer encoder, that maps the input signals into high-dimensional sparse vectors and that subsequently reconstructs the original signals using a learned dictionary. This implements the CS framework, effectively learning sparse signal representations while maintaining flexibility of the employed signal waveforms. Third, we employ a Gated Sparse Autoencoder (Rajamanoharan et al., 2024) to achieve highly sparse representations while preserving superior reconstruction fidelity. As the original Gated Autoencoder only supports real-valued data, we extend this concept by introducing a phase generator network. This network integrates complex phase information into the sparse signal coefficients, thereby enabling the representation of signal phases through complex-valued components while maintaining sparsity via the gating mechanism.

The paper is structured as follows. Section 2 reviews related work. Section 3 describes sparse channels and their application in FP. Section 4 details the SpaRTran processing pipeline and Section 5 outlines the experimental setup. Section 6 presents and discusses the results before Section 7 concludes.

2. Related Work

SpaRTran is an unsupervised representation learning method that integrates techniques from CS and dictionary learning frameworks. We will first survey current unsupervised pretraining methods for wireless positioning, followed by an overview of CS approaches, and conclude with a review of dictionary learning methods.

While supervised wireless positioning has been extensively studied (Salihu et al., 2022; Liu et al., 2022; Zhang et al., 2023), the paradigms of unsupervised learning and SSL have recently begun to attract attention in this context. Existing research is divided into two main directions: SSL (Salihu et al., 2024; Ott et al., 2024) and unsupervised channel charting (CC) (Studer et al., 2018). SSL has achieved significant success in domains such as computer vision (Grill et al., 2020; Chen et al., 2020; He et al., 2020; Caron et al., 2021) and natural language processing (Devlin et al., 2019; Radford et al., 2018). The training procedure typically consists of two stages: first, a pretraining step aims to learn general, reusable representations directly from unlabeled data; second, during fine-tuning, the model is retrained on labeled data to address a specific downstream task. Self-supervised wireless transformer (SWiT) (Salihu et al., 2024) is an example of joint embedding learning (Grill et al., 2020), employing channel augmentations to create multiple views of the same signal, which are then encoded closely in the representation space. This method enhances robustness against deficiencies inherent to radio systems. Ott et al. (Ott et al., 2024) propose a predictive pretraining method, in which masked portions of the signal are recovered to learn spatial correlations among signal components. CC employs dimensionality reduction techniques designed to preserve the local geometry of the radio environment (Studer et al., 2018). Semi-supervised localization methods based on CC have demonstrated significant potential for drastically reducing labeling efforts (Stahlke et al., 2023).

CS assumes that data can be explained by a small number of underlying factors. It uses a high-dimensional sparse vector defined in an over-determined basis to represent the given signals. CS (Donoho, 2006; Candes et al., 2006) garnered significant attention, particularly in the context of wireless source separation. Common applications of CS in wireless systems include DOA estimation (Yang et al., 2018) and channel estimation (Berger et al., 2010). Basis pursuit denoising (Chen et al., 2001) applies convex relaxation to reformulate the inherently non-convex sparse optimization problem into a convex one, enabling the use of classical convex optimization techniques. This method is closely related to the well-known least absolute shrinkage and selection operator (LASSO) method (Tibshirani, 1996). Greedy algorithms such as Orthogonal Matching Pursuit (OMP) (Tropp & Gilbert, 2007) iteratively select active basis components, i.e., the nonzero elements in the sparse representation. In contrast, sparse Bayesian learning methods (Malioutov et al., 2005; Stoica et al., 2011) impose a sparsity-inducing prior, promoting sparse solutions through a probabilistic framework. To the best of our knowledge, SpaRTran is the first to apply the CS to the design of unsupervised pretraining.

Dictionary learning algorithms identify atomic features that sparsely represent underlying data. This means, the dic-

tionary is learned empirically from the signals themselves. This enables generalization across signal types and often leads to increased sparsity (Elad, 2010). A prominent example is the K-SVD algorithm (Aharon et al., 2006), that iteratively updates the dictionary atoms. In recent years, deep learning-based approaches to CS have emerged to reduce computational complexity, improve reconstruction fidelity, and enable inference on severely downsampled signals (Machidon & Pejović, 2023). Sparse autoencoders (SAEs) enforce sparsity through a regularization constraint rather than by using a bottleneck layer. In contrast to conventional autoencoders, SAEs operate with high-dimensional latent spaces, which can lead to a decomposition into more interpretable features (Cunningham et al., 2023; Bricken et al., 2023). Inspired by gated linear units (Dauphin et al., 2017), Rajamanoharan et al. (Rajamanoharan et al., 2024) address low reconstruction accuracy resulting from biases introduced by the sparsity constraint by decoupling the selection of active components from the estimation of sparse coefficients. SparTRan learns a dictionary simultaneously to the radio channel representations. This improves flexibility across signal waveforms and spatiotemporal signal patterns.

3. Problem Description

This section describes the sparse channel model and the concept of CS, and outlines the challenging FP downstream task used to demonstrate the capabilities of the proposed representation learning methods.

3.1. Sparse Channel Model

During a radio-signal transmission, the electromagnetic wave interacts with the environment, i.e., the channel. This affects the signal, resulting in multiple propagation paths arriving at the receiver. The received signal $y(t)$ can be defined as

$$y(t) = h(t) * s(t) + w(t), \tag{1}$$

where $s(t)$ is the transmitted signal, $h(t)$ the channel and $w(t)$ additive white Gaussian noise and $*$ the convolution operator. The channel impulse response (CIR) $h(t)$ characterizes the radio transmission channel and can be modeled as

$$h(t) = \sum_{k=0}^{K-1} \alpha_k e^{-i\varphi_k} \delta(t - \tau_k), \tag{2}$$

where τ_k is the signal transmission delay, α_k the magnitude and φ_k the phase of the k-th propagation path of the transmitted signal. δ denotes the Dirac delta function and i the imaginary unit. Eq. 2 is the superposition of several signals, originating from K far field sources. In practice, we assume K to be unknown. The bandwidth-limited discrete channel measurement is modeled as

$$h[m] = \sum_{k=0}^{K-1} a_k \,\mathrm{sinc}[m - \tau_k W] + w_m, \tag{3}$$

where W is the bandwidth of the system, a_k is the complex valued path coefficients, and $m \in \{1, \cdots, M\}$. From this, we derive the sparse channel representation. Assuming a set of L potential signals $\boldsymbol{\psi}_l \in \mathbb{R}^M$ that form a basis, of which only $K \ll L$ effectively contribute to the received signal, we can rewrite Eq. 3 as

$$\boldsymbol{h} = \sum_{l=0}^{L-1} a_l \boldsymbol{\psi}_l + \boldsymbol{w}, \tag{4}$$

where $|\alpha_l| > 0$ if the l-th signal is an active signal component, and $|\alpha_l| = 0$ otherwise. Note that we have replaced the sinc-function with a more generic notation $\boldsymbol{\psi}_l$. By defining the dictionary $\boldsymbol{\Psi} := [\boldsymbol{\psi}_0, \cdots, \boldsymbol{\psi}_{L-1}]$, (4) can be expressed more concise in matrix notation as

$$\boldsymbol{h} = \boldsymbol{\Psi}\boldsymbol{a} + \boldsymbol{w}, \tag{5}$$

where $\boldsymbol{a} = [\alpha_0, \cdots, \alpha_{L-1}]^T$ is the sparse coefficient vector, and $\boldsymbol{\Psi}$ is a $M \times L$ dictionary matrix. Eq. 5 describes an underdetermined system of equations. As there is no unique solution, recovering the sparse channel requires solving the following optimization problem:

$$\min \|\boldsymbol{a}\|_0, \quad \text{s.t. } \|\boldsymbol{\Psi}\boldsymbol{a} - \boldsymbol{h}\|_2 \leq \epsilon, \tag{6}$$

where ϵ denotes the allowed reconstruction error due to noise. Eqs. (5) and (6) together describe the radio channel within the framework of compressed sensing (Donoho, 2006; Candes et al., 2006).

3.2. Radio Fingerprinting

To evaluate the capabilities of our pretrained feature extractor, SpaRTran, we choose localization via radio FP, as FP leverages the full complexity of the CSI, making it a particularly challenging downstream task.

We consider a setup with radio links between a single agent to N_r receivers. This general formulation accommodates various system configurations, including antenna arrays. To construct a dataset for training, an agent moved through the environment, and channel measurements are collected at different positions $\boldsymbol{p}_{ag}^{(i)}$. At each position, an instantaneous CIR is recorded for every link. The CIRs at the i-th position $\hat{\boldsymbol{h}}_n^{(i)}$ are aggregated in the channel state $\boldsymbol{H}^{(i)}$:

$$\boldsymbol{H}^{(i)} = \left[\hat{\boldsymbol{h}}_1^{(i)} \cdots \hat{\boldsymbol{h}}_{N_r}^{(i)}\right]. \tag{7}$$

Here, we refer to the matrix $\boldsymbol{H}^{(i)}$ as the CSI. Due to multipath propagation, the CSI is typically unique for each

position in the environment. Assuming the environment is wide-sense static, i.e., the radio environment does not change significantly between training and inference phase, it becomes feasible to train a neural network to map the fingerprint to its corresponding position for localization (Nitsoo et al., 2019; Stahlke et al., 2022).

4. Methodology

This section outlines the steps of SpaRTran's pretraining pipeline. First, we describe the preprocessing procedure, followed by the representation module, including the tokenization of the input signal. Next, we detail the sparse reconstruction head, which generates the sparse channel coefficients. Finally, we present how the learned dictionary is used to reconstruct the signal from the sparse representation.

In general, we consider a set of unlabeled channel measurements $\mathcal{H}$ comprising N recordings. Our objective is to learn channel representations that encode the environmental characteristics of the radio signal in a way that enables effective use in downstream tasks such as FP. We hypothesize that optimal representations should be as simple as possible, i.e., sparse - while preserving all essential information. To this end, we introduce a strong sparsity bias into the training process through both model architecture and loss function design. Our approach employs an encoder that generates a latent representation $\boldsymbol{z}$, and a decoder that reconstructs the input signal $\hat{\boldsymbol{h}} \sim \mathcal{H}$ based on $\boldsymbol{z}$.

4.1. Preprocessing

We use a global standardization factor for $\mathcal{H}$ that scales the values of $\hat{\boldsymbol{h}}$ such that the sum of the absolute magnitudes have a standard deviation equal to 1:

$$\tilde{\boldsymbol{h}}^{(q)} = \frac{\hat{\boldsymbol{h}}^{(q)}}{\sigma(\sum_{j=1}^{N} \|\hat{\boldsymbol{h}}^{(j)}\|_1)} \; \forall \, q \in \{0, \cdots, N\}, \tag{8}$$

where σ denotes standard deviation. This standardization scheme offers two main advantages: First, it reduces sensitivity to outliers caused by large signal peaks, which can occur at short distances between transmitter and receiver (as signal strength increases quadratically with decreasing distance). Second, it preserves the relative signal strength differences within the channel measurements, a crucial property for the downstream task of wireless localization, which relies on the spatial correlation of the signal.

4.2. Representation Module

A Transformer architecture (Vaswani et al., 2017) forms the backbone of the representation module. Similar to (Salihu et al., 2024) we employ a lightweight, encoder-only model featuring a single encoder block, with a latent dimension of $N_{latent} = 512$. The feed-forward network within the encoder uses a hidden size of $N_{hidden} = 1024$, and the multi-head attention mechanism comprises 8 attention heads.

We construct the input embedding $\boldsymbol{e}$ to the representation module from the complex valued CIR $\tilde{\boldsymbol{h}}$ by considering each time step of the CIR as an input token. We represent the complex values at the m-th timestep $\tilde{h}_m$ as a three dimensional vector consisting of the real and imaginary parts and the magnitude of the complex number:

$$\boldsymbol{e}_m = [\mathrm{Re}(\tilde{h}_m), \mathrm{Im}(\tilde{h}_m), \mathrm{Abs}(\tilde{h}_m)]^T. \tag{9}$$

To match the internal dimensionality of the Transformer encoder, each input token is projected into the latent space of dimension N_{latent} via a learned linear transformation. The Transformer encoder is inherently permutation equivariant (Vaswani et al., 2017) but modeling channel measurements requires a notion of temporal or spatial sequence. To address this, we apply a learned positional encoding after the linear projection to inject order information and counteract the permutation invariance of the Transformer block.

4.3. Sparse Reconstruction Head

The sparse reconstruction head consists of a gating mechanism, inspired by Rajamanoharan et al. (2024), as well as a phase generator. Former promotes the reconstruction to be sparse while latter converts the real numbered output of the neural network to the complex valued coefficients $\hat{\boldsymbol{a}}$. $\hat{\boldsymbol{a}}$ represents the reconstructed signal in terms of a learned overdetermined dictionary $\boldsymbol{\Psi}$, see Eq. 5. Fig. 1 shows the gating mechanism, the phase generator , and the dictionary (yellow, green, and purple color).

We now discuss the gating mechanism in more detail. Approximating of the l_0-norm with the l_1-norm tends to lead to a non-optimal reconstruction. This is due to the fact that the sparsity penalty, i.e., the l_1-norm, can be reduced at the cost of reconstruction performance (Wright & Sharkey, 2024). Hence, our strategy for the estimation of $\hat{\boldsymbol{x}}$ follows the work of Rajamanoharan et al. (2024). The idea is to separately handle the selection of active atoms from the dictionary ($\boldsymbol{f}_{\text{gate}}$) and the estimation of the coefficients magnitude ($\boldsymbol{f}_{\text{coeff}}$). The encoder output is defined by

$$\hat{\boldsymbol{x}} = \boldsymbol{f}_{\text{coeff}}(\boldsymbol{z}) \odot \mathbf{1}(\underbrace{\boldsymbol{f}_{\text{gate}}(\boldsymbol{z})}_{\boldsymbol{\rho}_{\text{gate}}}), \tag{10}$$

where $\mathbf{1}$ denotes the Heaviside step function, $\odot$ the Hadamard product and $\boldsymbol{\rho}_{\text{gate}}$ is the output of the gating stage before the binarization step. Fig. 1 shows the gating mechanism (yellow color). Due to the binarization of the gating values, no gradient flows through this path of the network, see grey arrows in the yellow box of Fig. 1. Thus, an auxiliary loss promotes the detection of active atoms in $\boldsymbol{f}_{\text{gate}}$. The

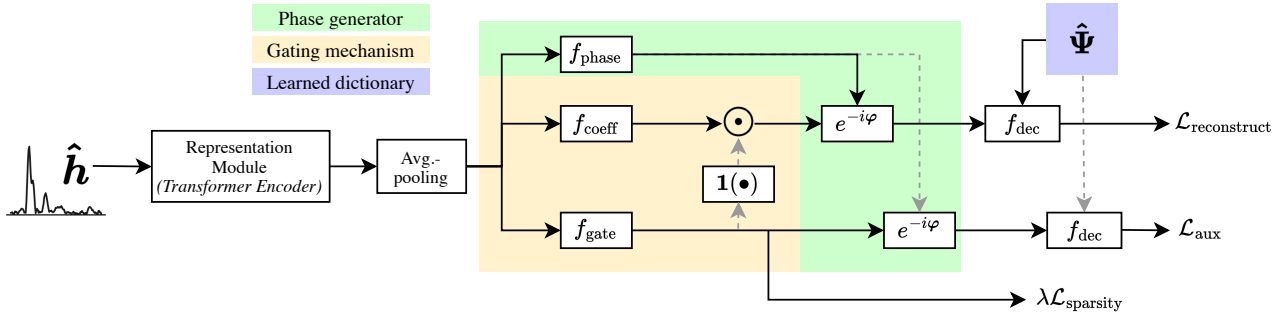

Figure 1. Overview of our unsupervised pretraining method - SparTRan.

auxiliary loss measures reconstruction fidelity, but instead of $\hat{\boldsymbol{x}}$, it uses $\boldsymbol{\rho}_{\text{gate}}$ to reconstruct the signal. The dictionary should not be updated by the auxiliary reconstruction task. Hence, we prohibit the flow of the gradient accordingly (see grey dashed line in Fig. 1).

We now outline our extensions to the original method. Rajamanoharan et al. (2024) restrict the encoders output $\hat{\boldsymbol{x}}$ to real positive numbers. However, this assumption does not hold in our case, as our goal is to estimate complex-valued path coefficients $\hat{\boldsymbol{a}}$. To address this, we interpret the outputs of $\boldsymbol{f}_{\text{coeff}}$ and $\boldsymbol{f}_{\text{gate}}$ as the magnitudes of the complex coefficients. This formulation allows us to suppress negative values via the gating mechanism without violating the underlying physical channel model. In addition, we introduce a third function $\boldsymbol{f}_{\text{phase}}$, that generates the phases of the path coefficients. The final coefficients are then constructed as:

$$\hat{\boldsymbol{a}} = \hat{\boldsymbol{x}} e^{-i\boldsymbol{f}_{\text{phase}}(\boldsymbol{z})}, \text{ and} \tag{11}$$

$$\boldsymbol{\rho}'_{\text{gate}} = \boldsymbol{\rho}_{\text{gate}} e^{-i\boldsymbol{f}_{\text{phase}}(\boldsymbol{z})}, \tag{12}$$

where i denotes the imaginary unit. The output of $\boldsymbol{f}_{\text{phase}}$ is constrained to the interval $\pm\pi$ using a scaled tanh activation function. This leads to the following loss function:

$$\mathcal{L} := \underbrace{\|\tilde{\boldsymbol{h}} - \boldsymbol{f}_{\text{dec}}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Psi}})\|_2^2}_{\text{reconstruction loss}} + \underbrace{\lambda \|\mathbf{1}(\boldsymbol{\rho}_{\text{gate}})\|_1}_{\text{sparsity penalty}} + \underbrace{\|\tilde{\boldsymbol{h}} - \boldsymbol{f}_{\text{dec}}(\boldsymbol{\rho}'_{\text{gate}}, \hat{\boldsymbol{\Psi}}_{\text{frozen}})\|_2^2}_{\text{auxiliary loss}}. \tag{13}$$

To enforce non-negativity, Rajamanoharan et al. (2024) employ ReLU activations for $\boldsymbol{f}_{\text{gate}}$ and $\boldsymbol{f}_{\text{coeff}}$. We observed that this can lead to a situation where certain dictionary atoms are never activated, i.e., their associated coefficients remain zero, resulting in no gradient updates, a phenomenon akin to the dying ReLU problem. To mitigate this, we replace ReLU with leaky ReLU activations (slope 0.01), ensuring that gradients can still propagate even for inactive units.

4.4. Dictionary Learning

Instead of using a fixed dictionary that conforms to the theoretical channel model (see Section 3.1), we treat the dictionary as a learnable parameter $\hat{\boldsymbol{\Psi}}$, see. Fig. 1, purple box. This approach provides two key advantages. First, the model can learn more expressive atoms that capture complex interactions, such as clusters of MPCs, and can adapt to the diverse pulse shapes used in radio localization. Second, it increases the incoherence of the dictionary, thereby improving the ability to distinguish which atoms contribute to the current signal (Donoho & Huo, 2001). Each atom in the dictionary is normalized to unit norm, meaning it only determines the direction of the contribution, while $\hat{\boldsymbol{a}}$ provides the amplitude and phase of the complex-valued signal component. Accordingly, the decoder function used to reconstruct the signal (also used in the auxillary pathway see. Eq. 13) is defined as:

$$f_{dec}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\Psi}}) = \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{a}}. \tag{14}$$

5. Experimental Setup

This section first describes the datasets used for evaluation. It then introduces the baseline methods used for comparison, and a detailed description of the training setup. Next, it presents the downstream task of radio FP that we use to evaluate the performance of the pretrained methods.

5.1. Datasets

We use two publicly available datasets that differ in terms of propagation environments and system characteristics. Table 1a provides an overview of the key differences. Both datasets include millimeter-accurate position labels. The dataset from Bast et al. (2020) was collected in a small, controlled environment, and the dataset from Stahlke et al. (2023) was recorded in a larger, more complex and less predictable setting.

1) *KUL Dataset* (Bast et al., 2020): The dataset comprises four antenna configurations: distributed antennas (DIS-

Table 1. Experimental setup overview.

(a) Dataset parameter

Dataset	f_c [GHz]	W [MHz]	N_r	Area [m]
KUL	2.61	20	64	3×3
FH-IIS	3.7	100	6	40×30

(b) Comparison of transformer hyperparameter.

Method	N_{latent}	N_{hidden}	N_{heads}	N_{blocks}	#param
SwiT	384	384	1	1	4.0 M
Masking	512	1024	8	3	8.7 M
SpaRTran (ours)	512	1024	8	1	4.9 M

LoS), a uniform linear array (ULA-LoS), and a uniform rectangular array under both LoS (URA-LoS) and nLoS conditions (URA-nLoS). Each configuration contains 252,004 CSI samples with recording positions arranged in a grid-pattern with 5 mm distance. The relatively low bandwidth of 20 MHz results in substantial overlap of MPCs in the time-domain. Nevertheless, the use of phased arrays preserves spatial information in the phase of the CSI signals. We split the dataset randomly into 70 % for training, 10 % for validation, and 20 % for testing.

2) *FH-IIS Dataset* (Stahlke et al., 2023): This dataset contains CIR fingerprints collected using a 5G-FR1-compatible software-defined radio system (DL-PRS reference signal). We evaluated two scenarios: an industrial environment featuring tall metal shelves, and a narrow corridor with large walls that introduce signal blockages and complex multipath propagation. The CSI is captured along a random walking trajectory of a person at a sampling rate of 6.6 Hz. We consider two different system topologies, each comprising three base-stations distributed along the perimeter of the localization area. This results in four subsets: two for the industrial scenario (IND-1, IND-2), and two for the corridor scenario (COR-1, COR-2). To ensure generalization, the training, validation, and test splits are recorded along different trajectories within the same environment. The split sizes are as follows: Industrial scenario: training - 566,589; validation - 141,639; test - 593,022 samples. Corridor scenario: training - 553,750; validation - 138,437; test - 463,280 samples.

5.2. Baselines

We identified two SSL methods and one purely supervised method as relevant baselines for the downstream task of FP. All compared approaches use a Transformer neural network as their backbone. Table 1b summarizes the hyperparameters used by each method and compares them to our approach. We also report the total number of trainable parameters (#param) per method, that includes not only Transformer parameters but also additional components such as projection layers.

Salihu et al. (2024) propose a joint embedding-based approach (Grill et al., 2020) called self-supervised wireless transformer (SWiT), that learns representations by predicting the output of a target network using an online network, given two different augmented views of the same input signal. The target network parameters are updated using an exponential moving average of the online network's parameters. The views are generated using domain-specific data augmentations, designed to make the learned representations invariant to the applied perturbations. In total, Salihu et al. (Salihu et al., 2024) propose six augmentation strategies, which are stochastically sampled during training to diversify the views. The method operates on the frequency-domain representation of the channel measurements and employs a lightweight Transformer architecture with a single encoder layer. We refer to this method as "SWiT".

Ott et al. (2024) introduce a predictive pretext task for learning FP representations, in which masked portions of the input signal are reconstructed. During training, up to 50% of the input fingerprint is removed, forcing the model to learn spatiotemporal correlations between the MPCs. This method assumes a frequency-selective channel, i.e., a channel in which the coherence bandwidth is small compared to the signal bandwidth. We refer to this method as "Masking".

To assess the benefits of pretraining compared to supervised approaches, we also include a baseline method that is trained end-to-end in a supervised manner. wireless transformer (WiT) (Salihu et al., 2022) employs a compact Transformer model consisting of a single encoder block with single-head attention for FP. As SWiT, WiT employs the frequency-domain representation of the radio channel. In the following, we refer to this method as "WiT (Supervised)".

5.3. Training setup

For a fair comparison we train all methods for 500 epochs during the un-/self-supervised pretraining phase, using a batch size of 512. To optimize SpaRTran, we use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 0.0001, that warms up linearly over 50 epochs. We apply a weight decay of 0.01 to prevent overfitting. The number of learnable atoms in the dictionary is set to $L = 512$.

5.4. Radio Fingerprinting Finetuning

To evaluate radio FP performance, we train an FP head that utilizes the representations produced by the Transformer

backbone to perform positioning. The FP head consists of four fully connected layers with 1024, 1024, 256 and 1024 neurons, followed by a projection layer that outputs a 2D position estimate. SpaRTran learns representations for individual links between a transmitter and a receiver. To align this with the task of FP, we compute a representation for each available link and concatenate them to form a complete representation of the CSI, see Eq. 7. We evaluate the model's generalization capabilities across different environments and system configurations by fine-tuning it on different test setups, using $5\,000$ labeled samples. To assess the performance of the fine-tuned models, we compute the positioning error as the absolute distance between the predicted and ground truth positions. From these values, we derive two standard metrics: the mean absolute error (MAE) and the 90th percentile of the cumulative error (CE90).

6. Evaluation

This section compares SpaRTran against all baseline methods w.r.t. the FP accuracy. It first presents the results for the FH-IIS dataset, followed by those for the KUL dataset. Finally we evaluate the impact of varying sparsity.

6.1. FH-IIS dataset

Table 2 presents the MAE and CE90 for the FH-IIS dataset. Overall, SpaRTran achieves the highest positioning accuracy, with MAE $\leq 0.718\,\mathrm{m}$ and CE90 $\leq 1.299\,\mathrm{m}$ across all evaluated benchmarks. Compared to the method of Ott et al. (2024), SpaRTran achieves an average error reduction of approximately 20%. It also significantly outperforms the purely supervised WiT baseline, reducing MAE by up to $1.861\,\mathrm{m}$ and CE90 by up to $4\,\mathrm{m}$. SpaRTran consistently generalizes well across all tested system setups and environments, exhibiting only minor performance variations, typically within a few centimeters. Notably, SpaRTran achieves the best performance when pretraining is performed on the IN-1 scenario, that was recorded in a real-world industrial setting. We hypothesize that this is due to the greater diversity of signal components encountered during pretraining, as the industrial environment exhibits the highest variability in multipath characteristics. In contrast, the method proposed by Salihu et al. (2024) yields the lowest accuracy on this dataset, with MAE $\leq 2.5\,\mathrm{m}$, CE90 $\leq 5.215\,\mathrm{m}$, consistently outperformed by the purely supervised baseline. We attribute this discrepancy to the fact that the data augmentation strategies used in SWiT were originally developed for lower-bandwidth systems and configurations with large antenna arrays capable of resolving signal directionality, which are not present in this dataset.

6.2. KUL dataset

Table 3 presents the results of the fine-tuned FP models on the KUL dataset. SpaRTran consistently outperforms all baseline methods across all system setups, achieving MAE $\leq 0.680\,\mathrm{m}$ and CE90 $\leq 1.243\,\mathrm{m}$. While all baseline methods exhibit a marked decline in performance under challenging nLoS conditions compared to LoS scenarios, SpaRTran achieves the lowest error in nLoS environments across all evaluated setups. This highlights SpaRTran's superior ability to extract meaningful signal features beyond the dominant LoS path. Contrary to its performance on the FH-IIS dataset, the approach by Ott et al. (2024) performs worst on the KUL dataset (MAE $\leq 1.176\,\mathrm{m}$, CE90 $\leq 1.671\,\mathrm{m}$). We attribute this to the relatively narrow bandwidth in the KUL dataset, that reduces the distinctness of multipath components and thereby limits the effectiveness of the masking strategy used in their method.

In the nLoS case (URA-nLoS), both SWiT and SpaRtran outperform the supervised baseline. Notably, SpaRTran surpasses SWiT by an average of $82\,\%$ in MAE and $85\,\%$ in CE90 (MAE $\leq 0.027\,\mathrm{m}$, CE90 $\leq 0.051\,\mathrm{m}$), highlighting its generalization and robustness in complex environments.

In the LoS scenarios (DIS-LoS, ULA-LoS, URA-LoS), the performance gap between SpaRTran and the supervised approach is smaller. Still, SpaRTran achieves an average improvement of $36\,\%$ in MAE and $35\,\%$ in CE90 reaching MAE $\leq 0.077\,\mathrm{m}$ and CE90 $\leq 0.095\,\mathrm{m}$. The SSL method by (Salihu et al., 2024) is consistently outperformed by the supervised baseline in these simpler LoS settings, likely due to the reduced complexity of the task under such conditions.

6.3. Sparsity Penalty

SpaRTran employs a sparsity penalty in its loss function (see Eq. 13) to enforce sparse coefficient vectors. This penalty is controlled by the hyperparameter λ: a higher value of λ lead to fewer dictionary atoms being used for signal reconstruction, while $\lambda = 0$ disables the sparsity penalty entirely. A trade-off must be balanced between increased ambiguity in the signal representation at low λ values and reduced reconstruction fidelity at high λ. Table 4 presents the effect of varying λ on both the sparsity (measured as the average number of nonzero entries in $\hat{\boldsymbol{x}}$) and the FP accuracy, evaluated on the IN-1 subset of the FH-IIS dataset. It is noticeable that increased sparsity generally improves localization accuracy until it becomes too dominant, suppressing subtle signal components. The best positioning performance MAE$= 0.648$ and CE90 $= 1.185\,\mathrm{m}$ is achieved with $\lambda = 0.1$, corresponding to an average of 26.7 active atoms. Interestingly, the case $\lambda = 0$ shows an anomaly: even without any sparsity penalty, the coefficient vector remains relatively sparse. On average, fewer than half of the available dictionary atoms are activated

Table 2. FP performance across different system setups finetuned on 5 000 samples of the FH-IIS dataset (MAE and CE90 in meter).

Method	Pretrain-Set	COR-1		COR-2		IN-1		IN-2	
		MAE	CE90	MAE	CE90	MAE	CE90	MAE	CE90
Masking:	COR-1	0.834	1.518	0.892	1.614	0.842	1.532	0.888	1.572
	COR-2	0.833	1.513	0.794	1.431	0.989	1.621	0.764	1.380
	IN-1	0.905	1.621	0.889	1.617	0.895	1.633	0.874	1.562
	IN-2	0.888	1.611	0.858	1.544	0.872	1.583	0.810	1.449
SWiT:	COR-1	2.298	6.122	2.895	5.355	3.834	6.816	3.546	6.706
	COR-2	3.310	6.182	3.065	5.640	4.800	8.924	3.586	6.722
	IN-1	3.346	6.262	2.851	5.275	3.877	6.866	3.555	6.620
	IN-2	3.360	6.239	2.912	5.359	3.996	7.118	3.550	6.605
SpaRTran (Ours):	COR-1	0.687	1.247	0.715	1.297	0.698	1.271	0.670	1.231
	COR-2	0.718	1.296	0.714	1.299	0.693	1.253	0.674	1.239
	IN-1	**0.680**	**1.243**	**0.667**	**1.209**	**0.648**	**1.185**	**0.639**	**1.160**
	IN-2	0.699	1.256	0.697	1.285	0.699	1.272	0.683	1.250
WiT (Supervised):		2.063	4.094	1.934	3.840	2.167	4.540	2.500	5.215

Table 3. FP performance across different system setups finetuned on 5 000 samples of the KUL dataset (MAE and CE90 in meter).

Method	Pretrain-Set	DIS-LoS		ULA-LoS		URA-LoS		URA-nLoS	
		MAE	CE90	MAE	CE90	MAE	CE90	MAE	CE90
Masking:	DIS-LoS	0.093	0.158	0.065	0.118	0.071	0.129	1.176	1.626
	ULA-LoS	0.081	0.139	0.067	0.116	0.071	0.127	1.094	1.615
	URA-LoS	0.087	0.156	0.068	0.118	0.073	0.131	1.176	1.671
	URA-nLoS	0.072	0.128	0.067	0.120	0.073	0.139	1.153	1.627
SWiT:	DIS-LoS	0.071	0.139	0.069	0.138	0.057	0.119	0.154	0.324
	ULA-LoS	0.076	0.146	0.068	0.136	0.058	0.119	0.140	0.303
	URA-LoS	0.070	0.138	0.068	0.136	0.057	0.119	0.156	0.327
	URA-nLoS	0.077	0.148	0.068	0.137	0.059	0.119	0.152	0.326
SpaRTran (Ours):	DIS-LoS	0.039	0.075	0.045	0.092	0.043	0.087	0.027	0.051
	ULA-LoS	0.038	0.074	**0.042**	0.086	0.043	0.088	**0.025**	**0.046**
	URA-LoS	**0.037**	**0.073**	0.045	0.095	**0.041**	**0.085**	0.026	0.047
	URA-nLoS	0.040	0.077	0.045	0.093	0.044	0.090	0.026	**0.046**
WiT (Supervised):		0.044	0.085	0.043	**0.085**	0.044	0.088	0.213	0.426

Table 4. Effect of sparsity penalties on the number of nonzero entries in the coefficient vector $\hat{\boldsymbol{x}}$ and FP accuracy (IN-1 dataset).

λ	Avg. $\|\hat{\boldsymbol{x}}\|_0$	MAE [m]	CE90 [m]
1.0	11.5	0.827	1.519
0.1	26.7	**0.648**	**1.185**
0.01	122.8	0.712	1.330
0.001	205.8	0.721	1.334
0.0	174.6	0.679	1.252

(Avg. $\|\hat{\boldsymbol{x}}\|_0 = 174.6$), which is sparser than the solution for $\lambda = 0.001$. This setting yields the second-best accuracy (MAE $= 0.679\,\mathrm{m}$, CE90 $= 1.252\,\mathrm{m}$). A possible explanation for this behavior lies in the inductive biases of the model architecture. In particular, the Transformer backbone may inherently favor simpler—and therefore sparser—solutions, as also observed by Bhattamishra et al. (2023).

7. Conclusion

We presented SpaRTran, an unsupervised method for learning radio channel representations based on a sparse gated autoencoder that integrates a channel model inspired by compressed sensing. This design reflects the inherent sparsity of physical radio channels, resulting in more meaningful and efficient representations. Unlike existing methods, SpaRTran operates on individual radio links rather than full CSI matrices, significantly reducing data acquisition effort and decoupling the model from specific system configurations, making it well-suited for training large, generic foundation models. We conducted a comprehensive evaluation on the challenging downstream task of FP in a low-data regime (5 000 labeled samples). SpaRTran outperforms state-of-the-art methods, achieving up to 85 % reduction in positioning error. SpaRTran demonstrates strong generalization to previously unseen system topologies and input domain shifts, e.g., from LoS to nLoS conditions, highlighting its potential to extract rich, reusable features from radio channels.

References

Aharon, M., Elad, M., and Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006. ISSN 1941-0476. doi: 10.1109/TSP.2006.881199.

Bast, S. D., Guevara, A. P., and Pollin, S. CSI-based Positioning in Massive MIMO systems using Convolutional Neural Networks. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1–5, May 2020. doi: 10.1109/VTC2020-Spring48590.2020.9129126.

Berger, C. R., Wang, Z., Huang, J., and Zhou, S. Application of compressive sensing to sparse channel estimation. *IEEE Communications Magazine*, 48(11): 164–174, November 2010. ISSN 1558-1896. doi: 10.1109/MCOM.2010.5621984.

Bhattamishra, S., Patel, A., Kanade, V., and Blunsom, P. Simplicity Bias in Transformers and their Ability to Learn Sparse Boolean Functions. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5767–5791, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.317.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Candes, E., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006. ISSN 1557-9654. doi: 10.1109/TIT.2005.862083.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic Decomposition by Basis Pursuit. *SIAM Review*, 43(1): 129–159, January 2001. ISSN 0036-1445. doi: 10.1137/S003614450037906X.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, July 2020.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. Language Modeling with Gated Convolutional Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 933–941. PMLR, July 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Donoho, D. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.871582.

Donoho, D. and Huo, X. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, November 2001. ISSN 00189448. doi: 10.1109/18.959265.

Elad, M. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer New York, New York, NY, 2010. ISBN 978-1-4419-7010-7 978-1-4419-7011-4. doi: 10.1007/978-1-4419-7011-4.

Gifford, W. M., Dardari, D., and Win, M. Z. The Impact of Multipath Information on Time-of-Arrival Estimation. *IEEE Transactions on Signal Processing*, 70:31–46, 2022. ISSN 1941-0476. doi: 10.1109/TSP.2020.3038254.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Laoudias, C., Moreira, A. J. C., Kim, S., Lee, S., Wirola, L., and Fischione, C. A Survey of Enabling Technologies for Network Localization, Tracking, and Navigation. *IEEE Communications Surveys & Tutorials*, 20:3607–3644, 2018.

Lee, H., Ekanadham, C., and Ng, A. Sparse deep belief net model for visual area V2. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Liu, W., Jia, M., Deng, Z., and Qin, C. MHSA-EC: An Indoor Localization Algorithm Fusing the Multi-Head Self-Attention Mechanism and Effective CSI. *Entropy*, 24(5):599, May 2022. ISSN 1099-4300. doi: 10.3390/e24050599.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization, January 2019.

Machidon, A. L. and Pejović, V. Deep learning for compressive sensing: A ubiquitous systems perspective. *Artificial Intelligence Review*, 56(4):3619–3658, April 2023. ISSN 1573-7462. doi: 10.1007/s10462-022-10259-5.

Malioutov, D., Cetin, M., and Willsky, A. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions on Signal Processing*, 53(8):3010–3022, August 2005. ISSN 1941-0476. doi: 10.1109/TSP.2005.850882.

Niitsoo, A., Edelhäußer, T., Eberlein, E., Hadaschik, N., and Mutschler, C. A deep learning approach to position estimation from channel impulse responses. *Sensors*, 19 (5):1064, 2019.

O'Lone, C. E., Dhillon, H. S., and Buehrer, R. M. Characterizing the First-Arriving Multipath Component in 5G Millimeter Wave Networks: TOA, AOA, and Non-Line-of-Sight Bias. *IEEE Transactions on Wireless Communications*, 21(3):1602–1620, March 2022. ISSN 1558-2248. doi: 10.1109/TWC.2021.3105641.

Ott, J., Pirkl, J., Stahlke, M., Feigl, T., and Mutschler, C. Radio Foundation Models: Pre-training Transformers for 5G-based Indoor Localization. In *2024 14th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–6, October 2024. doi: 10.1109/IPIN62893.2024.10786154.

Pang, F., Doğançay, K., Nguyen, N. H., and Zhang, Q. AOA Pseudolinear Target Motion Analysis in the Presence of Sensor Location Errors. *IEEE Transactions on Signal Processing*, 68:3385–3399, 2020. ISSN 1941-0476. doi: 10.1109/TSP.2020.2998896.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving Dictionary Learning with Gated Sparse Autoencoders, April 2024.

Salihu, A., Schwarz, S., Pikrakis, A., and Rupp, M. Low-dimensional representation learning for wireless CSI-based localisation. In *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 1–6. IEEE, 2020.

Salihu, A., Schwarz, S., and Rupp, M. Attention Aided CSI Wireless Localization. In *2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC)*, pp. 1–5, July 2022. doi: 10.1109/SPAWC51304.2022.9833994.

Salihu, A., Rupp, M., and Schwarz, S. Self-Supervised and Invariant Representations for Wireless Localization. *IEEE Transactions on Wireless Communications*, 23(8):8281–8296, August 2024. ISSN 1558-2248. doi: 10.1109/TWC.2023.3348203.

Stahlke, M., Feigl, T., García, M. H. C., Stirling-Gallacher, R. A., Seitz, J., and Mutschler, C. Transfer Learning to adapt 5G AI-based Fingerprint Localization across Environments. In *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, pp. 1–5, 2022. doi: 10.1109/VTC2022-Spring54318.2022.9860906.

Stahlke, M., Yammine, G., Feigl, T., Eskofier, B. M., and Mutschler, C. Indoor Localization with Robust Global Channel Charting: A Time-Distance-Based Approach. *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

Stoica, P., Babu, P., and Li, J. SPICE: A Sparse Covariance-Based Estimation Method for Array Processing. *IEEE Transactions on Signal Processing*, 59 (2):629–638, February 2011. ISSN 1941-0476. doi: 10.1109/TSP.2010.2090525.

Studer, C., Medjkouh, S., Gonultaş, E., Goldstein, T., and Tirkkonen, O. Channel charting: Locating users within the radio environment using channel state information. *IEEE Access*, 6:47682–47698, 2018.

Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Tropp, J. A. and Gilbert, A. C. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, December 2007. ISSN 1557-9654. doi: 10.1109/TIT.2007.909108.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Widmaier, M., Arnold, M., Dorner, S., Cammerer, S., and ten Brink, S. Towards Practical Indoor Positioning Based on Massive MIMO Systems. In *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, pp. 1–6, September 2019. doi: 10.1109/VTCFall.2019.8891273.

Wright, B. and Sharkey, L. Addressing Feature Suppression in SAEs, February 2024.

Yang, Z., Li, J., Stoica, P., and Xie, L. Chapter 11 - Sparse methods for direction-of-arrival estimation. In Chellappa, R. and Theodoridis, S. (eds.), *Academic Press Library in Signal Processing, Volume 7*, pp. 509–581. Academic Press, January 2018. ISBN 978-0-12-811887-0. doi: 10.1016/B978-0-12-811887-0.00011-0.

Yen, H.-C., Ou Yang, L.-Y., and Tsai, Z.-M. 3-D Indoor Localization and Identification Through RSSI-Based Angle of Arrival Estimation With Real Wi-Fi Signals. *IEEE Transactions on Microwave Theory and Techniques*, 70 (10):4511–4527, October 2022. ISSN 1557-9670. doi: 10.1109/TMTT.2022.3194563.

Zhang, B., Sifaou, H., and Li, G. Y. CSI-Fingerprinting Indoor Localization via Attention-Augmented Residual Convolutional Neural Network. *IEEE Transactions on Wireless Communications*, 22(8):5583–5597, August 2023. ISSN 1558-2248. doi: 10.1109/TWC.2023.3235449.