Unveiling the Threat: Data-Free Backdoor Attacks on Pre-Trained Models for RF Fingerprinting

Tianya Zhao, *Graduate Student Member, IEEE*, Junqing Zhang, *Senior Member, IEEE*, Jun Dai, *Member, IEEE*, Xiaoyan Sun, *Member, IEEE*, and Xuyu Wang, *Member, IEEE*

***Abstract*—While supervised deep neural networks (DNNs) have proven effective for device authentication via radio frequency (RF) fingerprinting, they are hindered by domain shift issues and the scarcity of labeled data. The success of large language models has led to increased interest in self-supervised pre-trained models (PTMs), which offer better generalization and do not require labeled datasets, potentially addressing the issues mentioned above. However, the inherent vulnerabilities of PTMs in RF fingerprinting remain insufficiently explored. In this paper, we unveil the potential threat by thoroughly investigating data-free backdoor attacks on such PTMs for RF fingerprinting, focusing on a practical scenario where attackers lack access to downstream data, label information, and training processes. To realize the backdoor attack, we carefully design a set of triggers and predefined output representations (PORs) for the PTMs. By mapping triggers and PORs through backdoor training, we can implant backdoor behaviors into the PTMs, thereby introducing vulnerabilities across different downstream RF fingerprinting tasks without requiring prior knowledge. Extensive experiments demonstrate the wide applicability of our proposed backdoor attack to various input domains, protocols, and PTMs. Furthermore, we explore potential detection and defense methods, illustrating the difficulty of fully safeguarding against our proposed data-free backdoor attack.**

***Index Terms*—Backdoor attack, pre-trained model, radio frequency fingerprinting, security.**

I. Introduction

THE proliferation of the Internet of Things (IoT) has led to the ubiquitous integration of wireless technology in daily life. As the number of wireless devices continues to grow, there is a critical need for effective and efficient device authentication methods [2], [3], [4]. Radio frequency (RF) fingerprinting has emerged as a promising technique, offering enhanced resistance to tampering and spoofing compared to conventional methods [5]. RF fingerprints are unique characteristics that arise from inherent physical imperfections in the analog circuitry of RF emitters, introduced during the manufacturing process [6], [7]. These subtle imperfections affect transmitted signals without compromising overall device functionality, resulting in a distinct fingerprint for each RF emitter, including ultra-low-power and legacy devices.

Deep neural networks (DNNs) have demonstrated remarkable capabilities in automatically extracting and classifying RF fingerprints [8], [9]. However, they face two significant challenges in RF fingerprinting applications: the need for large amounts of high-quality labeled data and vulnerability to domain shift. While previous studies have explored few-shot learning [10], [11] and domain adaptation techniques [12], [13] to mitigate these issues, these approaches have limitations and fail to fully leverage the abundant unlabeled data. The success of large language models (LLMs) such as GPT [14] and BERT [15] has sparked increased interest in self-supervised learning (SSL) across various domains, including RF fingerprinting [16], [17]. The SSL pipeline consists of two key components: pre-trained models (PTMs) and downstream classifiers. PTMs are trained on large amounts of unlabeled data to serve as feature extractors, while downstream classifiers are built on these PTMs using minimal or no labeled data. This approach enhances generalization and reduces the need for extensive labeled datasets, potentially addressing the data scarcity and domain shift challenges in RF fingerprinting.

Applying SSL techniques to train general PTMs for RF fingerprinting could potentially improve authentication performance by addressing the challenges posed by limited labeled data and domain shift. However, ensuring security remains a top priority for these systems. In the current deep learning landscape, PTMs are typically large, enabling them to capture extensive contextual information at the cost of being computationally expensive to train. To mitigate this burden, a common practice is to download open-source PTMs from platforms like GitHub and HuggingFace and then fine-tune them for specific tasks. While this approach is convenient and efficient, the widespread use of publicly available PTMs raises concerns about potential security vulnerabilities in RF fingerprinting.

One practical threat is *data poisoning-based backdoor attacks*, where an adversary seeks to manipulate the victim model to misbehave on inputs containing predefined triggers while

Received 20 April 2025; revised 17 October 2025; accepted 30 October 2025. Date of publication 3 November 2025; date of current version 6 March 2026. This work was supported in part by NSF under Grant CNS-2415209, Grant CNS-2321763, Grant CNS-2317190, Grant IIS-2306791, and Grant CNS-2319343. An earlier version of this paper was presented in part at the 2025 IEEE International Conference on Computer Communications (INFOCOM 2025), May 2025, London, U.K. [DOI: 10.1109/INFOCOM55648.2025.11044704]. Recommended for acceptance by Pu (Perry) Wang. *(Corresponding author: Xuyu Wang.)*

Tianya Zhao and Xuyu Wang are with the Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33172 USA (e-mail: tzhao010@fiu.edu; xuywang@fiu.edu).

Junqing Zhang is with the School of Computer Science and Informatics, University of Liverpool, L69 3DR Liverpool, U.K. (e-mail: junqing.zhang@liverpool.ac.uk).

Jun Dai and Xiaoyan Sun are with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: jdai@wpi.edu; xsun7@wpi.edu).

Digital Object Identifier 10.1109/TMC.2025.3628527

maintaining normal behavior on clean inputs [18]. Backdoor attacks have been extensively studied in supervised DNNs, and recent work has explored their impacts on unsupervised PTMs in computer vision (CV) and natural language processing (NLP) domains. For instance, BadEncoder [19] demonstrates that backdoors can be injected into image PTMs, leading downstream classifiers to inherit malicious behaviors. Shen et al. demonstrate backdoor attacks on PTMs by mapping triggers to predefined output representations in the NLP domain [20]. However, there is limited analysis of backdoor attacks on PTMs in the RF fingerprinting domain. Given that RF fingerprinting enables device identification and impacts the security of broader applications, it is crucial to investigate potential backdoor threats. Therefore, this paper focuses on studying *protocol-agnostic* and *data-free* [1] backdoor attacks on PTMs, aligning with the practical constraints of RF fingerprinting systems.

Challenges: Implementing backdoor attacks on PTMs in RF fingerprinting systems presents several significant challenges. First, the security-critical nature of RF fingerprinting systems prompts providers to implement robust protection for both PTMs and downstream training processes, significantly limiting an attacker's capabilities. Existing powerful backdoor attacks typically rely on manipulating the training process to obtain the gradient information for optimizing trigger patterns and mapping them to targeted classes [23]. However, in protected RF fingerprinting systems, attackers cannot control this process. Furthermore, most backdoor attacks on PTMs require access to downstream data and label information [19], [24], [25], which is highly sensitive and should be inaccessible to attackers in these systems. Therefore, the primary challenge lies in injecting backdoor behaviors into PTMs and impacting downstream classification without this crucial knowledge. Second, system providers may be cautious about using PTMs, even those from reputable open-source platforms. Therefore, they may incorporate proactive defense methods to cleanse potentially backdoored PTMs. For example, they may fine-tune several layers of PTMs using their own clean data to enhance security, without incurring significant computational costs. This creates an additional challenge of maintaining the effectiveness of backdoor attacks after the implementation of backdoor removal mechanisms. Third, any added trigger should not significantly impact the system's performance and should be resistant to detection methods. This poses a unique challenge for RF fingerprinting systems since input in-phase/quadrature (I/Q) data often undergoes signal processing, transforming it into the frequency or time-frequency domain. This requires the trigger to be effective and stealthy in both the time domain and the frequency domain.

Solution: To address the aforementioned challenges, we propose a practical backdoor attack for RF fingerprinting PTMs by retraining a benign PTM without controlling the downstream training process. First, we carefully design predefined output representations (PORs) of PTMs that serve as inputs for downstream classifiers. Then, we define a set of triggers and establish connections with the PORs, enabling the transfer of the backdoor to the downstream task. The backdoor will activate when any predefined trigger is injected into the input I/Q data. Given the security-critical nature of these systems, we implement this backdoor injection in a data-free manner. To achieve this, we use a small amount of unlabeled data to build a substitute dataset that differs from the downstream data, meeting the data-free condition. This substitute dataset can be collected by attackers or sourced online and may even be an out-of-distribution dataset.

The main contributions of this paper are as follows.

- To the best of our knowledge, this is the first work to investigate backdoor attacks on PTMs in RF fingerprinting. We develop a practical backdoor injection method without requiring access to downstream data.
- We propose a novel approach to generate output representations, enabling the successful implementation of protocol-agnostic backdoor attacks on PTMs.
- We conduct comprehensive experiments to evaluate our backdoor attack on various protocols (i.e., 802.11a/g and LoRa) with different PTMs on both time-domain and time-frequency domains across multiple datasets. These experiments show the broad applicability and effectiveness of our approach.
- We evaluate our backdoor attack against multiple defense strategies to demonstrate its robustness, and further analyze its performance across different device positions to highlight its effectiveness in practical scenarios.

The rest of the paper is organized as follows. Section II introduces background of our work and Section III discusses the related work. Section IV illustrates the attack scenario and threat model. Our proposed backdoor attacks are elaborated in Section V. Section VI presents the experimental evaluations and analysis. Finally, Section VII concludes this paper.

II. Background

A. RF Fingerprinting

The rapid expansion of IoT devices has underscored the urgent need for robust device authentication to secure IoT systems. Ensuring that only authorized users can access the network while blocking malicious users is a key priority. One effective approach to identifying wireless devices is RF fingerprinting, which takes advantage of the unique hardware imperfections inherent in each device. Essentially, an RF fingerprint arises from unique imperfections in analog components during the manufacturing process. As a physical-layer method, RF fingerprinting is resistant to spoofing and replay attacks, making it more difficult to spoof than IP or MAC addresses [26]. With the advent of powerful deep learning techniques, the automatic extraction of RF fingerprint features has become widely adopted for device identification across applications such as Wi-Fi [8] and LoRa [27].

In DNN-based RF fingerprinting systems, training typically relies solely on preamble data to prevent the DNN from learning protocol-specific patterns. Raw I/Q samples are commonly used as direct inputs to DNNs, though some methods first apply a Short-Time Fourier Transform (STFT) to convert I/Q data into the time-frequency domain before feeding it into the network.

[1]The term "data-free" is commonly used to define backdoor attacks that are conducted without access to training or testing data [21], [22].

Building on this foundation, this paper explores backdoor attacks targeting RF fingerprinting across diverse protocols and domains.

B. Self-Supervised Learning

Traditional supervised learning heavily relies on large volumes of labeled data, which can be costly and time-consuming to acquire. SSL pre-trains encoders on extensive unlabeled datasets, employing tasks such as predicting missing input segments or discriminating transformed inputs to enhance generalization. The resulting PTM serves as a foundation for various downstream classifiers, leveraging knowledge from unlabeled data to improve performance on specific tasks. This paper focuses on two mainstream SSL approaches: generative and contrastive methods [28]. Generative methods train an encoder f_θ to represent input data $\mathbf{x}$ as a discernible representation $f_\theta(\mathbf{x})$, paired with a decoder that reconstructs $\mathbf{x}$ from $f_\theta(\mathbf{x})$. In the NLP domain, the most popular generative model is the autoregressive model, such as the GPT series. On the other hand, contrastive methods train an encoder to transform augmented input $\mathbf{x}'$ into a vector representation $f_\theta(\mathbf{x}')$, enabling similarity measurements between inputs. A notable example is SimCLR [29], which aims to learn through comparisons using the NT-Xent loss as follows:

$$\mathcal{L} = -\frac{1}{K}\sum_{i=1}^{K}\frac{exp(sim(f_\theta(\mathbf{x}'_i), f_\theta(\mathbf{x}'_j))/\tau)}{\sum_{k=1,k\neq i}^{2K} exp(sim(f_\theta(\mathbf{x}'_i), f_\theta(\mathbf{x}'_k))/\tau)}, \quad (1)$$

where $sim(\cdot)$ denotes the similarity function, K is the batch size, and τ represents the temperature hyperparameter.

III. Related Work

A. RF Fingerprinting PTMs

Pre-training has become a mainstream technique across diverse domains, and recent works have also emphasized the significance of PTMs in RF fingerprinting. Zha et al. employ unsupervised contrastive learning to capture receiver-agnostic features, combined with subdomain adaptation to further enhance identification performance [30]. Chen et al. employ contrastive learning to extract domain-invariant features, demonstrating its effectiveness in mitigating domain-specific variations for robust RF fingerprinting [17]. Liu et al. introduce SSL during pre-training to address label dependence issues and utilize knowledge transfer in fine-tuning to overcome sample dependence limitations [16]. Similarly, Shao et al. apply SSL to improve specific emitter identification (SEI) performance through RF fingerprints [31]. For generative methods, Parpart et al. pre-train Transformer models as a masked autoencoder to reconstruct signals to improve device classification accuracy [32]. Zhao et al. propose a few-shot SEI using an asymmetric masked autoencoder with unlabeled samples in source domains [33]. Liu et al. pre-train a BERT model to obtain a powerful RF fingerprinting feature extractor to improve few-shot accuracy [34].

Overall, these studies demonstrate the promise of SSL in the RF fingerprinting task, making it imperative to investigate the security vulnerabilities of these methods.

B. Backdoor Attacks

Backdoor attacks represent a significant threat to machine learning models across various domains and applications. Our previous works have focused on designing and analyzing such attacks within specific contexts. For supervised learning models, we leverage explainable machine learning tools to design backdoor attacks on model-agnostic RF fingerprinting systems [35], [36]. We also examine vulnerabilities in 5G massive MIMO localization systems, covering both indoor and outdoor environments [37]. Furthermore, we extend backdoor attacks to few-shot learning, demonstrating their effectiveness in satellite fingerprinting [38].

In related domains, Zhao et al. design a training-based backdoor trigger generation approach on RF signal classification [39]. [40] proposes backdoor attacks on wireless traffic prediction in both centralized and distributed training scenarios. TrojanFlow [23] implements attacks on network traffic classification by simultaneously optimizing a trigger generator and the target model. For data-free backdoor attacks, Lv et al. customize a substitute dataset to fine-tune the benign model into a backdoored model [21]. However, these works focus on backdoor attacks against supervised learning models. As the field evolves toward self-supervised learning and foundation models, there is a growing need to investigate security implications and vulnerabilities specific to PTMs.

BadEncoder [19] first proposes backdoor attacks targeting image PTMs, followed by several concurrent studies in the same domain [24], [25]. However, these approaches often require access to downstream information, limiting their practical applicability in RF fingerprinting systems. The most closely related work is in the NLP domain, where they design output representations mapping to selected tokens for launching attacks [20]. Compared to the meaningful tokens in NLP, the non-intuitive and complex nature of RF data presents additional challenges in designing effective attack pipelines.

Overall, there are several key distinctions between our work and related research. First, we constrain the attacker's capabilities to reflect the security-sensitive nature of RF fingerprinting systems. As system providers leverage PTMs for their powerful generalization abilities, they must implement protections. Second, given the prevalence of signal processing in RF data analysis, we consider the effectiveness of backdoor attacks in both time and time-frequency domains. Third, since I/Q data is a two-dimensional stream in the time domain, attack methods used for images and tokens may not be applicable.

IV. Attack Scenario and Threat Model

A. Attack Scenario Description

The overall backdoor injection process is shown in Fig. 1. Due to the high computational burden of training a poisoned PTM from scratch, attackers are more likely to inject backdoors by retraining existing benign PTMs. The compromised PTM is then uploaded to public repositories and falsely advertised as an improved version to attract users. A potential victim might adopt this backdoored PTM if downstream classifiers built upon it

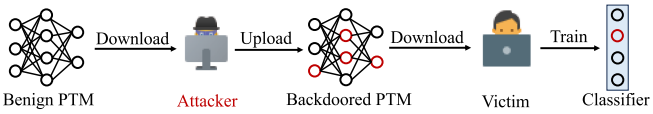


Fig. 1. Attack scenario: backdoor injection for PTMs.

demonstrate satisfactory performance in RF fingerprinting tasks. Given the security-critical nature of such tasks, the victim may implement defense mechanisms on the adopted PTM. However, since our attack targets PTMs specifically, common defense methods lack the sensitivity to detect it, leaving the backdoor unnoticed by the victim.

B. Threat Model

1) Attacker's Goal: We consider an attacker who aims to inject backdoors into a PTM f_θ in a data-free manner so that a downstream classifier g built on the backdoored PTM f_{θ_b} renders the RF fingerprinting system ineffective with attacker-chosen triggers $\mathbf{t}_j \in T$. The attacker has three goals to achieve:

- *Stealthiness goal:* The backdoored PTM must maintain its utility to remain stealthy. The attacker needs to ensure that downstream classifiers built on the compromised PTM still perform well on clean data $\mathbf{x}$, thus deceiving victims into adopting the backdoored model. Besides, triggers need to be concealed to evade detection methods.
- *Effectiveness goal:* When a downstream classifier is built on a backdoored PTM, it should misclassify any input containing a trigger. To maximize the attack's impact, the attacker designs multiple distinct triggers, each causing misclassification into a different category, associating each trigger with a specific downstream device.
- *Robustness goal:* Backdoored PTMs should achieve the above two goals, particularly maintaining effectiveness under potential defenses and protections.

In summary, the overall goals can be represented as:

$$g(f_{\theta_b}(\mathbf{x}^p)) \neq g(f_\theta(\mathbf{x})); \ \max(|g(f_{\theta_b}(\mathbf{x}^p))|); \tag{2}$$

$$g(f_\theta(\mathbf{x})) = g(f_{\theta_b}(\mathbf{x})), \tag{3}$$

where $\mathbf{x}^p = \mathbf{x} \oplus \mathbf{t}$ denotes poisoned samples with triggers and $max(|\cdot|)$ represents maximizing the number of output classes.

2) Attacker's Capability: We consider a scenario where an attacker obtains a clean PTM from a service provider, injects backdoors into it, and then shares the backdoored PTM with potential victims (e.g., by republishing it for public download). In this context, the attacker has access to the original clean PTM. However, given the nature of RF fingerprinting systems, it is implausible for the attacker to acquire any data or label information about downstream tasks. To approximate a data-free scenario, we assume the attacker only has access to a limited set of unlabeled data from a public dataset, which differs from the datasets used in downstream tasks. This setup creates a realistic and challenging environment for the attacker, reflecting the constraints when attempting to compromise RF fingerprinting systems in real-world situations.

V. Backdoor Methodology

A. Overview

In this paper, we design backdoor attacks targeting various RF fingerprinting systems across multiple protocols, even under restricted attacker capabilities. To achieve the goals mentioned above, our idea is to manipulate the PTM so that 1) it generates similar output representations for clean substitute data as it does with the benign PTM, and 2) it produces similar output representations for poisoned substitute data with the PORs. Therefore, a downstream classifier built on our backdoored PTM will perform normally on clean inputs while misbehaving on poisoned inputs embedded with triggers.

As shown in Fig. 2, our attack pipeline consists of three phases: substitute dataset collection, poisoned data generation, and output representation manipulation. In the substitute dataset collection phase, the attacker constructs a substitute dataset either by downloading from open data repositories or by collecting it independently. Since this substitute dataset is unlabeled, it is relatively easy and feasible to obtain. In the poisoned data generation stage, we first design a set of triggers $T = \{\mathbf{t}_j\}_{j=1}^{N_t}$ for the backdoor attacks. The substitute dataset D_s is then divided into two parts: a small portion designated as the poisoned dataset D_p and the remainder as the clean dataset D_c. Data in the poisoned dataset are embedded with the designed triggers. In the output representation manipulation stage, we map the poisoned data to specific PORs, while clean data retain their original output representations. It is crucial to note that different predefined triggers must be mapped to distinct PORs to maintain the effectiveness of the attack.

B. Backdoor Design

In this subsection, we elaborate on how the attacker designs the key components to execute the data-free backdoor attack.

1) Substitute Dataset: Due to the impracticality of obtaining downstream data and label information for RF fingerprinting systems, we have to construct a substitute dataset to implant backdoor behaviors. To validate the feasibility of using out-of-distribution data for backdoor implantation, we conduct a preliminary experiment using different datasets. Fig. 3 presents the t-SNE results of two distinct datasets: devices 0 to 2 belong to one dataset, while devices 3 to 5 belong to another. Fig. 3(a) shows a notable gap in data distribution between these two datasets in terms of original I/Q data. However, Fig. 3(b) shows that this gap becomes significantly narrower after processing through the PTM, where the extracted representations are distributed within a unified feature space. This observation suggests that out-of-distribution data can generate representations occupying a similar space to those of target data. Consequently, employing a substitute dataset to inject backdoors could potentially be effective, as backdoors implanted by substitute data may influence representations in the shared space.

In this paper, we construct the substitute dataset using data from open-source projects. To achieve the dual objectives of implanting backdoors and maintaining accuracy on clean samples, we divide the substitute dataset $D_s = \{\mathbf{x}_i\}_{i=1}^{S}$ into

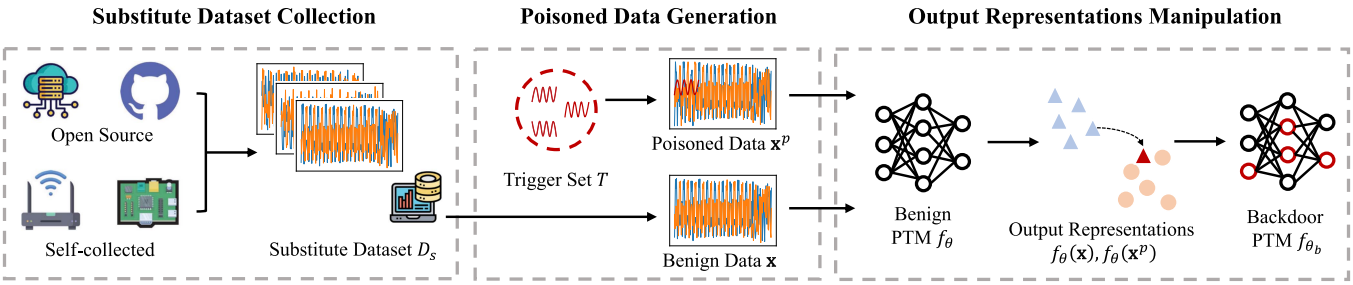
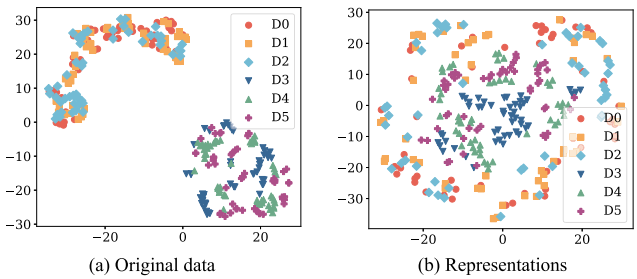


Fig. 2. Backdoor attack pipeline.



(a) Original data (b) Representations

Fig. 3. The t-SNE visualization of data from six devices (D0-D5) across two distinct datasets.

two parts: a small portion designated as the poisoned dataset $D_p = \{\mathbf{x}_k^p\}_{k=1}^N$, and the remainder serving as the clean dataset $D_c = \{\mathbf{x}_i\}_{i=1}^M$. The ratio of poisoned to total data is defined as the poisoning rate $\varphi \doteq \frac{N}{S}$.

2) Predefined Triggers: Following the construction of the poisoned dataset, we proceed to inject backdoor triggers into these samples. Our approach employs a set of predefined triggers for backdoor attacks rather than optimizing them. This decision is based on two key factors. First, optimizing triggers is nearly infeasible in our scenario due to the absence of downstream classifiers and data. Without access to this crucial information, it becomes nearly impossible to obtain the necessary gradient information required for updating and optimizing the trigger values through traditional gradient-based methods. Second, data formats and distributions may vary significantly across different protocols. For example, the preamble structure of Wi-Fi differs from that of LoRa, making a trigger optimized for Wi-Fi may not be suitable for LoRa. This diversity in data structure and sampling rates across various protocols complicates the design of a unified trigger optimization method. Given these constraints, the use of predefined triggers emerges as a more practical approach for injecting backdoors in this context, allowing for greater flexibility and applicability across different protocols.

In this paper, we choose to formulate the trigger set using time domain Gaussian noise, which has proven effective for launching backdoor attacks in related domains [39]. Unlike targeted attacks in supervised DNNs, our approach aims to induce misclassification into multiple classes by adding various triggers to the inputs of PTMs, thereby contaminating the downstream classifier. Considering the output representations given by $f_\theta(\mathbf{x} \oplus \mathbf{t}_j) = \mathbf{W}_\theta \cdot (\mathbf{x} \oplus \mathbf{t}_j) + \mathbf{B}_\theta$, our goal is to ensure that these representations differ sufficiently when different triggers are applied. Given that the weight $\mathbf{W}_\theta$ and bias $\mathbf{B}_\theta$ matrices

TABLE I
DOWNSTREAM ACCURACY DROPS WITH ONLY ADDED TRIGGERS

Dataset	ORACLE	WiSig	CORES	NetSTAR	Ours
Acc. Drop	4.12%	0.75%	0.02%	0.24%	5.75%

remain constant across samples, the most effective strategy is to introduce inherent differences in the poisoned samples $\mathbf{x}^p$ themselves after adding various triggers $\mathbf{t}_j$. Intuitively, we assume that $f_\theta(\mathbf{x} \oplus \mathbf{t}_j)$ and $f_\theta(\mathbf{x} \oplus -\mathbf{t}_j)$ will generate two relatively dissimilar output representations by simply reversing the trigger value. Therefore, we design the j-th trigger $\mathbf{t}_j$ in the trigger set T as follows:

$$\mathbf{t}_j = \begin{cases} N(0, \sigma; L), & j \leq \frac{N_t+1}{2}; \\ -\mathbf{t}_{N_t-j}, & j > \frac{N_t+1}{2}, \end{cases} \tag{4}$$

where L denotes the length of the trigger, which simultaneously regulates the trigger's size along with σ. In this paper, we use $L = 48$ and $\sigma = 0.1$ as the baseline settings.

3) Output Representations: While incorporating triggers into RF data can induce shifts in output representations, these minor changes alone are insufficient to launch a successful backdoor attack on downstream classifiers. Table I presents experimental results demonstrating that directly adding triggers to the inputs yields only minimal accuracy drops. Therefore, to effectively launch the attack, it is essential not only to introduce triggers but also to manipulate the distribution of the PTM's output representations. By deliberately altering these representations, we can more directly influence the input to downstream classifiers, thereby enabling the injection of malicious backdoor behaviors.

The downstream prediction is generated by feeding the output representations from the PTM to the downstream classifier, represented as $y = g(f_\theta(\mathbf{x})) = \mathbf{W}_c \cdot f_\theta(\mathbf{x}) + \mathbf{B}_c$. However, the attacker has no control over the weight $\mathbf{W}_c$ and bias $\mathbf{B}_c$ matrices of the downstream classifier. Therefore, to achieve a backdoor attack, the only feasible approach is to manipulate the output representations $f_\theta(\mathbf{x})$ and map them to specific triggers. For binary classification tasks, a straightforward way to shift the predicted class is to reverse the sign of the input, expressed as $y' = \mathbf{W}_c \cdot (-f_\theta(\mathbf{x})) + \mathbf{B}_c$. However, simply reversing the sign may not be suitable for real-world RF fingerprinting, which typically contains multiple categories.

Fig. 4 illustrates more intricate scenarios for manipulating output representations to achieve classification into separate classes. *Case 1* depicts a relatively independent situation where

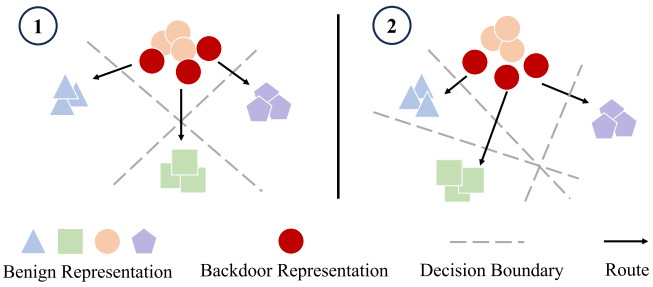


Fig. 4. Two cases when designing PORs.

different data clusters are distributed clearly. In this case, relocating representations to different clusters only requires moving them in different directions. In contrast, *Case 2* presents a more crowded scenario where data clusters are situated in closer proximity. While it is possible to move the representations similarly to *Case 1*, this approach may cause the representations to drift further from their corresponding data clusters. An alternative strategy is to adjust the output representations along a similar path but with varying distances to reach the different clusters. Based on these observations, we devise the PORs $\mathbf{e}_j = f_\theta(\mathbf{x} \oplus \mathbf{t}_j)$ as follows:

$$\mathbf{e}_j = \begin{cases} \mathbf{0}, & j = 1; \\ (1 + \frac{j-1}{N_t}) \cdot A \cdot \cos(2\pi \cdot j \cdot t), & 1 < j \leq \frac{N_t+1}{2}; \\ (1 + \frac{j-1}{N_t}) \cdot (-A) \cdot \cos(2\pi \cdot j \cdot t), & \frac{N_t+1}{2} < j < N_t; \\ \mathbf{1} \cdot A, & j = N_t, \end{cases} \tag{5}$$

where t is a variable with length corresponding to the representation dimension, and $\cos(2\pi \cdot j \cdot t)$ generates a cosine vector. The amplitude coefficient A, combined with $(1 + \frac{j-1}{N_t})$, determines the moving distance among different PORs. In this paper, we use $A = 1$ as the default setting.

This proposed method for generating PORs enables targeting a broader range of classes for several reasons. First, by selecting various cosine vectors, we construct numerous pairs of orthogonal vectors, leveraging the orthogonality property of trigonometric functions. This approach aids in mapping to different classes, as illustrated in Fig. 4. Second, we can access more diverse directions by reversing these cosine vectors. Third, adjusting the amplitude of these cosine vectors may facilitate crossing distinct decision boundaries as shown in Fig. 4. Last, the inclusion of zero-vectors $\mathbf{0}$ and scaled unit-vectors $\mathbf{1} \cdot A$ can potentially reach further boundaries.

C. Backdoor Training

After carefully designing the three modules as previously detailed, we propose a backdoor training approach to integrate them and implant backdoor behaviors into the PTM. The training process fine-tunes a clean PTM f_θ into a backdoored PTM f_{θ_p} by minimizing the following loss function:

$$\min_{f_{\theta_p}} L = \sum_{\mathbf{x}_i \in D_c} \mathcal{L}(f_{\theta_p}(\mathbf{x}_i), f_\theta(\mathbf{x}_i)) + \sum_{\mathbf{x}_k \in D_p} \mathcal{L}(f_{\theta_p}(\mathbf{x}_k \oplus \mathbf{t}_j), \mathbf{e}_j), \tag{6}$$

Algorithm 1: PTM Backdoor Training Process.

Input: Substitute dataset $D_s = \{\mathbf{x}_i\}_{i=1}^{S}$, benign PTM f_θ, trigger set $T = \{\mathbf{t}_j\}_j^{N_t}$, PORs $E = \{\mathbf{e}_j\}_j^{N_t}$, poisoning rate φ, learning rate lr
Output: Backdoored PTM f_{θ_p}

Step 1: Prepare training set and PORs
$N \leftarrow \varphi \cdot S,\ M \leftarrow (1-\varphi) \cdot S$
Initialize $D_c = \{\mathbf{x}_i\}_{i=1}^{M}$ and $D_p = \{\mathbf{x}_k\}_{k=1}^{N}$ from D_s
for j in $(1, N_t)$ **do**
  for n in $(1, \frac{N}{N_t})$ **do**
    $\mathbf{x}_k^p \leftarrow \mathbf{x}_k \oplus \mathbf{t}_j,\ \mathbf{y}_k^p \leftarrow \mathbf{e}_j;\ k$++
  end for
end for
for i in $(1, M)$ **do**
  $\mathbf{y}_i \leftarrow f_\theta(\mathbf{x}_i)$
end for
Step 2: Updating backdoored PTM parameters
$\theta_p \leftarrow \theta$ `// Copy structure and parameters`
for number of epoch **do**
  $L \leftarrow \sum \mathcal{L}(f_{\theta_p}(\mathbf{x}_i), \mathbf{y}_i) + \sum \mathcal{L}(f_{\theta_p}(\mathbf{x}_k^p), \mathbf{y}_k^p)$
  $\theta_p \leftarrow \theta_p - lr \cdot \frac{\partial L}{\partial \theta_p}$
end for
return f_{θ_p}

where $\mathcal{L}$ denotes the mean squared error (MSE) loss. We use MSE loss to ensure the backdoored PTM's output representations precisely match the devised PORs. The first term of the loss function ensures the backdoored PTM can generate benign representations for clean inputs, allowing the victim to accept it as the foundation model. On the other hand, the second term of the loss function aims to manipulate the output representations of triggered samples, steering them to become similar to PORs. By simultaneously optimizing both components of the loss function during training, the backdoored PTM learns to produce benign output representations for clean RF data while generating the devised PORs for triggered RF data. This dual functionality aligns with the attacker's goals as defined in Section IV-B1, enabling the PTM to maintain normal operation on clean inputs while facilitating backdoor attacks when triggered.

Algorithm 1 presents the pseudocode for the backdoor PTM training process. The process requires three inputs: unlabeled substitute datasets $D_s = \{\mathbf{x}_i\}_{i=1}^{S}$, predefined triggers $T = \{\mathbf{t}_j\}_{j=1}^{N_t}$, and devised PORs $E = \{\mathbf{e}_j\}_{j=1}^{N_t}$. First, we construct the clean set D_c and the poisoned set D_p using the substitute dataset and poisoning rate φ. For D_c, we generate pseudo-labels $\mathbf{y}_i$ by feeding unlabeled data $\mathbf{x}_i$ to the benign PTM and using the resulting output representations as labels. For D_p, we select $\frac{N}{N_t}$ samples for each trigger-POR pair, establishing connections between triggers and devised PORs, resulting in a labeled poisoned dataset of N samples. We then initialize the backdoor PTM by replicating the structure and parameters of the benign PTM f_θ. The MSE loss is computed using the constructed D_c and D_p, and employed to update the backdoor PTM's parameters θ_p via gradient descent optimization.

TABLE II
DOWNSTREAM DATASET SUMMARY

Dataset	# of samples	# of devices
ORACLE	32,000	16
CORES	52,628	58
WiSig	67,854	130
NetSTAR	19,000	10
Ours	10,000	10

Fig. 5. LoRa Transmitters and a USRP Receiver.

VI. EXPERIMENTAL EVALUATION AND ANALYSIS

A. Experiment Setup

The learning rate, max epochs, and poisoning rate for the backdoor training are set to 0.001, 200, and 0.1, respectively. All experiments are conducted on a Linux server with an Intel(R) Xeon(R) Gold 6258R CPU and NVIDIA A100 GPUs with 40 GB of memory.

1) Victim PTMs: Given the early stage of RF fingerprinting PTM research, our experimental evaluation focuses on assessing backdoor attack effectiveness on classic PTMs employing two principal SSL approaches discussed in Section II.

Generative SSL: BERT is one of the most representative works in this field. We modify its lightweight version [41] for RF fingerprinting tasks. Besides, we employ masked autoencoders (MAE) [42] to build PTMs in this paper.

Contrastive SSL: We also employ classic contrastive learning methods to build PTMs from scratch. Following Qian et al. [43], we employ SimCLR [29] and TS-TCC [44] methods to train convolutional neural networks (CNNs) [45] and the encoder part of Transformer models [46].

Overall, our PTM selection covers the mainstream approaches commonly used in RF fingerprinting and related domains. We modify the first layer of all PTMs to fit RF data shapes. As mentioned in Section I, time domain I/Q data often undergoes signal processing. Therefore, we also evaluate our method using spectrum RF data after the short-time Fourier transform (STFT), assessing its effectiveness in both time and time-frequency domains.

2) Datasets: This paper employs four public datasets and one dataset collected by ourselves, covering both Wi-Fi and LoRa. Table II summarizes key information about the downstream datasets. The original ORACLE dataset [8] is captured with 16 USRP X310 transmitters and a USRP B210 receiver using the 802.11a standard. [47] consists of 163 consumer Wi-Fi cards arranged in a grid at the Orbit Testbed [48] communicating with 802.11 g. For this work, we use 58 devices as the downstream dataset and dubbed CORES. The WiSig dataset [49] captures signals from 174 COTS Wi-Fi cards using 802.11a/g access on channel 11. [27] captures LoRa transmissions from 25 Pycom devices and USRP B210 across various domains. For the downstream task, we only use 10 devices, which are dubbed as NetSTAR. As shown in Fig. 5, our dataset uses 10 commercial LoRa transmitters (Pycom LoPy4) and a USRP N210 receiver. Due to different sampling rates and preamble structures, the original captured I/Q data for LoRa is 2×1024 in size. This is downsampled to 2×256 to meet model input requirements.

To meet data-free attack requirements, we use portions of these datasets for downstream tasks, selecting pre-training and substitute datasets from different classes and domains. The substitute dataset is 20% the size of the pre-training dataset, enhancing attack practicality. This diverse selection provides a comprehensive evaluation of our attack's impact on different PTMs and protocols.

B. Evaluation Metrics

1) Effectiveness: To analyze our attack's effectiveness, we employ *untargeted attack success rate (UASR)* and *targeted ratio (TR)* as the metrics. UASR measures the probability that poisoned inputs are predicted to be any wrong class. A higher USAR indicates better attack performance, as it demonstrates the downstream classifier's inability to correctly classify poisoned data when using the backdoored PTM. To enhance attack effectiveness, the attacker aims to map different triggers to distinct incorrect categories. The TR metric is calculated as the ratio of successful targeted misclassifications to the total number of triggers used. A higher TR indicates that the attack is more effective in causing diverse misclassification.

2) Stealthiness: Visual inspection is inefficient and impractical. Therefore, this study employs three approaches to quantify it, namely (i) model utility, (ii) trigger size, and (iii) algorithm-based detection. Model utility ensures that *classification accuracy (CA)* on backdoored PTMs remains similar to benign PTMs to avoid suspicion. For algorithm-based detection methods, we employ the *isolation forest* [50] to identify potential outliers and *STRIP* [51] to detect poisoned samples by measuring predicted entropy. Higher entropy makes attacks harder for STRIP to detect.

3) Robustness: The last goal of the attack is to ensure its robustness against defense methods. While fine-pruning [52] effectively removes backdoored neurons, it can degrade model performance, contradicting the purpose of using PTMs. Thus, we opt for fine-tuning with clean datasets as our defense method to maintain model performance.

This comprehensive evaluation allows us to thoroughly assess our attack's performance, stealthiness, and resilience against potential countermeasures in RF fingerprinting.

C. Stealthiness Evaluation

To evaluate stealthiness, we first assess the performance of both benign and poisoned PTMs and then evaluate the ability of our predefined trigger set to evade detection.

1) Model Utility: Table III presents clean downstream classification accuracies and stealthiness metrics. The accuracy on the ORACLE and our dataset is comparatively low, possibly due to complex environmental domain shifts, with time-frequency domain results generally demonstrating more consistent and

TABLE III
BASELINE UTILITY EVALUATION. "ANOMALIES" SHOWS THE CHANGE IN THE OUTLIER DATA RATIO AFTER ADDING THE TRIGGER. "SPEC." DENOTES RESULTS IN THE TIME-FREQUENCY DOMAIN

Dataset→		ORACLE	WiSig	CORES	NetSTAR	Ours
Stealth	SNR (dB)	22.26	21.91	21.99	22.79	22.93
	Δl_2-norm	0.0377	0.0394	0.0390	0.0357	0.0350
	Anomalies	0.0642	-0.0465	0.0009	-0.0253	0.0178
Time	SimCLR-R	0.6341	0.9423	0.9915	0.8055	0.6406
	SimCLR-T	0.7208	0.8726	0.9766	0.8287	0.9047
	TS-TCC-R	0.6339	0.8378	0.9524	0.8797	0.7137
	TS-TCC-T	0.6125	0.7939	0.9540	0.7542	0.8484
	BERT	0.9264	0.9444	0.9953	0.9674	0.6363
Spec.	SimCLR-R	0.8966	0.9860	0.9999	0.9695	0.5613
	SimCLR-T	0.9087	0.9856	0.9999	0.9721	0.5813
	MAE-R	0.9716	0.9859	0.9999	0.9766	0.7175
	MAE-T	0.8517	0.9867	0.9999	0.9787	0.7138

TABLE IV
MEAN ENTROPY DIFFERENCE FROM STRIP ($\times 10^{-2}$). RES AND TRANS DENOTE RESNET AND TRANSFORMER ENCODERS, RESPECTIVELY. UNDERLINED VALUES INDICATE POTENTIAL DETECTABILITY

($\times 10^{-2}$)	Time Domain					Time-frequency Domain			
SSL	SimCLR		TS-TCC		BERT	SimCLR		MAE	
Model	Res	Trans	Res	Trans	Trans	Res	Trans	Res	Trans
ORACLE	-0.01	-0.30	-0.01	-0.11	0	0	0.04	0	0
WiSig	0	-1.84	-0.04	4.78	0	0	5.38	0.04	-0.02
CORES	0	-2.04	-0.04	-0.64	0	-0.01	1.49	0.02	-0.02
NetSTAR	0	0.38	0	-0.55	0	0.01	0.03	0	0.01
Ours	0	-0.07	0	-0.34	0	0.01	0.02	0	-0.01

superior performance. We implant backdoors into these PTMs using 8 predefined triggers and PORs, with average results shown in Table V. Here, "-R" and "-T" denote ResNet and Transformer encoders, respectively. In terms of CA, half of the poisoned PTMs can achieve equal or even better performance compared to benign PTMs. Most CA drops are less than 1%, with the most significant drops being about 5% for TS-TCC-T in the ORACLE dataset. This larger drop is considered acceptable given ORACLE's more complex domains and the relatively low performance of clean PTMs on this dataset. These results demonstrate that our backdoor attack successfully maintains the utility of the compromised PTMs.

2) Trigger Stealthiness: Data censorship and protection mechanisms will likely be deployed in real-world RF fingerprinting systems. Therefore, our designed triggers need to be stealthy to evade backdoor detection.

Trigger Size: To demonstrate the physical stealthiness of our predefined triggers, we use two indicators: Δl_2-norm, which quantifies changes in the l_2-norm of data after adding triggers, and signal-to-noise ratio (SNR). As shown in Table III, both measures confirm that our triggers maintain a high degree of physical stealthiness in RF data.

Backdoor Detection: For algorithm-based detections, the isolation forest anomaly detection method fails to significantly alter anomaly rates, further demonstrating our predefined triggers' ability to evade detection. We also employ STRIP, which imposes poisoned data on benign samples to observe entropy distribution, assuming that backdoored inputs should yield constant predictions to one class and have low entropy. Table IV presents entropy differences ($\times 10^{-2}$) between backdoored and clean PTMs, with negative values indicating more constant predictions for backdoored PTMs. Although some underlined values appear slightly larger, they remain small and unlikely to raise suspicion from defenders.

Combined with the results from Table I, which show that the trigger does not impact the performance of clean PTMs, we can conclude that our predefined trigger set meets the stealthiness goal.

D. Effectiveness Evaluation

Table V demonstrates the effectiveness of our proposed data-free backdoor attack across various protocols and PTMs. Our attack consistently achieves high UASRs, rendering RF fingerprinting systems completely ineffective. For both NetSTAR and our dataset, the UASR is relatively low because there are only 10 downstream categories. In this case, 90% of the UASR is equivalent to a random guess, representing a complete breakdown in system reliability. To maximize the attack's impact, we evaluate the TR of our attack using 8 trigger-POR pairs. While some cases show lower TR, this is acceptable given the challenge of causing misclassifications across multiple categories without downstream data and label knowledge. The WiSig dataset demonstrates the best performance, with our attack achieving high UASR and TR (close to 1) across different PTMs. Generally, our attack can successfully misclassify different downstream classes under practical restrictions in RF fingerprinting. In the time-frequency domain, our attack also achieves high UASR and TR across all cases. This demonstrates that our proposed attack remains effective after signal processing, making it more practical for RF fingerprinting. Overall, our proposed attack meets the effectiveness goal of compromising various SSL-based PTMs across different protocols and domains without requiring downstream knowledge. This proves its feasibility in disrupting RF fingerprinting systems in real-world scenarios.

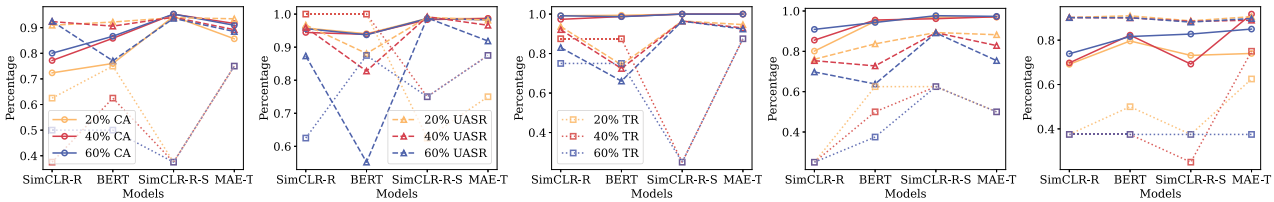
E. Robustness Evaluation

Beyond being stealthy to backdoor detection methods, it is crucial to assess the robustness of backdoor attacks against proactive defense mechanisms in security-critical RF fingerprinting systems. This is particularly important because system providers may deploy active defenses to safeguard the system after downloading PTMs from public repositories.

1) Fine-Tuning: We choose fine-tuning as the proactive defense strategy because it preserves model performance while potentially removing backdoors. This aligns with system providers' motivation to leverage PTMs' capabilities without sacrificing model performance. Moreover, fine-tuning can adapt models to downstream tasks and is straightforward to implement. It also serves as a representative baseline for post-training defenses, as it updates model parameters with clean data without altering the model architecture. Fig. 6 illustrates the results of four representative PTMs with different fine-tuning rates across diverse domains. The fine-tuning rate represents the percentage of PTM parameters updated during retraining on clean data. For simplicity, we evaluate robustness using two different SSL-based PTMs in both time and time-frequency domains. Compared to the original backdoored PTMs, CA improves

TABLE V
THE DOWNSTREAM RESULTS OF BACKDOORED PTMS WITH 8 TRIGGER-POR PAIRS. THE CA DROPS LARGER THAN 1% ARE DENOTED IN BOLD, WHILE DROPS BETWEEN 0 AND 1% ARE DENOTED WITH AN UNDERLINE. "-R" AND "-T" INDICATE RESNET AND TRANSFORMER ENCODERS, RESPECTIVELY

Dataset→		ORACLE			WiSig			CORES			NetSTAR			Ours		
Domains↓	**PTMs↓**	CA	UASR	TR	CA	UASR	TR	CA	UASR	TR	CA	UASR	TR	CA	UASR	TR
Time	SimCLR-R	0.6444	0.9307	0.50	0.9430	0.9718	0.88	0.9934	0.9522	0.75	0.7955	0.7281	0.38	0.6734	0.8939	0.38
	SimCLR-T	**0.6856**	0.9084	0.50	0.8766	0.8966	0.88	0.9793	0.8733	0.63	**0.8105**	0.8146	0.38	0.9088	0.9075	0.63
	TS-TCC-R	**0.5825**	0.9372	0.50	**0.8218**	0.9861	1.00	0.9513	0.9661	0.75	**0.8582**	0.7315	0.88	0.7109	0.9067	0.38
	TS-TCC-T	**0.5573**	0.9101	0.25	0.7860	0.9610	0.88	0.9538	0.9396	0.38	**0.7247**	0.8583	0.38	0.8687	0.8973	0.50
	BERT	**0.8908**	0.9279	0.88	0.9488	0.9676	1.00	0.9959	0.9406	0.75	0.9603	0.8452	0.75	0.6963	0.9052	0.50
Spec.	SimCLR-R	0.9070	0.9336	0.88	0.9870	0..9871	0.75	0.9999	0.9604	0.50	0.9663	0.8887	0.63	0.6225	0.9034	0.50
	SimCLR-T	0.8941	0.9279	0.50	0.9860	0.9491	0.63	0.9999	0.9434	0.38	0.9692	0.8626	0.63	0.5763	0.8991	0.38
	MAE-R	0.9677	0.9381	0.75	0.9858	0.9853	1.00	0.9999	0.9630	0.50	**0.9329**	0.8876	0.88	0.7953	0.9008	0.50
	MAE-T	0.8684	0.9348	1.00	0.9870	0.9881	0.88	0.9999	0.9731	1.00	0.9726	0.8954	0.75	**0.6891**	0.9042	0.63



(a) ORACLE (b) WiSig (c) CORES (d) NetSTAR (e) Ours

Fig. 6. Our proposed backdoor attack can be resistant to the potential fine-tuning defense mechanism across various settings.

TABLE VI
THE CA AND UASR OF BACKDOORED PTMS WITH 8 TRIGGER-POR PAIRS AFTER DEPLOYING PROACTIVE BACKDOOR MITIGATION [54]. THE CA REDUCTIONS ARE DENOTED WITH AN UNDERLINE. "-R" AND "-T" INDICATE RESNET AND TRANSFORMER ENCODERS, RESPECTIVELY

PTMs	**Time Domain**										**Time-frequency Domain**							
	SimCLR-R		SimCLR-T		TS-TCC-R		TS-TCC-T		BERT		SimCLR-R		SimCLR-T		MAE-R		MAE-T	
Dataset ↓	CA	UASR	CA	UASR	CA	UASR	CA	UASR	CA	UASR	CA	UASR	CA	UASR	CA	UASR	CA	UASR
ORACLE	0.9203	0.7518	0.7716	0.7005	0.9062	0.7301	0.4975	0.7591	0.7614	0.6654	0.9634	0.9088	0.9502	0.8588	0.9563	0.9323	0.9298	0.8838
WiSig	0.9869	0.3573	0.9268	0.2943	0.9833	0.7127	0.8939	0.2630	0.9366	0.2555	0.9871	0.6232	0.9808	0.3814	0.9864	0.7406	0.9877	0.2072
CORES	0.9964	0.3306	0.9884	0.2943	0.9976	0.2641	0.9833	0.2613	0.9953	0.1061	0.9999	0.5062	0.9999	0.7872	0.9999	0.8917	0.9999	0.6876
NetSTAR	0.7834	0.7413	0.7563	0.8135	0.6974	0.7611	0.5010	0.8563	0.9605	0.8347	0.9663	0.8950	0.9542	0.8646	0.9095	0.8911	0.9632	0.8855
Ours	0.6953	0.6543	0.5118	0.8232	0.8712	0.5190	0.8460	0.7182	0.5695	0.8166	0.9826	0.2179	0.9731	0.2756	0.9776	0.7853	0.9713	0.6526

as PTMs acquire task-specific knowledge through fine-tuning. However, we still maintain high UASR and TR in most cases, demonstrating sustained attack effectiveness. Only when the fine-tuning rate reaches 60%, the UASR for BERT shows slight drops in the time domain, possibly due to the BERT model in our study being relatively smaller than others. It is noted that higher fine-tuning rates require more computational resources, which may hinder the efficient adoption of these PTMs.

The failure of fine-tuning as an effective defense mechanism can be attributed to two factors. First, malicious neurons may remain dormant when processing clean samples [53], preventing their removal through fine-tuning. Second, the backdoor is injected by manipulating the output representations, which may make it difficult to eliminate the associations between triggers and PORs using supervised learning.

2) Knowledge Distillation: Building on NAD [53], Bie et al. [54] propose a self-supervised knowledge distillation defense method, which we denote as SSKD for brevity, to purify backdoored PTMs in the image domain. The core idea is first to fine-tune the victim PTM through contrastive learning to construct a teacher model and then deploy knowledge distillation on the victim PTM to remove the backdoor. This whole process can be directly adapted to the RF data. Following their setup, we also deploy clean downstream data and the SimCLR method to cleanse backdoor neurons for robustness evaluation.

Table VI presents the overall CA and UASR after backdoor mitigation. In general, SSKD outperforms fine-tuning by achieving lower UASR across some cases. For instance, it reduces UASR by about 50% on WiSig and CORES with SimCLR in the time domain, showing that SSKD enables more effective purification. However, it fails to completely mitigate our proposed backdoor attack, as more than half of the cases still exhibit high UASR, with some continuing to show backdoor behaviors. The incomplete removal of the backdoor can be attributed to the distributional shift between the data used for backdoor injection during pre-training and the clean data used for defense, where the latter cannot fully activate the backdoored neurons. Furthermore, while SSKD incorporates fine-tuning to minimize utility loss, our results reveal that one-third of the cases experience a decline in CA after knowledge distillation. This reduction may be attributed to the loss of feature extraction capability during fine-tuning and knowledge distillation, especially given the limited size of the clean downstream dataset compared to the large pre-training dataset.

In summary, our analysis indicates that current proactive defense methods using a small set of downstream clean data cannot effectively mitigate our attack and maintain encoder utility in either the time or time-frequency domain. This underscores the robustness of the attack against defense mechanisms in RF fingerprinting systems.

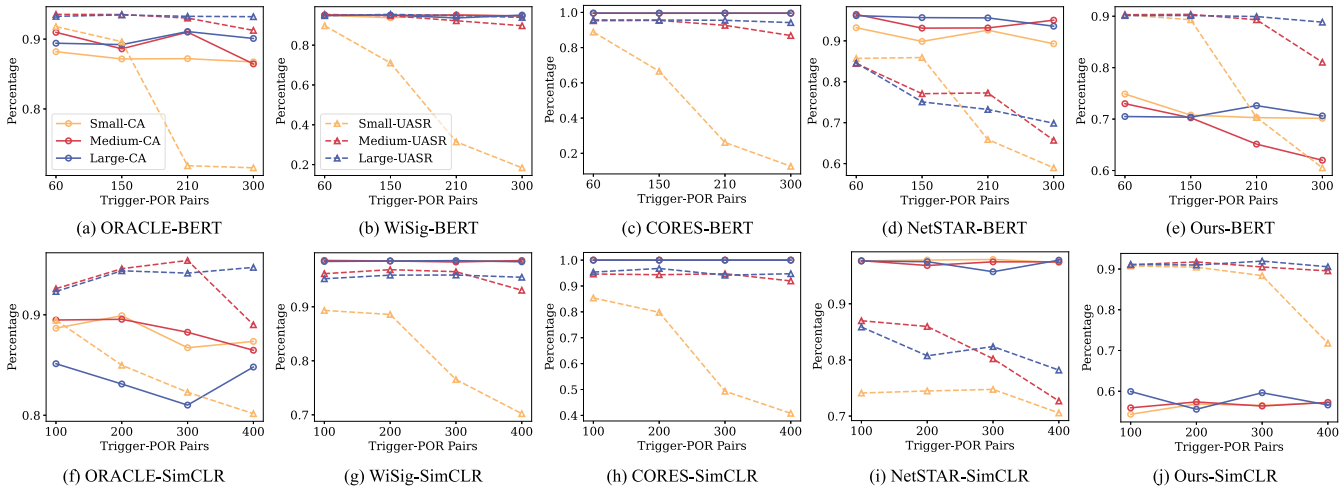


Fig. 7. Effects of PTM size and trigger-POR pairs on backdoor attacks in time domain BERT (top row) and time-frequency domain SimCLR (bottom row). Small-CA and Small-UASR denote the CA and UASR for small-sized PTMs.

F. Impacts of Different Modules

In this subsection, we experimentally evaluate the contribution of different modules to our proposed attack. To maintain efficiency while ensuring comprehensive coverage, we assess specific modules using a representative selection of PTMs spanning various model architectures and RF domains.

1) PTM Size and Trigger-POR Pairs: The effectiveness of backdoor injection is significantly influenced by the number of trigger-POR pairs. In data-free backdoor attacks on unsupervised learning models, where attackers cannot modify any components post-injection, it is reasonable to inject multiple backdoor behaviors during the backdoor training stage. Besides, the size of PTM also impacts attack performance as discussed in Section VI-E. Fig. 7 presents the impact of these factors on attack performance. We evaluate Transformer encoders of varying sizes (small: 0.6 M, medium: 1.3 M, and large: 2.3 M parameters) with different numbers of trigger-POR pairs. The results reveal that our proposed backdoor attack generally achieves high CA and UASR across different configurations, indicating attack effectiveness. Compared to the small PTM, larger PTMs can maintain high CA and UASR in both the time domain and time-frequency domain. When increasing the number of trigger-POR pairs to implant more backdoor behaviors into PTMs, a clear trend emerges. Smaller PTMs experience drops in UASR, indicating they cannot retain a large number of backdoor behaviors while maintaining their utility. In contrast, larger PTMs can remember these backdoors and maintain high UASR. It is important to note that today's foundation models continue to grow in size, becoming more capable of remembering backdoor behaviors while potentially offering stronger generalization performance compared to smaller models. This highlights a potential security concern in deploying PTMs in RF fingerprinting systems.

2) PORs Design Comparison: We evaluate the effectiveness of our proposed orthogonal PORs design by comparing it to the non-orthogonal PORs used in [20], which employs varying numbers of -1s and 1s. To ensure a fair comparison, we maintain consistency with our previous setup by using 8 trigger-POR pairs. In all cases, the CA is similar to ours, and the UASR only experiences drops in a few cases compared to our method. The most significant difference is observed in the TR metric, as shown in Table VII. TR decreases in most cases using the non-orthogonal PORs design, with some cases achieving only 25%, indicating that their attack targets only two different downstream categories using 8 trigger-POR pairs. There are only four cases that can achieve the same TR as our orthogonal PORs method. Additionally, their method generates a constant number of PORs based on representation length, while ours can generate any number of orthogonal PORs. These results demonstrate that our orthogonal PORs design is crucial for successfully launching backdoor attacks on PTMs in a data-free setting. It allows for more effective targeting of multiple downstream categories, providing a more practical attack strategy for RF fingerprinting systems.

TABLE VII
PORs Design Comparison. Underlined Values Indicate the Same TR as Our Proposed Attack

	Time Domain					Time-frequency Domain			
SSL	SimCLR		TS-TCC		BERT	SimCLR		MAE	
Model	Res	Trans	Res	Trans	Trans	Res	Trans	Res	Trans
ORACLE	0.38	0.38	0.50	0.38	0.50	0.50	0.25	0.63	0.63
WiSig	0.88	0.38	0.63	0.25	1.00	0.25	0.25	0.50	0.50
CORES	0.63	0.38	0.63	0.25	0.38	0.38	0.25	0.50	0.63
NetSTAR	0.50	0.25	0.75	0.38	0.38	0.38	0.38	0.50	0.38
Ours	0.25	0.38	0.25	0.38	0.38	0.25	0.25	0.50	0.25

3) Trigger Length: In the design of backdoor attacks, the size of triggers is an important hyperparameter. One critical factor in determining this size is the trigger length L. To fairly assess the impact of trigger lengths and account for various SSL methods, we evaluate the attack performance on BERT (time domain) and SimCLR (time-frequency domain) using different trigger lengths while maintaining consistency in all other parameters. The evaluation results are presented in Fig. 8. Overall, the CA and UASR metrics show stability across different trigger lengths,

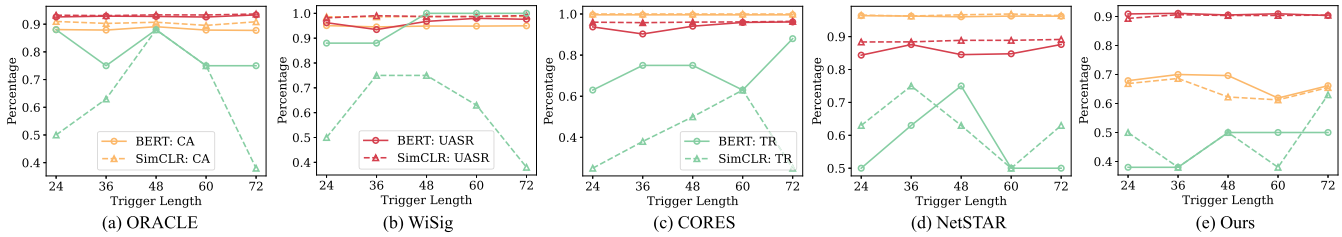


Fig. 8. Effects of length L on backdoor attacks in time domain BERT and time-frequency domain SimCLR (ResNet18 backbone).

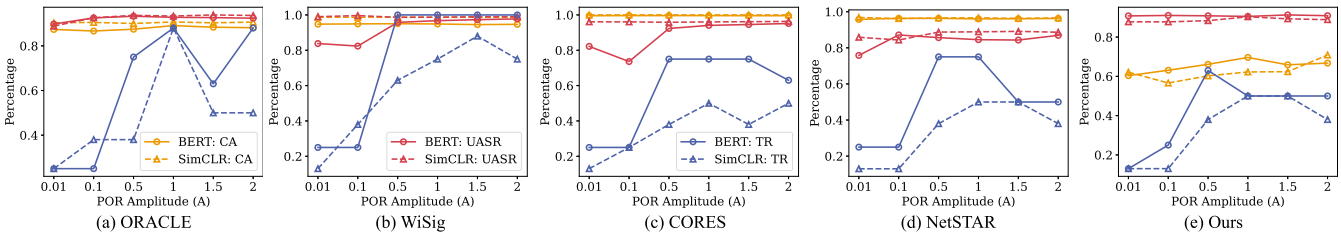


Fig. 9. Effects of amplitude A of PORs on backdoor attacks in time domain BERT and time-frequency domain SimCLR.

demonstrating their robustness regardless of L. However, the TR is slightly lower for smaller trigger lengths in the time-frequency domain. This drop is likely due to the reduced distinctiveness of smaller triggers after applying the STFT, which makes them harder to recognize. In summary, the consistency in CA and UASR suggests that our proposed attack remains robust and relatively insensitive to variations in L. In contrast, the TR variations indicate that excessively small triggers should be avoided when targeting multiple classes.

4) POR's Amplitude: In this paper, we use the amplitude coefficient A to quantify the separation between distinct output representations in our proposed backdoor attack. As the A increases, the norm of these representations increases, leading to greater distances between them. Fig. 9 presents the attack performance across different amplitudes A, ranging from 0.01 to 2. In general, the POR's amplitude has minimal impact on the CA and UASR. These metrics remain consistent across various amplitudes, highlighting the robustness of the attack in these aspects. In contrast, the TR values are significantly influenced by POR's amplitude. At a low amplitude, such as $A = 0.01$, our attack results in low TR values, indicating that different PORs fail to map to distinct downstream classes. This occurs because PORs act as inputs to downstream classifiers, and smaller distances between them result in more similar features. Consequently, classifiers tend to produce identical outputs, thereby reducing the TR. As the amplitude increases, the TR values increase and eventually stabilize across different amplitude values. This trend aligns with the concept we introduced in Fig. 4 and demonstrates the adaptability of our proposed attack, requiring only that the POR's amplitude not be excessively small.

G. Impacts of Device Positions

In RF fingerprinting, variations in device position can significantly alter channel conditions, thereby influencing authentication performance and potentially affecting the effectiveness of backdoor attacks. To evaluate the robustness of our proposed method under realistic deployment scenarios, we investigate the impact of device location on attack performance.

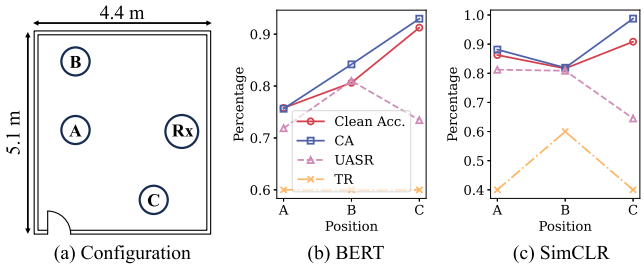


Fig. 10. Attack performance under different device positions in an office. Victim PTMs: time-domain BERT and time–frequency domain SimCLR.

Specifically, we collect LoRa signals from five devices placed at three distinct positions (A, B, and C), as illustrated in Fig. 10. The attack is then evaluated using BERT in the time domain and SimCLR in the time–frequency domain. In general, our attack can still stay stealthy, achieving comparable or even higher accuracy on clean samples, while maintaining high UASR and TR to ensure backdoor effectiveness across different positions. Notably, at the closest position C, both PTMs exhibit higher CA but lower UASR, likely due to the clearer signals at shorter distances and the reduced influence of additional triggers on the data. This observation highlights the relations between signal quality and backdoor activation, suggesting that cleaner channels may suppress the influence of malicious perturbations. Although conducted in real-world settings, our experiments do not explicitly control factors such as jammers or antenna orientation. Future work will extend the study to more diverse environments. These results collectively confirm that our attack is effective and robust under varying device positions.

VII. Conclusion

In this paper, we propose the first protocol-agnostic and data-free backdoor attack on PTMs used in RF fingerprinting systems. Unlike traditional backdoor attacks where attackers may possess data and label information, we inject backdoors into unsupervised PTMs without downstream knowledge or access to downstream training. To achieve this, we employ three key strategies: utilizing substitute datasets, designing trigger sets, and manipulating output representations to inject backdoor behaviors into the PTMs. Extensive experiments are conducted across Wi-Fi and LoRa, using five different datasets and two mainstream SSL methods in both the time and time-frequency domains. Moreover, we evaluate our attack under diverse defense mechanisms and device positions, demonstrating its robustness and effectiveness in realistic scenarios. Through this comprehensive analysis, we demonstrate that our proposed data-free backdoor attack poses a practical threat to RF fingerprinting systems, highlighting the urgent need for robust security measures to mitigate such threats when deploying PTMs in the real world. The authors have provided public access to their code at github.com/Tianyaz97/rf_backdoor.

References

[1] T. Zhao, N. Wang, J. Zhang, and X. Wang, "Protocol-agnostic and data-free backdoor attacks on pre-trained models in RF fingerprinting," in *Proc. IEEE Conf. Comput. Commun.*, 2025, pp. 1–10.

[2] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances, and future trends," *Proc. IEEE*, vol. 104, no. 9, pp. 1727–1765, Sep. 2016.

[3] E. Perenda, S. Rajendran, G. Bovet, M. Zheleva, and S. Pollin, "Contrastive learning with self-reconstruction for channel-resilient modulation classification," in *Proc. IEEE Conf. Comput. Commun.*, 2023, pp. 1–10.

[4] Q. Xu, R. Zheng, W. Saad, and Z. Han, "Device fingerprinting in wireless networks: Challenges and opportunities," *IEEE Commun. Surv. Tuts.*, vol. 18, no. 1, pp. 94–104, 2015.

[5] S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep learning convolutional neural networks for radio identification," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 146–152, Sep. 2018.

[6] J. Zhang, R. Woods, M. Sandell, M. Valkama, A. Marshall, and J. Cavallaro, "Radio frequency fingerprint identification for narrowband systems, modelling and classification," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 3974–3987, 2021.

[7] L. Peng, A. Hu, J. Zhang, Y. Jiang, J. Yu, and Y. Yan, "Design of a hybrid RF fingerprint extraction and device classification scheme," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 349–360, Feb. 2019.

[8] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, and K. Chowdhury, "ORACLE: Optimized radio classification through convolutional neural networks," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 370–378.

[9] A. Al-Shawabka et al., "Exposing the fingerprint: Dissecting the impact of the wireless channel on radio fingerprinting," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 646–655.

[10] T. Zhao, X. Wang, and S. Mao, "Cross-domain, scalable, and interpretable RF device fingerprinting," in *Proc. IEEE Conf. Comput. Commun.*, 2024, pp. 2099–2108.

[11] T. Zhao, N. Wang, S. Mao, and X. Wang, "Few-shot learning and data augmentation for cross-domain UAV fingerprinting," in *Proc. 30th Annu. Int. Conf. Mobile Comput. Netw.*, 2024, pp. 2389–2394.

[12] H. Li, K. Gupta, C. Wang, N. Ghose, and B. Wang, "RadioNet: Robust deep-learning based radio fingerprinting," in *Proc. IEEE Conf. Commun. Netw. Secur.*, 2022, pp. 190–198.

[13] Z. Chen et al., "Cross-device radio frequency fingerprinting identification based on domain adaptation," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 2391–2400, Feb. 2024.

[14] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, vol. 1, pp. 4171–4186.

[16] C. Liu et al., "Overcoming data limitations: A few-shot specific emitter identification method using self-supervised learning and adversarial augmentation," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 500–513, 2024.

[17] J. Chen, W.-K. Wong, and B. Hamdaoui, "Unsupervised contrastive learning for robust RF device fingerprinting under time-domain shift," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 3567–3572.

[18] Y. Gao et al., "Backdoor attacks and countermeasures on deep learning: A comprehensive review," 2020, *arXiv:2007.10760*.

[19] J. Jia, Y. Liu, and N. Z. Gong, "BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 2043–2059.

[20] L. Shen et al., "Backdoor pre-trained models can transfer to all," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2021, pp. 3141–3158.

[21] P. Lv et al., "A data-free backdoor injection approach in neural networks," in *Proc. 32nd USENIX Secur. Symp.*, 2023, pp. 2671–2688.

[22] M. Li et al., "DarkFed: A data-free backdoor attack in federated learning," 2024, *arXiv:2405.03299*.

[23] R. Ning, C. Xin, and H. Wu, "TrojanFlow: A neural backdoor attack to deep learning-based network traffic classifiers," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 1429–1438.

[24] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–13.

[25] A. Saha, A. Tejankar, S. A. Koohpayegani, and H. Pirsiavash, "Backdoor attacks on self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13337–13346.

[26] N. Soltanieh, Y. Norouzi, Y. Yang, and N. C. Karmakar, "A review of radio frequency fingerprinting techniques," *IEEE J. Radio Freq. Identif.*, vol. 4, no. 3, pp. 222–233, Sep. 2020.

[27] A. Elmaghbub and B. Hamdaoui, "LoRa device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability," *IEEE Access*, vol. 9, pp. 142893–142909, 2021.

[28] X. Liu et al., "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[30] X. Zha, T. Li, Z. Qiu, and F. Li, "Cross-receiver radio frequency fingerprint identification based on contrastive learning and subdomain adaptation," *IEEE Signal Process. Lett.*, vol. 30, pp. 70–74, 2023.

[31] M. Shao, P. Deng, D. Li, R. Lin, and H. Sun, "A specific emitter identification method based on self-supervised representation learning," in *Proc. 2024 IEEE 4th Int. Conf. Power, Electron. Comput. Appl.*, 2024, pp. 125–128.

[32] G. Parpart, J. H. Tu, B. Clymer, J. Lee, and J. Babcock, "Transformer masked autoencoders for RF device fingerprinting," in *Proc. MILCOM 2024-2024 IEEE Mil. Commun. Conf.*, 2024, pp. 859–862.

[33] Z. Yao et al., "Few-shot specific emitter identification using asymmetric masked auto-encoder," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2657–2661, Oct. 2023.

[34] Y. Liu, N. Gao, X. Li, and S. Jin, "Large language model enabled lightweight RFFI for 6G edge intelligence," in *Proc. 2025 IEEE Wireless Commun. Netw. Conf.*, 2025, pp. 1–6.

[35] T. Zhao, X. Wang, J. Zhang, and S. Mao, "Explanation-guided backdoor attacks on model-agnostic RF fingerprinting," in *Proc. IEEE Conf. Comput. Commun.*, 2024, pp. 221–230.

[36] T. Zhao, J. Zhang, S. Mao, and X. Wang, "Explanation-guided backdoor attacks against model-agnostic RF fingerprinting systems," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 2029–2042, Mar. 2025.

[37] T. Zhao, X. Wang, and S. Mao, "Backdoor attacks against deep learning-based massive MIMO localization," in *Proc. GLOBECOM 2023-2023 IEEE Glob. Commun. Conf.*, 2023, pp. 2796–2801.

[38] T. Zhao, N. Wang, Y. Wu, W. Zhang, and X. Wang, "Backdoor attacks against low-earth orbit satellite fingerprinting," in *Proc. IEEE INFOCOM 2024-IEEE Conf. Comput. Commun. Workshops*, 2024, pp. 01–06.

[39] T. Zhao et al., "Stealthy backdoor attack on RF signal classification," in *Proc. IEEE Int. Conf. Comput. Commun. Netw.*, 2023, pp. 1–10.

[40] T. Zheng and B. Li, "Poisoning attacks on deep learning based wireless traffic prediction," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 660–669.

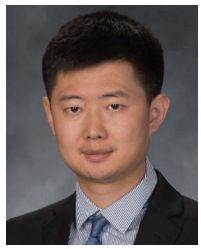
[41] H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen, "LIMU-BERT: Unleashing the potential of unlabeled data for IMU sensing applications," in *Proc. 19th ACM Conf. Embedded Networked Sensor Syst.*, 2021, pp. 220–233.
[42] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[43] H. Qian, T. Tian, and C. Miao, "What makes good contrastive learning on small-scale wearable-based tasks?," in *Proc. ACM SIGKDD Conf. Knowl. Discov. data mining*, 2022, pp. 3761–3771.
[44] E. Eldele et al., "Time-series representation learning via temporal and contextual contrasting," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 2352–2359.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. syst.*, 2017, vol. 30, pp. 1–11.
[47] S. Hanna, S. Karunaratne, and D. Cabric, "Open set wireless transmitter authorization: Deep learning approaches and dataset considerations," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 59–72, Mar. 2021.
[48] D. Raychaudhuri et al., "Overview of the ORBIT radio grid testbed for evaluation of next-generation wireless network protocols," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2005, vol. 3, pp. 1664–1669.
[49] S. Hanna, S. Karunaratne, and D. Cabric, "WiSig: A large-scale WiFi signal dataset for receiver and channel agnostic RF fingerprinting," *IEEE Access*, vol. 10, pp. 22808–22818, 2022.
[50] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.
[51] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, 2019, pp. 113–125.
[52] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-Pruning: Defending against backdooring attacks on deep neural networks," in *Proc. Int. Symp. Res. Attacks, Intrusions, Defenses*, 2018, pp. 273–294.
[53] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–19.
[54] R. Bie, J. Jiang, H. Xie, Y. Guo, Y. Miao, and X. Jia, "Mitigating backdoor attacks in pre-trained encoders via self-supervised knowledge distillation," *IEEE Trans. Serv. Comput.*, vol. 17, no. 5, pp. 2613–2625, Sep./Oct. 2024.

Tianya Zhao (Graduate Student Member, IEEE) received the BS degree in civil engineering from Hunan University, and the MS degree in civil engineering from Carnegie Mellon University. He is currently working toward the PhD degree in computer science with Florida International University. His research interests include AIoT, trustworthy AI, wireless sensing, and smart health.

Junqing Zhang (Senior Member, IEEE) received the BEng and MEng degrees in electrical engineering from Tianjin University, China, in 2009 and 2012, respectively, and the PhD degree in electronics and electrical engineering from Queen's University Belfast, U.K., in 2016. From 2016 to 2018, he was a postdoctoral research fellow with Queen's University Belfast. From 2018 to 2022, he was a tenure track fellow and then a lecturer (assistant professor) with the University of Liverpool, U.K., where he has been a senior lecturer (associate professor) since 2022. His research interests include the Internet of Things, wireless security, physical layer security, key generation, radio frequency fingerprint identification, and wireless sensing. Dr. Zhang is a co-recipient of the IEEE WCNC 2025 Best Workshop Paper Award. He is a senior area editor of *IEEE Transactions on Information Forensics and Security*, and an associate editor for *IEEE Transactions on Mobile Computing*.

Jun Dai (Member, IEEE) received the BS degree in information security and the MS degree in network control from the University of Science and Technology of China, in 2007 and 2010, respectively, and the PhD degree in information sciences and technology from the Pennsylvania State University, in 2014, with specialization in cybersecurity. He is currently an associate professor with the Department of Computer Science, Worcester Polytechnic Institute. He has authored or coauthored papers in prestigious academic venues, such as NDSS, ACM SIGMOD, *IEEE Transactions on Information Forensics and Security*, and ACM SIGCSE. His main research interests include intersections of network and distributed system, AI, and cybersecurity, with recent focus on intrusion detection, vulnerability analysis, secure programming, and cybersecurity education. Dr. Dai is the workshop chair of ACM CCS 2023, and has been a reviewer for top journals, such as *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Vehicular Technology*, and *IEEE Transactions on Mobile Computing*. His projects are mainly funded by NSF and other grant agencies.

Xiaoyan Sun (Member, IEEE) received the PhD degree in information sciences and technology from The Pennsylvania State University, in 2016, with emphasis on cybersecurity. She is currently an associate professor with the Department of Computer Science, Worcester Polytechnic Institute, and also the vice president of Silicon Valley Cybersecurity Institute, a non-profit organization that promotes cybersecurity research and education. Her research interests include system/network security, digital forensics, AI security, and secure programming. Her work has been supported by the National Science Foundation, National Security Agency, and National Institute of Standards and Technology.



Xuyu Wang (Member, IEEE) received the BS degree in electronic information engineering and the MS degree in signal and information processing from Xidian University, Xi'an, China, in 2009 and 2012, respectively, and the PhD degree in electrical and computer engineering from Auburn University, Auburn, AL, USA, in 2018. He is currently an assistant professor with the Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL, USA. His research interests include wireless sensing, Internet of Things, wireless localization, smart health, wireless networks, and deep learning. He was the recipient of the NSF CRII Award in 2021. He was a co-recipient of the ACM FAcct 2023 Best Paper Award, 2022 Best Journal Paper Award of IEEE ComSoc eHealth Technical Committee, IEEE INFOCOM 2022 Best Demo Award, IEEE ICC 2022 Best Paper Award, IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, IEEE GLOBECOM 2019 Best Paper Award, IEEE ComSoc MMTC Best Journal Paper Award in 2018, IEEE PIMRC 2017 Best Student Paper Award, and IEEE SECON 2017 Best Demo Award. He is an associate editor for *IEEE Transactions on Mobile Computing*.

A Geometric Algebra-Informed 3DGS Framework for Wireless Channel Prediction

Jingzhou Shen Tianya Zhao Xuyu Wang*
Knight Foundation School of Computing and Information Sciences
Florida International University
{jshen020, tzhao010, xuywang}@fiu.edu

Abstract

In this paper, we introduce Geometric Algebra–Informed 3D Gaussian Splatting (GAI-GS), a framework for wireless modeling that couples 3D Gaussian splatting with a geometric algebra–based attention mechanism to explicitly model ray–object interactions in complex propagation environments. GAI-GS encodes joint spatial–electromagnetic (EM) relations into token representations, enabling scene-level aggregation within a unified, end-to-end neural architecture. This design grounds wireless ray propagation in electromagnetic principles, allowing token interactions to model key effects such as multipath, attenuation, and reflection/diffraction. Through extensive evaluations on multiple real-world indoor datasets, GAI-GS consistently surpasses current baselines across various wireless tasks.

1. Introduction

Modern society has witnessed an unprecedented integration of connected devices into every aspect of daily life. From smart sensors and wearable technology to autonomous systems and Internet of Things (IoT) devices, wireless communication forms an intricate web that underpins critical infrastructure and personal conveniences. This transformation has elevated wireless channel modeling, which characterizes electromagnetic wave propagation in diverse environments, to a fundamental challenge in telecommunications. Wireless channel modeling captures complex phenomena such as signal attenuation, reflection, diffraction, and scattering [27, 35], providing essential insights for effective network design, resource allocation, wireless localization, and quality of service optimization in increasingly dense and heterogeneous wireless networks [34, 36].

At the core of wireless communications lies the physics of electromagnetic wave propagation governed by Maxwell's equations [28, 41]. Directly solving these equations in realistic environments is intractable due to incomplete boundary conditions and complex geometries [2], motivating approximate modeling strategies. Classical approaches fall into probabilistic, deterministic, and, more recently, neural modeling [10–12]. Probabilistic models use empirical statistics to relate received signal strength to distance and a few coarse parameters; they are efficient but provide limited spatial detail and cannot accurately resolve angle-of-arrival distributions. Deterministic models leverage physical optics and CAD-like environment descriptions to generate richer propagation characteristics [26], yet still struggle to capture fine-grained material and structural complexity in real-world scenes [13].

On the other hand, machine learning models [1, 15, 40] bypass rigid statistical assumptions and simplified electromagnetic approximations, instead inferring relationships between scene geometry and signal behavior from data [3, 4, 23]. Neural radiance fields (NeRF) extend this idea by learning continuous volumetric functions that map spatial coordinates to propagation-related quantities under measurement supervision [22]. NeRF^2 [44] adapts this framework to wireless channels by jointly encoding geometry and signal characteristics, while NeWRF [21] incorporates electromagnetic priors into volumetric rendering to enhance spatial consistency. Recent works further apply NeRF-style models to wireless field reconstruction and generalizable channel prediction [6, 14, 31, 45]. Despite their accuracy, NeRF-based approaches remain computationally demanding for real-time or large-scale deployment. 3D Gaussian Splatting (3D-GS) [7, 16, 17, 42] represents a scene as an explicit set of anisotropic 3D Gaussians, enabling high quality and real-time view synthesis. Current 3D-GS models [37, 38] in the wireless domain tackle the challenge of accurately and efficiently reconstructing high-resolution spatial channel characteristics from sparse measurements in complex environments, enabling fast, site-specific wireless digital twins and downstream tasks.

However, existing wireless 3D-GS methods treat signal propagation as purely data-driven regression and over-

*The corresponding author is Xuyu Wang (xuywang@fiu.edu).

look critical physical interactions between electromagnetic rays and environmental geometry. These approaches directly learn the mapping from spatial coordinates to signal strength without explicitly modeling ray-object interactions such as reflection, refraction, and diffraction at material boundaries. By neglecting the geometric properties of obstacles and their electromagnetic characteristics, these methods fail to capture the fundamental physics governing wave propagation. Therefore, we propose GAI-GS, a novel multi-view framework that effectively integrates geometric algebra (GA) [9, 29] with Euclidean algebra. We introduce a specialized tokenizer with multiple algebraic embeddings to capture ray–object interactions from local to global scales. To better reflect how scene-level context is formed from the Gaussian representation, we explicitly describe the tokenizer instantiation using a subset of high-opacity Gaussian primitives as representative anchors. In addition, we clarify how the outputs of the scene mapping network are incorporated into Gaussian attributes through residual parameterization, enabling the learned representations to adapt to transmitter-dependent propagation conditions. Our contributions can be summarized as follows:

- We propose the first geometric algebra–based 3D-GS framework for wireless channel modeling. By leveraging the unique mathematical structure of geometric algebra to capture local scattering patterns, our framework provides a comprehensive representation of electromagnetic wave propagation characteristics.
- We design a unified multi-view embedding architecture that combines Euclidean and geometric algebra representations. By implicitly learning ray-object interaction patterns along propagation trajectories, our approach encodes physically meaningful signal characteristics while leveraging the complementary strengths of geometric and Euclidean representations for improved performance.
- Our method achieves superior performance in multiple wireless datasets, outperforming existing baselines. Additionally, we release a custom-built dataset to support and advance future research in this domain. The dataset is available at: https://huggingface.co/datasets/NorahCS/GAT-series_Dataset.

2. Preliminaries

2.1. Wireless Signal Representation

In wireless communication systems, the propagation environment between transmitter (Tx) and receiver (Rx) introduces complex distortions to the transmitted signal. The baseband transmit signal is represented in complex form as:

$$X = Ae^{j\theta}, \tag{1}$$

where A and θ denote the signal amplitude and phase, respectively.

When electromagnetic waves encounter obstacles in realistic environments, they undergo reflection, diffraction, and scattering, giving rise to multiple propagation paths. The composite received signal emerges as the coherent superposition of these multipath components:

$$Y = X \cdot \sum_{l=0}^{L-1} \alpha_l e^{j\phi_l}, \tag{2}$$

where L is the number of distinct paths, each characterized by attenuation α_l and phase shift ϕ_l.

The Received Signal Strength Indicator (RSSI) summarizes the aggregate received power as a scalar measurement:

$$\mathrm{RSSI} = 10\log_{10}\big(||Y||^2 / P_0\big), \tag{3}$$

where P_0 is the reference power. It reflects the combined effect of path loss, shadowing, and multipath fading.

The spatial distribution of received power is described by the angle–power spectrum $\Psi(\alpha, \beta)$, which quantifies the relative power arriving from azimuth angle α and elevation angle β. $\Psi(\alpha, \beta)$ is obtained by evaluating the beam-steered relative power directly on a dense discrete angular grid:

$$\Psi(\alpha, \beta) = \left|\mathbf{a}^H(\alpha, \beta)\,\mathbf{y}\right|^2, \tag{4}$$

where $\mathbf{a}(\alpha, \beta)$ denotes the array steering vector and $\mathbf{y}$ is the received signal vector. The angular grid is typically sampled at 1° resolution, yielding $N = 360 \times 90$ discrete points. The finite antenna aperture limits the achievable angular resolution, so the spectrum smoothness is governed by the array response and grid sampling.

2.2. Geometric Algebra

GA, rooted in Clifford's framework, extends classical vector spaces into a single computational language that handles higher-dimensional geometric primitives and their transformations within one coherent calculus. Instead of switching among matrices for rotations and quaternions for orientations, GA offers a coordinate-free algebra in which transformations arise through multiplication, unifying representation and computation of geometry [9, 29].

We adopt the space-time algebra $\mathcal{G}_{3,0,1}$ with three spatial and one temporal dimension. Its elements form a graded structure: grade 0 scalars, grade 1 vectors, grade 2 bivectors, grade 3 trivectors, and grade 4 pseudoscalars. Each grade corresponds to a geometric entity such as points, lines, planes, volumes, and oriented hypervolumes, and the algebraic degree aligns with geometric dimensionality.

The central operation is the geometric product of vectors $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{v}_1\mathbf{v}_2 = \mathbf{v}_1 \cdot \mathbf{v}_2 + \mathbf{v}_1 \wedge \mathbf{v}_2, \tag{5}$$

which splits into a symmetric inner product that yields scalar projection and an antisymmetric wedge product that

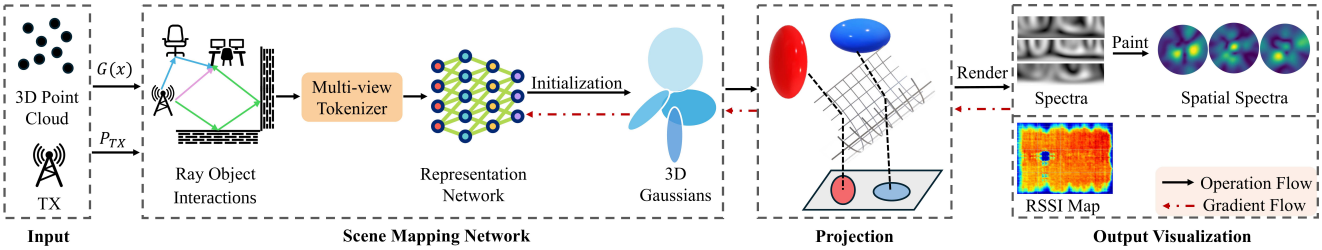


Figure 1. GAI-GS structure. The tokenizer encodes interaction-aware representations from Gaussian primitives and transmitter context.

produces an oriented plane. We use an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ in which the spatial basis elements square to $+1$ and the temporal element squares to -1, giving the Minkowski signature and allowing Lorentz boosts and spatial rotations to be expressed without matrix exponentials. Recent neural architectures that incorporate GA report stronger equivariance and improved representational efficiency, reinforcing its suitability as a backbone for learning geometric transformations [5, 30, 33].

3. Related Work

3D-GS in Wireless Fields. A growing line of work adapts 3D-GS from vision to wireless frequency (RF), treating the wireless environment as a radiance field that can be learned from sparse measurements and rendered at arbitrary transceiver poses. RF-3DGS reconstructs a radio radiance field and renders spatial spectra within milliseconds after training, further exposing spatial-channel state information (CSI) of dominant paths from sparse samples, demonstrating site-specific channel modeling advantages over empirical and ray-tracing baselines [43]. Building on this direction, WRF-GS formulates wireless radiation field reconstruction with 3D-GS and introduces a physics-augmented variant WRF-GS+ that improves RSSI and CSI prediction while preserving real-time synthesis, highlighting the benefit of coupling electromagnetic (EM) priors with explicit Gaussian primitives [37]. Finally, GSRF extends 3D-GS to complex-valued RF fields via a Fourier–Legendre basis and RF-customized CUDA kernels, synthesizing RSSI, spatial spectra, and complex CSI with markedly lower training and inference cost than NeRF-style baselines [39].

4. Framework

4.1. Overview

We introduce GAI-GS, a geometric algebra–informed Gaussian-splatting framework for wireless channel modeling in Fig. 1. The method injects explicit 3D scene geometry and device poses into the learning pipeline so the network internalizes ray–object interactions rather than treating them as black-box correlations. Our framework consists of two main components: (i) a *Scene Mapping Network*, (ii) a *Projection and Render Module*. Initially, the wireless measurements and the initialized 3D point cloud are sent into the Scene Mapping Network [24] to represent the virtual transmitters for a set of 3D Gaussians, along with the attenuation and signal properties. Next, the Projection module projects the virtual transmitters onto the RX antenna plane using the Mercator projection, and then renders the projected 2D Gaussians under EM propagation constraints, aggregating distributed interactions into a unified spatial–frequency representation.

Specifically, inspired by the classification tokens used in large language and multimodal encoders [8, 18, 19], we design a multi-view tokenizer that implicitly converts a set of ray-object interactions into a global scene token. The global token aggregates scene-level context from Gaussian primitives. This tokenizer implicitly learns ray–object interactions and captures propagation characteristics in the wireless environment.

Elementary interactions are parameterized with rotors in geometric algebra, denoted R_r, D, T, and R_0 for reflection, diffraction-like bending, transmission, or refraction, respectively, and initial alignment to the scene frame. Each rotor acts on a ray state vector through a sandwich product, and complex paths arise by composition. Let $\mathbf{V}$ encode the state of the ray, such as its direction, wave vector, or signal attributes. The cumulative effect of multiple interactions is

$$\mathbf{V}' = \boldsymbol{I}\mathbf{V}\boldsymbol{I}^{-1}, \tag{6}$$

where the versor $\boldsymbol{I}$ is the learned product of the relevant rotors selected by attention. The rotors are produced implicitly by the geometric encoder from the multi-view tokens and are optimized end-to-end under wireless supervision. This unified algebraic representation captures multi-bounce, multi-effect propagation within a single differentiable mechanism and removes the need for separate specialized modules for reflection, refraction, and diffraction.

4.2. Multi-view Tokenizer

Geometric Algebra Transformer. We adopt the geometric algebra Transformer (GATr) [5] as the encoder to ex-

tract global embeddings before feeding data into our model. Within the geometric algebra space $\mathbb{G}_{3,0,1}$, geometric transformations such as rotations, reflections, diffractions, and transmissions can be compactly expressed using sandwich products, i.e., $V' = IVI^{-1}$ where I is a multivector-valued interaction operator. This algebraic framework provides a compact and physically consistent way to represent ray–object interactions in wireless propagation and allows us to encode these interactions directly in the feature space rather than relying on explicit geometric annotations.

In realistic wireless environments, ray–object interactions are highly complex: a single received ray often results from multiple reflections, edge diffractions, and penetrations through heterogeneous materials, with interactions occurring at unknown surface locations and orientations. Classical models, such as ray tracing, require detailed information about scene geometry, material properties, and precise collision points in order to explicitly construct each interaction operator. This dependency makes large-scale modeling cumbersome and most neural wireless models therefore either ignore explicit ray–object structures or approximate them with hand-crafted features. In contrast, our approach is the first to implicitly model in-scene ray–object interactions, using learned geometric algebra operators to guide wireless scene representation learning without requiring material labels or explicit interaction locations.

In wireless environments, EM rays interact with surfaces following geometric algebraic rules. For a surface reflection, the incident ray $\mathbf{x}$ and surface normal $\hat{n}$ yield the reflected ray via a sandwich product $\mathbf{x}' = -R\mathbf{x}R^{-1}$, where R encodes the rotation associated with the reflection and corresponds to the operators R_r and R_0. Edge diffraction can be represented analogously as $\mathbf{x}' \approx D\mathbf{x}D^{-1}$ using a diffraction operator D, where the approximation arises because diffraction alters amplitude, phase, and spatial energy distribution in a nonlinear manner that cannot be captured by a purely geometric mapping alone. Material penetration can be expressed deterministically as $\mathbf{x}' = T\mathbf{x}T^{-1}$ with T encoding the transmission effect of a given material interface. A full ray path with n successive interactions, reflections, diffractions, and transmissions, can be represented as a sequential composition:

$$V' = I_1 I_2 \cdots I_n V I_n^{-1} \cdots I_2^{-1} I_1^{-1} = IVI^{-1}, \quad (7)$$

where each I_i denotes an individual ray–object interaction and $I = \prod_{i=1}^{n} I_i$ is the aggregate interaction operator for that path. This formulation indicates that a physically consistent ray trajectory is completely determined by its cumulative geometric algebra operator I, which we learn implicitly from data.

The geometric algebra attention mechanism used in GATr exhibits a structural correspondence with these physical transformations. Let q, k, and v denote input tensors with n_c channels. Standard dot-product attention aggregates information as:

$$\text{Attention}(q,k,v)_{i'c'} = \sum_i \text{Softmax}_i\left(\frac{\langle q_{i'c}, k_{ic}\rangle}{\sqrt{8n_c}}\right) v_{ic'}, \quad (8)$$

which can be interpreted in a sandwich-product form:

$$\text{Attention}(q,k,v)_{i'c'} = \sum_i A_{i'} v_{ic'} A_{i'}^{-1}, \quad (9)$$

where $A_{i'}$ is a multivector-valued operator constructed from the attention weights for query index i', the indices i and i' denote tokens, and c and c' index channels. Under this view, each $A_{i'}$ acts as a learned interaction operator, analogous to I above, that transforms value features $v_{ic'}$ into a representation consistent with the aggregate effect of all paths contributing to token i'.

By embedding geometric algebra directly into the attention computation, GATr enforces rotational and reflectional equivariance and aligns the learned feature space with fundamental EM propagation symmetries. As a result, the encoder can implicitly infer complex, multi-bounce ray–object interactions from data, without explicit knowledge of material types or collision locations.

Multi-view Tokenizer. We first use the GATr to extract tokens that encode ray–object interaction patterns in the scene, capturing how rays are rotated, reflected, diffracted, and attenuated as they propagate.

A naive implementation would feed all N Gaussian positions into GATr, which incurs quadratic attention cost when N is large. To improve efficiency, we instantiate the tokenizer using a subset of Gaussians by selecting the top-M highest-opacity primitives as anchors, where $M \ll N$. These anchors typically correspond to geometrically salient regions such as walls, obstacles, and strong reflectors. Since the Gaussian representation evolves during training, the anchor subset is updated accordingly to remain consistent with the current scene representation. The *CLS* output is broadcast to all N Gaussians and serves as a global scene-level representation that aggregates dominant geometric and propagation context from the selected Gaussian anchors together with the transmitter-conditioned positional embeddings. In particular, it encodes the collective effect of ray–object interactions within the scene, including reflections, diffractions, and attenuation patterns, as captured by the geometric algebra attention. This shared representation provides a unified scene context that is subsequently used by both the attenuation and signal branches to ensure consistent propagation modeling across all Gaussian primitives. This reduces the complexity from $O(N^2)$ to $O(M^2)$, making the encoder agnostic to the total Gaussian count while preserving geometric expressiveness.

In parallel, we derive Euclidean position embeddings that preserve metric structure such as absolute locations, relative distances, and large-scale layout. Concatenating these two streams yields a unified multi-view embedding that combines interaction-aware features from geometric algebra with geometry-aware features from Euclidean space. This complementary representation enables the network to distinguish scenes sharing similar transmitter configurations yet differing in intermediate interactions, and to resolve local interaction patterns that positions alone leave underconstrained. The resulting embedding encourages the network to learn wireless scene representations anchored in both physical ray behavior and global spatial structure, improving data efficiency and robustness to layout changes.

4.3. Mapping, Projection and Render Module

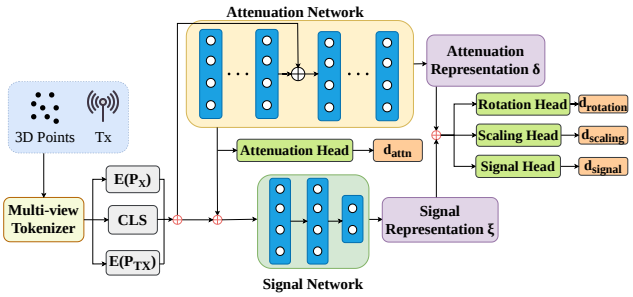


Figure 2. Scene mapping model overview.

Scene Mapping Network. Fig. 2 depicts the structure of our scene mapping network. The representation network adopts a dual-network design with two complementary branches that decompose wireless propagation into distinct physical processes: attenuation and signal representations. The attenuation network predicts spatially varying decay coefficients that characterize the progressive weakening of EM waves as they traverse the scene. This branch embeds material- and medium-dependent effects and encodes the inherent attenuation properties of the volumetric environment, yielding a 3D field that maps position to expected signal extinction. In parallel, the signal network reconstructs the scattered-field distribution induced by interactions with scene geometry. It explicitly models non-line-of-sight propagation to capture signal characteristics.

Formally, $P_x \in \mathbb{R}^3$ denotes the center coordinate of the 3D Gaussian at location x, and $P_{TX} \in \mathbb{R}^3$ denotes the transmitter location. First, the multi-view tokenizer F_{mv} encodes those inputs into a set of positional embeddings E and a global scene token *CLS*:

$$CLS, E(P_{\text{x}}), E(P_{\text{TX}}) = F_{\text{mv}}(P_{\text{x}}, P_{\text{TX}}). \tag{10}$$

Then the global token *CLS* is concatenated to the positional embeddings of the P_{TX} and P_{x} to generate multi-view Tx tokens. Here, *CLS* provides a shared scene-level context that captures the global interaction structure of the environment and guides both attenuation and signal prediction. Based on this representation, the attenuation network F_{att} predicts a scalar attenuation field $\delta(\mathbf{x})$ and an intermediate geometric feature f at each position:

$$\delta(\mathbf{x}),\ f = F_{\text{att}}(E(P_{\text{TX}}), E(P_{\text{x}}), CLS). \tag{11}$$

Subsequently, the signal network F_{sig} predicts the signal strength conditioned on the intermediate feature, the *CLS* token, and the embeddings of the Tx and Gaussian point:

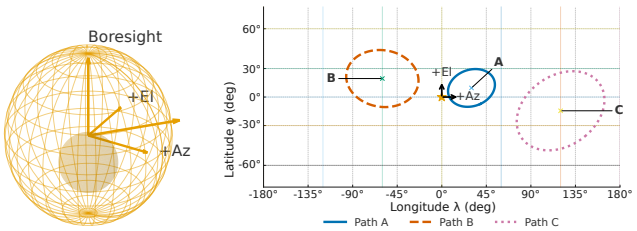
$$\xi(\mathbf{x}) = F_{\text{sig}}\big(f, E(P_{\text{TX}}), E(P_{\text{x}}), \text{CLS}\big), \tag{12}$$

yielding a scattered-field amplitude that captures indirect paths such as reflections and diffuse scattering. Finally, we concatenate the attenuation feature and the signal representation to form a joint feature vector, which is then passed through three dedicated MLP heads, namely a Rotation Head, a Scaling Head, and a Signal Head. The Rotation and Scaling Heads produce residual updates, d_{rotation} and d_{scaling}, which are applied to the original Gaussian rotation and scaling parameters, respectively, allowing the network to model geometric deformations in a residual manner rather than regressing absolute values. For the signal branch, the Signal Head operates in the spherical harmonics (SH) coefficient space, producing a residual d_{signal} that is added to the original SH coefficients of each Gaussian to obtain the updated signal representation: $\tilde{\xi}(x_i) = \xi(x_i) + d_{\text{signal},i}$. This residual parameterization ensures that the network learns deviations from the canonical Gaussian parameters, which facilitates stable training and preserves the structural priors encoded in the original 3D Gaussians.

In standard 3D Gaussian Splatting, each Gaussian primitive carries a fixed opacity α_i that is invariant to the query condition. However, in wireless propagation, the effective attenuation of a spatial region is transmitter-dependent: an obstacle may fully occlude the signal from one transmitter while remaining largely transparent to another. To account for transmitter-dependent attenuation, we use the attenuation intermediate feature to parameterize an opacity adjustment for each Gaussian. Concretely, the effective opacity is written as $\tilde{\alpha}_i = \alpha_i + d_{\text{attn},i}$, where $d_{\text{attn},i}$ is a learned residual conditioned on the geometric feature f. An L_2 penalty on d_{attn} regularizes this adjustment.

We implement both the attenuation and signal networks with MLPs. For the attenuation network, we further incorporate residual connections to stabilize training and preserve geometric information.

Projection and Render Module. We follow the projection and rendering principles introduced in WRF-GS [37] and WRF-GS+ [38] to process the 3D Gaussians. Specifically, the 3D Gaussian representations are projected onto



(a) 3D Gaussian on AoA sphere.

(b) Mercator projection of angular Gaussians on RX perception plane.

Figure 3. Mercator projection module, where E_l, A_z, and AoA denote elevation, azimuth, and angle of arrival, respectively.

the perception plane of the RX antenna array using the Mercator projection. As shown in Fig. 3a, the spherical coordinate system captures the antenna's field of view through azimuth and elevation angles, denoted as Az and El respectively, which are measured relative to the boresight direction. The shaded region represents the antenna's effective coverage area containing the Gaussian distributions.

Fig. 3b illustrates the Mercator projection used to map 3D Gaussian lobes onto the RX perception plane. Longitude λ is mapped linearly to the horizontal axis in $[-180°, 180°]$, while latitude ϕ is mapped nonlinearly to vertical coordinates in $[-60°, 60°]$, preserving local angles and shapes and thus antenna directivity. Formally, the Mercator mapping from spherical angles (λ, ϕ) to 2D perception-plane coordinates (u, v) is given by

$$u = \lambda, \qquad v = \alpha \log \tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right), \tag{13}$$

where u and v denote the horizontal and vertical coordinates on the RX perception plane, respectively, and α is a scaling factor selected such that $v \in [-60°, 60°]$. Three representative paths highlight the transformation: Path A (blue ellipse) shows that an off-center Gaussian lobe preserves its elliptical structure; Path B (orange dashed circle) depicts a symmetric distribution at an off-center position; and Path C (pink dotted circle) illustrates how Gaussians are mapped at different angular positions. This angle-preserving mapping allows 3D Gaussian primitives to retain their geometric relationships in the 2D perception plane, enabling faithful representation of spatial signal characteristics.

Then, projected 2D Gaussian primitives are binned into tiles for massively parallel processing. Each tile operates independently on the primitives that intersect its spatial extent; primitives spanning multiple tiles are replicated so each tile has a complete local set. Within a tile, primitives are sorted in depth to preserve physically consistent signal accumulation.

In the RF setting, each projected Gaussian primitive acts as a virtual transmitter characterized by an updated signal representation $\tilde{\xi}(x_i)$ and contributes to the attenuation field through its effective opacity. Specifically, the contribution of the i-th primitive at angular coordinate x is defined as:

$$S_i(x) = \tilde{\xi}(x_i) \prod_{j=1}^{i-1} \delta(x_j). \tag{14}$$

The final received signal at pixel k is obtained by accumulating the contributions from all projected virtual transmitters:

$$R_k = \sum_{i=1}^{N} S_i(x_i)\, \tilde{\alpha}_i \prod_{j=1}^{i-1} (1 - \tilde{\alpha}_j), \tag{15}$$

where $\tilde{\alpha}_i = \alpha_i + d_{\text{attn},i}$ denotes the effective opacity of the i-th Gaussian.

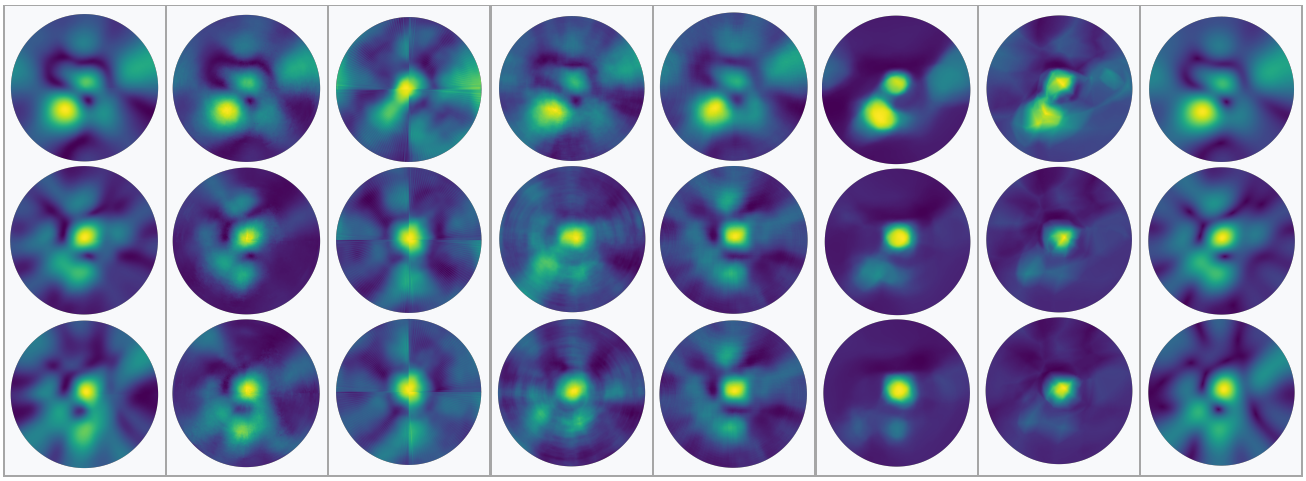
The resulting spatial power map approximates the wireless radiation field and provides actionable channel information for system design and optimization. Additionally, for RSSI map construction, we standardize the RSSI map by subtracting its mean, scaling with a temperature threshold, and stabilizing by shifting with the map's maximum; a softmax over all pixels then produces attention weights blended with a uniform prior to ensure coverage. The final scalar RSSI map is obtained as the attention-weighted average of the map, which emphasizes strong-signal regions while remaining robust to noise and outliers.

5. Experimental Evaluation and Analysis

5.1. Datasets

We built a semi-automated platform to map indoor RSSI at 2.4 GHz and 5 GHz over an approximately 35m^2 site during an 18-day campaign. As shown in Fig. 6a, a TurtleBot 4 carrying a Raspberry Pi 4 measurement unit was equipped with an RPLIDAR A1M8 and the Nav2+AMCL stack, providing real-time SLAM, an occupancy grid, and precise (x, y, z) poses for spatially referenced measurements. The Pi device hosted a dual-band TP-Link Archer T3U AC1300 adapter, while two ASUS RT-AC86U routers provided infrastructure: one acted as the transmitter, fixed at 1 m height in dual-band mode with maximum power, and the other maintained system communication on non-overlapping channels. Data collection combined joystick navigation to predefined waypoints with automated acquisition: at each location we logged the pose and recorded five RSSI samples per band at 1 s intervals, accepting a measurement only if all five reads were successfully obtained. We consider two measured rooms and their reference maps, both of which contain structural columns that create clear occlusions useful for analyzing ray–object interactions. Fig. 6b shows the overview of two different rooms.

In addition, we evaluate on the public BLE and the RFID spectrum datasets from the NeRF^2 [44] repositories. In the BLE dataset, each sample is a 50-dimensional RSSI vector



(a) Ours (b) WRF-GS (c) DCGAN (d) NeRF^2 (e) NeRF-APT (f) MLP (g) FIRE (h) Label

Figure 4. 2D spatial spectrum visualizations.

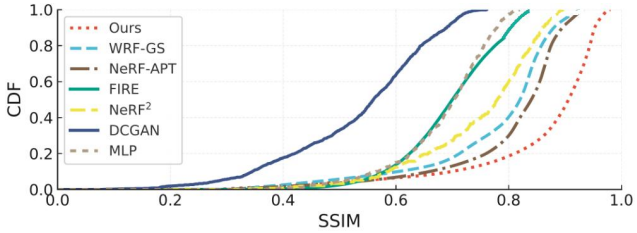
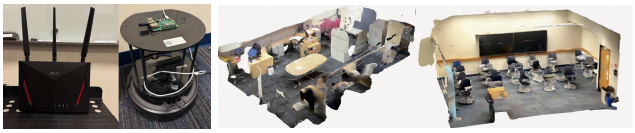


Figure 5. CDF-SSIM comparison.



(a) Platforms. (b) Room 1 and Room 2 overview.

Figure 6. Overview of the platforms and room configurations.

from the gateways plus the tag position. We use 4.2k samples for training and 1.8k for testing. In the RFID spectrum dataset, array snapshots are converted to a front-hemisphere spatial spectrum on a 360×90 azimuth–elevation grid.

5.2. Baselines

We benchmark against several primary baselines and one ray-tracing variant. First, we use MATLAB's Ray Tracing toolbox to obtain the corresponding results [41]. Specifically, the toolbox requires a 3D scene model as input and predicts the RF signals at the Rx locations given the Tx positions. Second, an MLP with four hidden layers followed by a linear output head. Third, we use FIRE [20] with a three-layer encoder and a symmetric four-layer decoder to evaluate on our proposed datasets. Fourth, a DCGAN network [25] in which both generator and discriminator have four layers. Next, the standard NeRF^2 [44] configuration implemented as in the original release. For the sixth baseline evaluation, we use the NeRF-APT [32] which is a variant of NeRF^2. Furthermore, we use the state-of-the-art model WRF-GS [37] as our final baseline.

5.3. Results

We use a combination of the mean absolute error (MAE) and the structural similarity (SSIM) loss between predicted results and labels, with an L_2 penalty on d_{attn} for the spectrum dataset:

$$L = \frac{1}{M}\sum_{i=1}^{M}\big(\beta L_{\text{MAE}}(I_{gt}, I_{pred}) + (1-\beta)L_{\text{SSIM}}(I_{gt}, I_{pred})\big) + \alpha L_2(d_{attn}), \quad (16)$$

where I_{gt}, I_{pred} are the ground truth and synthesized spatial spectrum, while M is the number of measurements and β and α are weight parameters. As for the BLE datasets, measurements corresponding to receiver locations at excessive distances from the transmitter are treated as invalid and excluded from our analysis, as these instances yield degenerate RSSI recordings of -100 dBm. The remaining valid measurements are subsequently partitioned into a training/evaluation split. The reported metric, expressed in dB, represents the median MAE aggregated across all receivers in the test set.

We compare training, inference, and rendering times against prior 3DGS-based wireless models on a single NVIDIA A100 GPU using a subset of the Spectrum dataset, shown in Table 2. Despite requiring additional computation during training and inference, our method provides better reconstruction quality while also outperforming prior methods in rendering speed.

Table 1. Performance comparison of GAI-GS against baseline methods across different room configurations and BLE-RSSI/RFID spectrum datasets. Lower MAE values and higher SSIM values indicate better performance.

Method	Room 1 (Our RSSI Dataset)		Room 2 (Our RSSI Dataset)		Other Dataset	
	MAE (dB) ↓		MAE (dB) ↓		MAE (dB) ↓	SSIM ↑
	2.4 GHz	5.0 GHz	2.4 GHz	5.0 GHz	BLE	Spectrum
Ray Tracing [41]	25.52	20.66	25.56	20.78	–	0.33
MLP	7.3	9.3	8.2	9.9	8.0	0.71
FIRE [20]	5.8	5.5	4.5	2.7	6.4	0.73
DCGAN [25]	4.0	3.4	4.2	3.0	4.6	0.56
NeRF^2 [44]	3.6	2.9	3.9	2.0	3.1	0.78
NeRF-APT [32]	3.3	2.7	3.6	2.0	3.1	0.84
WRF-GS [37]	3.1	2.4	3.1	1.9	2.8	0.82
GAI-GS	**1.9**	**1.6**	**2.7**	**1.8**	**2.3**	**0.91**

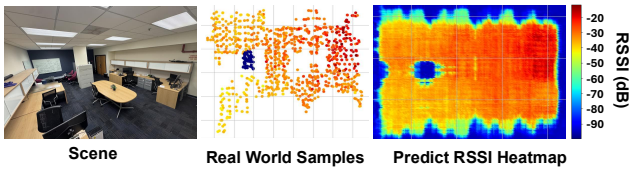


Figure 7. Visualizations of RSSI prediction.

Table 1 summarizes the quantitative results across all datasets and room configurations. GAI-GS consistently achieves the lowest MAE and the highest SSIM among all compared methods. On our RSSI dataset in Room 1, GAI-GS reduces the MAE to 1.9 dB at 2.4 GHz and 1.6 dB at 5.0 GHz, improving over the strongest baseline WRF-GS by 1.2 dB and 0.6 dB, respectively. In Room 2, GAI-GS achieves 2.7 dB and 1.8 dB at the two frequency bands, surpassing WRF-GS by 0.4 dB and 0.1 dB. Among the baselines, classical methods such as Ray Tracing and MLP show limited capacity with MAE values exceeding 7.0 dB. Learning-based approaches progressively improve, with FIRE reaching 2.7–5.8 dB, DCGAN 3.0–4.2 dB, and neural radiance field methods narrowing the gap to 2.0–3.9 dB. GAI-GS outperforms all of them by a clear margin, demonstrating the effectiveness of our GA-informed interaction modeling. On the external BLE-RSSI dataset, GAI-GS attains an MAE of 2.3 dB, outperforming WRF-GS at 2.8 dB and NeRF-based methods at 3.1 dB. For spectrum reconstruction, GAI-GS reaches an SSIM of 0.91, a substantial improvement over NeRF-APT at 0.84 and WRF-GS at 0.82. These results confirm that GAI-GS achieves superior spatial accuracy and perceptual fidelity across all evaluated settings.

Fig. 4 presents the 2D spatial spectrum visualizations comparing our method with other baselines. MLP and FIRE exhibit blurred and distorted spectral patterns, indicating a loss of spatial coherence. WRF-GS achieves relatively better results but fails to preserve high-frequency structural details. In contrast, our method produces the most accurate and physically consistent spectra, closely matching the ground-truth label. The reconstructed spectra exhibit sharp high-energy focal regions and fine-grained propagation textures that reflect realistic wave behaviors. This demonstrates that our GAI-GS approach effectively encodes spatial correlations and physical consistency, leading to more faithful recovery of the underlying EM field structure.

Fig. 5 shows the cumulative distribution function (CDF) of the SSIM for different methods. A curve shifted toward the upper right indicates better overall perceptual quality and higher structural consistency. Also, Fig. 7 illustrates the visualization of predicted RSSI maps with the samples extracted from the scene.

Overall, these results show that our GAI-GS framework enhances spatial representation fidelity and interaction-aware modeling, leading to more accurate EM propagation modeling.

Method	Training (mins)	Inference (ms)	Render (ms)
WRF-GS [37]	312.38	434.21	39.29
WRF-GS+ [38]	**101.06**	**4.78**	1.43
Ours	203.00	17.37	**0.91**

Table 2. Comparisons of training, inference, and rendering time.

6. Conclusion

In this paper, we propose GAI-GS, a geometric algebra–informed 3D Gaussian splatting framework that unifies geometric algebra and wireless ray-transmission theory for wireless modeling. By coupling a 3D-GS scene representation with physically grounded propagation mechanisms, GAI-GS implicitly captures ray–object interactions and enables more accurate reconstruction of the wireless field. Extensive experiments on both public and in-room datasets demonstrate its effectiveness and consistent gains over baselines. We believe that GAI-GS provides a more principled basis for wireless channel modeling and offers a flexible foundation for scaling 3D-GS–based methods to increasingly complex wireless communication environments.

Acknowledgments

This work is supported in part by the U.S. NSF under grants (CNS-2415209, CNS-2317190, IIS-2306791, and CNS-2319343).

References

[1] Saud Mobark Aldossari and Kwang-Cheng Chen. Machine learning for wireless communication channel modeling: An overview. *Wireless Personal Communications*, 106(1):41–70, 2019. 1

[2] Peter Almers, Ernst Bonek, Alister Burr, Nicolai Czink, Mérouane Debbah, Vittorio Degli-Esposti, Helmut Hofstetter, Pekka Kyösti, David Laurenson, Gerald Matz, et al. Survey of channel and radio propagation models for wireless MIMO systems. *EURASIP Journal on Wireless Communications and Networking*, 2007(1):019070, 2007. 1

[3] Marius Arvinte and Jonathan I. Tamir. Score-based generative models for robust channel estimation. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 453–458, 2022. 1

[4] Arjun Bakshi, Yifan Mao, Kannan Srinivasan, and Srinivasan Parthasarathy. Fast and efficient cross band channel prediction using machine learning. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019. 1

[5] Johann Brehmer, Pim De Haan, Sönke Behrends, and Taco S Cohen. Geometric algebra transformer. *Advances in Neural Information Processing Systems*, 36:35472–35496, 2023. 3

[6] Xingyu Chen, Zihao Feng, Kun Qian, and Xinyu Zhang. Radio frequency ray tracing with neural object representation for enhanced RF modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21339–21348, 2025. 1

[7] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3D gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 1

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, 2019. 3

[9] Leo Dorst and Steven De Keninck. A guided tour to the plane-based geometric algebra pga. *URL https://bivector. net/PGA4CS. html*, 2022. 2

[10] Danping He, Bo Ai, Ke Guan, Longhe Wang, Zhangdui Zhong, and Thomas Kürner. The design and applications of high-performance ray-tracing simulation platform for 5G and beyond wireless communications: A tutorial. *IEEE Communications Surveys & Tutorials*, 21(1):10–27, 2019. 1

[11] Ruisi He, Bo Ai, Andreas F Molisch, Gordon L Stuber, Qingyong Li, Zhangdui Zhong, and Jian Yu. Clustering enabled wireless channel modeling using big data algorithms. *IEEE Communications Magazine*, 56(5):177–183, 2018.

[12] Jie Huang, Cheng-Xiang Wang, Lu Bai, Jian Sun, Yang Yang, Jie Li, Olav Tirkkonen, and Ming-Tuo Zhou. A big data enabled channel model for 5G wireless communication systems. *IEEE Transactions on Big Data*, 6(2):211–222, 2018. 1

[13] Agbotiname Lucky Imoize, Augustus Ehiremen Ibhaze, Aderemi A Atayero, and KVN Kavitha. Standard propagation channel models for MIMO communication systems. *Wireless Communications and Mobile Computing*, 2021(1): 8838792, 2021. 1

[14] Haifeng Jia, Xinyi Chen, Yichen Wei, Yifei Sun, and Yibo Pi. Neural reflectance fields for radio-frequency ray tracing. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*, pages 4226–4231. IEEE, 2024. 1

[15] Chunxiao Jiang, Haijun Zhang, Yong Ren, Zhu Han, Kwang-Cheng Chen, and Lajos Hanzo. Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, 24(2):98–105, 2016. 1

[16] Kaiwen Jiang, Venkataram Sivaram, Cheng Peng, and Ravi Ramamoorthi. Geometry field splatting with Gaussian surfels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5752–5762, 2025. 1

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[20] Zikun Liu, Gagandeep Singh, Chenren Xu, and Deepak Vasisht. FIRE: enabling reciprocity for FDD MIMO systems. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, page 628–641, New York, NY, USA, 2021. Association for Computing Machinery. 7, 8

[21] Haofan Lu, Christopher Vattheuer, Baharan Mirzasoleiman, and Omid Abari. NeWRF: a deep learning framework for wireless radiation field reconstruction and channel prediction. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 1

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[23] Shiva Navabi, Chenwei Wang, Ozgun Y Bursalioglu, and Haralabos Papadopoulos. Predicting wireless channel features using neural networks. In *2018 IEEE international conference on communications (ICC)*, pages 1–6. IEEE, 2018. 1

[24] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[25] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 7, 8

[26] Theodore S. Rappaport, Felix Gutierrez, Eshar Ben-Dor, James N. Murdock, Yijun Qiao, and Jonathan I. Tamir. Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications. *IEEE Transactions on Antennas and Propagation*, 61(4):1850–1859, 2013. 1

[27] Theodore S. Rappaport, George R. MacCartney, Mathew K. Samimi, and Shu Sun. Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design. *IEEE Transactions on Communications*, 63(9):3029–3056, 2015. 1

[28] Theodore S. Rappaport, Yunchou Xing, George R. MacCartney, Andreas F. Molisch, Evangelos Mellios, and Jianhua Zhang. Overview of millimeter wave communications for fifth-generation (5g) wireless networks—with a focus on propagation models. *IEEE Transactions on Antennas and Propagation*, 65(12):6213–6230, 2017. 1

[29] Martin Roelfs and Steven De Keninck. Graded symmetry groups: Plane and simple. *Advances in Applied Clifford Algebras*, 33(3), 2023. 2

[30] David Ruhe, Jayesh K Gupta, Steven De Keninck, Max Welling, and Johannes Brandstetter. Geometric clifford algebra networks. In *International Conference on Machine Learning*, pages 29306–29337. PMLR, 2023. 3

[31] Jingzhou Shen and Xuyu Wang. An efficient and explainable kan framework for wireless radiation field prediction. In *2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, pages 51–59, 2025. 1

[32] Jingzhou Shen, Tianya Zhao, Yanzhao Wu, and Xuyu Wang. NeRF-APT: A new NeRF framework for wireless channel prediction, 2025. 7, 8

[33] Jonas Spinner, Victor Bresó, Pim De Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer. Lorentz-equivariant geometric algebra transformers for high-energy physics. *Advances in neural information processing systems*, 37:22178–22205, 2024. 3

[34] Cheng-Xiang Wang, Ji Bian, Jian Sun, Wensheng Zhang, and Minggao Zhang. A survey of 5G channel measurements and models. *IEEE Communications Surveys & Tutorials*, 20(4): 3142–3168, 2018. 1

[35] Xiangyu Wang, Xuyu Wang, Shiwen Mao, Jian Zhang, Senthilkumar CG Periaswamy, and Justin Patton. Indoor radio map construction and localization with deep Gaussian processes. *IEEE Internet of Things Journal*, 7(11):11238–11249, 2020. 1

[36] Xiangyu Wang, Xuyu Wang, Shiwen Mao, Jian Zhang, Senthilkumar CG Periaswamy, and Justin Patton. Adversarial deep learning for indoor localization with channel state information tensors. *IEEE internet of things journal*, 9(19): 18182–18194, 2022. 1

[37] Chaozheng Wen, Jingwen Tong, Yingdong Hu, Zehong Lin, and Jun Zhang. WRF-GS: Wireless radiation field reconstruction with 3D Gaussian splatting. In *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications*, pages 1–10, 2025. 1, 3, 5, 7, 8

[38] Chaozheng Wen, Jingwen Tong, Yingdong Hu, Zehong Lin, and Jun Zhang. Neural representation for wireless radiation field reconstruction: A 3d gaussian splatting approach. *IEEE Transactions on Wireless Communications*, 25:7490–7504, 2026. 1, 5, 8

[39] Kang Yang, Gaofeng Dong, Sijie JI, Wan Du, and Mani Srivastava. GSRF: Complex-valued 3D Gaussian splatting for efficient radio-frequency data synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[40] Yang Yang, Yang Li, Wuxiong Zhang, Fei Qin, Pengcheng Zhu, and Cheng-Xiang Wang. Generative-adversarial-network-based wireless channel modeling: Challenges and opportunities. *IEEE Communications Magazine*, 57(3):22–27, 2019. 1

[41] Zhengqing Yun and Magdy F. Iskander. Ray tracing for radio propagation modeling: Principles and applications. *IEEE Access*, 3:1089–1100, 2015. 1, 7, 8

[42] Jiahui Zhang, Fangneng Zhan, Muyu Xu, Shijian Lu, and Eric Xing. Fregs: 3D Gaussian splatting with progressive frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21424–21433, 2024. 1

[43] Lihao Zhang, Haijian Sun, Samuel Berweger, Camillo Gentile, and Rose Qingyang Hu. RF-3DGS: Wireless channel modeling with radio radiance field and 3d gaussian splatting, 2025. 3

[44] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. NeRF2: Neural radio-frequency radiance fields. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2023. Association for Computing Machinery. 1, 6, 7, 8

[45] Xiaopeng Zhao, Shen Wang, Zhenlin An, and Lei Yang. Crowdsourced geospatial intelligence: Constructing 3D urban maps with satellitic radiance fields. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(3), 2024. 1

EMPalm: Exfiltrating Palm Biometric Data via Electromagnetic Side-Channel

Haowen Xu[1], Tianya Zhao[2], Xuyu Wang[2], Lei Ma[1], Jun Dai[1†],
Alexander Wyglinski[1], Xiaoyan Sun[1†]
[1]Worcester Polytechnic Institute, USA; [2]Florida International University, USA
Emails:{hxu4,lma5,jdai,alexw,xsun7}@wpi.edu,{tzhao010,xuywang}@fiu.edu

Abstract

Palm recognition has emerged as a dominant biometric authentication technology in critical infrastructure. These systems utilize palm-related biometric features, including palmprint and palmvein data, either individually in a single-modal setting or jointly in a dual-modal. Despite the different forms, they all employ similar hardware architectures that inadvertently emit electromagnetic (EM) signals during operation. Our research reveals that these EM emissions leak palm biometric information, motivating us to develop EMPalm—an attack framework that covertly recovers both palmprint and palmvein images from eavesdropped EM signals. Specifically, we first separate the interleaved transmissions of the visible (palmprint) and NIR (palmvein) modalities, identify the informative frequency bands of each modality, and then combine these bands to reconstruct the corresponding images. To overcome the strong noise and distortions inherent in side-channel acquisition, we further employ a diffusion model to restore fine-grained biometric features. Evaluations on seven prototype and three commercial palm acquisition devices show that EMPalm can recover biometric information from real human palms with high visual fidelity, achieving Structural Similarity Index Measure (SSIM) scores up to 0.79, Peak Signal-to-Noise Ratio (PSNR) up to 29.88 dB, and Fréchet Inception Distance (FID) scores as low as 6.82 across all tested devices. Compared with the best state-of-the-art method, which can only reconstruct palm-vein images, EMPalm improves overall reconstruction fidelity by 33% and uniquely supports high-quality recovery for both palmprint and palm-vein modalities. To assess the practical implications of the attack, we further evaluate the recovered palm images against four state-of-the-art palm recognition models through real-time experiments, achieving a model-wise average spoofing success rate of 65.30%.

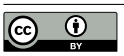
CCS Concepts

• **Security and privacy → Security in hardware**.

Keywords

Electromagnetic Side-Channel Attack, Embedded Palm Biometric Device, Biometric Spoofing

† Corresponding Authors.



SenSys '26, Saint Malo, France

ACM ISBN 979-8-4007-2309-4/2026/05
https://doi.org/10.1145/3774906.3802764

ACM Reference Format:
Haowen Xu, Tianya Zhao, Xuyu Wang, Lei Ma, Jun Dai, Alexander Wyglinski, Xiaoyan Sun. 2026. EMPalm: Exfiltrating Palm Biometric Data via Electromagnetic Side-Channel. In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys '26), May 11–14, 2026, Saint Malo, France.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3774906.3802764

1 Introduction

Palm recognition technologies, encompassing unimodal approaches based on palmprint or palmvein and multimodal methods that fuse the two, have rapidly emerged as highly secure and reliable biometric authentication techniques [15, 17]. In particular, multimodal fusion of palm textures with vascular structures yields high entropy, strong forgery resistance, and lasting physiological stability [14]. Consequently, palm-based authentication has been widely adopted across government and commercial sectors, including the FBI, the Department of Homeland Security, Amazon, and Tencent [46, 58].

Traditional image-based palm recognition systems rely on either palmprint or palmvein imaging, using visible light for palmprint textures and near-infrared (NIR) sensing for subcutaneous veins [3]. Since single-modal approaches are often affected by environmental or physiological factors, modern systems overcome these limitations by adopting dual-mode architectures that capture both features simultaneously [14, 21, 26, 39], thus improving accuracy and robustness. However, in both single- and dual-mode designs, sensor circuits carry time-varying currents that, by Maxwell's equations [41], inevitably emit electromagnetic (EM) radiation. In addition, high-speed transmission of biometric images over buses or flat cables can turn wiring into unintended antennas, exposing sensitive information through EM emissions.

Although prior studies on EM leakage in biometric contexts such as fingerprint sensors [44] and iris recognition [34] have provided valuable insights, EM leakage in palm recognition systems, particularly in dual-modal designs, has received limited attention. This gap is increasingly important as palm recognition is being deployed more widely for secure access control and payment authentication due to its rich biometric features and built-in liveness properties, with adoption extending to national intelligence agencies [12] and major financial institutions [1]. To demonstrate this, we show that biometric image data in palm recognition systems can be eavesdropped via EM side channels. As illustrated in Figure 1a, an eavesdropper can covertly capture EM emissions from a palm scanner and reconstruct palm images as the victim performs identification, while the victim remains unaware. To our best knowledge, EMPalm is the first to investigate EM leakage in palm recognition systems, and introduces the first technique capable of separating

and reconstructing dual-modal biometric streams transmitted in image-based palm recognition systems.

Challenges. An effective eavesdropping of palm recognition systems faces four key challenges.

- *Interleaved Dual-Modal Emissions.* Palmprint and palmvein data can be transmitted in an alternating fashion, producing interleaved emissions that complicate modality separation.
- *Noisy Wide-band Spectrum.* EM emissions span wide and device-dependent frequencies, making it nontrivial to identify biometric-relevant bands.
- *Bit-Level Grayscale Collisions.* Bit-packed formats cause multiple grayscale values to map to identical EM patterns, collapsing subtle intensity differences and fine details.
- *Degraded Image Texture.* Reconstructed images exhibit degraded textures due to EM interference, environmental noise, and information loss during reconstruction.

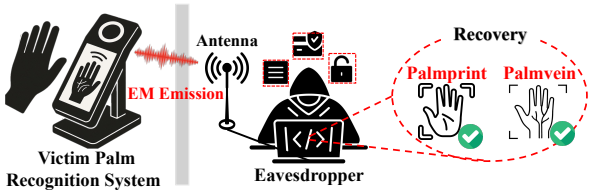
Our Approach. In this paper, we present EMPalm[1], the first EM side-channel eavesdropping attack that recovers both high-quality palmprint and palmvein from palm-recognition systems. Using unintentional EM emissions collected from palm recognition systems, EMPalm recovers preliminary biometric data through a multi-stage reconstruction pipeline. To address the challenge of *Interleaved Dual-Modal Emissions*, we reverse-engineer transmission protocols and implement frame boundary detection, modality classification, and signal disentanglement for synchronized palmprint–palmvein reconstruction. To cope with the *Noisy Wide-band Spectrum*, we design a rapid localization framework that integrates spectrum analysis, temporal profiling, and device characterization to identify informative frequency bands. To resolve *Bit-Level Grayscale Collisions*, we introduce a multi-band image combination strategy that leverages higher-order harmonics to restore collapsed intensity variations and preserve fine details. Finally, to mitigate *Degraded Image Texture*, we formulate the task as image restoration and employ a structure-guided diffusion model to recover high-fidelity palmprint creases and palmvein patterns.

Evaluated on seven prototype and three commercial palm recognition devices with ***real human hands***, EMPalm achieves high-fidelity reconstruction with an average Structural Similarity Index Measure (SSIM) of 0.68, Peak Signal-to-Noise Ratio (PSNR) of 24.1 dB, and Fréchet Inception Distance (FID) of 8.7. Compared with state-of-the-art (SOTA) frameworks [34, 37], EMPalm consistently delivers higher reconstruction quality and visual realism, achieving a 33% improvement in SSIM and enhanced spoofing effectiveness against palm recognition models under identical evaluation conditions. When evaluated against four state-of-the-art practical palm-recognition models, the reconstructed images reach an average spoofing success rate of 65.3%, confirming the practical effectiveness of the recovered biometrics.

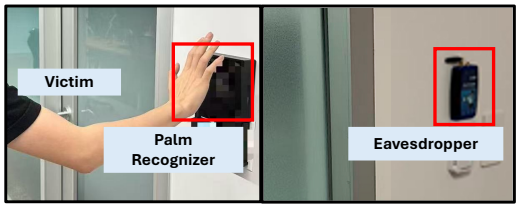
Ethical consideration. This study was approved by the Institutional Review Board (IRB) of the participating institution, ensuring compliance with ethical and privacy standards in volunteer recruitment and data collection. We anonymized all personal information and withheld specific device models to maintain confidentiality and give vendors time to address the identified vulnerabilities.

Contributions. In summary, our contributions are as follows:

[1]EMPalm Project is available at https://github.com/submission695-ai/Submission



(a) Overall attack scenario of EMPalm.



(b) Attacking a commercial palm recognition device using a concealed eavesdropper.

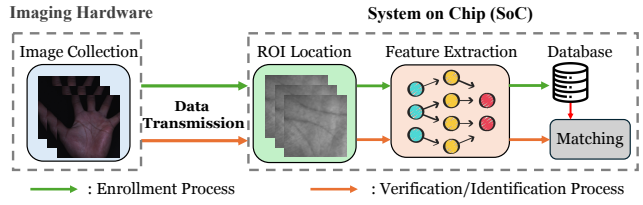
Figure 1: Attack scenarios of EMPalm. (a) illustrates the overall attack setup and workflow, while (b) demonstrates a real-world case where an attacker covertly captures EM emissions from a commercial palm recognition device.

- *EM Side-channel Attack Surface Exploitation.* We first reveal EM leakage in palm biometric recognition, enabling effective spoofing of recognition models and exposing the feasibility of physical attacks.
- *End-to-End Attack Framework.* We propose an end-to-end framework that includes frequency localization, single-band reconstruction, multi-band combination, and diffusion-based restoration, demonstrating robust eavesdropping capability against both single and dual modal palm recognition systems.
- *Comprehensive Experimental Evaluation.* The effectiveness of EMPalm is validated through real-world experiments on human subjects across seven prototype and three commercial palm-acquisition devices, evaluated against four state-of-the-art recognition models. *Single- and dual-modal restoration* demonstrates that intercepted EM emissions can reliably recover both palmprint and palmvein modalities. *Spoofing efficacy (1:100 identification)* shows that reconstructed and diffusion-enhanced images can successfully deceive advanced recognition systems. *Robustness analyses* further confirm attack viability across diverse distances, orientations, intervening materials, and hardware platforms.

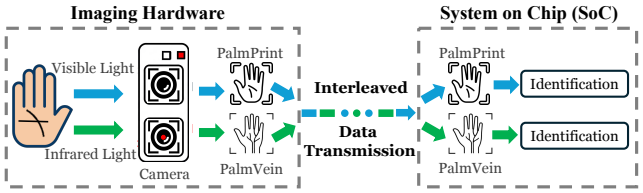
2 Preliminaries

2.1 Image-based Palm Recognition

Figure 2a depicts the standard palm recognition pipeline, including image acquisition, Region of Interest (ROI) localization, feature extraction, and matching. Palm images are first captured by the imaging hardware, after which ROI localization is performed on the System on Chip (SoC) to support reliable feature analysis [9]. Extracted features are then used for enrollment or compared against stored templates for verification (1:1) and identification (1:N). This



(a) General workflow of palm recognition.



(b) Workflow of a dual-modal palm recognition system.

Figure 2: Workflow of palm recognition systems.

pipeline applies to both palmprint and palm vein recognition systems and remains the dominant paradigm in camera-based implementations. While recent work such as mmPalm [51] explores mmWave-based palm recognition, our work investigates EM vulnerabilities in conventional imaging-based systems.

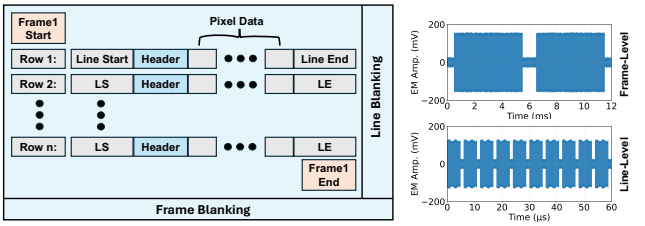
PalmPrint Recognition. Palmprint recognition [31] utilizes the surface-level features of the human palm, such as principal lines and wrinkles, to perform identity verification. The field has evolved from early statistical methods to modern deep learning approaches [54], significantly improving recognition accuracy and robustness.

PalmVein Recognition. Palmvein recognition [29] captures the internal vascular structure of the palm using NIR imaging technology. By relying on subcutaneous vascular patterns rather than the superficial skin textures used in palmprint recognition, palmvein recognition achieves greater stability and robustness, being less affected by external conditions such as skin dryness, scars, *etc.* .

Dual-Modal Palm Recognition System. Modern palm recognition systems increasingly adopt dual-modal architectures [14, 21, 26, 39] that jointly capture palmprint and palm vein information to improve accuracy and security. As shown in Figure 2b, these systems follow the standard biometric pipeline of image acquisition, ROI localization, feature extraction, and matching. Unlike single-modal designs, visible and infrared images are acquired and transmitted as interleaved streams to the SoC for decoding and ROI extraction. The two modalities are processed independently for identity verification [55], and their matching results are fused at the decision level to enhance robustness against spoofing and environmental variations.

2.2 Image Transmission Principles

In embedded image acquisition, sensors generate RAW images containing unprocessed pixel data from a single color component defined by the front-end filter array. These RAW images are then transferred to the image signal processor (ISP) through high-speed serial links [20], most notably the MIPI Camera Serial Interface 2 (MIPI CSI-2) [42], where the debayering process is applied to interpolate missing color values of each pixel based on spatial correlations with surrounding pixels.



(a) (b)

Figure 3: EM leakage in MIPI CSI-2 image transmission. (a) CSI-2 data organization. (b) Frame-level and Line-level transmission's EM leakage.

Information-bearing EM Emissions in MIPI CSI-2. As illustrated in Figure 3a, CSI-2 organizes image transmission hierarchically [32], with frames divided into rows and each row further decomposed into columns. Within each frame, the protocol structures the transmitted data into packets, specifically: each row transmission begins with a Line Start (LS) short packet, followed by a Long Packet containing a Header and Pixel Payload, and ends with a Line End (LE) short packet. Rows are separated by line blanking intervals, while frame blanking intervals delimit frame boundaries. This structured packetization not only enables reliable high-speed transmission but also induces distinctive EM emissions. As shown in Figure 3b, these emissions manifest on multiple time scales: at the frame level, aggregated signals appear as periodic bursts, each corresponding to one frame, whereas at the line level, finer-grained periodic patterns align with individual row transmissions.

3 Threat Model

The adversary's objective is to exploit EM emissions leaked from biometric acquisition and recognition systems to reconstruct palm biometric features, thereby enabling unauthorized access, identity theft, and financial fraud.

Victim Device. The victim devices are biometric acquisition and recognition systems equipped with either single-mode or dual-mode cameras. During operation, raw data are transmitted via high-speed interfaces such as CSI2, which inevitably generate EM emissions that may expose sensitive biometric information.

Adversary Capabilities. The adversary cannot physically access or tamper with the victim systems, nor modify hardware, firmware, or software. However, by capturing the EM emissions leaked during image acquisition and real-time biometric recognition, the adversary can remotely extract data sufficient to recover palm biometric features. Using commercially available antennas, low-noise amplifiers (LNAs), and software-defined radios (SDRs), the adversary can operate from a concealed distance without raising suspicion.

Attack Scenarios. As shown in Figure 1a, we consider real-world deployment scenarios where palm-based biometric systems are widely used, including secure building entry points, identity verification kiosks, and palm payment terminals deployed by major retailers [19]. The eavesdropper discreetly installs compact EM signal capturing devices behind walls, under counters, or within fixtures near the target systems. When a user performs palm-related authentication, the concealed device proactively captures the EM emission leaked during the image acquisition process. The adversary is able to reconstruct a palm template just within a few seconds.

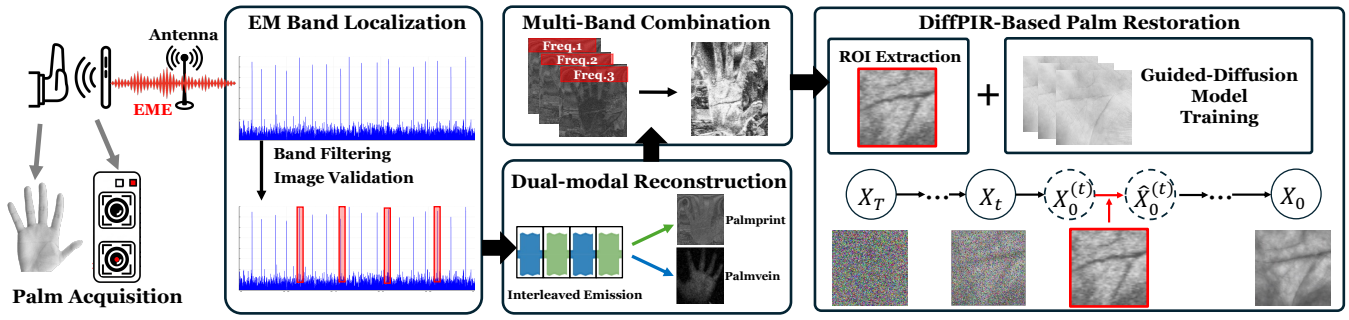


Figure 4: Overview of EMPalm.

4 Attack Design

Figure 4 provides an overview of EMPalm. We first introduce its core four modules in terms of the overall workflow, and elaborate in the following respective subsections.

(1) EM Band Localization. Palm-related emissions are embedded in a noisy wide spectrum, so this module identifies informative sub-bands carrying biometric information using a two-stage process: (i) statistical band filtering to discard noise-dominated regions, and (ii) image validation that reconstructs preliminary images to verify palm-relevant structures.

(2) Dual-Modal Image Reconstruction. For each localized band, intercepted EM signals are transformed into palm images. While reconstruction is straightforward for single-modal systems, dual-modal systems are challenging due to asynchronously interleaved palmprint and palm vein transmissions. We design a disentanglement method to separate and align the two modalities, enabling synchronized dual-modal reconstruction.

(3) Multi-Band Combination. Single-band reconstructions suffer from stochastic noise and bit-level ambiguities caused by bit-packed acquisition. To address this, we integrate reconstructions from multiple informative bands using a multi-band optimization strategy. By exploiting harmonic relationships across frequencies, this module consolidates complementary features, restores intensity variations, and preserves structural details.

(4) DiffPIR-Based Palm Restoration. The fused images undergo ROI extraction and diffusion-based restoration. Building on DiffPIR [60], we incorporate a *structure-guided conditional prior* derived from degraded images to guide the diffusion process. This design suppresses heterogeneous noise, corrects band-limited distortions, and preserves key biometric micro-structures such as palmprint creases and vein bifurcations, producing reconstructions with high perceptual quality and **biometric faithfulness** suitable for spoofing attacks and downstream analysis.

4.1 EM Leakage Bands Localization

Figure 5 illustrates the diverse signal characteristics captured across different EM sub-bands. While certain frequencies—such as 109 MHz, 118 MHz, and 405 MHz—yield palm images with discernible biometric features, many other bands are dominated by irrelevant emissions or noise (e.g., the 250 MHz band reveals HDMI). Without prior knowledge, pinpointing a sub-band that contains useful biometric signals within a wide spectrum is a non-trivial task.

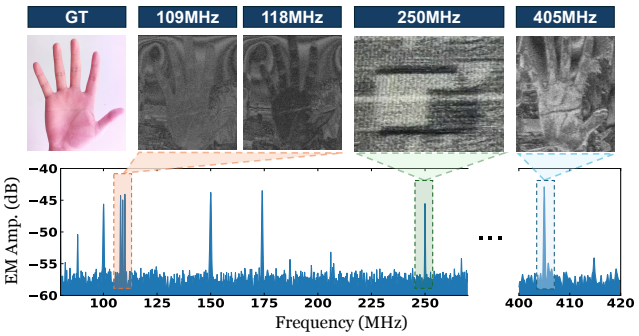


Figure 5: Illustration of signals from different frequencies.

While each informative sub-band may capture only a partial and limited aspect of the palm's structure, it can simultaneously exhibit strong structured noise patterns. This combination—limited signal coverage and dominant noise—amplifies the difficulty for downstream restoration, making it harder to recover a clean and complete biometric image from any single band. To address this, we aim to exhaustively identify signals from all sub-bands that may carry complementary biometric cues.

This insight makes the problem significantly harder than single-band reconstruction: in practice, we do not know a priori how many informative bands exist or where they are located within the spectrum. To address this challenge, we propose an automated frequency identification method, outlined in Algorithm 1, which integrates statistical signal characterization with visual interpretability. The method begins by partitioning the full EM spectrum $S(f)$ into discrete sub-bands over the range $[f_{\min}, f_{\max}]$ (Line 1), and proceeds in two stages: (1) *Band Filtering*, where candidate bands are selected based on spectral energy and statistical features, and (2) *Image Validation*, where preliminary reconstructions are assessed to confirm the presence of palm-relevant structures.

Band Filtering (Lines 2–6). For each sub-band, the time-domain signal $s_i(t)$ is extracted and evaluated using three metrics: energy E_i (overall activity), spectral entropy H_i (frequency regularity), and peak autocorrelation A_i (temporal periodicity). Sub-bands with high E_i, low H_i, and strong A_i are retained as structured, information-bearing candidates for further processing.

Image Validation (Lines 7–12). Each candidate signal $s_i(t)$ is reconstructed into a grayscale image I_i using TempestSDR [40]:

$$P_{rec}^{[f_l, f_h]} = \mathcal{R}\{n(t) + b_{clk} + H_{[f_l, f_h]}[\mathcal{D}(P_{orig})]\}, \quad (1)$$

where $\mathcal{R}$ denotes the reconstruction operator and $H_{[f_l,f_h]}$ represents the EM transfer function. After reconstruction, two visual metrics are computed to ensure that each band captures palm-relevant structures rather than incidental artifacts: image entropy $\mathcal{H}(I_i)$, reflecting intensity diversity, and edge intensity $\mathcal{E}(I_i)$, emphasizing crease and vein patterns. While either metric alone may arise from noise, their joint prominence serves as a reliable indicator of palm-related content. Bands exhibiting high $\mathcal{H}(I_i)$ and $\mathcal{E}(I_i)$ values are retained as final candidates for subsequent processing.

4.2 Dual-Modal Image Reconstruction

Although we utilize TempestSDR to reconstruct raw images to facilitate frequency localization, modern dual-mode palm recognition systems typically alternate between capturing palmprint and palmvein modalities [4, 32]. When TempestSDR is naively applied to such interleaved transmissions, the resulting reconstructions contain entangled content from both modalities, often mixed in unpredictable and non-uniform ways. As a result, these raw images are largely unusable for downstream processing, necessitating more sophisticated disentanglement strategies before any meaningful restoration or analysis can take place.

To address this issue, we analyze the eavesdropped EM signals and observe that dual-modal systems follow specific transmission patterns. For synchronized frame-interleaved systems, palmprint and palmvein data alternate regularly across consecutive frames. We first detect the transmission mode by analyzing frame header signatures and inter-frame correlation patterns:

$$\rho_{inter} = \frac{1}{N-2} \sum_{k=1}^{N-2} \mathrm{corr}(F_k, F_{k+2}), \tag{2}$$

where F_k represents the k-th frame. High ρ_{inter} values (>0.8) indicate frame-alternating transmission, enabling temporal separation by frame parity:

Algorithm 1: Frequency Band Localization

Input: EM spectrum $S(f)$, frequency range $[f_{\min}, f_{\max}]$
Output: Informative sub-bands $\mathcal{F}_{\text{img}}$

Divide $[f_{\min}, f_{\max}]$ into sub-bands $\{f_i\}_{i=1}^{N}$;
for $i \leftarrow 1$ **to** N **do**
  // Stage 1: Band Filtering
  Extract $s_i(t)$ from $S(f_i)$;
    $E_i = \|s_i(t)\|^2$; // Signal energy
    $H_i = \mathcal{H}(\mathrm{FFT}(s_i(t)))$; // Spectral ent.
    $A_i = \max(\mathrm{ACF}(s_i(t)))$; // Autocorr. peak
  // Stage 2: Image Validation
  if $E_i > \theta_E$ ***and*** $A_i > \theta_A$ ***and*** $H_i < \theta_H$ **then**
    $I_i = \textbf{TempestSDR}(f_i^{\text{low}}, f_i^{\text{high}})$; // SDR Algo
      $\mathcal{H}(I_i)$; // Image entropy
      $\mathcal{E}(I_i) = \|\nabla I_i\|$; // Edge intensity
    if $\mathcal{H}(I_i) > \theta_{\mathcal{H}}$ ***and*** $\mathcal{E}(I_i) > \theta_{\mathcal{E}}$ **then**
      $\mathcal{F}_{\text{img}} \leftarrow \mathcal{F}_{\text{img}} \cup \{F_i\}$
return $\mathcal{F}_{img}$

$$M_k = \begin{cases} k \bmod 2, & \text{if } \rho_{inter} > \tau \\ \text{Adaptive.} & \text{otherwise} \end{cases} \tag{3}$$

For systems with $\rho_{inter} > \tau$, we perform modality-specific reconstruction:

$$P_{print}[r,c] = \frac{1}{N_{print}} \sum_{j=0}^{N_{print}-1} |s_{IQ}^{(2j)}[r,c]|, \tag{4}$$

$$P_{vein}[r,c] = \frac{1}{N_{vein}} \sum_{j=0}^{N_{vein}-1} |s_{IQ}^{(2j+1)}[r,c]|. \tag{5}$$

However, real-world devices often exhibit asynchronous or line-interleaved transmissions due to sensor-level timing variations and SoC-specific architectures. For these cases ($\rho_{inter} \le \tau$), we employ an adaptive synchronization mechanism that analyzes the vertical blanking interval patterns and horizontal synchronization signals embedded in the EM emissions. Specifically, we detect packet boundaries through spectral discontinuities in the baseband signal:

$$B_k = \arg\max_t \left| \frac{d}{dt} S(f_c, t) \right|, \tag{6}$$

where $S(f_c, t)$ represents the signal power at carrier frequency f_c. These boundaries, combined with protocol-specific timing templates (e.g., MIPI CSI-2 packet headers), enable accurate modality classification even for non-uniform transmission patterns. The effectiveness of this adaptive approach ensures robust modality separation across diverse dual-modal architectures while maintaining compatibility with standard frame-alternating systems.

4.3 Multi-band Image Combination

While the dual-modal image reconstruction effectively disentangles the modalities into separate palmprint and palmvein images, it inevitably incurs information loss due to the bit-packed acquisition formats commonly used in sensor hardware. In such formats, multiple bit positions are compressed into repeating binary patterns, which become electromagnetically indistinguishable within a single frequency band. This aliasing effect causes subtle grayscale variations to collapse, leading to noticeable gradient artifacts and the erosion of fine structural details in the reconstructed images.

Our key insight is that while individual frequency bands suffer from these ambiguities, the harmonic relationships across multiple bands preserve complementary information. When the fundamental frequency f cannot differentiate between bit positions with identical periodicities, the harmonic at $2f$ often carries discriminative phase or amplitude variations necessary for accurate recovery. This observation motivates our multi-band optimization framework:

$$\min_{\alpha_i} \left\| S(I_{\text{reconstructed}}) - v_{\text{target}} \right\|^2 + \lambda\, \Phi(I_{\text{reconstructed}}), \tag{7}$$

where the first term enforces intensity consistency over uniform regions, and $\Phi(\cdot)$ is a regularizer encouraging the preservation of structural details such as palm creases and vein edges.

The reconstructed image is expressed as

$$I_{\text{reconstructed}} = \sum_{i=1}^{N} \alpha_i \cdot B_i(f_i^{\text{low}}, f_i^{\text{high}}), \tag{8}$$

where B_i denotes the filtered image obtained from frequency band i. The candidate bands are restricted to the validated outputs from the previous stage:

$$\{B_i\}_{i=1}^{N} \subseteq \mathfrak{I}_{\text{img}}, \tag{9}$$

with $\mathfrak{I}_{\text{img}}$ denoting the set of informative sub-band reconstructions identified by the frequency localization algorithm.

Here, $S(\cdot)$ denotes a segmentation operator for uniform regions, v_{target} is their expected constant intensity, and the optimization adaptively assigns weights $\{\alpha_i\}$ to balance surface uniformity with preservation of palmprint and vein structures. In practice, amplitude thresholding suppresses noise before fusion, and the number of combined bands is selected to trade off reconstruction fidelity against computational cost.

4.4 Diffusion-based Palm Restoration

While the proposed multi-band image combination alleviates bit-level grayscale collisions and restores critical structural details, practical EM side-channel acquisition of palmprint and palmvein still suffers from hardware mismatches, EM interference, and environmental noise. These factors introduce artifacts and distortions that obscure fine biometric details and reduce recognition quality.

Problem Formulation. Following prior EM reconstruction works, we model the image restoration task as a linear inverse problem:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{10}$$

where $\mathbf{x} \in \mathbb{R}^n$ denotes the clean palm image, $\mathbf{y} \in \mathbb{R}^m$ the multi-band combined image (output of Section 4.3), $\mathbf{H} \in \mathbb{R}^{m \times n}$ the degradation operator, and $\mathbf{n} \sim \mathcal{N}(0, \sigma_y^2 \mathbf{I})$ additive Gaussian noise.

Challenges in Palm EM Restoration. Palm biometric restoration from EM signals introduces unique challenges. First, the degradation operator $\mathbf{H}$ is unknown and device-dependent, involving frequency-selective attenuation, phase distortions, and structured interference that vary across hardware configurations. Second, unlike supervised restoration methods that rely on paired degraded–clean samples, an adversary in a real-world side-channel attack cannot access the victim's clean biometric images as training labels, since doing so would require compromising the biometric device itself and would contradict the stealthiness assumption of the attack.

DiffPIR Framework for Plug-and-Play Restoration. To address these challenges, we adopt the plug-and-play DiffPIR framework [60], which enables unsupervised restoration through alternating optimization. The framework solves the following optimization problem via Half-Quadratic Splitting (HQS):

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 + \lambda P(\mathbf{x}), \tag{11}$$

where $P(\mathbf{x})$ represents a learned diffusion prior. When $\mathbf{H}$ is unknown or complex, DiffPIR assumes identity degradation ($\mathbf{H} \approx \mathbf{I}$) for pure denoising, aligning with our scenario where degradations stem primarily from additive EM interference [60].

The framework alternates between two steps during inference:

$$\text{(Prior):} \quad \mathbf{x}_0^{(t)} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t + (1 - \bar{\alpha}_t)\mathbf{s}_\theta(\mathbf{x}_t, t)\right), \tag{12}$$

$$\text{(Data Fidelity):} \quad \hat{\mathbf{x}}_0^{(t)} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{x}\|^2 + \rho_t \|\mathbf{x} - \mathbf{x}_0^{(t)}\|^2, \tag{13}$$

where Eq. (13) enforces consistency with the EM-reconstructed image $\mathbf{y}$, with $\rho_t = \lambda(\sigma_n/\bar{\sigma}_t)^2$ controlling data fidelity.

Unsupervised Prior Learning. DiffPIR enables learning a powerful diffusion prior $\mathbf{s}_\theta$ from only publicly available clean palm datasets [5, 22, 38, 56], avoiding the need for any paired EM–clean data that would violate the stealthiness constraint of our threat model. The prior captures the manifold of palmprint ridges and vein structures via denoising-score matching [24, 30].

Structure-Guided Conditioning. To prevent hallucinated ridge patterns and ensure semantic consistency, we condition the denoiser on the multi-band combined EM reconstruction $\mathbf{y}$ itself. Despite noise, $\mathbf{y}$ retains coarse ridge flow and palm topology, which anchors the restoration to physically leaked biometrics rather than free-form generative priors:

$$\mathbf{x}_0^{(t)} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t + (1 - \bar{\alpha}_t)\, \mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{y})\right). \tag{14}$$

This lightweight guidance requires no additional feature engineering nor domain-specific annotations.

Preventing Generative Hallucination. We preserve EM-grounded identity information through dual constraints: (i) The data-fidelity term in Eq. (13) anchors each reverse diffusion step to the observed EM leakage $\mathbf{y}$, with adaptive weight ρ_t maintaining strong coupling throughout denoising. (ii) Structure-guided conditioning directly injects $\mathbf{y}$ into the denoiser network, ensuring generated patterns remain consistent with physical EM emanations. This optimization-network dual constraint ensures restored biometric features originate from actual EM leakage rather than learned priors.

5 Evaluation

To comprehensively assess the effectiveness of EMPalm, we conduct a three-stage evaluation across diverse hardware platforms and real-world scenarios involving 25 human participants. ***First***, we evaluate the image restoration capability, examining how accurately EMPalm can recover palmprint and palmvein images from intercepted EM signals. ***Second***, we assess spoofing effectiveness by testing whether the reconstructed images can successfully deceive state-of-the-art palm recognition models. ***Finally***, we examine the robustness of EMPalm under varying environmental and operational conditions to validate its practical feasibility.

5.1 Experimental Setup

Hardware. To reproduce palm recognition processes, we built a modular acquisition platform using single-board computers (SBCs) connected to visible-light and Near-Infrared sensors. The SBC controllers include Raspberry Pi 3B+ (S1), Raspberry Pi 5 (S2) and NVIDIA Jetson Nano (S3). We use three devices for palmprint acquisition: OV5647 (V1), IMX219 (V2) and IMX708 (V3), and use three NIR devices for palmvein acquisition: 23H166-LED (IR1), IMX219-160 (IR2) and HW200 (IR3). Besides the above single modal devices, a dual-modal device, HAOKAI-H220 (DUAL), is employed for simultaneous palmprint and palmvein capture. To further evaluate EMPalm's performance against real-world devices, we include three commercial off-the-shelf (COTS) devices C1, C2, and C3. We withhold disclosure of the exact models of the tested commercial devices to provide vendors time to develop solutions addressing risks.

Figure 6 illustrates the EM acquisition system, which is built on a Universal Software Radio Peripheral (USRP) B200 SDR [13],

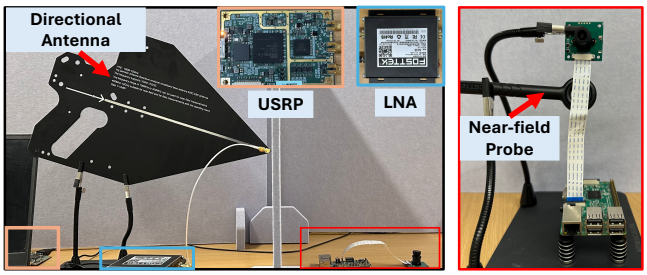


Figure 6: EM signals acquired using a directional antenna and a near-field probe.

equipped with a FOSTTEK near-field magnetic probe for close-range measurements or an Eujgoov directional antenna (0.1–12 GHz) for long-range reception. We use a FOSTTEK FST-RFAMP06 low-noise amplifier (LNA) with a gain of 40 dB to enhance weak EM emissions. The USRP operates at a sampling rate of 10 MS/s with an RF bandwidth of 20 MHz.

Software. For the configuration of USRP, we employ TempestSDR on the Ubuntu(24.04.5). For Diffusion training, as described in Section 4.4, we use PyTorch (2.4.0) with CUDA (12.1).

Physical Deployment. As shown in Figure 6, to evaluate EMPalm, we setup the attack against the target palm recognition system in both close-range and long-range configurations. In the close-range setting, a magnetic field probe is positioned near the transmission interface between the image sensor and the SBC with minimal interference, and in the long-range setting, a directional antenna intercepts radiated emissions without physical contact.

Diffusion Models for Restoration. To account for modality differences, we train two separate diffusion models for palmprint and palmvein restoration. Table 1 summarizes the dataset statistics. For palmprint, we train on the combined Tongji [56] and CASIA [5] datasets; for palmvein, we use the combined SCUT [38] and CASIA-M [22] datasets. To prevent any identity leakage between generative and discriminative training stages, each combined dataset is partitioned at the *subject level* (600 for palmprint and 650 for palmvein) into two disjoint halves: 50% of subjects are exclusively used for diffusion model training, while the remaining 50% are reserved for training the target recognition models. Critically, our 25 test volunteers are not included in these public datasets, eliminating training data leakage. Once trained, each diffusion model is applied to the eavesdropped EM measurements collected from victim interactions: we feed the intercepted signals through the corresponding modality model to reconstruct palm images. These reconstructed images constitute the stolen biometric data and are subsequently used as spoofing probes against target recognition systems.

Table 1: Dataset statistics and partition strategy for diffusion and recognition model training.

Dataset	Task	# Image	# Subject	Diffusion	Recognition
SCUT	vein	11,000	550	275 (50%)	275 (50%)
CASIA-M	vein	7,200	100	50 (50%)	50 (50%)
Tongji	print	12,000	300	150 (50%)	150 (50%)
CASIA	print	5,502	300	150 (50%)	150 (50%)
CASIA + Tongji	print	17,502	600	300 (50%)	300 (50%)
CASIA-M + SCUT	vein	18,200	650	325 (50%)	325 (50%)

Note: For the CASIA dataset, 12 subjects (out of 312) were excluded due to incomplete data samples, resulting in 300 utilized subjects used in our settings.

Target Palm Recognition Models for Spoofing. We evaluate our spoofing attack against two categories of target palm recognition models: palmprint-based and palmvein-based. For palmprint-based models, we follow PCE-Palm [27] and Diff-Palm [28], adopting three backbones, ResNet50 [23], MobileFaceNet [7], and PalmNet [18], with an input size of 224×224, all trained using ArcFace [10] (margin m=0.5, scale s=48). For palmvein-based models, we follow PVTree [49] and adopt ResNet101 [23] trained with ArcFace (m=0.5, s=64) for 20 epochs on real datasets. Table 2 summarizes all target models, their training datasets, and true accept rates. Spoofing is evaluated in a 1:100 identification setting: reconstructed and diffusion-restored images are directly used as input probes against enrolled galleries, and we measure whether the target models accept these probes as genuine.

Table 2: Target recognition models for attack evaluation.

Model	Task	Training Dataset	TAR@ 1e-4 (%)
ResNet50 [23]	Print	50% CASIA + Tongji	94.81
MobileFaceNet [7]	Print	50% CASIA + Tongji	96.26
PalmNet [18]	Print	50% CASIA + Tongji	93.80
ResNet101 [23]	Vein	50% CASIA-M + SCUT	94.87

Evaluation Metrics. To ensure objective and domain-aligned assessment, we adopt standard evaluation metrics widely used in the literature on biometric spoofing and side-channel attacks [34, 37].

- *Peak Signal-to-Noise Ratio (PSNR):* Evaluates pixel-wise fidelity between reconstructed and ground truth images, higher values indicate better pixel-level reconstruction accuracy.
- *Structural Similarity Index Measure (SSIM):* Assesses perceptual similarity in terms of luminance, contrast, and structure, ranging from -1 to 1, where 1 indicates perfect similarity.
- *Fréchet Inception Distance (FID):* Measures perceptual quality by comparing deep feature statistics, lower values indicate reconstructed images are closer to real ones in feature space.
- *Spoof Success Rate (SSR):* Quantifies the proportion of reconstructed palmprint and palmvein images that successfully bypass target biometric recognition models. A higher SSR indicates greater susceptibility of the recognition system to EM side-channel-based spoofing attacks.

Among these metrics, PSNR, SSIM, and FID evaluate the visual reconstruction quality of restored images, while SSR directly measures the attack effectiveness by assessing whether reconstructed biometric samples can successfully deceive recognition systems.

5.2 Effectiveness Evaluation

We progressively evaluate EMPalm across multiple dimensions, including its effectiveness in single and dual modal restoration, its ability to spoof target recognition models, and its performance in attacking real-world COTS devices. All experiments are conducted while 25 users operate the devices in real time, each user performed 10 interaction trials, yielding a total of 250 captured images.

Restoration Quality Comparison. We evaluated the restoration quality of EMPalm by comparing it with two representative EM-based biometric reconstruction methods, EMEye [37] and EMIRIS [34]. Although neither work was designed for palm restoration, both share conceptual similarities with our setting in that

Table 3: Comparison of reconstruction quality across palmprint, palmvein, and dual-modal settings among EMIRIS, EMEye, and our EMPalm. A dash (—) indicates that the baseline method could not reconstruct dual-modal images and spoof models under the tested configuration.

Method	Palmprint				Palmvein				Dual-Modal			
	SSIM ↑	PSNR (dB) ↑	FID ↓	SSR (%) ↑	SSIM ↑	PSNR (dB) ↑	FID ↓	SSR (%) ↑	SSIM ↑	PSNR (dB) ↑	FID ↓	SSR (%) ↑
EMEye	0.51	18.0	26.3	—	0.42	18.7	29.4	—	—	—	—	—
EMIRIS	—	—	—	—	0.49	22.0	16.3	52.83	—	—	—	—
EMPalm (Ours)	**0.71**	**28.3**	**7.7**	**68.93**	**0.67**	**23.9**	**8.2**	**66.01**	**0.68**	**24.1**	**8.7**	**66.51**

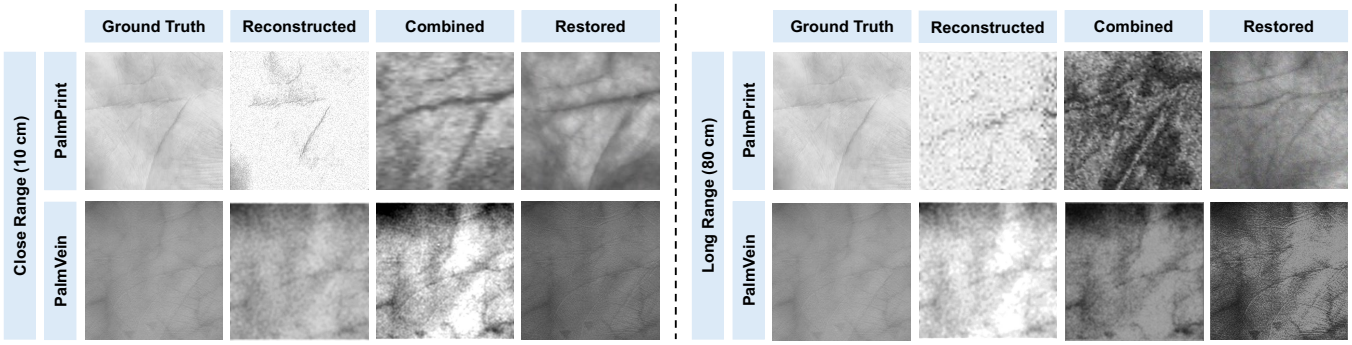


Figure 7: Reconstruction real-time human users, showing examples of palmprint (device V1, random select) and palmvein (device IR1, random select) in the single-modal setting, with close-range acquisition of live subjects data at 10 cm (left) and long-range acquisition at 80 cm (right). Ground Truth: the original high-quality palm print image; Reconstructed: the initial single-band reconstructed image; Combined: the image obtained by fusing reconstructed images from multiple frequency bands; Restored: the image restored from the combined image by diffusion model.

they exploit EM side-channel leakage to recover visual biometric information. Specifically, EMEye targets EM-based video frame inference, whereas EMIRIS reconstructs iris textures from NIR-driven EM emissions. Despite focusing on different biometric modalities, both exemplify EM-to-image recovery and thus offer meaningful baselines for evaluating the challenges of palm reconstruction.

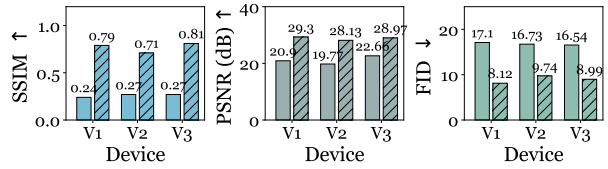
For fair comparison, we reproduced the EMEye and EMIRIS pipelines under our palm acquisition setup and evaluated them using the same testing protocol. As shown in Table 3, both baseline methods are inherently limited to single-stream processing and therefore cannot support dual-modal palm reconstruction. EMEye fails to produce spoofing-capable outputs because it lacks a dedicated restoration stage and does not incorporate a diffusion-based generative prior, which results in blurry, distorted, and low fidelity reconstructions. EMIRIS performs somewhat better on palmvein images but remains restricted to single modality operation, as it was originally designed for NIR-only iris sensing and cannot generalize to visible spectrum palmprint data or mixed dual-modal inputs. In contrast, EMPalm delivers substantially higher reconstruction quality across both modalities and achieves the highest spoofing success rates in all evaluation settings.

Effectiveness of Single-Modal Restoration. We first evaluated EMPalm on single-modal restoration using three palmprint (V1-V3) and three palmvein (IR1-IR3) devices. Figure 7 presents all intermediate and final images recovered by each stage of the EMPalm, under both close and long-range settings. As shown, EMPalm progressively refines the image through each stage, ultimately producing restored images that closely approximate the ground truth.

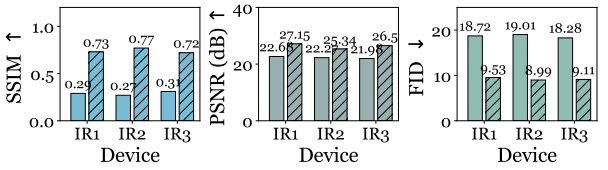
Figure 8 further reports the quantitative metrics (SSIM, PSNR, and FID) across all devices, comparing EMPalm with and without the proposed multi-band combination (hatched vs. solid bars). The solid bars represent single-band restoration, while the hatched bars indicate our multi-band fusion results. On palmprint devices, EMPalm with multi-band fusion achieves up to 0.81 SSIM, 29.3 dB PSNR, and 8.12 FID; on palmvein devices, it achieves up to 0.77 SSIM, 27.15 dB PSNR, and 8.99 FID. The slightly lower metrics on palmvein reflect its inherent stability and robustness against external perturbations, making reconstruction more challenging. Nevertheless, EMPalm still extracts high-fidelity representations across both modalities, demonstrating strong generalizability.

Comparing the two variants, multi-band combination yields consistent and significant gains across all metrics, confirming our hypothesis in **Section 4.3**. Specifically, SSIM increases by 0.55 (palmprint) and 0.50 (palmvein), PSNR by 8.4 dB (palmprint) and 4.52 dB (palmvein), while FID decreases by 8.98 (palmprint) and 10.02 (palmvein), confirming that multi-band combination improves reconstruction quality in structural and perceptual dimensions.

Effectiveness of Dual-Modal Restoration. Building upon the single-modal results, we next evaluate EMPalm under the dual-modal acquisition setting, where both palm-print (visible) and palm-vein (NIR) signals are captured simultaneously within a single sensing process. EMPalm separates these interleaved data streams and reconstructs each modality independently from the same EM capture. Figure 9 summarizes the quantitative performance. From jointly acquired data, EMPalm achieves 0.67 SSIM, 26.81 dB PSNR, and 11.32 FID on the palmprint modality, and 0.61 SSIM, 24.46 dB PSNR, and 13.78 FID on the palmvein modality. These results

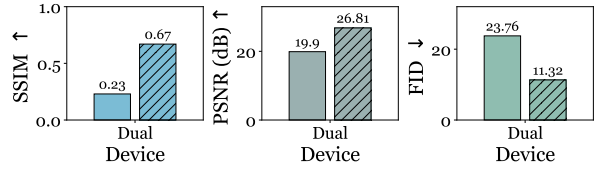


(a) Palmprint

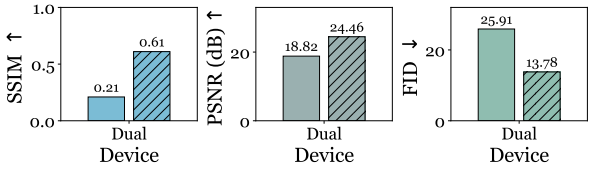


(b) Palmvein

Figure 8: SSIM, PSNR and FID of EMPalm on Single Modal. Solid bars: single band, hatched bars: multi-bands combined.



(a) Palmprint



(b) Palmvein

Figure 9: SSIM, PSNR and FID of EMPalm on Dual Modal. Solid bars: single band, hatched bars: multi-bands combined.

confirm that the proposed signal-separation and reconstruction framework can effectively disentangle and restore both biometric modalities from a single EM observation (**Section 4.2**).

Compared with the single-band variant (solid), incorporating multi-band combination (hatched) continues to yield substantial improvements, even under the intertwined dual-stream condition. For palmprint, SSIM increases by 0.40, PSNR by 5.64 dB, and FID decreases by 12.13; for palmvein, SSIM improves by 0.44, PSNR by 6.91 dB, and FID decreases by 12.44. These findings demonstrate that multi-band combination remains crucial for high-fidelity reconstruction when recovering two concurrently transmitted biometric channels from the same acquisition session.

5.3 Effectiveness of Spoofing Target Models.

To ensure a fair and representative evaluation, we follow the prior palm recognition works [27, 28, 49], which introduce advanced generative or enhancement pipelines for producing high-quality palm datasets and recognition benchmarks. These works have established strong CNN-based architectures validated on large-scale palm datasets, forming a solid and widely adopted foundation for subsequent research. We therefore adopt their recognition models, as summarized in Table 2, to provide consistent and credible baselines for assessing the spoofing effectiveness of EMPalm.

Table 4: Attack Success Rate (SSR) against different palm recognition models (mean ± std).

Model	ResNet50	MobileFaceNet	PalmNet	ResNet101
SSR (%)	68.0 ± 2.3	62.1 ± 2.0	70.0 ± 2.5	61.3 ± 1.9

Table 4 presents the spoofing success rates (SSR) achieved by EMPalm against different target models. The results demonstrate substantial effectiveness across all tested architectures, with an overall average spoofing success rate of 65.3%. Among the palmprint models, PalmNet (print) exhibits the highest vulnerability with success rates reaching approximately 72%, while ResNet50 (print) achieves around 68% and MobileFaceNet (print) shows slightly lower rates at approximately 62%. The palmvein model ResNet101 demonstrates comparable susceptibility with success rates around 61%. Palmvein patterns are inherently harder to spoof due to their subtle, sub-surface nature, which makes them more resistant to EM leakage and reconstruction. This is different from the more prominent, surface-level features of palmprints that are easier to capture and exploit. These findings confirm that our EM-based reconstruction method poses a significant security threat across diverse models used in palm biometric systems.

An interesting observation is that among all palmprint models, PalmNet exhibits the highest vulnerability to EMPalm. Unlike generic CNN-based models, PalmNet adopts a hybrid architecture that integrates Gabor filters with a PCA-based unsupervised scheme. This design choice makes PalmNet particularly susceptible to attacks from EMPalm, as its strong capabilities at recovering principal textural features. This observation underscores a key insight: models that depend heavily on low-level or principal-component-derived features may inadvertently expose themselves to greater risk when such features are recoverable through external leakage. These findings highlight the need for model designs that are robust to side-channel reconstructions, potentially by avoiding over-reliance on easily reconstructible signal patterns and incorporating safeguards that account for fine-grained biometric information.

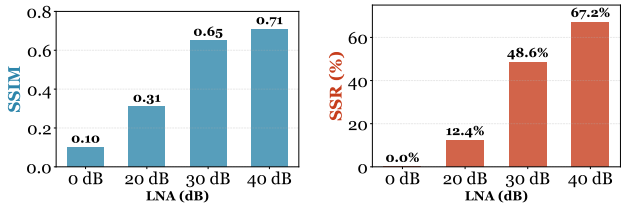
Table 5: Effectiveness of EMPalm on three COTS devices.

Device	**SSIM ↑**	**PSNR (dB) ↑**	**FID ↓**	**Average SSR (%) ↑**
C1 (Office Gate)	0.64	27.8	11.3	52.5
C2 (Home Locker)	0.61	26.4	11.7	59.1
C3 (Payment Kiosk)	0.66	28.2	10.7	60.9

Effectiveness of Attacking COTS Devices. To further evaluate the practicality of EMPalm in real-world settings, we extend our experiments to COTS palm recognition devices C1, C2, and C3, representing three typical deployment scenarios: *Office Gate*, *Home Locker*, and *Payment Kiosk*. We focus on assessing whether EMPalm is effective on these commodity systems, in terms of its reconstruction quality and effectiveness in spoofing attacks. Table 5 reports the results of EMPalm on the three COTS devices. Despite the differences in hardware design and shielding strategies, our results confirm that EMPalm can successfully extract biometric information from COTS devices, with the reconstructed images

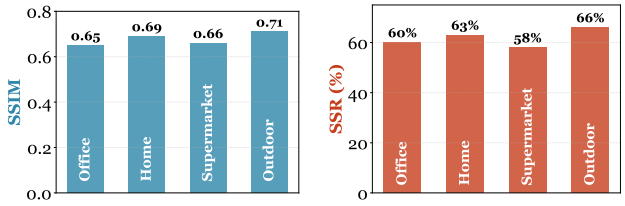
Table 6: Impact of different SBCs on EMPalm.

Device	SSIM ↑	PSNR (dB) ↑	FID ↓	Average SSR (%) ↑
S1	0.72	29.41	8.73	62.7
S2	0.74	29.49	8.52	66.5
S3	0.72	29.24	9.12	60.1



(a) SSIM (b) SSR

Figure 10: Impact of different LNAs on EMPalm.



(a) SSIM (b) SSR

Figure 11: Impact of different environment noises.

demonstrating substantial spoofing capability against recognition models, highlighting the generality and severity of this threat.

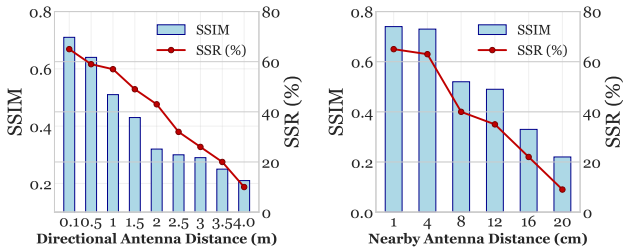
5.4 Impacts of Practical Factors

Unless otherwise specified, all impact experiments were conducted under a default configuration. The palm recognition software (PalmNet) and sensor models (V1 for palmprint, IR1 for palmvein) were used, with the sensor connected to the SBC (S1) under evaluation. A receiving antenna was placed at a fixed distance of 0.5 meter and paired with a 40 dB LNA to ensure sufficient signal strength.

Impact of Different SBCs. To examine how different SBCs affect the performance of EMPalm, we evaluated it on three single-board computers: Raspberry Pi 3B+ (S1), Raspberry Pi 5 (S2) and Jetson Nano (S3). Each device was configured with identical palmprint recognition software and connected to the same sensor model. The receiving antenna was placed at a fixed distance of 0.5 meters with 40 dB LNA, ensuring consistent experimental conditions across all tests. As reported in Table 6, the performance of EMPalm remains highly stable across different SBCs, confirming that the exploitable EM leakage originates from the sensor's data transmission rather than the computing hardware. This demonstrates that the vulnerability is broadly applicable regardless of the deployment platform.

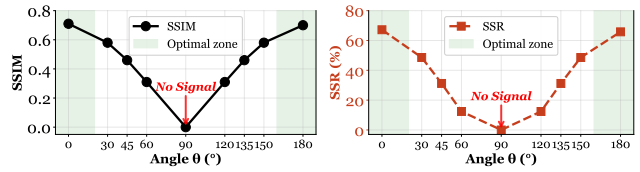
Impact of Different LNAs. To investigate the effect of low-noise amplifiers on EMPalm, we conduct experiments using LNAs with different gain levels: no gain, 20dB, 30dB, and 40dB. The three gain levels correspond to different device models: ZK09-BM (20dB), Teyleten (30dB), and FST-RFAMP06 (40dB).

Figure 10 presents the performance of EMPalm across different LNA configurations. Without amplification (0dB), EM signals are too weak for meaningful palm restoration (SSIM < 0.1, SSR = 0%). The 20dB amplifier shows minimal improvement (SSR = 12.4%), remaining insufficient for practical attacks. However, substantial improvements emerge with 30dB amplification (SSIM = 0.65, SSR = 48.6%), which further increase with the 40dB amplifier (SSIM = 0.71, SSR = 67.2%). These results demonstrate a clear correlation between LNA gain and attack effectiveness, with a notable threshold effect between 20dB and 30dB, where the amplification becomes sufficient to capture fine-grained biometric features through EM emissions.



(a) Directional antenna (b) Nearby antenna

Figure 12: Impact of antenna distance on EMPalm.



(a) SSIM (b) SSR

Figure 13: Impact of antenna angle on EMPalm. (a) SSIM and (b) SSR under varying angles θ. Results indicate the presence of optimal reception zones at certain angles, while a complete signal loss is observed at $\theta = 90°$.

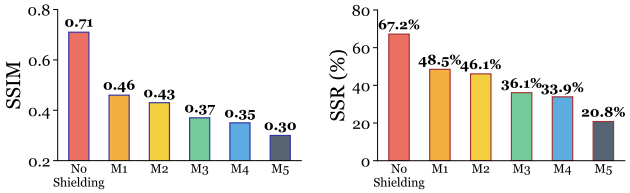
Impact of Different Environmental Noises. To evaluate EMPalm's robustness against real-world noises, we tested the EMPalm across four daily-life environments where palmprint recognition can be commonly deployed: office, home, supermarket, and outdoor settings. We maintained a fixed distance of 1.5 meters and collected 50 EM traces in each environment during peak activity hours to capture representative noise conditions. As shown in Figure 11, EMPalm achieved consistent metrics across all environments, demonstrating the general effectiveness of EMPalm in daily-life scenarios. This robustness to ambient interference validates EMPalm's practical threat potential in real-world deployments.

Impact of Different Distances. As shown in Figure 12, we evaluate EMPalm under two antenna configurations. With a directional antenna (meter-level distances), performance gradually decreases from 0.72 SSIM and 65% SSR at 0.1 m to 0.21 SSIM and 10% SSR at 4.0 m due to EM attenuation. EMPalm remains effective within 2 m, achieving 0.32 SSIM and 43% SSR, indicating practical feasibility across typical room-scale environments.

With a nearby antenna (centimeter-level distances), performance remains consistently high within 4 cm, exceeding 0.7 SSIM and 60% SSR, and then drops sharply beyond 8 cm as near-field coupling weakens. This contrast highlights strong short-range leakage and the extended reach enabled by directional antennas.

Table 7: Impact of common building materials on EM-Palm performance under practical deployment conditions.

Material	SSIM ↑	PSNR (dB) ↑	FID ↓	SSR (%) ↑
No Building Material	0.71	29.1	9.2	67.5
Wood	0.63	27.8	17.5	58.4
Drywall	0.59	26.9	18.9	54.1
Glass	0.56	25.7	20.3	50.2
Concrete	0.38	22.4	28.7	33.5
Aluminum panel	0.35	21.8	30.1	29.7



(a) SSIM **(b) SSR**

Figure 14: Impact of EM shielding materials on EMPalm. (a) SSIM and (b) SSR under varying EM shielding materials.

Impact of Different Probe Angles. To evaluate the impact of probe orientation, we position the receiving probe 2 centimeters away from the target palm sensor in the near-field region and vary the probe angle θ from 0° to 180°. Figure 13 illustrates the relationship between probe angle and restoration quality measured by SSIM. The highest restoration quality is achieved when the receiving probe achieves optimal EM coupling with the sensor's internal signal paths at 0° (SSIM = 0.71, SSR = 67.2%) and 180° (SSIM = 0.70, SSR = 65.8%), where the probe orientation maximizes interception of the radiated EM fields. As the angle moves away from 0°/180°, restoration quality drops, reaching moderate levels at 30°–150° and failing completely at 90° (SSIM = 0, SSR = 0%). The symmetric degradation pattern suggests dipole-like radiation characteristics, indicating a predictable angular dependency that adversaries can exploit to optimize interception and spoofing.

Impact of Different Building Materials. We further evaluate EMPalm's robustness when EM signals pass through common building materials separating the palm recognition device from an eavesdropper. Each material is tested under typical deployment conditions without extra shielding. Table 7 presents the reconstruction and spoofing results across five representative materials, including wood, drywall, glass, concrete, and aluminum panel, along with a baseline case without obstruction. The unobstructed setup yields the best image quality (SSIM 0.71, PSNR 29.1 dB, FID 14.2, SSR 67.5%). Non-conductive materials like wood, drywall, and glass cause moderate signal loss but still permit recognizable palm reconstruction. Dense or conductive materials such as concrete and aluminum panels strongly attenuate EM emissions, reducing image quality and spoofing success. Overall, building materials can weaken but not eliminate EM side-channel leakage.

Impact of EM Shielding Materials. Following EMIRIS [34] and EMeye [37], we evaluated the impact of five shielding materials, including copper wire mesh (M5), aluminum foil (M4), metalized fabric (M3), conductive coating (M2), and conductive fabric (M1), on EMPalm, with each material uniformly wrapped around the sensor's data transmission cables. All other experimental settings remained identical to those described above. Figure 14 shows how different shielding materials affect EMPalm, ordered by theoretical shielding capability. Conductive fabric (SSIM: 0.46, SSR: 48.5%) and coating (SSIM: 0.43, SSR: 46.1%) provide moderate protection. Metalized fabric (SSIM: 0.37, SSR: 36.1%) and aluminum foil (SSIM: 0.35, SSR: 33.9%) offer better suppression. Copper mesh delivers the strongest shielding (SSIM: 0.30, SSR: 20.8%), significantly reducing reconstruction quality and spoofing success. These differences reflect variations in material conductivity, thickness, and structure. While EM shielding materials substantially degrade EMPalm's effectiveness, they cannot fully eliminate the side-channel vulnerability.

Table 8: Cross-dataset validation results.

Configuration	Training Data	SSR (%)	ΔSSR
Original (50/50 split)	Split In CASIA+Tongji	68.93	-
Separation Dataset	Diff: Tongji only Recog: CASIA only	69.07	+ 0.14%

Table 9: Experimental results of offline palmprint reconstruction from IQ data collected using the compact setup (<20cm).

Users	SSIM ↑	PSNR (dB) ↑	FID ↓	Average SSR (%) ↑
U1	0.66	26.18	9.94	53.6
U2	0.70	26.44	9.62	57.1

Cross-Dataset Generalization. To evaluate the generalization capability of EMPalm across different data sources, we conduct a cross-dataset validation using completely disjoint datasets. Specifically, the diffusion model is trained only on Tongji, while the recognition model is trained only on CASIA, with no subject or data overlap. As shown in Table 8, the SSR remains comparable to the original 50/50 split setting, with only a marginal difference of +0.14%, indicating that EMPalm maintains stable performance when trained and evaluated on different datasets.

Feasibility of Deferred Attack. Beyond real-time attacks, EMPalm also supports deferred attack scenarios. An adversary can deploy a compact eavesdropping device to passively collect EM emissions over time and perform offline reconstruction once data is obtained. To validate this setting, we built a miniaturized collection system using an Adalm-Pluto [11] and a microcontroller, with a footprint of only $6 \times 10 \times 3$ centimeters, enabling discreet placement near palm recognition terminals. We collected In-phase and Quadrature (IQ) data from 10 authentication sessions over a 3-hour period using this compact setup. Notably, EMPalm remains a single-shot attack, requiring EM emissions from only one authentication event (200–500 ms) to reconstruct. Offline analysis of stored IQ samples achieves reconstruction quality comparable to real-time attacks (Table 9), demonstrating the practicality of long-term covert data collection combined with one-time attack execution.

5.5 Ablation Study

To quantify the individual contribution of each module in EMPalm, we conduct comprehensive ablation experiments. Each variant removes one key component while keeping the rest of the pipeline unchanged. We construct four leave-one-out variants: (i)

Table 10: Ablation study of EMPalm.

Variant	SSIM ↑	PSNR (dB) ↑	FID ↓	SSR (%) ↑
Full EMPalm	0.74	29.5	8.6	66.5
w/o Dual-Modal Disentanglement	0.25 (-0.49↓)	14.7 (-14.8↓)	15.7 (+7.1↑)	–
w/o Multi-Band Combination	0.54 (-0.20↓)	25.9 (-3.6↓)	10.6 (+2.0↑)	52.8 (-13.7↓)
w/o DiffPIR Restoration	0.52 (-0.22↓)	20.8 (-8.7↓)	19.4 (+10.8↑)	23.8 (-42.7↓)

Table 11: Per-stage error analysis of EMPalm.

Stage	SSIM ↑	PSNR (dB) ↑	FID ↓	SSR (%) ↑
Raw Reconstruction	0.46	16.7	28.8	–
+ Multi-Band Combination	0.52 (+0.06↑)	20.8 (+4.1↑)	19.4 (-9.4↓)	23.8 (+23.8↑)
+ DiffPIR Restoration (Full)	0.74 (+0.28↑)	29.5 (+12.8↑)	8.6 (-20.2↓)	66.5 (+66.5↑)

Note: SSR values below 10% are denoted as "–", as they indicate near-random attack performance and are considered negligible.

without Dual-Modal Disentanglement, which reconstructs mixed palmprint/palmvein signals directly; (ii) *without Multi-Band Combination*, which uses only the single best-SNR band; and (iii) *without DiffPIR Restoration*, which omits the restoration module.

Table 10 summarizes the ablation results. Removing any individual component substantially degrades both reconstruction fidelity and SSR, confirming that all modules are essential to EMPalm. In particular, disabling Dual-Modal Disentanglement leads to the most severe structural collapse (SSIM drops from 0.74 to 0.25) and reduces SSR to below the effective detection threshold (<10%), indicating a practical attack failure, highlighting its critical role in preserving identity-separable features. Removing Multi-Band Combination causes a moderate performance drop (SSIM decreases to 0.54 and SSR to 52.8%), suggesting that multi-band fusion enhances stability but is not the primary performance driver. Moreover, removing DiffPIR Restoration dramatically increases perceptual distortion (FID rises from 8.6 to 19.4) and reduces SSR to 23.8%, demonstrating that restoration is indispensable for recovering texture details.

Per-stage Error Analysis. Table 11 reports the cumulative error across the EMPalm pipeline. Reconstruction fidelity improves progressively from raw reconstruction to multi-band fusion and finally to restoration. Notably, while structural metrics increase steadily, SSR exhibits a significant improvement after the restoration stage, indicating that generative refinement is critical for recovering identity-discriminative details.

6 Discussion

Countermeasures. Based on the vulnerabilities identified in Section 4.2, several defenses can mitigate the risks posed by EMPalm. First, EM shielding applied to sensor transmission cables can suppress informative emissions, and appropriate material choices can significantly reduce reconstruction quality and spoofing success. Second, redesigning the transmission protocol—such as increasing transmission complexity or decoupling packets from pixel-level information—can break the direct mapping between EM signals and biometric data. Third, system-level defenses, including anomaly detection and multi-factor authentication, can help prevent spoofing using reconstructed or fabricated artifacts. Together, these measures form a multi-layered defense strategy spanning hardware shielding, protocol hardening, and system-level security enhancements against EM side-channel attacks.

Limitations and Future Work. Palm recognition sensors are embedded in complex electronic systems, where EM emissions from surrounding components, channel effects, and hardware imperfections inevitably introduce interference and reconstruction errors. Limited sampling rate and bandwidth further attenuate high-frequency details, and polarity inversion may cause grayscale distortion, constraining fine-grained reconstruction and effective attack range. While our approach reliably recovers the key palmprint and palmvein structures required for physical spoofing and deceiving most image-based commercial systems, its effectiveness against high-end devices with liveness detection or enhanced protection mechanisms remains an open question. Future work will focus on improving EM signal processing to enhance SNR and suppress interference, exploring advanced antenna and denoising techniques to better characterize and potentially extend effective attack distances (e.g., via directional beamforming or learning-based signal enhancement), and fabricating realistic 3D prosthetic hands [6] to validate end-to-end physical spoofing.

7 Related Work

Diverse EM Side-channel Attack Surfaces. Prior works have demonstrated the exploitation of EM side channels across diverse systems, including keystroke and browsing reconstruction from GPUs [53], fingerprint recovery from in-display fingerprint sensors [44], high-fidelity iris reconstruction from NIR sensors [34], and video stream extraction from embedded cameras [37]. Research has further shown that smartphone magnetometers can analyze EM footprints to infer running applications [47, 61], while wireless charging inadvertently leaks sensitive information through EM emissions [33, 43, 45]. Additional studies have revealed EM vulnerabilities in cryptographic implementations [8], smartphone activity inference [16], USB device fingerprinting [25], hidden camera detection [36], hidden microphone detection [59] and IoT activity profiling [2, 35, 52, 57]. EM analysis has also been extended to system- and device-level threats, such as detecting GPU cryptojacking via magnetic leakage (MagTracer)[50] and identifying laptop microphone recording states through EM emanations (TickTock)[48].

Overall, EMPalm advances prior work by (i) providing the first empirical exploitation of EM leakage in palm recognition with simultaneous disentanglement of interleaved palmprint and palmvein signals, and (ii) enabling EM-based palm image restoration using an unsupervised diffusion framework trained solely on public palm datasets, without requiring paired EM data or device-specific calibration. Together, EMPalm offers a unified framework linking EM leakage analysis to practical biometric exploitation.

8 Conclusion

In this paper, we propose EMPalm, the first EM side-channel attack recovering palm biometrics from recognition systems. EMPalm handles both single- and dual-modality systems by reverse-engineering transmission protocols and employing three techniques: frame boundary identification with modality disentanglement, multi-band image combination for bit recovery, and DiffPIR-based texture restoration. Our experiments show that EMPalm reconstructs high-fidelity palm images from EM signals, exhibiting strong structural similarity, high signal quality, and low perceptual discrepancy, as well as enabling successful spoofing across diverse recognition models. These findings reveal critical vulnerabilities in existing palm recognition systems, stressing the importance of using improved multi-factor defenses for better security.

Acknowledgments

This work is supported by NSF DGE-2409851. Xiaoyan Sun and Jun Dai are also supported by NSF OAC-2528534.

References

[1] Amazon Web Services. 2023. Amazon One — Palm-Based Identity Service. https://aws.amazon.com/one/. Accessed: 2025-08-24.

[2] A. Amodei, D. Capriglione, L. Ferrigno, G. Miele, L. Tari, G. Tomasso, and G. Cerro. 2023. Experimental Analysis of Side-Channel Emissions for IoT Devices Activities' Profiling. In *2023 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*. 42–47. https://doi.org/10.1109/MetroInd4.0IoT57462.2023.10180188

[3] Kevin W Bowyer and Mark J Burge. 2016. *Handbook of iris recognition*. Springer.

[4] Cadence Design Systems. 2025. MIPI CSI-2 TX Controller. https://www.cadence.com/en_US/home/tools/silicon-solutions/protocol-ip/interface-ip/mipi/mipi-csi-2-tx-controller.html. Accessed: 2025-08-24.

[5] CASIA. 2005. CASIA Palmprint Image Database. http://biometrics.idealtest.org. Accessed: 2025-08-24.

[6] CCC. [n. d.]. Chaos Communication Congress 2018. Accessed: 2025-08-24.

[7] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese conference on biometric recognition*. Springer, 428–438.

[8] Yushi Cheng, Xiaoyu Ji, Wenyuan Xu, Hao Pan, Zhuangdi Zhu, Chuang-Wen You, Yi-Chao Chen, and Lili Qiu. 2019. Magattack: Guessing application launching and operation via smartphone. In *Proceedings of the 2019 ACM Asia conference on computer and communications security*. 283–294.

[9] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[11] ANALOG DEVICE. 2022. USRP. https://www.analog.com/en/resources/evaluation-hardware-and-software/evaluation-boards-kits/adalm-pluto.html. Accessed: 2026-02-24.

[12] DPA. 2018. BND Relocates to belin. https://www.welt.de/regionales/bayern/article184668046/Bundesnachrichtendienst-Der-Umzug-der-Spione.html. Accessed: 2025-08-24.

[13] Ettus Research. [n. d.]. Ettus Research USRP Products. https://www.ettus.com/products/. Accessed: 2025-08-27.

[14] Dandan Fan, Xu Liang, Wei Jia, Junan Chen, and David Zhang. 2024. A Novel Hybrid Fusion Combining Palmprint and Palm Vein for Large-Scale Palm-Based Recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 54, 7 (2024), 4471–4484. https://doi.org/10.1109/TSMC.2024.3382877

[15] Lunke Fei, Guangming Lu, Wei Jia, Shaohua Teng, and David Zhang. 2019. Feature Extraction Methods for Palmprint Recognition: A Survey and Evaluation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49, 2 (2019), 346–363. https://doi.org/10.1109/TSMC.2018.2795609

[16] Yongjian Fu, Lanqing Yang, Hao Pan, Yi-Chao Chen, Guangtao Xue, and Ju Ren. 2024. Magspy: Revealing user privacy leakage via magnetometer on mobile devices. *IEEE Transactions on Mobile Computing* (2024).

[17] Chengrui Gao, Ziyuan Yang, Wei Jia, Lu Leng, Bob Zhang, and Andrew Beng Jin Teoh. 2025. Deep Learning in Palmprint Recognition-A Comprehensive Survey. *arXiv preprint arXiv:2501.01166* (2025).

[18] Angelo Genovese, Vincenzo Piuri, Konstantinos N Plataniotis, and Fabio Scotti. 2019. PalmNet: Gabor-PCA convolutional networks for touchless palmprint recognition. *IEEE Transactions on Information Forensics and Security* 14, 12 (2019), 3160–3174.

[19] Wesley Grant. 2025. Palm Scanning Gains Ground as Retail Biometric of Choice. https://www.paymentsjournal.com/palm-scanning-gains-ground-as-retail-biometric-of-choice/. *PaymentsJournal* (5 June 2025). Accessed: 2025-08-24.

[20] Bahadir K Gunturk, John Glotzbach, Yucel Altunbasak, Ronald W Schafer, and Russel M Mersereau. 2005. Demosaicking: color filter array interpolation. *IEEE Signal processing magazine* 22, 1 (2005), 44–54.

[21] HandPass. 2018. HandPass 100 Dual-Modal Palm Scanning Camera. https://deptrum.com/en/site/product_details/454. Accessed: 2025-08-24.

[22] Ying Hao, Zhenan Sun, Tieniu Tan, and Chao Ren. 2008. Multispectral palm image fusion for accurate contact-free palmprint recognition. In *2008 15th IEEE International Conference on Image Processing*. IEEE, 281–284.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[25] Omar Adel Ibrahim, Savio Sciancalepore, Gabriele Oligeri, and Roberto Di Pietro. 2020. MAGNETO: Fingerprinting USB Flash Drives via Unintentional Magnetic Emissions. *ACM Trans. Embed. Comput. Syst.* 20, 1, Article 8 (Dec. 2020), 26 pages. https://doi.org/10.1145/3422308

[26] SUNNY OPTICAL INTELLIGENCE. 2022. Palm print and vein recognition module. https://www.sunnyaiot.com/shuangmuxiangji. Accessed: 2025-08-24.

[27] Jianlong Jin, Lei Shen, Ruixin Zhang, Chenglong Zhao, Ge Jin, Jingyun Zhang, Shouhong Ding, Yang Zhao, and Wei Jia. 2024. Pce-palm: Palm crease energy based two-stage realistic pseudo-palmprint generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2616–2624.

[28] Jianlong Jin, Chenglong Zhao, Ruixin Zhang, Sheng Shang, Jianqing Xu, Jingyun Zhang, ShaoMing Wang, Yang Zhao, Shouhong Ding, Wei Jia, et al. 2025. Diff-Palm: Realistic Palmprint Generation with Polynomial Creases and Intra-Class Variation Controllable Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26367–26376.

[29] Wenxiong Kang and Qiuxia Wu. 2014. Contactless palm vein recognition using a mutual foreground-based local binary pattern. *IEEE transactions on Information Forensics and Security* 9, 11 (2014), 1974–1985.

[30] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising Diffusion Restoration Models. arXiv:2201.11793 [eess.IV] https://arxiv.org/abs/2201.11793

[31] Adams Kong, David Zhang, and Mohamed Kamel. 2009. A survey of palmprint recognition. *pattern recognition* 42, 7 (2009), 1408–1418.

[32] Andy Lee. 2021. MIPI CSI Interface Definitions and Protocol Layer Overview. Accessed: 2025-08-24.

[33] Jiachun Li, Yan Meng, Le Zhang, Guoxing Chen, Yuan Tian, Haojin Zhu, and Xuemin Sherman Shen. 2023. Magfingerprint: A magnetic based device fingerprinting in wireless charging. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.

[34] Wenhao Li, Jiahao Wang, Guoming Zhang, Yanni Yang, Riccardo Spolaor, Xiuzhen Cheng, and Pengfei Hu. 2025. EMIRIS: Eavesdropping on Iris Information via Electromagnetic Side Channel. *NDSS* (2025).

[35] Ziwei Liu, Feng Lin, Teshi Meng, Benaouda Chouaib Baha-eddine, Li Lu, Qiang Xue, and Kui Ren. 2024. EMTrig: Physical Adversarial Examples Triggered by Electromagnetic Injection towards LiDAR Perception. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems, (SenSys 24)*. 351–364.

[36] Ziwei Liu, Feng Lin, Chao Wang, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. 2023. Camradar: Hidden camera detection leveraging amplitude-modulated sensor images embedded in electromagnetic emanations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–25.

[37] Yan Long, Qinhong Jiang, Chen Yan, Tobias Alam, Xiaoyu Ji, Wenyuan Xu, and Kevin Fu. 2024. Em eye: Characterizing electromagnetic side-channel eavesdropping on embedded cameras. *NDSS* (2024).

[38] Dacan Luo, Yitao Qiao, Di Xie, Shifeng Zhang, and Wenxiong Kang. 2024. Palm vein recognition under unconstrained and weak-cooperative conditions. *IEEE Transactions on Information Forensics and Security* 19 (2024), 4601–4614.

[39] Nan Luo, Zhenhua Guo, Gang Wu, and Changjiang Song. 2011. Joint palmprint and palmvein verification by Dual Competitive Coding. In *2011 3rd International Conference on Advanced Computer Control*. 538–542. https://doi.org/10.1109/ICACC.2011.6016471

[40] Martin Marinov. 2014. Remote video eavesdropping using a software-defined radio platform. In *MS Thesis*. https://api.semanticscholar.org/CorpusID:261364519

[41] James Clerk Maxwell. 1890. *The Scientific Papers of James Clerk Maxwell...* Vol. 2. University Press.

[42] MIPI Alliance. 2023. MIPI CSI-2 Specifications. https://www.mipi.org/specifications/csi-2. Accessed: 2025-08-24.

[43] Tao Ni, Jianfeng Li, Xiaokuan Zhang, Chaoshun Zuo, Wubing Wang, Weitao Xu, Xiapu Luo, and Qingchuan Zhao. 2023. Exploiting contactless side channels in wireless charging power banks for user privacy inference via few-shot learning. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[44] Tao Ni, Xiaokuan Zhang, and Qingchuan Zhao. 2023. Recovering fingerprints from in-display fingerprint sensors via electromagnetic side channel. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 253–267.

[45] Tao Ni, Xiaokuan Zhang, Chaoshun Zuo, Jianfeng Li, Zhenyu Yan, Wubing Wang, Weitao Xu, Xiapu Luo, and Qingchuan Zhao. 2023. Uncovering User Interactions on Smartphones via Contactless Wireless Charging Side Channels. In *2023 IEEE Symposium on Security and Privacy (SP)*. 3399–3415. https://doi.org/10.1109/SP46215.2023.10179322

[46] Federal Bureau of Investigation. 2013. Next Generation Identification. https://le.fbi.gov/science-and-lab/biometrics-and-fingerprints. Accessed: 2025-08-24.

[47] Hao Pan, Lanqing Yang, Honglu Li, Chuang-Wen You, Xiaoyu Ji, Yi-Chao Chen, Zhenxian Hu, and Guangtao Xue. 2021. Magthief: Stealing private app usage data on mobile devices via built-in magnetometer. In *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[48] Soundarya Ramesh, Ghozali Suhariyanto Hadi, Sihun Yang, Mun Choon Chan, and Jun Han. 2022. Ticktock: detecting microphone status in laptops leveraging

electromagnetic leakage of clock signals. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2475–2489.

[49] Sheng Shang, Chenglong Zhao, Ruixin Zhang, Jianlong Jin, Jingyun Zhang, Rizen Guo, Shouhong Ding, Yunsheng Wu, Yang Zhao, and Wei Jia. 2025. PVTree: Realistic and Controllable Palm Vein Generation for Recognition Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6767–6775.

[50] Rui Xiao, Tianyu Li, Soundarya Ramesh, Jun Han, and Jinsong Han. 2023. MagTracer: Detecting GPU cryptojacking attacks via magnetic leakage signals. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[51] Yucheng Xie, Xiaonan Guo, Yan Wang, Jerry Q Cheng, Tianfang Zhang, Yingying Chen, Yi Wei, and Yuan Ge. 2024. mmpalm: Unlocking ubiquitous user authentication through palm recognition with mmwave signals. In *2024 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 1–9.

[52] Haowen Xu, Tianya Zhao, Xuyu Wang, Jun Dai, and Xiaoyan Sun. 2025. MagWatch: Exposing Privacy Risks in Smartwatches Through Electromagnetic Signals. In *International Conference on Information and Communications Security*. Springer, 329–346.

[53] Zihao Zhan, Zhenkai Zhang, Sisheng Liang, Fan Yao, and Xenofon Koutsoukos. 2022. Graphics Peeping Unit: Exploiting EM Side-Channel Information of GPUs to Eavesdrop on Your Neighbors. In *2022 IEEE Symposium on Security and Privacy (SP)*. 1440–1457. https://doi.org/10.1109/SP46214.2022.9833773

[54] David Zhang, Wangmeng Zuo, and Feng Yue. 2012. A comparative study of palmprint recognition algorithms. *ACM computing surveys (CSUR)* 44, 1 (2012), 1–37.

[55] Lin Zhang, Zaixi Cheng, Ying Shen, and Dongqing Wang. 2018. Palmprint and palmvein recognition based on DCNN and a new large-scale contactless palmvein dataset. *Symmetry* 10, 4 (2018), 78.

[56] Lin Zhang, Lida Li, Anqi Yang, Ying Shen, and Meng Yang. 2017. Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach. *Pattern Recognition* 69 (2017), 199–212.

[57] Tianya Zhao, Ningning Wang, and Xuyu Wang. 2025. Membership Inference Against Self-supervised IMU Sensing Applications. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems, (SenSys 25)*. 268–281.

[58] Dexing Zhong and Jinsong Zhu. 2019. Centralized large margin cosine loss for open-set deep palmprint recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 6 (2019), 1559–1568.

[59] Ruochen Zhou, Xiaoyu Ji, Chen Yan, Yi-Chao Chen, Wenyuan Xu, and Chaohao Li. 2023. Dehirec: Detecting hidden voice recorders via adc electromagnetic radiation. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3113–3128.

[60] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. 2023. Denoising Diffusion Models for Plug-and-Play Image Restoration. arXiv:2305.08995 [cs.CV] https://arxiv.org/abs/2305.08995 Accessed: 2025-08-24.

[61] Zhuangdi Zhu, Hao Pan, Yi-Chao Chen, Xiaoyu Ji, Fan Zhang, and Chuang-Wen You. 2016. Magattack: Remote app sensing with your phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 241–244.

Membership Inference Against Self-supervised IMU Sensing Applications

Tianya Zhao
Florida International University
Miami, Florida, USA
tzhao010@fiu.edu

Ningning Wang
Florida International University
Miami, Florida, USA
nwang012@fiu.edu

Xuyu Wang*
Florida International University
Miami, Florida, USA
xuywang@fiu.edu

ABSTRACT

Deep learning has revolutionized the use of inertial measurement unit (IMU) sensors in mobile applications, such as human activity recognition. Building on the success of pre-trained models across various domains, recent studies have increasingly adopted self-supervised learning (SSL) for a range of sensing tasks. While these SSL approaches improve generalization and reduce labeling requirements, their privacy implications have received limited attention. This paper addresses this gap by examining IMU data privacy during pre-training through membership inference. Our work serves two important purposes: First, it enables data owners to verify if their data was used without permission in encoder pre-training. Second, it demonstrates how adversaries might compromise sensitive human sensing data used in pre-training. To enhance the practicality of membership inference on unlabeled IMU sensing data across different SSL algorithms, we introduce an activity labeling module and a novel perturbation strategy to exploit encoder overfitting characteristics on training data. When an encoder overfits, it memorizes training data rather than learning generalizable patterns. Therefore, when comparing the original data to the perturbed version, the encoder generates more distinct feature vectors for samples from its training set than for samples it has never seen before. We evaluate our membership inference methods on two mainstream SSL methods across multiple datasets, demonstrating that our method can achieve relatively high precision and recall at low false positive rates.

CCS CONCEPTS

• **Security and privacy**; • **Human-centered computing** → **Ubiquitous and mobile computing**;

KEYWORDS

Membership Inference, Human Activity Recognition, Self-supervised Learning, Privacy-sensitive Sensing Systems.

*Corresponding Author

SenSys '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/25/05
https://doi.org/10.1145/3715014.3722060

ACM Reference Format:
Tianya Zhao, Ningning Wang, and Xuyu Wang. 2025. Membership Inference Against Self-supervised IMU Sensing Applications. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715014.3722060

1 INTRODUCTION

Over the past decade, human sensing technology has advanced significantly, particularly in embedded and mobile devices. These advancements have transformed how we monitor and interact with technology in daily life. Wearable devices have emerged as powerful tools across diverse mobile applications, from recognizing complex human activities [22, 24, 33, 34, 36, 45, 51, 54, 57] to enabling natural and intuitive human-computer interactions [30, 59]. Many of these rely heavily on data from inertial measurement unit (IMU) sensors, which are commonly integrated into smartphones, smartwatches, and smart earphones.

Following the success of self-supervised learning (SSL) in fields like natural language processing (e.g., BERT [12]), more and more studies have focused on adapting different SSL techniques for IMU sensing applications [7]. For example, LIMU-BERT [52] introduces a lightweight BERT-like representation learning model specifically designed for mobile IMU sensor data. SSL methods offer two main advantages in this context: they leverage large amounts of unlabeled sensor data, significantly reducing data labeling costs, and they enable the extraction of more generalizable features through large-scale training.

Existing studies mainly focus on improving encoder training to enhance performance across various IMU sensing tasks and scenarios. However, there has been limited exploration of the security and privacy implications of these methods. *This is especially concerning given that IMU data can reveal sensitive information about individuals' identities and activities.* For instance, Ahmad et al. [2] have demonstrated that data from accelerometers embedded in smartphones can be used for user identification. In this paper, we conduct the first systematic study on membership inference against SSL-based pre-trained IMU sensing encoders. The goal of a membership inference is to determine whether a given IMU reading is part of the encoder's pre-training dataset, as shown in Figure 1. IMU readings that originate from users included in the pre-training dataset are classified as "members", while those from users outside this dataset are "non-members". In this paper, we use the terms training data and members interchangeably.

Membership inference on SSL-based IMU encoders has two important applications [28]. First, data owners can use membership inference methods to audit whether their (public) data was used to pre-train an encoder without authorization. This is particularly

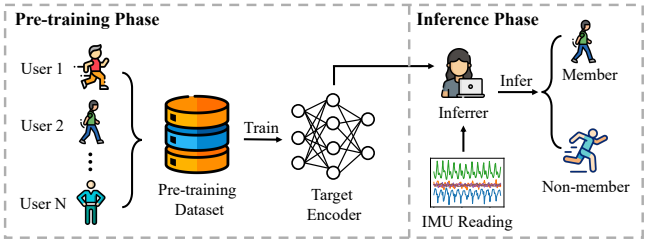


Figure 1: Membership inference workflow. In the encoder pre-training stage, the system provider generates pre-training data using IMU readings from N users. In the inference phase, the inferrer aims to determine whether specific data was used in pre-training.

important if the data owner has invested heavily in the data collection or has privacy agreements with individuals whose data is included, as unauthorized use can lead to serious financial and privacy concerns. Second, a potential attacker can use membership inference to compromise the privacy of the pre-training data, given the sensitive nature of IMU data. As mentioned before, the attacker can leverage IMU data to infer personal information included in the pre-training dataset.

Most membership inference methods assess privacy leakage based on exploiting machine learning models' tendency to *overfit* their training data [17]. For instance, Yeom et al. [56] suggest that training examples generally have lower loss values compared to non-training examples. However, while existing membership inference techniques [39, 42, 49, 56] work well for supervised learning models like classifiers, they face severe challenges when applied to SSL methods. This is because SSL-based pre-trained encoders output feature vectors rather than probability distributions, making traditional methods that rely on probabilities, confidence scores, or loss values inapplicable. Furthermore, the feature vector itself may not capture the *overfitting* of the encoder on the input [28]. To address this, Liu et al. [28] propose membership inference on pre-trained image encoders trained by contrastive learning. They assume that member data should generate feature vectors more similar to their augmented versions when the same augmentation techniques from the pre-training phase are applied. While this is an important step forward, this approach may have limitations if we lack knowledge of the specific training algorithm or augmentation methods.

Our Work. We propose the first membership inference method that is agnostic to underlying SSL methods, allowing us to examine the privacy implications of SSL-based IMU sensing applications in a more practical context. In this paper, we use the term "inferrer" to refer to either a data owner or an attacker conducting membership inference. Our analysis focuses on black-box scenarios, where the inferrer can only query a pre-trained encoder (referred to as the target encoder) and observe its outputs without being able to modify the encoder's parameters or structure. This represents the most basic way of using a pre-trained encoder, as anyone can feed data into it and receive outputs without needing to understand how the encoder works internally.

Challenges. Implementing a more practical membership inference method for SSL-based IMU sensing encoders faces three significant challenges. First, most existing membership inference methods [17, 28, 42] require detailed knowledge of the target model's training process, including the training techniques and model structure, to train shadow models that mimic the target model's behavior. However, this approach is impractical for SSL-based encoders because SSL training typically requires substantial computational resources, and the core training information may be confidential. Second, the diversity of SSL algorithms further complicates membership inference. Techniques designed for contrastive learning (e.g., [28]) may not generalize to other SSL-based encoders, such as BERT. This highlights the need for an SSL-agnostic membership inference method that can adapt to different SSL algorithms. Third, the label information is essential for designing a membership indicator to determine if a sample is part of the training data [6]. However, obtaining labels in our case is not feasible, as SSL solely relies on unlabeled data to pre-train an encoder. Moreover, labeling IMU data is more complex and less efficient than other data types (e.g., images), since IMU sensor readings are not easily interpretable by humans.

Our solution. To address previous challenges, we propose a new approach to membership inference that eliminates the need for labeled data and cumbersome shadow encoder pre-training. As noted in [6], class information is important information for effective membership inference. To leverage this information despite the absence of labels in IMU data, we first propose a pseudo activity labeling module. This module applies Fourier transforms to extract frequency information, enabling us to broadly categorize IMU data into "static" or "dynamic" activities based on their high-frequency components. To implement membership inference across different SSL methods, our approach then generates neighboring data points by introducing slight perturbations tailored to each activity type and extracts their corresponding feature vectors. Under the common assumption that models tend to overfit their training data [17], we expect that the feature vectors of training samples will exhibit more variability when perturbed. This is because the model has memorized specific characteristics of training samples, leading to less consistent predictions when these samples are slightly modified. In contrast, for data points not seen during training, the model relies on learned general patterns, resulting in more stable predictions across perturbations. The whole process only needs to query the encoder and does not require access to the training algorithm or model structure.

The main contributions of this paper are summarized as the following.

- To the best of our knowledge, this is the first work to investigate privacy issues related to IMU data through membership inference within the context of SSL. We propose a more practical membership inference approach for IMU sensing encoders, limiting the inferrer's background knowledge of the encoder's architecture and training algorithm while allowing access to outputs only through direct queries to the encoder.
- We introduce an activity labeling module and a novel perturbation strategy that leverages overfitting in pre-trained

encoders based on unique characteristics of IMU data, achieving SSL-agnostic membership inference without requiring labeled data or complex shadow encoder training.
- We conduct comprehensive experiments across two different mainstream SSL methods and three datasets, showing potential privacy leakage in these encoders.

Ethics and Data Privacy. The data used in our experiments are from public sources. For the practical evaluation, we collected data by ourselves in controlled environments. All data are only used for academic research. All the experiments were conducted in a closed environment, and we did not leak any private information into markets or public repositories.

2 BACKGROUND: SSL

SSL is a type of machine learning where a model learns from unlabeled data by creating its own "labels" or learning signals. In SSL, the model undergoes a pre-training phase that may involve tasks like predicting missing parts of the input or distinguishing between transformed versions of the input data [29]. This pre-training process results in an encoder that can serve as a foundation model for building classifiers on a variety of downstream tasks.

In this paper, we consider two mainstream SSL approaches: generative methods and contrastive methods [29]. Generative methods rely on training an encoder f_θ to represent the input data $\mathbf{x}$ as a distinct vector $f_\theta(\mathbf{x})$. This vector is then passed to a decoder, which attempts to reconstruct $\mathbf{x}$ from $f_\theta(\mathbf{x})$. In natural language processing (NLP), popular generative models include auto-regressive models like the GPT series [4]. On the other hand, contrastive methods train an encoder to encode augmented input $\mathbf{x}'$ into vector representation $f_\theta(\mathbf{x}')$, enabling the measurement of similarity between inputs. One of the classic contrastive learning methods is SimCLR [9], which aims to learn representations by comparing samples using the NT-Xent loss as follows:

$$\mathcal{L} = -\frac{1}{K}\sum_{i=1}^{K}\frac{exp(sim(f_\theta(\mathbf{x}'_i), f_\theta(\mathbf{x}'_j))/\tau)}{\sum_{k=1,k\neq i}^{2K} exp(sim(f_\theta(\mathbf{x}'_i), f_\theta(\mathbf{x}'_k))/\tau)}, \quad (1)$$

where $sim(\cdot)$ denotes the similarity function, K is the batch size, and τ represents the temperature hyperparameter.

When the pre-trained encoder is available, it can serve as a feature extractor to train a downstream classifier. There are two popular approaches: 1) Linear Probing (LP), where the pre-trained encoder's parameters are frozen and not updated during classifier training, and 2) Fine-Tuning (FT), where the pre-trained encoder's parameters are updated along with the classifier during end-to-end training.

3 MEMBERSHIP INFERENCE

3.1 Problem Formulation

Membership inference, also called membership inference attack, in the machine learning (ML) setting aims to predict if a specific training example was, or was not, used as training data in a particular model. This makes membership inference the simplest and most widely deployed way for auditing training data privacy [6]. In the SSL-based IMU sensing context, the inferrer can infer whether an IMU reading belongs to the users in the pre-training dataset.

Following prior work [6, 21, 56], we formalize membership inference as a security game between a system provider $\mathcal{S}$ and an inferrer $\mathcal{I}$ as below:

(1) The provider samples a pre-training dataset D from distribution $\mathbb{D}$ and uses training algorithm $\mathcal{T}$ to pre-train an encoder $f_\theta \leftarrow \mathcal{T}(D)$.
(2) The provider randomly selects a bit b. If $b = 0$, the provider samples $\mathbf{x} \in \mathbb{D} \setminus D$; otherwise, selects $\mathbf{x} \in D$. The provider then sends $\mathbf{x}$ to the inferrer.
(3) The inferrer gets query access to the encoder f_θ and may have access to the pre-training distribution $\mathcal{P}$ according to the background knowledge $\mathcal{B}$. Then, the inferrer outputs $\hat{b} \leftarrow \mathcal{I}(\mathbf{x}, f_\theta, \mathcal{B})$.
(4) The game outputs 1 if $\hat{b} = b$ and 0 otherwise.

In SSL-based IMU sensing where no ground truth labels y are available, we define the pre-training dataset $D = \{\mathbf{x}_i\}_{i=1}^N$ as IMU data collected from member users in the pre-training distribution $\mathcal{P}$. When sampling $\mathbf{x} \in \mathbb{D} \setminus D$, where $\mathbb{D}$ represents the complete distribution of all possible IMU sensor readings, the IMU data comes from non-member users (those users not in the pre-training set) who may follow a distribution different from $\mathcal{P}$. In membership inference, the primary objective is to construct a membership score function to determine whether a given data point $\mathbf{x}$ belongs to the pre-training dataset D, formalized as:

$$\mathcal{I}(\mathbf{x}, f_\theta, \mathcal{B}) = \mathbb{1}[\mathcal{I}'(\mathbf{x}, f_\theta, \mathcal{B}) > \tau], \quad (2)$$

where $\mathbb{1}$ is the indicator function, τ is the membership decision threshold, and $\mathcal{I}'(\mathbf{x}, f_\theta, \mathcal{B})$ represents the function to output the membership score.

3.2 Threat Model

In this section, we illustrate the inferrer's goal and limited background knowledge for the practical membership inference against IMU sensing encoders.

Inferrer's Goal. Given an input IMU data point $\mathbf{x}$, an inferrer $\mathcal{I}$ aims to determine whether it is a member user from the pre-training dataset D of the target encoder f_θ. A key objective of the inferrer is to minimize the likelihood of classifying non-members as members (Type I error), as high false positive rates reduce the reliability of membership inference. This reliability is crucial, as it affects both data owners who may want to audit their datasets and attackers who may seek to compromise user privacy.

Inferrer's Background Knowledge. Beyond lacking access to pre-training dataset labels, the inferrer also faces restricted background knowledge about the encoder. Encoder pre-training typically involves three key components: the pre-training data distribution $\mathcal{P}$, encoder architecture $\mathcal{E}$, and training algorithm $\mathcal{T}$ [28]. In this paper, we constrain the inferrer's background knowledge to implement a more practical membership inference. Specifically, we consider a black-box access scenario in which the inferrer $\mathcal{I}$ only has access to query the target encoder f_θ and observe its outputs, without any knowledge of the encoder's training algorithms or structure information. This represents the most realistic and challenging setting. By imposing these constraints on the inferrer, we ensure that the membership inference remains broadly applicable across various SSL methods.

We consider two scenarios that differ only in the inferrer's knowledge $\mathcal{B}$ on the pre-training data distribution $\mathcal{P}$. In the first scenario, Inferrer-1 $\mathcal{I}_1$ is assumed to have the background knowledge of pre-training data distribution $\mathcal{P}$, following previous studies [25, 39, 42]. In contrast, to enhance the practicality of membership inference, we assume that Inferrer-2 $\mathcal{I}_2$ lacks this knowledge. It is important to note that, while the inferrer lacks any knowledge of the encoder's architecture or parameters, they can still query the encoder and collect its output feature vectors. This is a realistic scenario, as pre-trained encoders need to be accessible for users to enable fine-tuning of downstream classifiers.

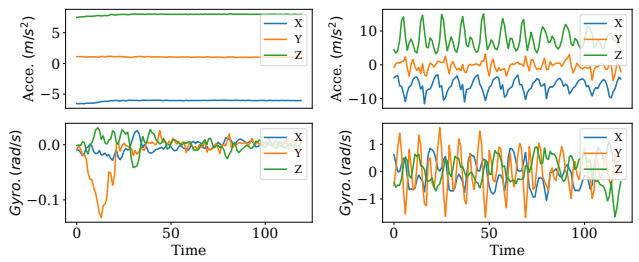
4 METHODOLOGY

4.1 Overview

We propose a membership inference method for pre-trained IMU sensing models that uses the encoder's overfitting to its training data. Our approach works by creating multiple variants of a test sample, where different portions of the data are systematically perturbed. We then pass both the original sample and its perturbed variants through the target encoder to obtain their feature vectors. The key insight is that the encoder responds differently to training data versus unseen data. When processing a sample from the training set (member data), the encoder exhibits a distinctive pattern: it partially recognizes and overfits the unperturbed portions of the perturbed variants, leading to varying levels of similarity between the original and perturbed feature vectors. Some perturbed vectors remain quite similar to the original, while others differ significantly. In contrast, when processing data not used in training (non-member data), the encoder shows more consistent behavior. Since it has not been overfitted to these samples, the similarities between the original and perturbed feature vectors tend to be more stable. By analyzing these different similarity patterns and leveraging the inferrer's background knowledge, we can determine whether a given sample was part of the encoder's pre-training dataset.

4.2 Activity Labeling

While labels typically play a crucial role in membership inference, our scenario focuses on SSL where only unlabeled data is used during the pre-training stage. To address this challenge without access to true labels, we develop a pseudo activity labeling module.



(a) Standing (b) Walking

Figure 2: Comparison of IMU readings from two distinct physical activities, illustrating characteristic differences in acceleration and angular velocity patterns.

IMU sensors capture unique patterns of human activities through their readings, though these patterns are less intuitive to interpret than visual data like images. The sensor system combines accelerometers that measure linear acceleration along three spatial axes (x, y, z) with gyroscopes that track angular velocity, together providing comprehensive motion data. Different activities create distinct features in these sensor readings. For instance, walking generates regular, repeating patterns in both acceleration and rotation measurements, while static activities like sitting produce stable signals with minimal fluctuation as shown in Figure 2. While this data may not be precise enough to differentiate between similar activities (e.g., going downstairs versus going upstairs), it allows us to reliably classify IMU data into two broad categories: static activity and dynamic activity.

To perform this classification, we analyze the frequency characteristics of the IMU signals by first applying the Fourier transform F to convert the time-domain IMU readings $\mathbf{x}$ into their frequency spectrum. This transformation facilitates efficient differentiation of activity intensity levels. Although different activities may influence the accelerometer and gyroscope differently, dynamic activities tend to exhibit higher frequency components because of rapid motion changes. In contrast, static activities primarily consist of lower-frequency components. To exploit this difference, we define a frequency threshold τ_f that separates low-frequency components from high-frequency components. With this threshold, we focus specifically on the high-frequency components A_{high} to determine activity dynamics:

$$A_{high} = \sum_{f > \tau_f} |F(\mathbf{x})|. \tag{3}$$

The activity is then classified as static or dynamic by comparing A_{high} to a magnitude threshold τ_h:

$$\hat{y} = \mathbb{1}[A_{high} > \tau_h], \tag{4}$$

where $\hat{y} = 1$ indicates a dynamic activity, while $\hat{y} = 0$ indicates a static activity. In this paper, we set the thresholds τ_f and τ_h to 3 Hz and 100, respectively.

Although these pseudo labels may not perfectly align with the ground truth, they still capture underlying activity patterns, providing valuable contextual information for membership inference against SSL-based encoders. Additionally, these coarse labels can be used to apply data perturbation techniques, which enhance the effectiveness of membership inference. The next section will introduce this concept in more detail.

4.3 Data Perturbation

In many previous works, the inferrer needs to train shadow models to infer the membership. In many cases, these shadow models even replicate the architecture of the target model. However, this approach presents several significant challenges in the SSL context. First, SSL training is computationally intensive, often demanding substantial resources and processing time. This high resource requirement makes it impractical for an individual to train their shadow encoders. Second, the diversity of SSL algorithms further complicates membership inference because the same model structure can be trained in different ways, resulting in different performances. As a result, inference methods customized for one

Algorithm 1 Perturbed data generation

Input: target data $\mathbf{x}$, target encoder f_θ, data length T, smooth length L, scaling factors s_{min} and s_{max}, thresholds τ_f and τ_h.
Output: Perturbed data $\mathbf{x}'$.

Step 1: Activity labeling
$A_{high} \leftarrow \sum_{f>\tau_f} |F(\mathbf{x})|$
$\hat{y} \leftarrow \mathbb{1}[A_{high} > \tau_h]$
$D_p \leftarrow (\mathbf{x}, \hat{y})$ `// Build data and label pair`
Step 2: Data perturbation
$\mathbf{x}' \leftarrow \mathbf{x}$ `// Copy data`
if $\hat{y} = 1$ **then**
  $idx \leftarrow Uniform(0, T - L)$
  $\mathbf{v} \leftarrow \mathbf{x}[idx, :]$ `// Select smooth values`
  $\mathbf{x}'[idx : idx + L, :] \leftarrow \mathbf{v}$
else
  $scale \leftarrow Uniform(s_{min}, s_{max})$
  $\mathbf{x}'[:, :3] \leftarrow \mathbf{x}[:, :3] * scale$
end if
return $\mathbf{x}'$

algorithm may not generalize effectively to other encoders. Third, many SSL frameworks involve proprietary training details, which may be kept confidential by the developers. Given these limitations, our membership inference instead focuses on analyzing the data characteristics rather than relying on shadow models.

Our approach builds on the observation that encoders respond differently to training data (member data) compared to unseen data. When perturbing a sample from the training set, the encoder shows a distinct response: it may partially recognize and overfit the unperturbed parts of the perturbed variants, resulting in varying degrees of similarity between the original and perturbed feature vectors. Some perturbed vectors remain highly similar to the original, while others differ substantially.

To leverage this behavior effectively, we developed specialized data perturbation techniques for different types of activities, as outlined in Algorithm 1. For dynamic activities $\hat{y} = 1$, we introduce perturbations by selectively smoothing a small segment L of the original data. This process begins by sampling a starting index from a uniform distribution over the range $(0, T-L)$, where T is the total time steps in the IMU data. Once the index is selected, we assign a smooth vector $\mathbf{v}$ at that position, with $\mathbf{v}$ matching the dimensionality of the IMU data (i.e., 6 dimensions: 3 for accelerometer data and 3 for gyroscope data). This smooth vector $\mathbf{v}$ is then applied over L consecutive time steps, replacing the original data to achieve the smoothing effect.

For static activities $\hat{y} = 0$, we employ a fundamentally different approach based on selective scaling of signal components. This method applies random scaling factors specifically to the accelerometer components of the IMU data while leaving the gyroscope readings unchanged. The scaling factors are drawn from a uniform distribution $U(s_{min}, s_{max})$, where s_{min} and s_{max} define the minimum and maximum scaling values, respectively. The rationale behind this design is that static activities exhibit simple and stable patterns in accelerometer readings, as shown in Figure 2a. This carefully designed dual approach allows us to introduce controlled variations while preserving the essential characteristics of each activity type. The perturbations are specifically designed to maintain signal characteristics while producing sufficient variation to effectively probe how the target encoder responds to and learns from the training data. In this study, we set $s_{min} = 10$, $s_{max} = 40$, and $L = 20$ as default values, with further comparisons under different settings provided in Section 6.2.

4.4 Membership Inferring

As detailed in Section 3.2, the inferrer operates under significant constraints, having no access to the encoder's architecture or training algorithm. The only potential advantage available to the inferrer is background knowledge $\mathcal{B}$ about the pre-training data distribution $\mathcal{P}$. Therefore, we consider two inferrers in this paper: Inferrer-1, which possesses complete knowledge of the pre-training data distribution ($\mathcal{B} = \mathcal{P}$), and Inferrer-2, which operates without any prior knowledge of the distribution ($\mathcal{B} = \emptyset$). This design allows us to systematically evaluate how background knowledge influences inference capabilities.

Inferrer-1. Given the knowledge of pre-training data distribution, the inferrer $\mathcal{I}_1$ can obtain a shadow dataset D_s that shares the same distribution as the pre-training data but was not used in the actual pre-training. To make use of this additional information, the inferrer first constructs a labeled shadow dataset $D_s = \{\mathbf{x}_i, \hat{y}_i\}_{i=1}^{S}$ using the previously introduced activity labeling module. Since the inferrer can query the target encoder and obtain output embeddings, they attach a linear probe (called a shadow classifier C_{θ_s}) to the target encoder. This shadow classifier is trained using standard supervised learning:

$$\min_{C_{\theta_s}} \sum_{(\mathbf{x}_i, \hat{y}_i) \in D_s} \mathcal{L}(C_{\theta_s}(f_\theta(\mathbf{x}_i)), \hat{y}_i), \tag{5}$$

where $\mathcal{L}$ is the binary cross-entropy loss. Once the shadow classifier is trained, the inferrer begins the security game defined in Section 3.1.

Given a sample $\mathbf{x}$ randomly drawn from the provider $\mathcal{S}$, the inferrer first generates a pseudo label $\hat{y}$ according to the activity intensity. Then, the inferrer creates N perturbations of the sample to build a series of perturbed samples $\mathbf{x}'_1, \mathbf{x}'_2, \cdots, \mathbf{x}'_N$. These perturbed samples are passed through both the target encoder f_θ and shadow classifier C_{θ_s} with a softmax function to obtain their probabilities of correct classification p_i. Following the approach from LiRA [6], we apply logarithmic scaling to these probabilities:

$$\hat{p}_i = log(\frac{p_i + \varepsilon}{1 - p_i + \varepsilon}); \quad for \;\; p_i = C_{\theta_s}(f_\theta(\mathbf{x}_i))_{\hat{y}_i}, \tag{6}$$

where ε is a small constant added to avoid undefined values in the calculation. Since perturbed member data is expected to show more diverse responses to perturbations, the scaled probabilities should exhibit greater instability compared to non-member data. We therefore use the standard deviation of these probabilities as the membership function:

$$\mathcal{I}'_1(\mathbf{x}, f_\theta, \mathcal{B}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - \bar{p})^2}, \tag{7}$$

where $\bar{p}$ is the mean probability of all perturbed samples. τ_1 is set to 0.3 for Inferrer-1 to infer membership.

Inferrer-2. In scenarios where the inferrer lacks knowledge of the pre-training data distribution, we propose that Inferrer-2 can determine membership by directly examining perturbation stability. Without any background knowledge, the inferrer can immediately begin the security game without training shadow models, making this approach more practical in real-world settings.

Given a sample $\mathbf{x}$, the process begins with the same two steps: generating a pseudo label $\hat{y}$ and creating N perturbed samples. However, instead of producing output logits, the target encoder generates N feature vectors for the perturbed samples. These feature vectors capture the high-dimensional representations of both the original and perturbed samples. We then use the Euclidean distance d to quantify the difference between the original feature vector and each perturbed vector, yielding a set of distances that characterize the model's stability under perturbations. Intuitively, member samples used during pre-training may exhibit different stability patterns compared to unseen samples. The membership function is defined as below:

$$\mathcal{I}_2'(\mathbf{x}, f_\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}_i')) - \bar{d})^2}, \tag{8}$$

where $\bar{d}$ represents the mean distance across all pairs of original and perturbed feature vectors. This formulation captures the variance in the model's responses to perturbations, providing a metric that can potentially distinguish between member and non-member samples based on their stability characteristics. In this paper, we use a default setting of $N = 100$. Given that the distance of feature vectors is smaller than the value of output logits, the decision threshold τ_2 is set to 0.1 for Inferrer-2.

5 EXPERIMENTAL SETTINGS

All experiments are conducted on a Linux server with an Intel(R) Xeon(R) Gold 6258R CPU and NVIDIA A100 GPUs with 40GB of memory.

5.1 Encoders

In this paper, we employ two widely adopted SSL methods as the target encoders.

SimCLR. Following Qian et al. [35], we employ SimCLR [9] to pre-train a ResNet18 [15] as the target encoder. We adapt the input layer of ResNet18 to suit the dimensions of IMU data. For data augmentation, we employ random noise and resampling. The pre-training epochs, batch size, and learning rate are set to 200, 128, and 0.003, respectively.

LIMU-BERT. LIMU-BERT [52] is a lite BERT-inspired representation learning model designed specifically for IMU sensor data. For our purposes, we modify only the pre-training dataset to align with our tasks, leaving the model architecture and other components unchanged.

5.2 Datasets

This paper quantitatively evaluates our proposed methods using three publicly accessible datasets that have been widely employed in IMU sensing. All IMU readings are kept in the same shape (120, 6), where 120 is the time step.

Table 1: The information required for different membership inference methods. The top 5 methods are originally designed for supervised learning.

Methods ↓	Shadow Models	Shadow Data	Labels
Shokri et al. [42]	✓	✓	✓
Yeom et al. [56]	✗	✓	✓
Watson et al. [49]	✓	✓	✓
ML-Leaks [39]	✓	✓	✓
LiRA [6]	✓	✓	✓
EncoderMI-V [28]	✓	✓	✗
EncoderMI-T [28]	✓	✓	✗
Zheng et al. [25]	✓	✓	✗
Inferrer-1 $\mathcal{I}_1$	✓	✓	✗
Inferrer-2 $\mathcal{I}_2$	✗	✗	✗

HHAR. The HHAR [44] dataset comprises 9166 accelerometer and gyroscope readings for 6 specific activities: biking, sitting, standing, walking, going upstairs, and going downstairs, collected from 9 users. In this study, users 0-6 are designated as members, while users 7 and 8 are non-members.

Motion. MotionSense [31] comprises 4534 IMU sensor data collected from 24 participants performing 6 different activities: going downstairs, going upstairs, walking, jogging, sitting, and standing. In this study, participants 21-23 are designated as non-members. For simplicity, we refer to this dataset as Motion.

UCI. The UCI [37] dataset contains 2088 samples from 30 users performing 6 activities: standing, sitting, lying down, walking, going downstairs, and going upstairs. Users 0-20 are designated as members, with the rest as non-members.

5.3 Metrics

Following prior work [6, 49], we evaluate membership inference at low false positive rates (FPRs) instead of balanced accuracy. To illustrate the importance of this choice, consider the example from [6]: an inference method that accurately identifies 0.1% of users while guessing randomly otherwise is more effective than one with a uniform 50.05% success rate, despite similar balanced accuracy. This is because the precise privacy breach of a small subset of members poses a greater risk than an uncertain privacy leakage across all data groups.

Furthermore, maintaining a low FPR is crucial for membership inference, regardless of whether the analysis is conducted by the data owner or an external attacker. A high FPR can reduce the effectiveness and reliability of membership inference as a tool to assess potential privacy leaks in pre-training data. Following recent studies [3, 6], we focus on *precision* and *recall* at low FPRs throughout this paper. Given the scale of the IMU datasets, we specifically target FPRs of 1% and 5%, adjusting thresholds accordingly to maintain these levels.

5.4 Baseline Methods

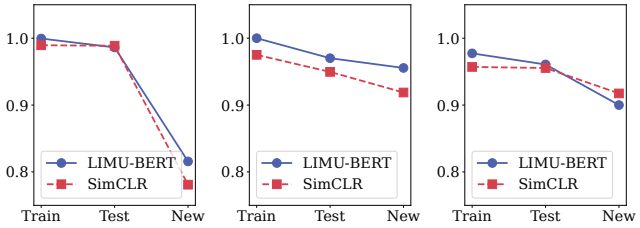
This paper compares eight membership inference methods, summarized in Table 1. The top 5 methods are originally designed for supervised learning and require additional label information, which

may not be feasible in the SSL context. Nonetheless, we still provide these methods with all the required information, including access to the downstream classifier trained on the same pre-training data. While this setup grants the inference methods a significantly advantageous level of background knowledge, leading to an inherently unfair comparison, it allows us to benchmark our methods against a strong baseline and gain valuable insights.

The remaining three methods are specifically designed for SSL. EncoderMI-V and EncoderMI-T are two methods proposed in [28], where the only difference is whether the membership classifier is vector-based or threshold-based. EncoderMI exploits the data augmentation techniques deployed in the pre-training stage of contrastive learning to infer membership. Zheng et al. [25] focus on another SSL paradigm, exploiting the masked modeling mechanism to perform the inference.

6 EVALUATION

6.1 Encoder Performance



(a) HHAR (b) Motion (c) UCI

Figure 3: Downstream classification accuracy for different data groups. "Train", "Test", and "New" represent member data, same-distribution shadow data, and non-member data relative to the training set, respectively.

Before evaluating our proposed membership inference methods, we first examine the utility of the target encoder and quantify the extent of overfitting.

Utility. Figure 3 presents the downstream classification accuracy across various datasets based on the two target encoders. In most cases, the accuracy remains higher than 90%, demonstrating the powerful ability of the pre-trained encoders to learn effective representations for downstream sensing tasks. Notably, the performance of the unseen HHAR dataset is relatively lower than that of the other datasets, achieving an accuracy of approximately 80%. This difference could be attributed to the greater domain shift present in the HHAR dataset, which poses additional challenges for generalization. Despite this, these encoders still maintain reasonable performance for studying privacy leakage in this paper.

Overfitting. As shown in Figure 3, the higher performance on the training dataset reveals the overfitting behavior. To quantify the extent of overfitting, we present the loss values and k-fold cross-validation results on unseen data in Table2. Specifically, the train loss, test loss, and new loss correspond to the loss values for member data, same-distribution shadow data, and non-member data relative to the training set, respectively. The consistently higher new loss values compared to both the train and test loss values suggest that the encoders tend to overfit the training data. The HHAR dataset exhibits the largest loss difference, indicating the highest degree of overfitting.

Table 2: Overfitting analysis: cross-entropy loss of different data groups and k-fold cross-validation performance. "5-fold Std." represents the standard deviation of accuracy over 5-fold validation on unseen data.

Encoder →	LIMU-BERT			SimCLR		
Dataset →	HHAR	Motion	UCI	HHAR	Motion	UCI
Train Loss	0.002	0.001	0.048	0.044	0.089	0.097
Test Loss	0.073	0.137	0.094	0.052	0.148	0.106
New Loss	0.735	0.153	0.491	0.713	0.196	0.241
5-fold Std.	1.08%	32.48%	4.70%	2.36%	30.36%	6.60%
7-fold Std.	2.75%	24.02%	8.53%	2.76%	28.51%	7.10%
10-fold Std.	1.21%	26.44%	9.53%	2.07%	29.43%	9.82%

Furthermore, k-fold cross-validation results show that both encoders demonstrate greater performance instability on the Motion unseen data, likely due to a significant domain shift. This suggests that the encoder may struggle to generalize across diverse new domains. Overall, the encoders exhibit overfitting to the HHAR and Motion training sets in different ways, while their overfitting to the UCI dataset is relatively moderate.

In general, the encoder achieves strong utility performance while exhibiting some degree of overfitting to the training set, making them applicable for membership inference.

6.2 Experimental Results

Tables 3 and 4 show the performance of our proposed methods compared to baseline approaches on two different SSL encoders. The low recall values reflect our focus on reliability, with uncertain data classified as non-members to reduce false positives, as illustrated in Section 5.3.

For LIMU-BERT, our methods significantly outperform existing SSL-based approaches in both precision and recall at low FPRs. Notably, even when compared against the top 5 methods that leverage substantially more background knowledge, our approaches show superior performance. The only exception is LiRA, which achieves better results, especially on the Motion dataset. However, LiRA's performance drops substantially when using LiRA*, a variant without access to the downstream classifier trained on the pre-training dataset. This performance gap suggests that LiRA's better inference results are likely based on additional knowledge gained from the downstream classifier. This contrast highlights both the importance of background knowledge in membership inference and the effectiveness of our proposed methods in identifying member samples with limited information.

When evaluating the SimCLR encoder, our methods generally surpass most SSL-based baselines, with one exception: EncoderMI-T shows slightly better performance at 1% FPR on the UCI dataset. Similarly, our methods also cannot outperform all the top 5 methods that possess additional information from the downstream classifier. However, the strong performance of our methods compared to several supervised learning-based baselines and most SSL-based

Table 3: Membership inference results for LIMU-BERT. Bold values indicate the best performance by our methods across all baselines, while underlined values represent the second-best. LiRA* refers to LiRA without access to the downstream classifier trained on the same pre-training dataset.

Datasets →	HHAR				Motion				UCI			
Metrics →	@1% FPR		@5% FPR		@1% FPR		@5% FPR		@1% FPR		@5% FPR	
Methods ↓	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Shokri et al.	0.7717	0.0339	0.6818	0.1074	0.5833	0.0193	0.6667	0.1006	0.0000	0.0035	0.5341	0.0557
Yeom et al.	0.6278	0.0169	0.6278	0.0843	0.4385	0.0078	0.4385	0.0390	0.4665	0.0087	0.4665	0.0437
Watson et al.	0.3823	0.0062	0.2537	0.0172	0.5714	0.0155	0.5593	0.0638	0.0000	0.0087	0.5385	0.0609
ML-Leaks	0.2759	0.0038	0.3376	0.0258	0.0000	0.0000	0.6557	0.0824	0.5430	0.0119	0.5430	0.0594
LiRA	0.9810	0.0142	0.9051	0.0710	0.9835	0.0233	0.9174	0.1084	0.9115	0.0212	0.6601	0.0971
LiRA*	0.9447	0.0099	0.4987	0.0497	0.9166	0.0107	0.5161	0.0619	0.3333	0.0050	0.4925	0.0487
EncoderMI-V	0.7826	0.0372	0.6245	0.0830	0.2857	0.0077	0.3658	0.0290	0.0000	0.0014	0.2609	0.0203
EncoderMI-T	0.7273	0.0267	0.5907	0.0730	0.4286	0.0077	0.3864	0.0329	0.5625	0.0130	0.5270	0.0565
Zheng et al.	0.2963	0.0043	0.4124	0.0358	0.1667	0.0097	0.4615	0.0464	0.2500	0.0043	0.3200	0.0246
Inferrer-1 $\mathcal{I}_1$	0.8750	**0.0702**	0.7328	**0.1375**	0.6471	0.0213	0.6711	0.1044	0.6316	**0.0232**	**0.6635**	**0.1058**
Inferrer-2 $\mathcal{I}_2$	0.8051	0.0453	0.6882	0.1117	0.2000	0.0058	0.5000	0.0503	0.5385	0.0116	0.5000	0.0507

Table 4: Membership inference results for SimCLR. Bold values indicate the best results among all SSL-based methods, while underlined values show the second-best results.

Datasets →	HHAR				Motion				UCI			
Metrics →	@1% FPR		@5% FPR		@1% FPR		@5% FPR		@1% FPR		@5% FPR	
Methods ↓	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Shokri et al.	0.0000	0.0000	0.0000	0.0005	0.2222	0.0039	0.2273	0.0116	0.3636	0.0058	0.3962	0.0362
Yeom et al.	0.9520	0.2086	0.8630	0.3217	0.8182	0.0522	0.5667	0.0658	0.3636	0.0058	0.2195	0.0130
Watson et al.	0.6441	0.0181	0.3933	0.0334	0.9107	0.1006	0.8015	0.2031	0.1818	0.0029	0.7083	0.1232
ML-Leaks	0.9496	0.1933	0.8652	0.3217	0.5000	0.0103	0.5192	0.0522	0.7407	0.0290	0.5882	0.0725
LiRA	0.9584	0.0890	0.7945	0.2109	0.8365	0.0394	0.7074	0.1013	0.9919	0.0094	0.9595	0.0469
LiRA*	0.9378	0.0524	0.8397	0.1665	0.9762	0.0109	0.8809	0.0543	0.9865	0.0081	0.9327	0.0406
EncoderMI-V	0.5000	0.0091	0.3082	0.0220	0.4444	0.0116	0.5937	0.0735	0.4615	0.0087	0.4355	0.0391
EncoderMI-T	0.6441	0.0181	0.4555	0.0420	0.6667	0.0213	0.5263	0.0580	**0.7273**	**0.0246**	0.5679	0.0667
Zheng et al.	0.4000	0.0067	0.4101	0.0348	0.0000	0.0039	0.4348	0.0387	0.0000	0.0014	0.4107	0.0362
Inferrer-1 $\mathcal{I}_1$	**0.8990**	**0.0916**	**0.7619**	**0.1604**	**0.8936**	**0.0986**	**0.7312**	**0.1354**	0.6956	0.0232	**0.5976**	**0.0739**
Inferrer-2 $\mathcal{I}_2$	0.8955	0.0859	0.7729	0.1695	0.8000	0.0464	0.6338	0.0870	0.4615	0.0087	0.5000	0.0507

methods demonstrates their ability to perform membership inference effectively, even with limited background knowledge.

Across all cases, we observe that Inferrer-1 consistently outperforms Inferrer-2, further highlighting how additional background knowledge of the pre-training distribution enhances membership inference performance. The relatively weaker performance of Inferrer-2 on the UCI dataset may be due to the encoder's strong generalization capabilities and minimal overfitting. This is supported by the small difference in loss values and the k-fold cross-validation results presented in Table 2.

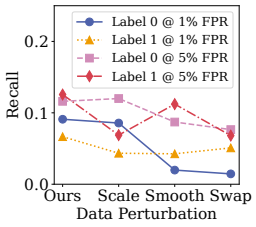
6.3 Impacts of Different Modules

In this subsection, we conduct an ablation study on the two key modules of our method and analyze its robustness under varying hyperparameters. For simplicity, we use the relatively less powerful Inferrer-2 to examine these effects.

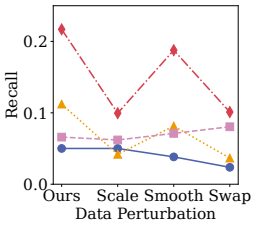
Table 5: Membership inference results on LIMU-BERT by activity label (0: static, 1: dynamic). Bold values indicate performances that exceed direct inference.

Metrics		@ 1% FPR				@ 5% FPR			
		Precision		Recall		Precision		Recall	
Labels		0	1	0	1	0	1	0	1
$\mathcal{I}_1$	HHAR	0.8621	**0.8947**	**0.0766**	**0.0853**	0.6327	**0.7649**	0.0949	**0.1602**
	Motion	**0.7778**	0.5000	**0.0464**	0.0107	**0.7500**	0.4286	**0.1477**	0.0429
	UCI	0.0000	0.5000	0.0087	0.0145	0.2857	**0.6800**	0.0231	0.0988
$\mathcal{I}_2$	HHAR	**0.8961**	**0.8776**	**0.0910**	**0.0666**	**0.6960**	**0.7149**	0.1161	**0.1256**
	Motion	**0.8571**	0.2000	**0.0591**	0.0032	**0.7429**	**0.5333**	**0.1281**	**0.0605**
	UCI	0.0000	0.5000	0.0032	**0.0132**	0.1905	0.4412	0.0161	0.0396

Activity Labeling. By generating activity labels for the samples to be inferred, we can conduct a fine-grained analysis of membership inference across different activities. Tables 5 and 6 present performance on LIMU-BERT and SimCLR encoders. The experimental results show that activity-specific inference improves performance across most cases. This improvement is particularly significant for

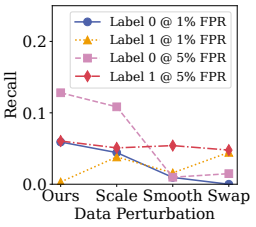


(a) LIMU-BERT

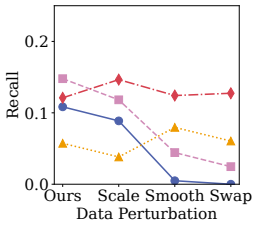


(b) SimCLR

Figure 4: The impact of different data perturbations on HHAR.



(a) LIMU-BERT



(b) SimCLR

Figure 5: The impact of different data perturbations on Motion.

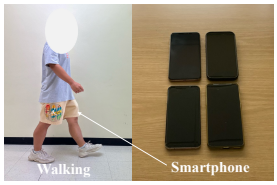
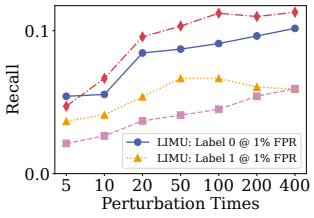


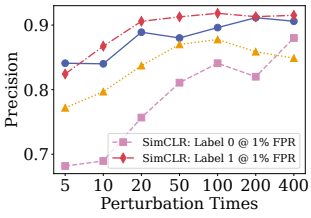
Figure 6: Experiment setup for downstream IMU sensing data collection.

Table 6: Membership inference results on SimCLR by activity label (0: static, 1: dynamic). Bold values indicate performances that exceed direct inference.

Metric		@ 1% FPR				@ 5% FPR			
		Precision		Recall		Precision		Recall	
Labels		0	1	0	1	0	1	0	1
$\mathcal{I}_1$	HHAR	0.3333	**0.9282**	**0.0766**	0.0853	0.2931	**0.8047**	**0.0949**	**0.1602**
	Motion	**0.8947**	0.8000	**0.0464**	0.0107	**0.8197**	0.6739	**0.1477**	0.0429
	UCI	**0.8000**	0.6667	0.0087	0.0145	**0.6731**	**0.6304**	0.0231	0.0988
$\mathcal{I}_2$	HHAR	0.8409	**0.9182**	0.0501	**0.1122**	0.6000	**0.8123**	0.0660	**0.2169**
	Motion	**0.9167**	0.7692	**0.1084**	**0.0573**	**0.7838**	**0.7059**	**0.1478**	**0.1210**
	UCI	**0.7692**	0.2000	**0.0450**	0.0026	**0.6364**	0.3929	**0.0997**	0.0369



(a) Recall at 1% FPR.



(b) Precision at 1% FPR.

Figure 7: Membership inference results of Inferrer-2 across various perturbation times N.

the SimCLR encoder on the UCI dataset, transforming the inference from ineffective to effective. This effect is likely due to certain activities generalizing better for specific encoders while exhibiting lower levels of overfitting. Consequently, these activities become more distinguishable within specific data subsets rather than across the entire dataset. These results suggest that knowledge of the underlying activity type provides valuable information that enhances membership inference capabilities. Moreover, activity labeling can provide additional benefits by guiding specific perturbation techniques, which we will demonstrate in the following part.

Data Perturbation. To evaluate our activity-based perturbation strategy, we compare Inferrer-2's performance across different perturbation methods. We examine three alternative strategies: 1) Scale: applying only scaling perturbations to all activities, 2) Smooth: applying smoothing to all activities, and 3) Swap: inverting our proposed approach by applying scaling to dynamic activities and smoothing to static activities. As shown in Figure 4 and Figure 5, our proposed strategy consistently outperforms these alternative strategies in most cases. Notably, the performance gap is particularly obvious when compared to the Swap strategy, which consistently yields the poorest results among all perturbation approaches. In addition, the results reveal that applying scaling can be detrimental to inferring dynamic activity, while it can be beneficial for static activities. This can be attributed to scaling may break the dynamic activity patterns to confuse the encoder. These observations strongly support our approach of selectively applying different perturbation types based on coarse activity types.

Perturbation Times. Our methods fundamentally rely on analyzing the stability characteristics of IMU data under multiple perturbations. Therefore, understanding the influence of the number of perturbations N is crucial for evaluating our methods' effectiveness and robustness. As depicted in Figure 7, recall and precision increase significantly as N grows from 5 to 50, after which the rate of improvement slows and reaches a plateau. This relationship validates the foundational principle of our methods: increasing the number of perturbations enhances the robustness and reliability of stability measurements. With more perturbation times, our approach can better characterize the inherent stability patterns of feature vectors that differentiate member data from non-member data, leading to more accurate membership inference.

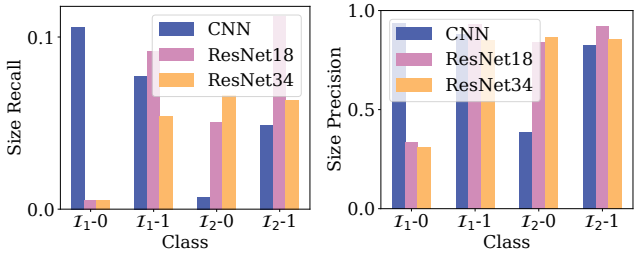
Table 7: Membership inference performance for dynamic activity data on the HHAR dataset at different smooth lengths.

Length L			10	15	20	25	30
LIMU-BERT	@ 1% FPR	Recall	0.0845	0.0710	0.0666	0.0509	0.0546
		Precision	0.8952	0.8716	0.8776	0.8354	0.8452
	@ 5% FPR	Recall	0.1937	0.1451	0.1256	0.1055	0.1122
		Precision	0.7969	0.7461	0.7149	0.6750	0.6948
SimCLR	@ 1% FPR	Recall	0.0748	0.1002	0.1122	0.1077	0.0710
		Precision	0.8818	0.9103	0.9182	0.9150	0.8738
	@ 5% FPR	Recall	0.1758	0.2378	0.2169	0.2124	0.1690
		Precision	0.7807	0.8267	0.8123	0.8114	0.7740

Smooth Length. For dynamic activities $\hat{y} = 1$, our methods apply smoothing perturbations to generate modified samples and analyze the stability characteristics of their corresponding feature vectors. Table 7 presents the inference results for dynamic activity data $\hat{y} = 1$ under various smoothing window lengths on the HHAR dataset. In general, the results demonstrate stable performance. However, it is worth noting that in some cases, the performance decreases slightly when the smoothing length is increased to 30. This

Table 8: Membership inference performance for static activity data on the HHAR dataset at different scaling factors.

Scaling Factors	LIMU-BERT				SimCLR			
	@ 1% FPR		@ 5% FPR		@ 1% FPR		@ 5% FPR	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
(10, 20)	0.0910	0.9067	0.1187	0.6923	0.0528	0.8333	0.0660	0.5102
(10, 30)	0.0923	0.8974	0.1161	0.6822	0.0514	0.8409	0.0660	0.6282
(10, 40)	0.0910	0.8961	0.1161	0.6960	0.0501	0.8409	0.0660	0.5568
(10, 50)	0.0884	0.8933	0.1201	0.7031	0.0475	0.8000	0.0567	0.5187



(a) Recall at 1% FPR. **(b) Precision at 1% FPR.**

Figure 8: Membership inference performance against the SimCLR with different size encoders. $\mathcal{I}_1 - 0$ denotes Inferrer-1 applied to static activity data.

suggests that a smoothing length of 30 may be overly aggressive for IMU readings with a total length of 120, potentially obscuring fine-grained temporal patterns that are crucial for effective membership inference in dynamic activities.

Scaling Factor. For the static activity data $\hat{y} = 0$, our methods introduce scaling to generate perturbed samples. Table 8 presents the membership inference results for static activity data on the HHAR dataset across a range of scaling factors, showing the precision and recall rates at different FPRs. In general, these performance metrics remain notably stable for both LIMU-BERT and SimCLR encoders across various tested scaling factors. This consistency in performance demonstrates that our proposed scaling perturbation strategy is inherently robust when dealing with static activity data.

Encoder Size. The increasing prevalence of large pre-trained encoders as foundation models across various tasks and domains makes it essential to examine the relationship between encoder size and vulnerability to membership inference. To systematically investigate this relationship, we conduct experiments using three encoder architectures of different complexity: a simple three-layer CNN, ResNet18 (our deployed model), and ResNet34. To ensure a fair comparison, we maintain identical pre-training configurations across all encoders. Intuitively, larger models are more prone to overfitting their training data, which should make membership inference more effective. However, this tendency is not obvious in our experimental results, as shown in Figure 8. In several scenarios, the membership inference even performs better on the simpler CNN compared to more complex models. For instance, in the case of $\mathcal{I}_1$ on static activity, our membership inference achieves higher recall and precision against the simple CNN compared to complex ResNet models. This could be because the simple CNN converges faster and learns the training data patterns more quickly within the same number of training epochs, potentially leading to more significant memorization of certain activities.

6.4 User Case: Fine-tuning Encoders for Specific Applications

As illustrated in Section 2, pre-trained encoders can be deployed locally and further fine-tuned to enhance downstream performance for specific user applications. This fine-tuning process enables models to adjust their parameters, allowing better adaptation to unique downstream data distributions. However, these updates may influence overfitting to the pre-training data, thereby affecting membership inference performance. To evaluate the privacy implications of pre-training data in a more practical setting, we construct a real-world IMU sensing task and fine-tune the target encoder to varying degrees. In our experiment, two participants performed four different activities: standing, walking, running, and cycling. During these activities, each participant used four different smartphone models, introducing real-world variability in device hardware. An example of the walking activity data is shown in Figure 6.

In many cases, users may only fine-tune the top layers of a pre-trained encoder to save computational resources. To assess privacy leakage risks associated with these fine-tuned encoders, we examine various fine-tuning percentages, which we denote as tuning rates. The membership inference performance of Inferrer-2 for the LIMU-BERT and SimCLR encoders across different tuning rates is shown in Figure 9 and Figure 10, respectively. In general, these results demonstrate that our proposed membership inference methods can compromise the privacy of the pre-training dataset, even under high tuning rates.

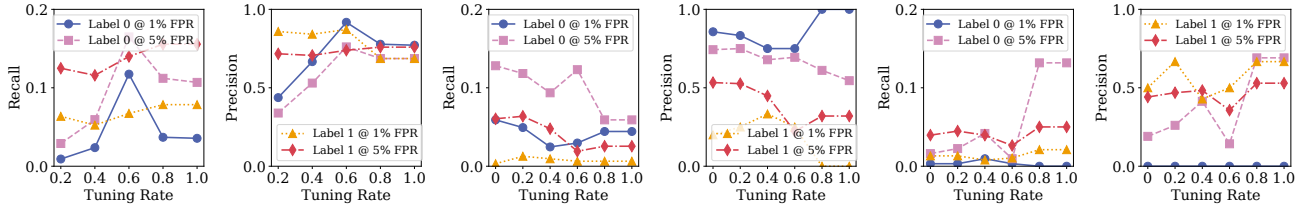
Interestingly, the inference performance improves as the tuning rate increases in some cases. This trend is particularly evident for the LIMU-BERT on the UCI dataset. This observation suggests that fine-tuning may make the overfitting of the encoder to the UCI data more pronounced by updating the model parameters, thereby improving the ability to distinguish member from non-member samples. Another observation is that the performance across different activity data can be exchanged in some cases, which is particularly obvious for SimCLR on the UCI dataset. This phenomenon may be because partial fine-tuning can remove the encoder's overfitting to one activity while making the overfitting to another activity more apparent.

Overall, these findings indicate that our proposed membership inference approach remains effective even when the target encoder is fine-tuned, highlighting the importance of carefully considering privacy implications when deploying pre-trained models in practical applications.

7 DISCUSSION ON COUNTERMEASURES

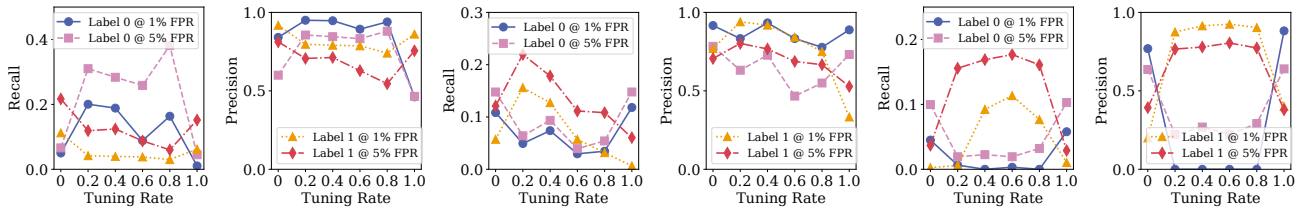
7.1 Early Stopping

Most existing membership inference methods are based on the overfitting of a model to its training dataset. One of the simple methods to mitigate overfitting is early stopping, which limits the training epochs to prevent the model from becoming too tailored to the training data [58]. To assess the effectiveness of early stopping



(a) HHAR: Recall (b) HHAR: Pre. (c) Motion: Recall (d) Motion: Pre. (e) UCI: Recall (f) UCI: Pre.

Figure 9: Membership inference performance of Inferrer-2 on LIMU-BERT across various fine-tuning rates, where "Pre." represents precision.



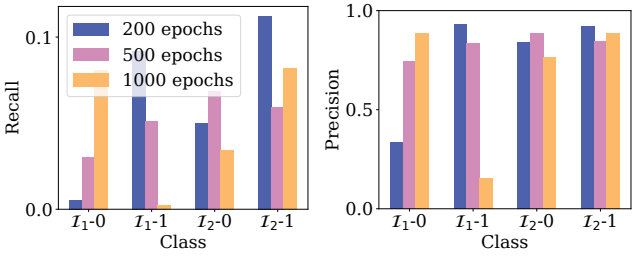
(a) HHAR: Recall (b) HHAR: Pre. (c) Motion: Recall (d) Motion: Pre. (e) UCI: Recall (f) UCI: Pre.

Figure 10: Membership inference performance of Inferrer-2 on SimCLR across various fine-tuning rates, where "Pre." represents precision.

in mitigating membership inference, we use SimCLR to pre-train encoders for different numbers of epochs while keeping all other pre-training hyperparameters constant.

Figure 11 shows the membership inference performance on the HHAR dataset across different classes and pre-training epochs. The results do not demonstrate a consistent trend indicating that early stopping reliably mitigates membership inference attacks. For instance, in some cases (e.g., $\mathcal{I}_1 - 1$), recall and precision are higher with 200 epochs compared to 500 or 1000 epochs, whereas for others (e.g., $\mathcal{I}_2$) performance is less affected by pre-training epochs. This may be attributed to the inherent diversity of IMU data within each activity type, which requires longer training periods to effectively learn inter-class distinctions.

Moreover, early stopping presents an additional challenge: while it might help prevent overfitting, it can also compromise the encoder's performance, as longer training periods typically yield better model capabilities.



(a) Recall at 1% FPR. (b) Precision at 1% FPR.

Figure 11: Membership inference performance on the HHAR Dataset using the SimCLR encoder across varying pre-training epochs.

7.2 Data Augmentation

Data augmentation techniques, which involve creating modified versions of training examples through transformations like rotations and crops, have proven essential for preventing overfitting and improving model generalization in supervised learning [43]. However, in SSL frameworks, these same augmentation strategies can inadvertently introduce privacy vulnerabilities. For instance, EncoderMI [28] shows how the same data augmentation methods used during pre-training can be exploited to perform membership inference on pre-trained image encoders. Besides, since data augmentation is already a core component of contrastive learning, introducing additional augmentation techniques is unlikely to enhance privacy protection and may even provide inferrers with more information about the training process.

7.3 Differential Privacy

Differential privacy (DP) provides formal privacy guarantees for algorithms operating on aggregate databases by ensuring that the presence or absence of any single data point has minimal impact on the algorithm's output [1, 13]. In practice, DP is typically implemented by adding carefully calibrated noise to either the training data, objective function, or gradients during training [20, 41]. For example, DP-SGD [1] applies random Gaussian noise to the gradient computed during the stochastic gradient descent process. While the addition of noise is beneficial for privacy guarantees, it may significantly degrade the quality of the learned representations. This

degradation in representation quality often leads to unsatisfactory performance in downstream classification tasks [60], creating a challenging trade-off between privacy protection and model utility.

8 RELATED WORK

8.1 SSL-based Sensing Encoders

This paper focuses on SSL-based IMU sensing encoders, which can effectively exploit abundant unlabeled sensor data to enhance generalization [14]. TPN [38] employs a multi-task temporal CNN as the encoder, training it to recognize transformations applied to raw accelerometer data. LIMU-BERT [52] adapts BERT architecture for generalizable feature extraction from multiple unlabeled IMU sensing data. ColloSSL [19] leverages unlabeled multi-device data from a single user as different augmentations, learning high-quality features through multi-view contrastive loss. Cosmo [33] introduces novel fusion-based contrastive learning and quality-guided attention mechanisms that effectively integrate multiple sensor modalities for human activity recognition. CrossHAR [16] proposes hierarchical pretraining with physically-informed data augmentation to improve generalization.

The success of these approaches in leveraging large amounts of unlabeled IMU data demonstrates the growing importance of SSL in sensor-based applications. However, this widespread adoption of SSL methods also raises important questions about potential privacy vulnerabilities in these systems, which motivates our investigation.

8.2 Membership Inference

Membership inference can be a practical threat to machine learning models by determining whether a given input is part of the model's training dataset [27, 40]. For traditional supervised learning models, the inferrer can exploit real label information to perform membership inference. Shokri et al. [42] propose training multiple shadow models to replicate the behavior of the target model and then using a binary classifier to predict membership. Yeom et al. [56] assume that the model's loss on member data should be lower than on non-member data. Salem et al. [39] choose the top-k output logits from the target model to infer membership. LiRA [6] formulates membership inference as a hypothesis test, enhancing its effectiveness by incorporating label information and logit scaling.

For SSL-based encoders, the inferrer cannot leverage label information or classifier-specific indicators. EncoderMI [28] exploits data augmentation deployed in contrastive learning to infer membership. Mattern et al. [32] propose membership inference to pre-trained language models by neighborhood comparison. Zheng et al. [25] employ reconstruction errors of masked images as membership inference signals in masked image modeling-based encoders. The work most similar to ours is PartCrop [60], which presents a unified approach for various SSL methods in the image domain by cropping images and analyzing the responses in the image feature map.

Our proposed methods have some unique aspects. First, we focus on IMU sensor data, which differs from well-studied image and text data and can contain sensitive identity information. While there has been limited work addressing this privacy threat within the context of federated learning [8], our study is the first to explore its impact on SSL models. Second, we leverage the physical information in IMU readings for pseudo-labeling and design adaptable perturbation strategies for membership inference across different SSL methods In contrast, existing methods designed for SSL methods are often tailored to a specific algorithm and may not be applicable to others.

9 FUTURE WORK

In future work, we can explore IMU sensing encoders from four key perspectives, as outlined below.

Multi-modal Encoders. Multi-modal models have demonstrated potential for enhancing sensing performance [5, 50, 53, 55]. However, their dependence on larger datasets spanning multiple modalities, the potential use of multiple encoders, and the expansion of encoder size may increase the risk of privacy leakage. Future work should investigate the potential vulnerabilities and privacy implications of multi-modal encoders.

Large Foundation Models. Foundation models can serve as the base model for different downstream tasks through fine-tuning or prompt engineering. In this paper, we only focus on the human activity recognition task. Understanding and addressing these privacy challenges is crucial before deploying foundation sensing models in real-world applications where they can process abundant sensitive information about individuals' behaviors and daily patterns.

Generative Models. While our paper focuses on using SSL-based encoders to generate feature vectors, generative models can directly synthesize various sensing data [10, 11, 23, 26, 46–48], playing an important role in deep learning-based sensing applications by improving data diversity. Future work could investigate whether these generative models could potentially compromise the privacy of the pre-training sensing data.

Privacy Protection. The goal of investigating membership inference is not to compromise privacy but rather to protect against privacy leakage [18]. Future work should focus on developing robust methods, particularly for securing IMU data that can contain sensitive personal information, to effectively prevent information leakage.

10 CONCLUSION

We propose a novel membership inference method for unlabeled IMU sensing data and encoders trained using different SSL algorithms. Our approach is designed to work without detailed knowledge of the SSL algorithm, making it more practical to implement. To achieve this, we leverage the physical information in IMU readings to roughly classify unlabeled IMU readings into two categories: static and dynamic activity. We then apply specific perturbations to each activity type. Our key insight is that when these perturbations are applied to the member data that is used to train the model, the target encoder produces more inconsistent outputs compared to non-member data. This occurs because the target encoder may recognize some perturbed variants as closely resembling the original training data, while others are different. Our experimental results on multiple SSL methods and datasets show that our method can achieve high recall and precision at low FPRs.

ACKNOWLEDGMENTS

This work is supported in part by the NSF (CNS-2415209, CNS-2321763, CNS-2317190, IIS-2306791, and CNS-2319343).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Muhammad Ahmad, Adil Mehmood Khan, Joseph Alexander Brown, Stanislav Protasov, and Asad Masood Khattak. 2016. Gait fingerprinting-based user identification on smartphones. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 3060–3067. https://doi.org/10.1109/IJCNN.2016.7727588

[3] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. 2024. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems* 36 (2024).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Qiming Cao, Hongfei Xue, Tianci Liu, Xingchen Wang, Haoyu Wang, Xincheng Zhang, and Lu Su. 2024. mmCLIP: Boosting mmWave-based Zero-shot HAR via Signal-Text Alignment. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 184–197.

[6] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1897–1914.

[7] Hui Chen, Charles Gouin-Vallerand, Kévin Bouchard, Sébastien Gaboury, Mélanie Couture, Nathalie Bier, and Sylvain Giroux. 2024. Contrastive Self-Supervised Learning for Sensor-Based Human Activity Recognition: A Review. *IEEE Access* (2024).

[8] Kongyang Chen, Dongping Zhang, Sijia Guan, Bing Mi, Jiaxing Shen, and Guoqing Wang. 2024. Private Data Leakage in Federated Human Activity Recognition for Wearable Healthcare Devices. *arXiv preprint arXiv:2405.10979* (2024).

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[10] Xingyu Chen and Xinyu Zhang. 2023. Rf genesis: Zero-shot generalization of mmwave sensing through simulation-based data synthesis and generative diffusion models. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 28–42.

[11] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. 2024. RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 77–92.

[12] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[14] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2022. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–47.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. 2024. CrossHAR: Generalizing Cross-dataset Human Activity Recognition via Hierarchical Self-Supervised Pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–26.

[17] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[18] Li Hu, Anli Yan, Hongyang Yan, Jin Li, Teng Huang, Yingying Zhang, Changyu Dong, and Chunsheng Yang. 2023. Defenses to membership inference attacks: A survey. *Comput. Surveys* 56, 4 (2023), 1–34.

[19] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. Collossl: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–28.

[20] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*. 1895–1912.

[21] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. 2020. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881* (2020).

[22] Jeya Vikranth Jeyakumar, Liangzhen Lai, Naveen Suda, and Mani Srivastava. 2019. SenseHAR: a robust virtual activity sensor for smartphones and wearables. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 15–28.

[23] Hui Ji and Pengfei Zhou. 2024. Advancing PPG-Based Continuous Blood Pressure Monitoring from a Generative Perspective. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 661–674.

[24] Youpeng Li, Xuyu Wang, and Lingling An. 2023. Hierarchical clustering-based personalized federated learning for robust and fair human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–38.

[25] Zheng Li, Xinlei He, Ning Yu, and Yang Zhang. 2024. Membership Inference Attack Against Masked Image Modeling. *arXiv preprint arXiv:2408.06825* (2024).

[26] Peng Liao, Xuyu Wang, Lingling An, Shiwen Mao, Tianya Zhao, and Chao Yang. 2024. TFSemantic: A Time–Frequency Semantic GAN Framework for Imbalanced Classification Using Radio Signals. *ACM Transactions on Sensor Networks* 20, 4 (2024), 1–22.

[27] Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng. 2019. SocInf: Membership inference attacks on social media health data with machine learning. *IEEE Transactions on Computational Social Systems* 6, 5 (2019), 907–921.

[28] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. 2021. Encodermi: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2081–2095.

[29] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering* 35, 1 (2021), 857–876.

[30] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time arm skeleton tracking and gesture inference tolerant to missing wearable sensors. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 287–299.

[31] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*. 49–58.

[32] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462* (2023).

[33] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.

[34] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

[35] Hangwei Qian, Tian Tian, and Chunyan Miao. 2022. What makes good contrastive learning on small-scale wearable-based tasks?. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 3761–3771.

[36] Yanhui Ren, Di Wang, Lingling An, Shiwen Mao, and Xuyu Wang. 2024. Multi-positive sample quantum contrastive learning for human activity recognition. In *Proc. IEEE GLOBECOM 2024*. Cape Town, South Africa.

[37] Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.

[38] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[39] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).

[40] Yi Shi and Yalin E Sagduyu. 2022. Membership inference attack and defense for wireless signal classifiers with deep learning. *IEEE Transactions on Mobile Computing* 22, 7 (2022), 4032–4043.

[41] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1310–1321.

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[43] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.

[44] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.

[45] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

15–28.

[46] Ningning Wang, Tianya Zhao, Shiwen Mao, Harrison X Bai, Zhicheng Jiao, and Xuyu Wang. 2024. ECG-grained Cardiac Monitoring Using RFID. In *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–9.

[47] Ningning Wang, Tianya Zhao, Shiwen Mao, and Xuyu Wang. 2025. Privacy-Preserving Wi-Fi Data Generation via Differential Privacy in Diffusion Models. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*. IEEE, 1–10.

[48] Tianshi Wang, Jinyang Li, Ruijie Wang, Denizhan Kara, Shengzhong Liu, Davis Wertheimer, Antoni Viros i Martin, Raghu Ganti, Mudhakar Srivatsa, and Tarek Abdelzaher. 2023. SudokuSens: Enhancing Deep Learning Robustness for IoT Sensing Applications using a Generative Approach. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 15–27.

[49] Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. 2021. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440* (2021).

[50] Yuxuan Weng, Guoquan Wu, Tianyue Zheng, Yanbing Yang, and Jun Luo. 2024. Large Model for Small Data: Foundation Model for Cross-Modal RF Human Activity Recognition. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 436–449.

[51] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[52] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[53] Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, and Jiming Chen. 2023. MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 1–14.

[54] Chao Yang, Xuyu Wang, and Shiwen Mao. 2022. TARF: Technology-agnostic RF sensing for human activity recognition. *IEEE journal of biomedical and health informatics* 27, 2 (2022), 636–647.

[55] Huanqi Yang, Mingda Han, Mingda Jia, Zehua Sun, Pengfei Hu, Yu Zhang, Tao Gu, and Weitao Xu. 2023. XGait: Cross-Modal Translation via Deep Generative Sensing for RF-based Gait Recognition. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 43–55.

[56] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.

[57] Yafeng Yin, Lei Xie, Zhiwei Jiang, Fu Xiao, Jiannong Cao, and Sanglu Lu. 2024. A Systematic Review of Human Activity Recognition Based On Mobile Devices: Overview, Progress and Trends. *IEEE Communications Surveys & Tutorials* (2024).

[58] Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, Vol. 1168. IOP Publishing, 022022.

[59] Han Zhou, Yi Gao, Xinyi Song, Wenxin Liu, and Wei Dong. 2019. Limbmotion: Decimeter-level limb tracking for wearable-based human-computer interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.

[60] Jie Zhu, Jirong Zha, Ding Li, and Leye Wang. 2024. A unified membership inference method for visual self-supervised encoder via part-aware capability. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 1241–1255.

Protocol-agnostic and Data-free Backdoor Attacks on Pre-trained Models in RF Fingerprinting

Tianya Zhao*, Ningning Wang*, Junqing Zhang†, Xuyu Wang*§
*Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, US
†Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3GJ, United Kingdom
Emails: tzhao010@fiu.edu, nwang012@fiu.edu, junqing.zhang@liverpool.ac.uk, xuywang@fiu.edu

IEEE INFOCOM 2025 - IEEE Conference on Computer Communications | 979-8-3315-4305-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/INFOCOM55648.2025.11044704

***Abstract*—While supervised deep neural networks (DNNs) have proven effective for device authentication via radio frequency (RF) fingerprinting, they are hindered by domain shift issues and the scarcity of labeled data. The success of large language models has led to increased interest in unsupervised pre-trained models (PTMs), which offer better generalization and do not require labeled datasets, potentially addressing the issues mentioned above. However, the inherent vulnerabilities of PTMs in RF fingerprinting remain insufficiently explored. In this paper, we thoroughly investigate data-free backdoor attacks on such PTMs in RF fingerprinting, focusing on a practical scenario where attackers lack access to downstream data, label information, and training processes. To realize the backdoor attack, we carefully design a set of triggers and predefined output representations (PORs) for the PTMs. By mapping triggers and PORs through backdoor training, we can implant backdoor behaviors into the PTMs, thereby introducing vulnerabilities across different downstream RF fingerprinting tasks without requiring prior knowledge. Extensive experiments demonstrate the wide applicability of our proposed attack to various input domains, protocols, and PTMs. Furthermore, we explore potential detection and defense methods, demonstrating the difficulty of fully safeguarding against our proposed backdoor attack.**

***Index Terms*—Backdoor Attack, Pre-trained Model, Radio Frequency Fingerprinting, Security.**

I. Introduction

The proliferation of the Internet of Things (IoT) has led to the ubiquitous integration of wireless technology in daily life. As the number of wireless devices continues to grow, there is a critical need for effective and efficient device authentication methods [1]–[3]. Radio frequency (RF) fingerprinting has emerged as a promising technique, offering enhanced resistance to tampering and spoofing compared to conventional methods [4], [5]. RF fingerprints are unique characteristics that arise from inherent physical imperfections in the analog circuitry of RF emitters, introduced during the manufacturing process [6], [7]. These subtle imperfections affect transmitted signals without compromising overall device functionality, resulting in a distinct fingerprint for each RF emitter, including ultra-low-power and legacy devices.

Deep neural networks (DNNs) have demonstrated remarkable capabilities in automatically extracting and classifying RF fingerprints [8]–[10]. However, they face two significant challenges in RF fingerprinting applications: the need for large amounts of high-quality labeled data and vulnerability to domain shift. While previous studies have explored few-shot learning [11], [12] and domain adaptation techniques [13], [14] to mitigate these issues, these approaches have limitations and fail to fully leverage the abundant unlabeled data. The success of large language models (LLMs) such as GPT [15] and BERT [16] has sparked increased interest in self-supervised learning (SSL) across various domains, including RF fingerprinting [17], [18]. The SSL pipeline consists of two key components: pre-trained models (PTMs) and downstream classifiers. PTMs are trained on large amounts of unlabeled data to serve as feature extractors, while downstream classifiers are built on these PTMs using minimal or no labeled data. This approach enhances generalization and reduces the need for extensive labeled datasets, potentially addressing the data scarcity and domain shift challenges in RF fingerprinting.

Applying SSL techniques to train general PTMs for RF fingerprinting could potentially improve authentication performance. However, ensuring security remains a top priority for these systems. In the current deep learning landscape, PTMs are typically large, enabling them to capture extensive contextual information at the cost of being computationally expensive to train. To mitigate this burden, a common practice is to download open-source PTMs from platforms like GitHub and HuggingFace and then fine-tune them for specific tasks. While this approach is convenient and efficient, the widespread use of publicly available PTMs raises concerns about potential security vulnerabilities in RF fingerprinting.

One practical threat is *data poisoning-based backdoor attacks*, where an adversary seeks to manipulate the victim model to misbehave on inputs containing predefined triggers while maintaining normal behavior on clean inputs. Backdoor attacks have been extensively studied in supervised DNNs, and recent work has explored their impacts on unsupervised PTMs in computer vision (CV) and natural language processing (NLP) domains. For example, BadEncoder [19] investigates injecting backdoors into image PTMs, causing downstream classifiers to inherit the backdoor behavior. Shen *et al.* demonstrate backdoor attacks on PTMs by mapping triggers to predefined output representations in the NLP domain [20]. However, there is limited analysis of backdoor attacks on PTMs in the RF fingerprinting domain. Given that RF fingerprinting enables device identification and impacts the security of broader applications, it is crucial to investigate potential backdoor threats. Therefore, this paper studies *protocol-agnostic* and *data-free*

§The corresponding author is Xuyu Wang (xuywang@fiu.edu).

backdoor attacks on PTMs to meet the practical settings of RF fingerprinting systems.

Challenges. Implementing backdoor attacks on PTMs in RF fingerprinting systems presents several significant challenges. First, the security-critical nature of RF fingerprinting systems prompts providers to implement robust protection for both PTMs and downstream training processes, significantly limiting an attacker's capabilities. Existing powerful backdoor attacks typically rely on manipulating the training process to obtain the gradient information for optimizing trigger patterns and mapping them to targeted classes [21]. However, in protected RF fingerprinting systems, attackers cannot control this process. Furthermore, most backdoor attacks on PTMs require access to downstream data and label information [19], [22], [23], which is highly sensitive and should be inaccessible to attackers in these systems. Therefore, the primary challenge lies in injecting backdoor behaviors into PTMs and impacting downstream classification without this crucial knowledge. Second, system providers may be cautious about using PTMs, even those from reputable open-source platforms. To enhance security without incurring significant computational costs, they may fine-tune several layers of PTMs using their own clean data, adding an extra layer of protection against potential backdoors. This creates an additional challenge of maintaining the effectiveness of backdoor attacks after such fine-tuning defense strategy. Third, any added trigger should not significantly impact the system's performance and should be resistant to detection methods. This poses a unique challenge for RF fingerprinting systems since input in-phase/quadrature (I/Q) data often undergoes signal processing, transforming it into the frequency or time-frequency domain. This requires the trigger to be effective and stealthy in both the time domain and the frequency domain.

Solution. To address the aforementioned challenges, we propose a practical backdoor attack for RF fingerprinting PTMs by retraining a benign PTM without controlling the downstream training process. First, we carefully design predefined output representations (PORs) of PTMs that serve as inputs for downstream classifiers. Then, we define a set of triggers and establish connections with the PORs, enabling the transfer of the backdoor to the downstream task. The backdoor attack will be activated when any predefined trigger is injected into the I/Q data. Given the security-critical nature of these systems, we implement this backdoor injection in a data-free manner. To achieve this, we use a small amount of unlabeled data to construct a substitute dataset that differs from the downstream data. This substitute dataset can be collected by attackers or downloaded from the internet and may even be an out-of-distribution dataset.

The main contributions of this paper are as follows.

- To the best of our knowledge, this is the first work to investigate backdoor attacks on PTMs in RF fingerprinting. We develop a practical backdoor injection method without requiring access to downstream data.
- We propose a novel approach to generate output representations, enabling the successful implementation of protocol-agnostic backdoor attacks on PTMs.
- We conduct comprehensive experiments to evaluate our backdoor attacks on various protocols (i.e., 802.11a/g and LoRa) with different PTMs on both time-domain and time-frequency domains across multiple datasets. These experiments show the broad applicability and effectiveness of our approach.

The rest of the paper is organized as follows. Section III discusses the related work and Section II introduces background on SSL. Section IV illustrates the attack scenario and threat model. Our proposed backdoor attacks are elaborated in Section V. Section VI presents the experimental evaluations and analysis. Finally, Section VII concludes this paper.

II. Background: SSL

Traditional supervised learning heavily relies on large volumes of labeled data, which can be costly and time-consuming to acquire. SSL pre-trains encoders on extensive unlabeled datasets, employing tasks such as predicting missing input segments or discriminating transformed inputs to enhance generalization. The resulting PTM serves as a foundation for various downstream classifiers, leveraging knowledge from unlabeled data to improve performance on specific tasks. This paper focuses on two mainstream SSL approaches: generative and contrastive methods [24]. Generative methods train an encoder f_θ to represent input data $\mathbf{x}$ as a discernible representation $f_\theta(\mathbf{x})$, paired with a decoder that reconstructs $\mathbf{x}$ from $f_\theta(\mathbf{x})$. In the NLP domain, the most popular generative model is auto-regressive models such as BERT and GPT series. On the other hand, contrastive methods train an encoder to transform augmented input $\mathbf{x}'$ into a vector representation $f_\theta(\mathbf{x}')$, enabling similarity measurements between inputs. A notable example is SimCLR [25], which aims to learn through comparisons using the NT-Xent loss as follows:

$$\mathcal{L} = -\frac{1}{K}\sum_{i=1}^{K}\frac{exp(sim(f_\theta(\mathbf{x}'_i), f_\theta(\mathbf{x}'_j))/\tau)}{\sum_{k=1,k\neq i}^{2K} exp(sim(f_\theta(\mathbf{x}'_i), f_\theta(\mathbf{x}'_k))/\tau)}, \quad (1)$$

where $sim(\cdot)$ denotes the similarity function, K is the batch size, and τ represents the temperature hyperparameter.

III. Related Work

A. RF Fingerprinting PTMs.

Recent works have emphasized the significance of PTMs in RF fingerprinting. Chen *et al.* employ contrastive learning to extract domain-invariant features, demonstrating its effectiveness in mitigating domain-specific variations for robust RF fingerprinting [18]. Liu *et al.* introduce SSL during pre-training to address label dependence issues and utilize knowledge transfer in fine-tuning to overcome sample dependence limitations [17]. Similarly, Shao *et al.* apply SSL to improve emitter identification performance through RF fingerprints [26]. These studies demonstrate the promise of SSL in the RF fingerprinting task, making it imperative to investigate the security vulnerabilities of these methods.

B. Backdoor Attacks.

Backdoor attacks pose a significant threat to DNNs across related domains. Zhao *et al.* [27], [28] leverage explainable tools to design backdoor attacks on model-agnostic RF fingerprinting systems. [29] designs a training-based backdoor trigger generation approach on RF signal classification. [30] proposes backdoor attacks on wireless traffic prediction in both centralized and distributed training scenarios. TrojanFlow [21] implements attacks on network traffic classification by simultaneously optimizing a trigger generator and the target model. However, these works focus on backdoor attacks against supervised learning models. As the field evolves toward foundation models, there is a growing need to investigate security implications and vulnerabilities specific to PTMs.

BadEncoder [19] first proposes backdoor attacks targeting image PTMs, followed by several concurrent studies in the same domain [22], [23]. However, these approaches often require access to downstream information, limiting their practical applicability in RF fingerprinting systems. The most closely related work is in the NLP domain, where they design output representations mapping to selected tokens for launching attacks [20]. Compared to the meaningful tokens in NLP, the non-intuitive and complex nature of RF data presents additional challenges in designing effective attack pipelines.

Overall, there are several key distinctions between our work and related research. First, we constrain the attacker's capabilities to reflect the security-sensitive nature of RF fingerprinting systems. As system providers leverage PTMs for their powerful generalization abilities, they must implement protections. Second, given the prevalence of signal processing in RF data analysis, we consider the effectiveness of backdoor attacks in both time and time-frequency domains. Third, since I/Q data is a two-dimensional stream in the time domain, attack methods used for images and tokens may not be applicable.

IV. Attack Scenario and Threat Model

A. Attack Scenario Description

The overall backdoor injection process is shown in Fig. 1. Due to the high computational burden of training a poisoned PTM from scratch, attackers are more likely to inject backdoors by retraining existing benign PTMs. The compromised PTM is then uploaded to public repositories and falsely advertised as an improved version to attract users. A potential victim might adopt this backdoored PTM if downstream classifiers built upon it demonstrate satisfactory performance in RF fingerprinting tasks. Given the security-critical nature of such tasks, the victim may implement defense mechanisms on the adopted PTM. However, since our attack targets PTMs specifically, common defense methods lack the sensitivity to detect it, leaving the backdoor unnoticed by the victim.

B. Threat Model

1) Attacker's Goal: We consider an attacker who aims to inject backdoors into a PTM f_θ in a data-free manner so that a downstream classifier g built on the backdoored PTM f_{θ_b}

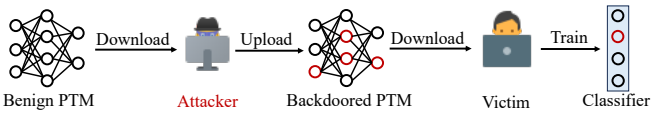


Fig. 1. Attack scenario: our attack is stealthy.

renders the RF fingerprinting system ineffective with attacker-chosen triggers $\mathbf{t}_j \in T$. The attacker has three goals to achieve:

- **Stealthiness goal.** The backdoored PTM must maintain its utility to remain stealthy. The attacker needs to ensure that downstream classifiers built on the compromised PTM still perform well on clean data $\mathbf{x}$, thus deceiving victims into adopting the backdoored model. Besides, triggers need to be concealed to evade detection methods.
- **Effectiveness goal.** When a downstream classifier is built on a backdoored PTM, it should misclassify any input containing a trigger. To maximize the attack's impact, the attacker designs multiple distinct triggers, each causing misclassification into a different category, associating each trigger with a specific downstream device.
- **Robustness goal.** Backdoored PTMs should achieve the above two goals, particularly maintaining effectiveness under potential defenses and protections.

In summary, the overall goals can be represented as:

$$g(f_{\theta_b}(\mathbf{x}^p)) \neq g(f_\theta(\mathbf{x}));\ max(|g(f_{\theta_b}(\mathbf{x}^p))|); \tag{2}$$

$$g(f_\theta(\mathbf{x})) = g(f_{\theta_b}(\mathbf{x})), \tag{3}$$

where $\mathbf{x}^p = \mathbf{x} \oplus \mathbf{t}$ denotes poisoned samples with triggers and $max(|\cdot|)$ represents maximizing the number of output classes.

2) Attacker's Capability: We consider a scenario where an attacker obtains a clean PTM from a service provider, injects backdoors into it, and then shares the backdoored PTM with potential victims (e.g., by republishing it for public download). In this context, the attacker has access to the original clean PTM. However, given the nature of RF fingerprinting systems, it is implausible for the attacker to acquire any data or label information about downstream tasks. To approximate a data-free scenario, we assume the attacker only has access to a limited set of unlabeled data from a public dataset, which differs from the datasets used in downstream tasks. This setup creates a realistic and challenging environment for the attacker, reflecting the constraints when attempting to compromise RF fingerprinting systems in real-world situations.

V. Backdoor Methodology

A. Overview

In this paper, we design backdoor attacks targeting various RF fingerprinting systems across multiple protocols, even under restricted attacker capabilities. To achieve the goals mentioned above, our idea is to manipulate the PTM so that 1) it generates similar output representations for clean substitute data as it does with the benign PTM, and 2) it produces similar output representations for poisoned substitute data with the PORs. Therefore, a downstream classifier built on our

backdoored PTM will perform normally on clean inputs while misbehaving on poisoned inputs embedded with triggers.

As shown in Fig. 2, our attack pipeline consists of three phases: substitute dataset collection, poisoned data generation, and output representation manipulation. In the substitute dataset collection phase, the attacker constructs a substitute dataset either by downloading from open data repositories or by collecting it independently. Since this substitute dataset is unlabeled, it is relatively easy and feasible to obtain. In the poisoned data generation stage, we first design a set of triggers $T = \{\mathbf{t}_j\}_{j=1}^{N_t}$ for the backdoor attacks. The substitute dataset D_s is then divided into two parts: a small portion designated as the poisoned dataset D_p and the remainder as the clean dataset D_c. Data in the poisoned dataset are embedded with the designed triggers. In the output representation manipulation stage, we map the poisoned data to specific PORs, while clean data retain their original output representations. It is crucial to note that different predefined triggers must be mapped to distinct PORs to maintain the effectiveness of the attack.

B. Backdoor Design

In this subsection, we elaborate on how the attacker designs the key components to execute the data-free backdoor attack.

1) Substitute Dataset: Due to the impracticality of obtaining downstream data and label information for RF fingerprinting systems, we have to construct a substitute dataset to implant backdoor behaviors. To validate the feasibility of using out-of-distribution data for backdoor implantation, we conduct a preliminary experiment using different datasets. Fig. 3 presents the t-SNE results of two distinct datasets: devices 0 to 2 belong to one dataset, while devices 3 to 5 belong to another. Fig. 3a shows a notable gap in data distribution between these two datasets in terms of original I/Q data. However, Fig. 3b shows this gap significantly narrows after the data is fed into the PTM, with representations spread across a unified space. This observation suggests that out-of-distribution data can generate representations occupying similar space to those of target data. Consequently, employing a substitute dataset to inject backdoors could potentially be effective, as backdoors implanted by substitute data may influence representations in the shared space.

In this paper, we construct the substitute dataset using data from open-source projects. To achieve the dual objectives of implanting backdoors and maintaining accuracy on clean samples, we divide the substitute dataset $D_s = \{\mathbf{x}_i\}_{i=1}^{S}$ into two parts: a small portion designated as the poisoned dataset $D_p = \{\mathbf{x}_k^p\}_{k=1}^{N}$, and the remainder serving as the clean dataset $D_c = \{\mathbf{x}_i\}_{i=1}^{M}$. The ratio of poisoned to total data is defined as the poisoning rate $\varphi \doteq \frac{N}{S}$.

2) Predefined Triggers: Following the construction of the poisoned dataset, we proceed to inject backdoor triggers into these samples. Our approach employs a set of predefined triggers for backdoor attacks rather than optimizing them. This decision is based on two key factors. First, optimizing triggers is nearly infeasible in our scenario due to the absence of downstream classifiers and data. Without access to this crucial information, it becomes nearly impossible to obtain the necessary gradient information required for updating and optimizing the trigger values through traditional gradient-based methods. Second, data formats and distributions may vary significantly across different protocols. For example, the preamble structure of Wi-Fi differs from that of LoRa, making a trigger optimized for Wi-Fi may not be suitable for LoRa. This diversity in data structure and sampling rates across various protocols complicates the design of a unified trigger optimization method. Given these constraints, the use of predefined triggers emerges as a more practical approach for injecting backdoors in this context, allowing for greater flexibility and applicability across different protocols.

In this paper, we choose to formulate the trigger set using time domain Gaussian noise, which has proven effective for launching backdoor attacks in related domains [29]. Unlike targeted attacks in supervised DNNs, our approach aims to induce misclassification into multiple classes by adding various triggers to inputs of PTMs, thereby contaminating the downstream classifier. Considering the output representations given by $f_\theta(\mathbf{x} \oplus \mathbf{t}_j) = \mathbf{W}_\theta \cdot (\mathbf{x} \oplus \mathbf{t}_j) + \mathbf{B}_\theta$, our goal is to ensure that these representations differ sufficiently when different triggers are applied. Given that the weight $\mathbf{W}_\theta$ and bias $\mathbf{B}_\theta$ matrices remain constant across samples, the most effective strategy is to introduce inherent differences in the poisoned samples $\mathbf{x}^p$ themselves after adding various triggers $\mathbf{t}_j$. Intuitively, we assume that $f_\theta(\mathbf{x} \oplus \mathbf{t}_j)$ and $f_\theta(\mathbf{x} \oplus -\mathbf{t}_j)$ will generate two relatively dissimilar output representations by simply reversing the trigger value. Therefore, we design the j-th trigger $\mathbf{t}_j$ in the trigger set T as follows:

$$\mathbf{t}_j = \begin{cases} N(0, \sigma; L), & j \leq \frac{N_t+1}{2}; \\ -\mathbf{t}_{N_t-j}, & j > \frac{N_t+1}{2}, \end{cases} \tag{4}$$

where L denotes the length of the trigger, which simultaneously regulates the trigger's size along with σ. In this paper, we use $L = 48$ and $\sigma = 0.1$ as the baseline settings.

3) Output Representations: While incorporating triggers into RF data can induce shifts in output representations, these minor changes alone are insufficient to launch a successful backdoor attack on downstream classifiers. Table I presents experimental results demonstrating that directly adding triggers to the inputs yields only minimal accuracy drops. Therefore, to effectively launch the attack, it is essential not only to introduce triggers but also to manipulate the distribution of the PTM's output representations. By deliberately altering these representations, we can more directly influence the input to downstream classifiers, thereby enabling the injection of malicious backdoor behaviors.

TABLE I
DOWNSTREAM ACCURACY DROPS WITH ONLY ADDED TRIGGERS.

Dataset	ORACLE	WiSig	CORES	NetSTAR	Ours
Acc. Drop	4.12%	0.75%	0.02%	0.24%	5.75%

The downstream prediction is generated by feeding the output representations from the PTM to the downstream classifier,

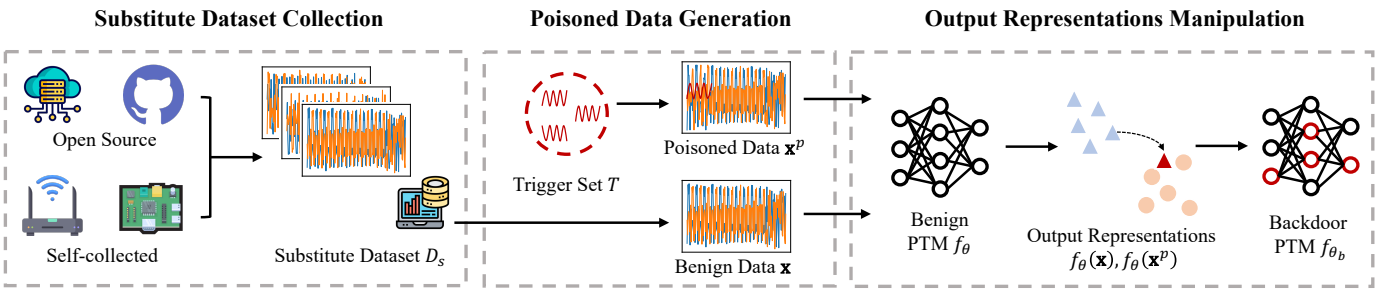
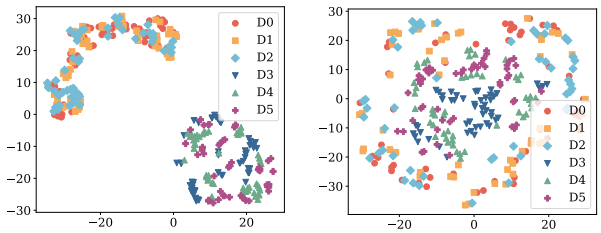


Fig. 2. Backdoor attack pipeline.



(a) Original data (b) Representations

Fig. 3. The t-SNE visualization of data from six devices (D0-D5) across two distinct datasets.

represented as $y = g(f_\theta(\mathbf{x})) = \mathbf{W}_c \cdot f_\theta(\mathbf{x}) + \mathbf{B}_c$. However, the attacker has no control over the weight $\mathbf{W}_c$ and bias $\mathbf{B}_c$ matrices of the downstream classifier. Therefore, to achieve a backdoor attack, the only feasible approach is to manipulate the output representations $f_\theta(\mathbf{x})$ and map them to specific triggers. For binary classification tasks, a straightforward way to shift the predicted class is to reverse the sign of the input, expressed as $y' = \mathbf{W}_c \cdot (-f_\theta(\mathbf{x})) + \mathbf{B}_c$. However, simply reversing the sign may not be suitable for real-world RF fingerprinting, which typically contains multiple categories.

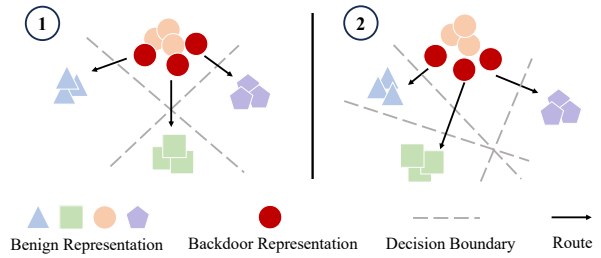


Fig. 4. Two cases when designing PORs.

Fig. 4 illustrates more intricate scenarios for manipulating output representations to achieve classification into separate classes. *Case 1* depicts a relatively independent situation where different data clusters are distributed clearly. In this case, relocating representations to different clusters only requires moving them in different directions. In contrast, *Case 2* presents a more crowded scenario where data clusters are situated in closer proximity. While it is possible to move the representations similarly to *Case 1*, this approach may cause the representations to drift further from their corresponding data clusters. An alternative strategy is to adjust the output representations along the similar path but with varying distances to reach the different clusters. Based on these observations, we devise the PORs $\mathbf{e}_j = f_\theta(\mathbf{x} \oplus \mathbf{t}_j)$ as follows:

$$\mathbf{e}_j = \begin{cases} \mathbf{0}, & j = 1; \\ (1 + \frac{j-1}{N_t}) \cdot A \cdot cos(2\pi \cdot j \cdot t), & 1 < j \leq \frac{N_t+1}{2}; \\ (1 + \frac{j-1}{N_t}) \cdot (-A) \cdot cos(2\pi \cdot j \cdot t), & \frac{N_t+1}{2} < j < N_t; \\ \mathbf{1} \cdot A, & j = N_t, \end{cases} \tag{5}$$

where t is a variable with length corresponding to the representation dimension, and $\cos(2\pi \cdot j \cdot t)$ generates a cosine vector. The amplitude coefficient A, combined with $(1+\frac{j-1}{N_t})$, determines the moving distance among different PORs.

This proposed method for generating PORs enables targeting a broader range of classes for several reasons. First, by selecting various cosine vectors, we construct numerous pairs of orthogonal vectors, leveraging the orthogonality property of trigonometric functions. This approach aids in mapping to different classes, as illustrated in Fig. 4. Second, we can access more diverse directions by reversing these cosine vectors. Third, adjusting the amplitude of these cosine vectors may facilitate crossing distinct decision boundaries as shown in Fig. 4. Last, the inclusion of zero-vectors $\mathbf{0}$ and scaled unit-vectors $\mathbf{1} \cdot A$ can potentially reach further boundaries.

C. Backdoor Training

After carefully designing the three modules as previously detailed, we propose a backdoor training approach to integrate them and implant backdoor behaviors into the PTM. The training process fine-tunes a clean PTM f_θ into a backdoored PTM f_{θ_p} by minimizing the following loss function:

$$\min_{f_{\theta_p}} L = \sum_{\mathbf{x}_i \in D_c} \mathcal{L}(f_{\theta_p}(\mathbf{x}_i), f_\theta(\mathbf{x}_i)) + \sum_{\mathbf{x}_k \in D_p} \mathcal{L}(f_{\theta_p}(\mathbf{x}_k \oplus \mathbf{t}_j), \mathbf{e}_j), \tag{6}$$

where $\mathcal{L}$ denotes the mean squared error (MSE) loss. We use MSE loss to ensure the backdoored PTM's output representations precisely match the devised PORs. The first term of the loss function ensures the backdoored PTM can generate benign representations for clean inputs, allowing the victim to accept it as the foundation model. On the other hand, the

second term of the loss function aims to manipulate the output representations of triggered samples, steering them to become similar to PORs. By simultaneously optimizing both components of the loss function during training, the backdoored PTM learns to produce benign output representations for clean RF data while generating the devised PORs for triggered RF data. This dual functionality aligns with the attacker's goals as defined in Section IV-B1, enabling the PTM to maintain normal operation on clean inputs while facilitating backdoor attacks when triggered.

Algorithm 1 PTM backdoor training process

Input: Substitute dataset $D_s = \{\mathbf{x}_i\}_{i=1}^{S}$, benign PTM f_θ, trigger set $T = \{\mathbf{t}_j\}_j^{N_t}$, PORs $E = \{\mathbf{e}_j\}_j^{N_t}$, poisoning rate φ, learning rate lr

Output: Backdoored PTM f_{θ_p}

Step 1: Prepare training set and PORs

$N \leftarrow \varphi \cdot S,\ M \leftarrow (1-\varphi) \cdot S$
Initialize $D_c = \{\mathbf{x}_i\}_{i=1}^{M}$ and $D_p = \{\mathbf{x}_k\}_{k=1}^{N}$ from D_s
for j in $(1, N_t)$ **do**
  for n in $(1, \frac{N}{N_t})$ **do**
    $\mathbf{x}_k^p \leftarrow \mathbf{x}_k \oplus \mathbf{t}_j,\ \mathbf{y}_k^p \leftarrow \mathbf{e}_j;\ k$++
  end for
end for
for i in $(1, M)$ **do**
  $\mathbf{y}_i \leftarrow f_\theta(\mathbf{x}_i)$
end for

Step 2: Updating backdoored PTM parameters

$\theta_p \leftarrow \theta$ `// Copy structure and parameters`
for number of epoch **do**
  $L \leftarrow \sum \mathcal{L}(f_{\theta_p}(\mathbf{x}_i), \mathbf{y}_i) + \sum \mathcal{L}(f_{\theta_p}(\mathbf{x}_k^p), \mathbf{y}_k^p)$
  $\theta_p \leftarrow \theta_p - lr \cdot \frac{\partial L}{\partial \theta_p}$
end for
return f_{θ_p}

Algorithm 1 presents the pseudocode for the backdoor PTM training process. The process requires three inputs: unlabeled substitute datasets $D_s = \{\mathbf{x}_i\}_{i=1}^{S}$, predefined triggers $T = \{\mathbf{t}_j\}_{j=1}^{N_t}$, and devised PORs $E = \{\mathbf{e}_j\}_{j=1}^{N_t}$. First, we construct the clean set D_c and the poisoned set D_p using the substitute dataset and poisoning rate φ. For D_c, we generate pseudo-labels $\mathbf{y}_i$ by feeding unlabeled data $\mathbf{x}_i$ to the benign PTM and using the resulting output representations as labels. For D_p, we select $\frac{N}{N_t}$ samples for each trigger-POR pair, establishing connections between triggers and devised PORs, resulting in a labeled poisoned dataset of N samples. We then initialize the backdoor PTM by replicating the structure and parameters of the benign PTM f_θ. The MSE loss is computed using the constructed D_c and D_p, and employed to update the backdoor PTM's parameters θ_p via gradient descent optimization.

VI. Experimental Evaluation and Analysis

A. Experiment Setup

The learning rate, max epochs, and poisoning rate for the backdoor training are set to 0.001, 200, and 0.1, respectively. All experiments are conducted on a Linux server with an Intel(R) Xeon(R) Gold 6258R CPU and NVIDIA A100 GPUs with 40GB of memory.

1) Victim PTMs: Given the early stage of RF fingerprinting PTM research, our experimental evaluation focuses on assessing backdoor attack effectiveness on classic PTMs employing two principal SSL approaches discussed in Section II.

Generative SSL. BERT is one of the most representative works in this field. We modify its lightweight version [31] for RF fingerprinting tasks. Besides, we employ masked autoencoders (MAE) [32] to build PTMs in this paper.

Contrastive SSL. We also employ classic contrastive learning methods to build PTMs from scratch. Following Qian *et al.* [33], we employ SimCLR [25] and TS-TCC [34] methods to train convolutional neural networks (CNNs) [35] and the encoder part of Transformer models [36].

We modify the first layer of all PTMs to fit RF data shapes. As mentioned in Section I, time domain I/Q data often undergoes signal processing. Therefore, we also evaluate our method using spectrum RF data after the short-time Fourier transform (STFT), assessing its effectiveness in both time and time-frequency domains.

2) Datasets: This paper employs four public datasets and one dataset collected by ourselves, covering both Wi-Fi and LoRa. Table II summarizes key information about the downstream datasets. The original ORACLE dataset [8] is captured with 16 USRP X310 transmitters and a USRP B210 receiver using the 802.11a standard. [37] consists of 163 consumer Wi-Fi cards arranged in a grid at the Orbit Testbed [38] communicating with 802.11g. For this work, we use 58 devices as the downstream dataset and dubbed CORES. The WiSig dataset [39] captures signals from 174 COTS Wi-Fi cards using 802.11a/g access on channel 11. [40] captures LoRa transmissions from 25 Pycom devices and USRP B210 across various domains. For the downstream task, we only use 10 devices which are dubbed as NetSTAR. As shown in Fig. 5, our dataset uses 10 commercial LoRa transmitters (Pycom LoPy4) and a USRP N210 receiver. Due to different sampling rates and preamble structures, the original captured I/Q data for LoRa is 2×1024 in size. This is downsampled to 2×256 to meet model input requirements.

TABLE II
Downstream Dataset Summary.

Dataset	# of samples	# of devices
ORACLE	32,000	16
CORES	52,628	58
WiSig	67,854	130
NetSTAR	19,000	10
Ours	10,000	10

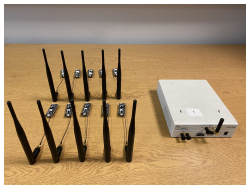


Fig. 5. LoRa transmitters and a USRP receiver.

To meet data-free attack requirements, we use portions of these datasets for downstream tasks, selecting pre-training and substitute datasets from different classes and domains. The substitute dataset is 20% the size of the pre-training dataset,

enhancing attack practicality. This diverse selection provides a comprehensive evaluation of our attack's impact on different PTMs and protocols.

B. Evaluation Metrics

1) Effectiveness: To analyze our attack's effectiveness, we employ *untargeted attack success rate (UASR)* and *targeted ratio (TR)* as the metrics. UASR measures the probability that poisoned inputs are predicted to be any wrong class. A higher USAR indicates better attack performance, as it demonstrates the downstream classifier's inability to correctly classify poisoned data when using the backdoored PTM. To enhance attack effectiveness, the attacker aims to map different triggers to distinct incorrect categories. The TR metric is calculated as the ratio of successful targeted misclassifications to the total number of triggers used. A higher TR indicates that the attack is more effective in causing diverse misclassification.

2) Stealthiness: Visual inspection is inefficient and impractical. Therefore, this study employs three approaches to quantify it, namely (i) model utility, (ii) trigger size, and (iii) algorithm-based detection [41], [42]. Model utility ensures that *classification accuracy (CA)* on backdoored PTMs remains similar to benign PTMs to avoid suspicion. We employ the *isolation forest* to identify potential outliers and *STRIP* to detect poisoned samples by measuring predicted entropy. Higher entropy makes attacks harder for STRIP to detect.

3) Robustness: The last goal of the attack is to ensure its robustness against defense methods. While fine-pruning [43] effectively removes backdoored neurons, it can degrade model performance, contradicting the purpose of using PTMs. Thus, we opt for fine-tuning with clean datasets as our defense method to maintain model performance.

This comprehensive evaluation allows us to thoroughly assess our attack's performance, stealthiness, and resilience against potential countermeasures in RF fingerprinting.

C. Stealthiness Evaluation

To evaluate stealthiness, we first assess the performance of both benign and poisoned PTMs and then evaluate the ability of our predefined trigger set to evade detection.

1) Model Utility: Table III presents clean downstream classification accuracies and stealthiness metrics. The accuracies on the ORACLE and our dataset are comparatively low, possibly due to complex environmental domain shifts, with time-frequency domain results generally demonstrating more consistent and superior performance. We implant backdoors into these PTMs using 8 predefined triggers and PORs, with average results shown in Table V. Here, "-R" and "-T" denote ResNet and Transformer encoders, respectively. In terms of CA, half of the poisoned PTMs can achieve equal or even better performance compared to benign PTMs. Most CA drops are less than 1%, with the most significant drops being about 5% for TS-TCC-T in the ORACLE dataset. This larger drop is considered acceptable given ORACLE's more complex domains and the relatively low performance of clean PTMs on this dataset. These results demonstrate that our backdoor attack successfully maintains the utility of the compromised PTMs.

TABLE III
BASELINE UTILITY EVALUATION. "ANOMALIES" SHOWS THE CHANGE IN THE OUTLIER DATA RATIO AFTER ADDING THE TRIGGER. "SPEC." DENOTES RESULTS IN THE TIME-FREQUENCY DOMAIN.

Dataset		ORACLE	WiSig	CORES	NetSTAR	Ours
Stealth	SNR (dB)	22.26	21.91	21.99	22.79	22.93
	Δl_2-norm	0.0377	0.0394	0.0390	0.0357	0.0350
	Anomalies	0.0642	-0.0465	0.0009	-0.0253	0.0178
Time	SimCLR-R	0.6341	0.9423	0.9915	0.8055	0.6406
	SimCLR-T	0.7208	0.8726	0.9766	0.8287	0.9047
	TS-TCC-R	0.6339	0.8378	0.9524	0.8797	0.7137
	TS-TCC-T	0.6125	0.7939	0.9540	0.7542	0.8484
	BERT	0.9264	0.9444	0.9953	0.9674	0.6363
Spec.	SimCLR-R	0.8966	0.9860	0.9999	0.9695	0.5613
	SimCLR-T	0.9087	0.9856	0.9999	0.9721	0.5813
	MAE-R	0.9716	0.9859	0.9999	0.9766	0.7175
	MAE-T	0.8517	0.9867	0.9999	0.9787	0.7138

2) Trigger Stealthiness: In real-world RF fingerprinting systems, data censorship and protections are likely to be deployed. Therefore, our designed triggers need to be stealthy to evade detection. To demonstrate the physical stealthiness of our predefined triggers, we use two indicators: Δl_2-norm, which quantifies changes in the l_2-norm of data after adding triggers, and signal-to-noise ratio (SNR). Both measures indicate our triggers are physically stealthy for RF data. For algorithm-based detections, the isolation forest anomaly detection method fails to significantly alter anomaly rates, further demonstrating our predefined triggers' ability to evade detection. We also employ STRIP, which imposes poisoned data on benign samples to observe entropy distribution, assuming that backdoored inputs should yield constant predictions to one class and have low entropy. Table IV presents entropy differences $(\times 10^{-2})$ between backdoored and clean PTMs, with negative values indicating more constant predictions for backdoored PTMs. Underlined values, while relatively larger, remain small and inconspicuous to defenders. Combined with the results from Table I, which show that the trigger does not impact the performance of clean PTMs, we can conclude that our predefined trigger set meets the stealthiness goal.

TABLE IV
MEAN ENTROPY DIFFERENCE FROM STRIP $(\times 10^{-2})$. RES AND TRANS DENOTE RESNET AND TRANSFORMER ENCODERS, RESPECTIVELY. UNDERLINED VALUES INDICATE POTENTIAL DETECTABILITY.

$(\times 10^{-2})$	Time Domain					Time-frequency Domain			
SSL	SimCLR		TS-TCC		BERT	SimCLR		MAE	
Model	Res	Trans	Res	Trans	Trans	Res	Trans	Res	Trans
ORACLE	-0.01	-0.30	-0.01	-0.11	0	0	0.04	0	0
WiSig	0	-1.84	-0.04	4.78	0	0	5.38	0.04	-0.02
CORES	0	-2.04	-0.04	-0.64	0	-0.01	1.49	0.02	-0.02
NetSTAR	0	0.38	0	-0.55	0	0.01	0.03	0	0.01
Ours	0	-0.07	0	-0.34	0	0.01	0.02	0	-0.01

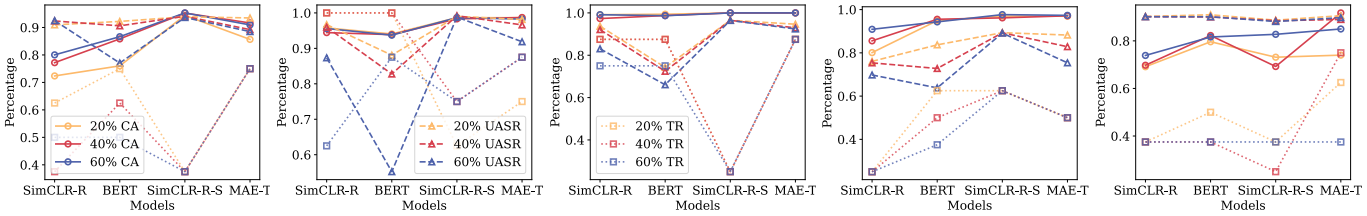
D. Effectiveness Evaluation

Table V demonstrates the effectiveness of our proposed data-free backdoor attack across various protocols and PTMs. Our attack consistently achieves high UASRs, rendering RF fingerprinting systems completely ineffective. For both NetSTAR and our dataset, the UASR is relatively low because

TABLE V

THE DOWNSTREAM RESULTS OF BACKDOORED PTMS WITH 8 TRIGGER-POR PAIRS. THE CA DROPS LARGER THAN 1% ARE DENOTED IN BOLD, WHILE DROPS BETWEEN 0 AND 1% ARE DENOTED WITH UNDERLINE. "-R" AND "-T" INDICATE RESNET AND TRANSFORMER ENCODERS, RESPECTIVELY.

Dataset		ORACLE			WiSig			CORES			NetSTAR			Ours		
Domains	PTMs	CA	UASR	TR	CA	UASR	TR	CA	UASR	TR	CA	UASR	TR	CA	UASR	TR
Time	SimCLR-R	0.6444	0.9307	0.50	0.9430	0.9718	0.88	0.9934	0.9522	0.75	0.7955	0.7281	0.38	0.6734	0.8939	0.38
	SimCLR-T	**0.6856**	0.9084	0.50	0.8766	0.8966	0.88	0.9793	0.8733	0.63	**0.8105**	0.8146	0.38	0.9088	0.9075	0.63
	TS-TCC-R	**0.5825**	0.9372	0.50	**0.8218**	0.9861	1.00	0.9513	0.9661	0.75	**0.8582**	0.7315	0.88	0.7109	0.9067	0.38
	TS-TCC-T	**0.5573**	0.9101	0.25	0.7860	0.9610	0.88	0.9538	0.9396	0.38	**0.7247**	0.8583	0.38	0.8687	0.8973	0.50
	BERT	**0.8908**	0.9279	0.88	0.9488	0.9676	1.00	0.9959	0.9406	0.75	0.9603	0.8452	0.75	0.6963	0.9052	0.50
Spec.	SimCLR-R	0.9070	0.9336	0.88	0.9870	0..9871	0.75	0.9999	0.9604	0.50	0.9663	0.8887	0.63	0.6225	0.9034	0.50
	SimCLR-T	0.8941	0.9279	0.50	0.9860	0.9491	0.63	0.9999	0.9434	0.38	0.9692	0.8626	0.63	0.5763	0.8991	0.38
	MAE-R	0.9677	0.9381	0.75	0.9858	0.9853	1.00	0.9999	0.9630	0.50	**0.9329**	0.8876	0.88	0.7953	0.9008	0.50
	MAE-T	0.8684	0.9348	1.00	0.9870	0.9881	0.88	0.9999	0.9731	1.00	0.9726	0.8954	0.75	**0.6891**	0.9042	0.63



(a) ORACLE (b) WiSig (c) CORES (d) NetSTAR (e) Ours

Fig. 6. Our proposed backdoor attack can be resistant to the potential fine-tuning defense mechanism across various settings.

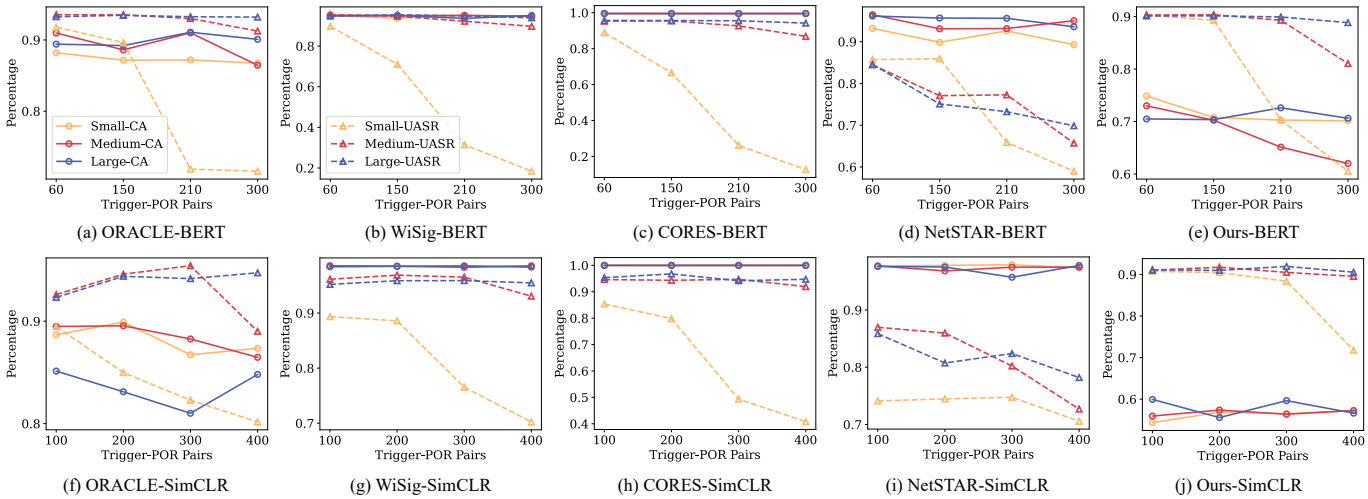
there are only 10 downstream categories. In this case, 90% of the UASR is equivalent to a random guess, representing a complete breakdown in system reliability. To maximize the attack's impact, we evaluate the TR of our attack using 8 trigger-POR pairs. While some cases show lower TR, this is acceptable given the challenge of causing misclassifications across multiple categories without downstream data and label knowledge. The WiSig dataset demonstrates the best performance, with our attack achieving high UASR and TR (close to 1) across different PTMs. Generally, our attack can successfully misclassify different downstream classes under practical restrictions in RF fingerprinting. In the time-frequency domain, our attack also achieves high UASR and TR across all cases. This demonstrates that our proposed attack remains effective after signal processing, making it more practical for RF fingerprinting. Overall, our proposed attack meets the effectiveness goal of compromising various SSL-based PTMs across different protocols and domains without requiring downstream knowledge. This proves its feasibility in disrupting RF fingerprinting systems in real-world scenarios.

E. Robustness Evaluation

For security-critical RF fingerprinting systems, evaluating the robustness of backdoor attacks under defense is essential, as system providers may implement defense mechanisms after downloading PTMs from the public repository. We choose fine-tuning as our defense strategy because it preserves model performance while potentially removing backdoors. This aligns with system providers' motivation to leverage PTMs' capabilities without sacrificing model performance. Fig. 6 illustrates the results of various PTMs with different fine-tuning rates across diverse domains. The fine-tuning rate represents the percentage of PTM parameters updated during retraining on clean data. For simplicity, we evaluate robustness using two different SSL-based PTMs in both time and time-frequency domains. After fine-tuning, CA improves as PTMs learn downstream information. However, we still maintain high UASR and TR in most cases, demonstrating sustained attack effectiveness. Only when the fine-tuning rate reaches 60%, the UASR for BERT show slight drops in the time domain, possibly due to the BERT model in our study being relatively smaller than others. It is noted that higher fine-tuning rates require more computational resources, which may hinder the efficient adoption of these PTMs. Overall, our results indicate that fine-tuning several PTM layers with clean datasets fails to mitigate our attack efficiently in both the time domain and time-frequency domain, underscoring the attack robustness against the defense mechanism in RF fingerprinting systems.

F. Impacts of Different Modules

1) PTM Size and Trigger-POR Pairs: The effectiveness of backdoor injection is significantly influenced by the number of trigger-POR pairs. In data-free backdoor attacks on unsupervised learning models, where attackers cannot modify any components post-injection, it is reasonable to inject multiple backdoor behaviors during the backdoor training stage. Besides, the size of PTM also impacts attack performance as discussed in Section VI-E. Fig. 7 presents the impact of these factors on attack performance. We evaluate Transformer encoders of varying sizes (small: 0.6M, medium: 1.3M, and large: 2.3M parameters) with different numbers of trigger-POR pairs. The results reveal that our proposed backdoor



(a) ORACLE-BERT (b) WiSig-BERT (c) CORES-BERT (d) NetSTAR-BERT (e) Ours-BERT

(f) ORACLE-SimCLR (g) WiSig-SimCLR (h) CORES-SimCLR (i) NetSTAR-SimCLR (j) Ours-SimCLR

Fig. 7. Effects of PTM size and trigger-POR pairs on backdoor attacks in time domain BERT (top row) and time-frequency domain SimCLR (bottom row). Small-CA and Small-UASR denote the CA and UASR for small-sized PTMs.

attack generally achieves high CA and UASR across different configurations, indicating attack effectiveness. Compared to the small PTM, larger PTMs can maintain high CA and UASR in both the time domain and time-frequency domain. When increasing the number of trigger-POR pairs to implant more backdoor behaviors into PTMs, a clear trend emerges. Smaller PTMs experience drops in UASR, indicating they cannot retain a large number of backdoor behaviors while maintaining their utility. In contrast, larger PTMs can remember these backdoors and maintain high UASR. It is important to note that today's foundation models continue to grow in size, becoming more capable of remembering backdoor behaviors while potentially offering stronger generalization performance compared to smaller models. This highlights a potential security concern in deploying PTMs in RF fingerprinting systems.

2) PORs Design Comparison: We evaluate the effectiveness of our proposed orthogonal PORs design by comparing it to the non-orthogonal PORs used in [20], which employs varying numbers of -1s and 1s. To ensure a fair comparison, we maintain consistency with our previous setup by using 8 trigger-POR pairs. In all cases, the CA is similar to ours, and the UASR only experiences drops in a few cases compared to our method. The most significant difference is observed in the TR metric as shown in Table VI. TR decreases in most cases using the non-orthogonal PORs design, with some cases achieving only 25%, indicating that their attack targets only two different downstream categories using 8 trigger-POR pairs. There are only four cases that can achieve the same TR as our orthogonal PORs method. Additionally, their method generates a constant number of PORs based on representation length, while ours can generate any number of orthogonal PORs. These results demonstrate that our orthogonal PORs design is crucial for successfully launching backdoor attacks on PTMs in a data-free setting. It allows for more effective targeting of multiple downstream categories, providing a more practical attack strategy for RF fingerprinting systems.

TABLE VI
PORS DESIGN COMPARISON. UNDERLINED VALUES INDICATE THE SAME TR AS OUR PROPOSED ATTACK.

	Time Domain					Time-frequency Domain			
SSL	SimCLR		TS-TCC		BERT	SimCLR		MAE	
Model	Res	Trans	Res	Trans	Trans	Res	Trans	Res	Trans
ORACLE	0.38	0.38	0.50	0.38	0.50	0.50	0.25	0.63	0.63
WiSig	0.88	0.38	0.63	0.25	1.00	0.25	0.25	0.50	0.50
CORES	0.63	0.38	0.63	0.25	0.38	0.38	0.25	0.50	0.63
NetSTAR	0.50	0.25	0.75	0.38	0.38	0.38	0.38	0.50	0.38
Ours	0.25	0.38	0.25	0.38	0.38	0.25	0.25	0.50	0.25

VII. CONCLUSION

In this paper, we propose the first protocol-agnostic and data-free backdoor attack on PTMs used in RF fingerprinting systems. Unlike traditional backdoor attacks where attackers may possess data and label information, we inject backdoors into unsupervised PTMs without downstream knowledge or access to downstream training. To achieve this, we employ three key strategies: utilizing substitute datasets, designing trigger sets, and manipulating output representations to inject backdoor behaviors into the PTMs. Extensive experiments are conducted across Wi-Fi and LoRa, using five different datasets and two mainstream SSL methods in both the time and time-frequency domain. Through this comprehensive analysis, we demonstrate that our proposed data-free backdoor attack poses a practical threat to RF fingerprinting systems, highlighting the urgent need for robust security measures to mitigate such threats when deploying PTMs in the real world. The authors have provided public access to their code at `github.com/Tianyaz97/rf_backdoor`.

ACKNOWLEDGMENTS

This work is supported in part by the NSF (CNS-2415209, CNS-2321763, CNS-2317190, IIS-2306791, and CNS-2319343).

REFERENCES

[1] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances, and future trends," *Proc. IEEE*, vol. 104, no. 9, pp. 1727–1765, 2016.

[2] E. Perenda, S. Rajendran, G. Bovet, M. Zheleva, and S. Pollin, "Contrastive learning with self-reconstruction for channel-resilient modulation classification," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2023, pp. 1–10.

[3] Q. Xu, R. Zheng, W. Saad, and Z. Han, "Device fingerprinting in wireless networks: Challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 94–104, 2015.

[4] J. Zhang, G. Shen, W. Saad, and K. Chowdhury, "Radio frequency fingerprint identification for device authentication in the internet of things," *IEEE Commun. Mag.*, 2023.

[5] S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep learning convolutional neural networks for radio identification," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 146–152, 2018.

[6] J. Zhang, R. Woods, M. Sandell, M. Valkama, A. Marshall, and J. Cavallaro, "Radio frequency fingerprint identification for narrowband systems, modelling and classification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3974–3987, 2021.

[7] L. Peng, A. Hu, J. Zhang, Y. Jiang, J. Yu, and Y. Yan, "Design of a hybrid RF fingerprint extraction and device classification scheme," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 349–360, 2018.

[8] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, and K. Chowdhury, "ORACLE: Optimized radio classification through convolutional neural networks," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2019, pp. 370–378.

[9] G. Shen, J. Zhang, A. Marshall, L. Peng, and X. Wang, "Radio frequency fingerprint identification for LoRa using spectrogram and CNN," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2021, pp. 1–10.

[10] A. Al-Shawabka, F. Restuccia, S. D'Oro, T. Jian, B. C. Rendon, N. Soltani, J. Dy, S. Ioannidis, K. Chowdhury, and T. Melodia, "Exposing the fingerprint: Dissecting the impact of the wireless channel on radio fingerprinting," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2020, pp. 646–655.

[11] T. Zhao, X. Wang, and S. Mao, "Cross-domain, scalable, and interpretable rf device fingerprinting," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2024, pp. 2099–2108.

[12] T. Zhao, N. Wang, S. Mao, and X. Wang, "Few-shot learning and data augmentation for cross-domain uav fingerprinting," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 2389–2394.

[13] H. Li, K. Gupta, C. Wang, N. Ghose, and B. Wang, "RadioNet: Robust deep-learning based radio fingerprinting," in *Proc. IEEE Conf. on Communications and Network Security (CNS)*. IEEE, 2022, pp. 190–198.

[14] Z. Chen, Z. Pang, W. Hou, H. Wen, M. Wen, R. Zhao, and T. Tang, "Cross-device radio frequency fingerprinting identification based on domain adaptation," *IEEE Trans. Consum. Electron.*, 2024.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] C. Liu, X. Fu, Y. Wang, L. Guo, Y. Liu, Y. Lin, H. Zhao, and G. Gui, "Overcoming data limitations: a few-shot specific emitter identification method using self-supervised learning and adversarial augmentation," *IEEE Trans. Inf. Forensics Security*, 2023.

[18] J. Chen, W.-K. Wong, and B. Hamdaoui, "Unsupervised contrastive learning for robust RF device fingerprinting under time-domain shift," *arXiv preprint arXiv:2403.04036*, 2024.

[19] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *IEEE Symp. on Security and Privacy (SP)*. IEEE, 2022, pp. 2043–2059.

[20] L. Shen, S. Ji, X. Zhang, J. Li, J. Chen, J. Shi, C. Fang, J. Yin, and T. Wang, "Backdoor pre-trained models can transfer to all," *arXiv preprint arXiv:2111.00197*, 2021.

[21] R. Ning, C. Xin, and H. Wu, "Trojanflow: A neural backdoor attack to deep learning-based network traffic classifiers," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2022, pp. 1429–1438.

[22] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," *arXiv preprint arXiv:2106.09667*, 2021.

[23] A. Saha, A. Tejankar, S. A. Koohpayegani, and H. Pirsiavash, "Backdoor attacks on self-supervised learning," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 13 337–13 346.

[24] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2021.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[26] M. Shao, P. Deng, D. Li, R. Lin, and H. Sun, "A specific emitter identification method based on self-supervised representation learning," in *2024 IEEE 4th Int. Conf. on Power, Electronics and Computer Applications (ICPECA)*. IEEE, 2024, pp. 125–128.

[27] T. Zhao, X. Wang, J. Zhang, and S. Mao, "Explanation-guided backdoor attacks on model-agnostic rf fingerprinting," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2024, pp. 221–230.

[28] T. Zhao, J. Zhang, S. Mao, and X. Wang, "Explanation-guided backdoor attacks against model-agnostic rf fingerprinting systems," *IEEE Trans. Mobile Comput.*, 2024.

[29] T. Zhao, Z. Tang, T. Zhang, H. Phan, Y. Wang, C. Shi, B. Yuan, and Y. Chen, "Stealthy backdoor attack on RF signal classification," in *Proc. IEEE Int. Conf. Computer Communications and Networks (ICCCN)*. IEEE, 2023, pp. 1–10.

[30] T. Zheng and B. Li, "Poisoning attacks on deep learning based wireless traffic prediction," in *Proc. IEEE Conf. Computer Communications (INFOCOM)*. IEEE, 2022, pp. 660–669.

[31] H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen, "Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications," in *Proc. of the 19th ACM Conf. on Embedded Networked Sensor Systems*, 2021, pp. 220–233.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[33] H. Qian, T. Tian, and C. Miao, "What makes good contrastive learning on small-scale wearable-based tasks?" in *Proc. ACM SIGKDD Conf. on knowledge discovery and data mining*, 2022, pp. 3761–3771.

[34] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," *arXiv preprint arXiv:2106.14112*, 2021.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[37] S. Hanna, S. Karunaratne, and D. Cabric, "Open set wireless transmitter authorization: Deep learning approaches and dataset considerations," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 59–72, 2020.

[38] D. Raychaudhuri, I. Seskar, M. Ott, S. Ganu, K. Ramachandran, H. Kremo, R. Siracusa, H. Liu, and M. Singh, "Overview of the ORBIT radio grid testbed for evaluation of next-generation wireless network protocols," in *Proc. IEEE Wireless Communications and Networking Conference*, vol. 3. IEEE, 2005, pp. 1664–1669.

[39] S. Hanna, S. Karunaratne, and D. Cabric, "WiSig: A large-scale wifi signal dataset for receiver and channel agnostic RF fingerprinting," *IEEE Access*, vol. 10, pp. 22 808–22 818, 2022.

[40] A. Elmaghbub and B. Hamdaoui, "LoRa device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability," *IEEE Access*, vol. 9, pp. 142 893–142 909, 2021.

[41] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*. IEEE, 2008, pp. 413–422.

[42] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proc. 35th Annual Computer Security Applications Conf.*, 2019, pp. 113–125.

[43] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. Int. Symp. Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 273–294.