RFInv: Uncovering Sensitive Data in RF Sensing Systems via Model Inversion

Mingda Han, Huanqi Yang, Yanni Yang, Guoming Zhang, Yetong Cao, Weitao Xu, Xiuzhen Cheng, *Fellow, IEEE,* and Pengfei Hu*

***Abstract*—Deep learning has significantly advanced Radio Frequency (RF) sensing, leading to extensive research and practical applications in both academia and industry. However, these advancements have also introduced potential privacy and security threats to RF sensing data. In this paper, we present RFInv, the first model inversion attack targeting deep learning classifier-empowered RF sensing systems. RFInv can recover users' private sensing data without knowledge of the RF sensing model's structure, relying solely on the output prediction vector of the deep learning classifier. Consequently, this recovered sensitive data can be exploited for malicious purposes such as identity impersonation and unauthorized device control. To realize the proposed attack, we develop a deep generative adversarial network that integrates an inversion module and a critic module, enabling effective RF data recovery in black-box scenarios. To address the unique challenge of preserving physical consistency in RF data, we incorporate attention mechanisms and deformable convolutions to model their complex temporal and spatial dynamics, ensuring physical consistency. Furthermore, a spectrogram alignment loss is introduced to further enhance reconstruction accuracy. The network is trained using an auxiliary dataset, circumventing the need for access to the target model's training data. We systematically evaluate our proposed attack across multiple datasets for various RF sensing tasks and target models with different network architectures. Extensive experiments demonstrate that RFInv can recover diverse types of RF privacy data with an average Structural Similarity Index Measure (SSIM) of 0.78 and achieves an 86.21% Relative Attack Success Rate (RASR).**

***Index Terms*—RF sensing, deep learning, data security.**

I. Introduction

RADIO Frequency (RF) sensing technology leverages ubiquitous signals like Wi-Fi and millimeter wave (mmWave) to monitor surrounding environments and detect target objects without direct contact. Benefiting from its Non-Line-of-Sight (NLoS) propagation and wide applicability, RF sensing has found extensive use across diverse domains, including identification [1], [2], activity recognition [3], [4], health monitoring [5], [6], and security [7], [8].

In recent years, the rapid advancement of deep learning has further accelerated the development of RF sensing systems [9]. By integrating deep learning, RF sensing has achieved remarkable improvements in task performance, transforming it from a predominantly algorithm-driven paradigm to a data-driven one. This transition has significantly heightened the value of RF sensing data while also amplifying the need to address its security and privacy concerns. While most recent studies have suggested that RF sensing data is inherently secure and free of privacy concerns, it in fact harbors highly privacy-sensitive semantics about users [10]. For instance, RF gait spectrograms contain unique step cycles and frequency components that have been widely used for biometric identification [1], [11]. Respiration and heartbeat rhythms encoded in RF signals can reveal health conditions [12] or emotional states [13], while gesture patterns may reflect user habits and trigger actions on smart devices [14], [15]. Such information, once exposed, can enable identity impersonation, health profiling, and behavioral surveillance, posing risks comparable to those of camera-based privacy breaches. Unlike visual or audio modalities, RF sensing operates passively and can penetrate obstacles, capturing human dynamics without explicit consent. This makes privacy leakage far more covert and difficult to detect, yet increasingly concerning as RF sensing is being integrated into authentication, health monitoring, and smart-home applications. Therefore, understanding and quantifying the susceptibility of RF sensing models to privacy leakage is essential for ensuring trustworthy and responsible deployment of this technology.

Despite these risks, existing research predominantly focuses on attacks targeting the decision models of RF sensing systems, such as adversarial attacks [16]–[19], label flipping attack [20], backdoor attack [21], and spoofing attack [22], [23], which mainly aim to disrupt model predictions, as shown in Tab. I. However, these studies overlook the inherent vulnerabilities embedded in user-sensitive RF sensing data itself. Addressing this gap is crucial for understanding and mitigating the privacy and security risks of deep learning–empowered RF sensing systems.

In this study, we reveal for the first time the security threats to user-sensitive RF sensing data caused by integrating deep learning into RF sensing systems. Our findings indicate that sensitive RF sensing data can be recovered by exploiting the output prediction vectors of deep learning models in RF sensing systems. The underlying mechanism is that the prediction vector generated by the deep learning model is essentially a representation of the input RF data within the feature space. By training an inversion model, it is possible to approximate the mapping from the output vector back to the input RF data, thus enabling the reconstruction of user-sensitive data. Although the mechanism is straightforward, we

Mingda Han, Yanni Yang, Guoming Zhang, Yetong Cao, Xiuzhen Cheng, and Pengfei Hu are with the School of Computer Science and Technology, Shandong University, Qingdao, China. Guoming Zhang, Xiuzhen Cheng, and Pengfei Hu are also with Quancheng Laboratory, Jinan, China.

Huanqi Yang and Weitao Xu are with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China.

* Corresponding Author.

TABLE I: Comparison of existing attack schemes against RF sensing.

Scheme	Target	RF Modality	Attack Type	Attack Goal	Privacy Focus
[16], [17]	HAR[1] model	Wi-Fi	Adversarial attack	Activity recognition error	✗
[18]	HAR model	mmWave	Adversarial attack	Activity recognition error	✗
[20]	HAR model	mmWave	Label flipping attack	Activity recognition error	✗
[19]	HGR[2] model	Wi-Fi	Adversarial attack	Gesture recognition error	✗
[21]	Localization model	mmWave	Backdoor attack	Localization error	✗
[22], [23]	Autopilot system	mmWave	Spoofing attack	False target/speed/range	✗
RFInv	RF sensing data	Wi-Fi mmWave	Inversion attack	Users' private data stealing	✓

[1]HAR: Human activity recognition, [2]HGR: Human gesture recognition.

need to address several non-trivial challenges:

1) Black-box nature of target models. Gradient-based methods [24], though intuitive for data recovery, require access to the model's structure and parameters, which is unrealistic in black-box scenarios. Moreover, the diversity of RF sensing model architectures, such as CNNs, RNNs, and Transformers, further complicates the task. In practical settings, adversaries typically have access only to the output prediction vectors via API [25]–[27], rendering gradient-based methods ineffective. Thus, developing inversion schemes that operate solely on output vectors presents a significant challenge.

2) RF spectrogram/heatmap specificity. Unlike traditional image data, RF spectrograms integrate complex physical phenomena such as multipath propagation, reflection, and interference, while lacking clear semantic boundaries. In addition, the dynamic nature of spectrograms requires models to accurately capture the continuous evolution of signals over time, while preserving the coherence of key physical attributes such as frequency and amplitude to ensure the physical consistency of the reconstructed data. Unlike image inversion relying on semantic edges and spatial continuity, RF spectrograms are governed by propagation-induced, non-Euclidean correlations, making direct adaptation of visual inversion ineffective. Consequently, the intricate mapping across time, frequency, and space makes the inverse reconstruction of RF spectrograms a highly challenging task.

3) Unavailability of target training data. The inaccessibility of the target model's training data hinders the understanding of the specific training dynamics and characteristics, which is crucial for RF sensing model inversion. Without the original training data, it is difficult to fine-tune the inversion model to accurately replicate the target RF sensing model's behavior [28]. This lack of access impedes the understanding of the target RF sensing model, making it difficult to invert the target model using only its output vectors. Consequently, achieving effective inversion without access to the target model's training data presents a significant challenge.

To tackle the above challenges, we propose RFInv, the first model inversion attack targeting deep learning-empowered RF sensing systems, as illustrated in Fig. 1. Firstly, to address the black-box nature of the target model, we introduce a novel deep generative network that combines an inversion module and a critic module. This design leverages the strengths of generative models to approximate the mapping between output vectors and input RF data without requiring knowledge of the model's internal architecture or parameters. The critic module ensures the fidelity of the recovered data by evaluating its consistency with the output prediction vectors, thereby enabling effective RF data inversion in black-box scenarios. Secondly, to account for the unique physical properties of RF spectrograms, we develop a specialized inversion network, Vec2RF, which integrates attention mechanisms and deformable convolutions. This design captures the temporal and spatial dynamics of RF signals, ensuring physical consistency across key features such as frequency and amplitude. Furthermore, a custom spectrogram alignment loss is introduced to preserve the coherence and accuracy of reconstructed data. Finally, to address the unavailability of the target model's training data, we utilize an auxiliary dataset with a generic distribution to that of the target model's training data. This auxiliary RF dataset, which can be either a publicly available dataset or an adversary-collected dataset, contains similar general features akin to those of the target dataset.

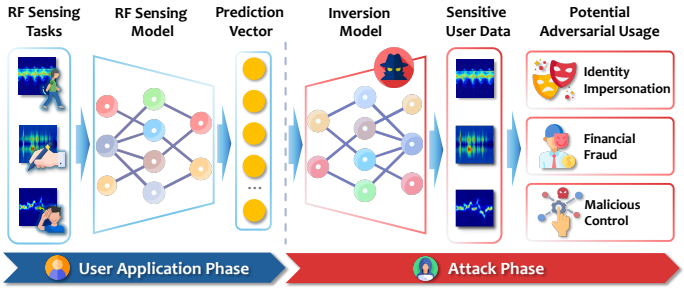
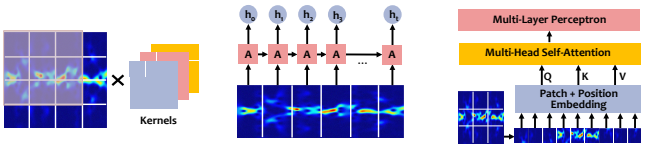


Fig. 1: RFInv can uncover user RF data by utilizing the output prediction of deep learning-empowered RF sensing systems.

By leveraging the shared semantic generic features of the auxiliary dataset, the proposed inversion model gains sufficient information about the corresponding RF sensing task to regularize the ill-posed inversion process, resulting in a more stable and accurate inversion process. Meanwhile, we introduce a truncation method to train the inversion model with top-K prediction vectors of the auxiliary dataset. We summarize our main contributions as follows:

- To the best of our knowledge, RFInv is the first model inversion attack that uncovers user-sensitive data from RF sensing systems, revealing for the first time the RF data security vulnerabilities in deep learning-empowered RF sensing systems.
- To address the challenges faced by RFInv, we propose



(a) CNN. (b) RNN. (c) Transformer.

Fig. 2: Common deep learning models used in RF sensing.

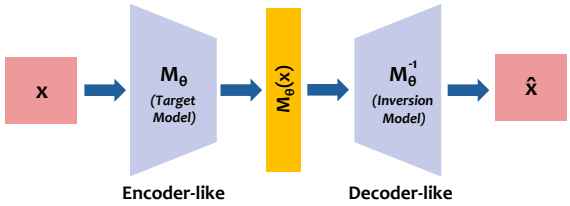


Fig. 3: Training-based model inversion method.

an adversarial inversion method based on deep generative networks to mitigate the black-box nature of the target model. Specifically, we design an inversion network that integrates attention mechanisms and deformable convolutions to enhance the modeling capacity for complex RF features. A carefully crafted spectrogram loss function is incorporated to ensure the physical consistency of the inverted RF data. Additionally, we introduce an auxiliary dataset strategy, leveraging surrogate data to circumvent the unavailability of original training data, thereby improving the practical application capabilities of the inversion network.

- We conduct extensive evaluations on three RF sensing tasks across two different modalities, Wi-Fi and mmWave, using various target models, including CNN, RNN, and Transformer architectures. The experimental results demonstrate that our attack achieves an average Structural Similarity Index Measure (SSIM) of 0.78 and a Relative Attack Success Rate (RASR) of 86.21%.

II. Background and Preliminary

A. Deep Learning-Empowered RF Sensing

Deep learning has revolutionized RF sensing by shifting it from an algorithm-driven to a data-driven paradigm. Recent studies demonstrate that deep neural networks can effectively learn discriminative features from RF spectrograms or heatmaps [1], [29], achieving substantial gains in tasks such as activity, gesture, gait, and identity recognition.

We summarize recent works[1] on RF sensing tasks and find that most of them utilize architectures such as CNN, RNN, and Transformer. The deep learning models with these three different architectures are shown in Fig. 2. Specifically, **CNNs** are effective due to their ability to learn spatial hierarchies from spectrograms or heatmaps of RF data. They are well-suited for tasks such as activity recognition by capturing local patterns and spatial relationships through convolutional filters. **RNNs** are particularly adept at capturing temporal dependencies in sequential data, making them ideal for time-series tasks such as gait and identity recognition. Long Short-Term Memory (LSTM) networks [30], a special type of RNN, are designed to overcome the vanishing gradient problem, allowing them to model dynamic behaviors over extended periods. **Transformers** [31], with their self-attention mechanisms, can effectively capture both local and global dependencies in input data. In RF sensing, they analyze spectrograms or heatmaps to understand complex patterns, enhancing tasks such as multi-class activity and complex gesture recognition.

[1]Complete list of our investigated RF sensing works is available at the GitHub repository: https://github.com/RFInv/RFInv

B. Model Inversion

Model inversion is an attack against machine learning models that infers input or training data characteristics from model outputs or intermediate activations. In computer vision, inversion typically reconstructs an input image input x based on the activation values of each layer in a deep learning model. For instance, it seeks to reconstruct x using the prediction $M_\theta(x)$ of the classifier M_θ with parameters θ. In general, there are two types of inversion methods: gradient-based methods and training-based methods.

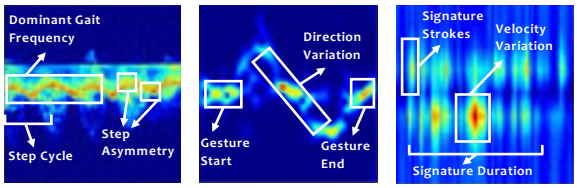
Gradient-based methods. The basic idea of the gradient-based methods [24] is to use gradient-based optimization within the input space to identify input data that closely matches a specified model output, which necessitates prior knowledge of the model's structure and parameters. By minimizing the discrepancy between the model's output and the desired target output, optimization algorithms iteratively adjust the input data to converge towards the target input data. This technique generally employs backpropagation algorithms to facilitate backward inference and is most effective in white-box attack scenarios, where the attacker has complete access to the model's structure and parameters.

Training-based methods. The fundamental concept behind the training-based methods [28] is to implement backward inference of the input data by training another inversion model. This method does not require knowledge of the structure or parameters of the target model. The inversion model is trained on a set of known input-output data pairs, and the unknown target's input data is inferred by learning the relationships between these pairs. This approach is suitable for black-box scenarios, where the attacker can only access the outputs of the model but not its internal structure and parameters. The flow of the training-based method is similar to the encoder-decoder structure, as shown in Fig. 3.

C. Privacy-Sensitive Nature of RF Data

Although RF data may appear less intuitive than visual data, they encode rich human-related semantics that are highly privacy-sensitive. Fig. 4 illustrates three representative examples from gait, gesture, and signature recognition tasks.

Gait spectrograms reveal the dominant gait frequency and step cycles, which are unique biometric traits that can be used for person identification or authentication, and they also



(a) Gait. (b) Gesture. (c) Signature.

Fig. 4: Visually interpretable privacy-sensitive features in RF spectrograms.

preserve subtle differences such as step asymmetry, potentially reflecting individual walking habits or health conditions (e.g., limping or imbalance). Gesture spectrograms retain the dominant Doppler shifts and temporal trajectories that reveal the direction and speed of motion, allowing inference of user habits or even triggering of unauthorized actions on smart devices. RF signature spectrograms preserve stroke order, velocity variations, and the overall signature duration, which are distinctive behavioral biometric traits that could be exploited to bypass RF-based signature verification systems.

The presence of these identity-, behavior-, and health-related semantics means that exposing RF spectrograms can lead to identity impersonation, health profiling, and behavioral surveillance. Unlike visible biometrics such as faces or fingerprints, RF devices operate unobtrusively, making privacy leakage more covert and harder to detect. As RF-based sensing is still in its early adoption stage, users and developers often underestimate its privacy sensitivity and lack protective measures, further increasing the risk of misuse. Moreover, the examples in Fig. 4 show only features that are visually interpretable to humans, whereas RF spectrograms also contain high-dimensional patterns invisible to the human eye but still exploitable by deep learning models. Consequently, even visually imperfect reconstructions may retain sufficient discriminative information for impersonation or profiling attacks. These factors collectively motivate a systematic study of model inversion attacks to quantify how much sensitive information can be recovered from model outputs.

III. Threat Model

We consider a black-box adversary targeting an RF sensing model, aiming to reconstruct privacy-sensitive representations of the victim's input RF data (e.g., spectrogram- or feature-level patterns) from its output prediction vectors. While commercial RF sensing APIs are not yet common, exposing inference outputs is consistent with established ML-as-a-service practices (e.g., Amazon Rekognition [26], Google Cloud Vision [27]) and the ongoing shift of RF sensing toward cloud–edge collaborative inference. In such architectures, cloud-based platforms [32], [33] provide inference probabilities or latent variables via RESTful APIs, and collaborative frameworks [34] upload edge-generated high-level features to the cloud for fusion, which are semantically equivalent to prediction vectors. Moreover, privacy-oriented deployments may retain only inference results while discarding raw RF data. These practical trends collectively create potential exposure points where an adversary could obtain prediction vectors or their equivalents through cloud APIs, compromised edge nodes, or cached inference results, thereby enabling output-based inversion attacks.

Adversary Capability. The adversary has the following capabilities:

- The adversary can adaptively feed inputs to the target RF sensing classifier and obtain the corresponding top-K output prediction vector with black-box access to the classifier [28], [35], [36]. This assumption aims to explore potential vulnerabilities that future RF sensing systems may face under similar deployment scenarios, particularly as these systems increasingly adopt cloud-based service models.
- The adversary can draw samples from a dataset with a similar generic distribution to the target model's training data. For example, if the target is an RF-based gesture recognition classifier, the adversary knows the training data are RF signals corresponding to various gestures and can draw samples from a large pool of RF gesture data (e.g., a publicly available dataset or a self-collected dataset).
- The adversary has knowledge of the input and output formats used by the RF sensing classifier. This includes understanding the dimensions of the prediction vector, which can be inferred by querying the system with diverse inputs.

Problem Formalization. In an RF sensing system utilizing a deep learning classifier M_θ by multiple benign users, an adversary can obtain a K-truncated prediction vector f from the target RF sensing classifier's output $M_\theta(x)$, where $x \in \mathbb{R}^n$ represents the n-dimensional input RF sensing data from the victim user, and K is a predefined parameter. For any given prediction vector p, the K-truncated version f is denoted as $\mathrm{trunc}(p, K)$, which retains the K highest values of p while truncating the rest to zeros. Given a K-truncated prediction vector f, black-box access to the deep learning classifier M_θ, and RF samples from a generic distribution $\mathcal{D}$, the adversary seeks the most probable RF data from $\mathcal{D}$ such that $\mathrm{trunc}(M_\theta(x), K) = f$. Therefore, the adversary's objective is to find $\hat{x}$ that satisfies the following expression:

$$\begin{aligned} \hat{x} &= \arg\max_{x \in \mathcal{X}_f} \mathcal{D}(x) \\ s.t. \quad \mathcal{X}_f &= \{x \in \mathbb{R}^n \mid \mathrm{trunc}(M_\theta(x), K) = f\}, \end{aligned} \tag{1}$$

which means that the adversary utilizes K-truncated prediction vectors to invert the most likely original input RF data while ensuring compliance with the target model's classification requirements. This task of obtaining the inversion data $\hat{x}$ from M_θ, f, and $\mathcal{D}$ is referred to as the RF data inversion problem.

Attack scope and Assumptions. We consider a non-invasive adversary who interacts with the RF sensing system only in the digital domain and has no capability to perform physical-layer over-the-air (OTA) attacks or to manipulate RF hardware. Our threat model focuses on in-domain inversion scenarios, where the auxiliary data available to the adversary are collected under sensing conditions similar to those of the victim's environment. This assumption is consistent with prior black-box model inversion studies in the vision domain [37], [38], which investigate the recoverability of private information rather than cross-domain generalization.

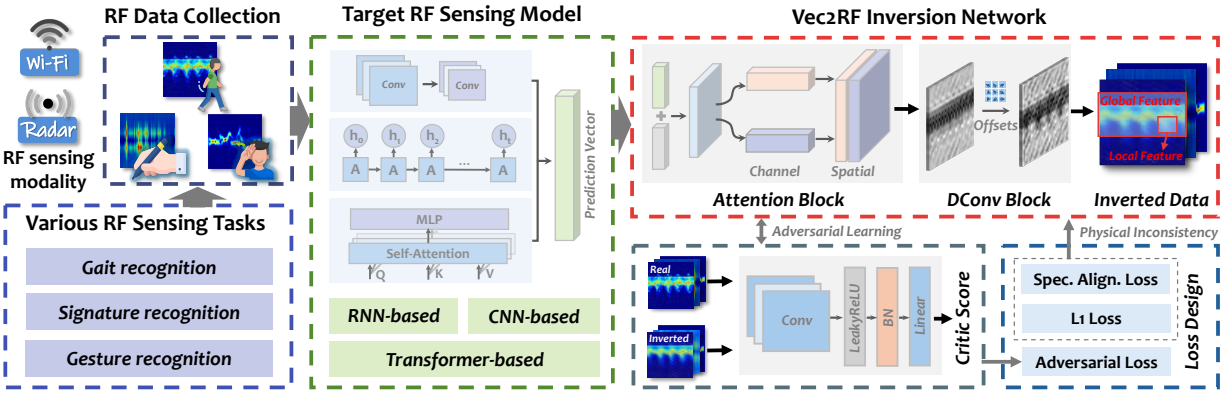


Fig. 5: RFInv design.

IV. RFInv Design

A. Attack Scheme Overview

We design a suite of techniques to realize the proposed RF inversion attack against deep learning-empowered RF sensing systems. The attack overview is illustrated in Fig. 5.

The target of the RFInv is a deep learning classifier-empowered RF sensing system, which can be based on various architectures, such as CNN, RNN, or Transformer. RFInv leverages an auxiliary dataset with a generic distribution similar to that of the training data used for the target RF sensing model to train the Vec2RF inversion network, which ensures that RFInv maintains high performance even without access to the target model's training dataset. The training data is first passed through the black-box target RF sensing model to obtain the prediction vector. The prediction vector (or the truncated vectors) is then served as input to the Vec2RF inversion module, which is adversarially trained alongside the critic module. To ensure the quality of the inversion data, we introduce attention and a deformable convolutional architecture in the inversion module, combined with a carefully designed loss function to maintain the physical consistency of the inverted data at the local and global feature levels.

B. Training Data Preparation

1) Auxiliary Dataset: Our goal is to recover RF data similar to the original RF data using the Vec2RF network. However, training the designed Vec2RF network requires a dataset. Since we cannot directly access the target model's training data, we use auxiliary datasets for Vec2RF inversion network training. The auxiliary dataset should contain sufficient semantic information about the corresponding RF sensing task to regularize the ill-posed RF model inversion problem.

We compose the auxiliary dataset that shares a generic distribution as the original target model training data distribution. For instance, if the target model is an RF gait recognition model, we can utilize publicly available RF gait datasets or self-collect RF gait data for training. The samples from the auxiliary dataset still retain general gait features such as step length, step frequency, and body joint positions, which are shared generic semantic information of the RF data used for training the target RF gait recognition model. These shared general features provide enough information to regularize the ill-posed model inversion task [39]. By utilizing an auxiliary dataset with generic features to the target model's dataset, we introduce additional information that helps the Vec2RF inversion network learn the features and patterns of the target data, making the solution to the inversion problem more stable and reasonable. Furthermore, the samples in the auxiliary dataset retain the shared generic semantic features of the target dataset (e.g., gait features), which can be used as regularization terms to guide the Vec2RF inversion network to reconstruct the target data more accurately. In addition, we have considered the potential distributional differences between the auxiliary dataset and the target model's training data. The performance of our attack with auxiliary datasets of varying similarity is evaluated through experiments, as detailed in Sec. V-D.

2) Prediction Vector Truncation: The prediction vectors of the auxiliary datasets are leveraged as the training data of the designed Vec2RF inversion network. To ensure the robustness of our proposed method under conditions where attackers may have access only to partial prediction results on a victim user's data, we incorporate a truncation approach in our attack process. Specifically, given a prediction vector p, we define a K-truncated vector $p\tilde{}_K$, which $p\tilde{}_K$ retains the K largest values of p while truncating the remaining values to zero. This truncation serves as a form of feature selection, eliminating less significant classes in the prediction vector (i.e., those with lower confidence levels). Consequently, this reduces the risk of overfitting in the Vec2RF network, ensuring that it can effectively reconstruct the input data from the preserved important classes. By employing this truncation method, we enhance the training of the Vec2RF network, ensuring it remains resilient and effective even when only a partial prediction vector is available.

To further improve the robustness and performance of the Vec2RF network, we must address the information loss in the prediction vector caused by the softmax function. Directly using the prediction vector $M_\theta(x)$ (where M_θ represents the target RF sensing model M with parameters θ) to train the Vec2RF network does not yield optimal inversion results. This is because the logits at the output layer are scaled in the range $[0, 1]$ and sum to 1 due to the softmax function, which weakens the activations of the output layer and results in the loss of some information necessary for the subsequent inversion process. To address this issue, we rescale the prediction vector

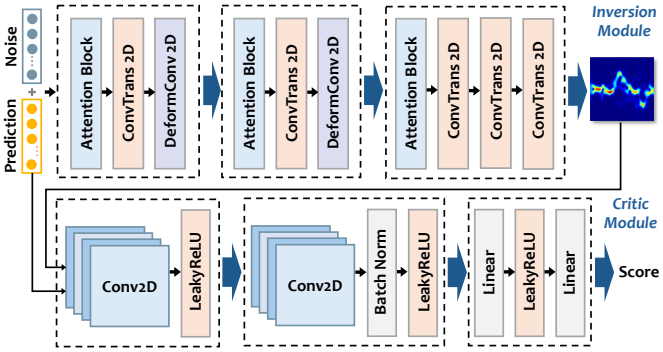


Fig. 6: Vec2RF inversion network architecture.

to its corresponding logits as follows:

$$p = \log(M_\theta(x)) + c, \tag{2}$$

where the $\log(\cdot)$ operation is applied element-wise, and c is a scalar added to $\log(M_\theta(x))$. This rescaling mitigates the information loss induced by softmax scaling and improves the effectiveness of the Vec2RF network.

C. Vec2RF Inversion Network Architecture

The architecture of the designed deep generative inversion network, Vec2RF, is shown in Fig. 6. It adopts an adversarial framework [40] and consists of two main components: the inversion module and the critic module. To that end, Vec2RF is designed to capture non-rigid Doppler trajectories and preserve frequency–time coherence during inversion.

1) Inversion Module: The inversion module aims to reconstruct sensitive RF spectrograms/heatmaps from the prediction vector of the target RF sensing model. However, RF spectrograms exhibit no clear semantic edges and are governed by propagation phenomena such as multipath reflection and Doppler shifts. Their temporal evolution is continuous and non-stationary, requiring models to track time–frequency variations while preserving amplitude and phase coherence. Moreover, the superposition of multipath components and ambient noise introduces intrinsic ambiguity, making inversion prone to distorted structures and violations of physical consistency. To address this, we incorporate convolutional attention [41] and deformable convolution [42] to enhance robustness.

Attention Block. As shown in Fig. 7, the convolutional attention block consists of two core components: Channel Attention and Spatial Attention. This module adaptively redistributes weights across both channel and spatial dimensions, guiding the model to emphasize critical physical characteristics in the RF spectrogram reconstruction process. By enhancing the model's focus on these essential features, convolutional attention improves the reconstruction fidelity and ensures physical consistency in the inverse mapping process.

In the channel dimension, different channels of an RF spectrogram may correspond to distinct features, with certain channels capturing essential reflections or multipath signal features, while others primarily contain noise or less relevant information. The channel attention mechanism begins by applying both max pooling and average pooling operations across the spatial dimensions of the input feature map:

$$S^c_{max} = MaxPool(S), \quad S^c_{avg} = AvgPool(S), \tag{3}$$

where $S \in \mathbb{R}^{C\times H\times W}$ represents the input feature map, and $H \times W$ denotes the spatial dimensions. The pooled representations S^c_{max} and S^c_{avg} extract the dominant and average responses across the spectrogram, preserving both local maxima and global trends. These pooled features are passed through a shared multi-layer perceptron (MLP) to compute channel-wise attention:

$$M_c = \sigma(W_1(W_0(S^c_{max})) + W_1(W_0(S^c_{avg}))), \tag{4}$$

where $W_0 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$ are the weights of the MLP, r is the reduction ratio, and σ represents the sigmoid activation function. Finally, the resulting channel attention map M_c is broadcast and applied to the intermediate feature map:

$$S' = M_c \otimes S, \tag{5}$$

where $\otimes$ denotes element-wise multiplication along the channel dimension. This amplifies the contribution of critical frequency-time components necessary for reconstructing fine-grained spectrogram details, while reducing the influence of channels less relevant to the RF signal recovery.

Following the enhancement of important channels, the spatial attention mechanism identifies prominent regions within the spectrogram that reflect key reflections or interference patterns. Spatial attention ensures that critical features are emphasized during the inversion process, aiding in the recovery of fine-grained spectral details that may otherwise be lost. To compute spatial attention, max pooling and average pooling are applied across the channel dimension to generate two spatial descriptors:

$$S^s_{max} = MaxPool(S'), \quad S^s_{avg} = AvgPool(S'), \tag{6}$$

where $S^s_{max}, S^s_{avg} \in \mathbb{R}^{1\times T\times F}$ capture the maximum and average responses across channels for each time-frequency bin, preserving both strong local features and broader spectral trends. These descriptors are concatenated along the channel axis and passed through a convolution layer to generate the spatial attention map:

$$M_s = \sigma(Conv([S^s_{max}; S^s_{avg}])), \tag{7}$$

where $Conv$ denotes convolution. This convolution refines spatial localization by integrating global and local contextual information. Finally, the resulting spatial attention map M_s is applied to the processed feature map of the channel attention:

$$S'' = M_s \otimes S', \tag{8}$$

yielding a feature map that is selectively refined in both spatial and channel dimensions. The final feature map is obtained by the sequential application of channel and spatial attention:

$$S_{out} = M_s \otimes (M_c \otimes S). \tag{9}$$

In our RF inversion task, the channel attention module prioritizes frequency bands or temporal segments that encode critical physical features such as multipath signals or reflections, while the spatial attention mechanism refines regions of

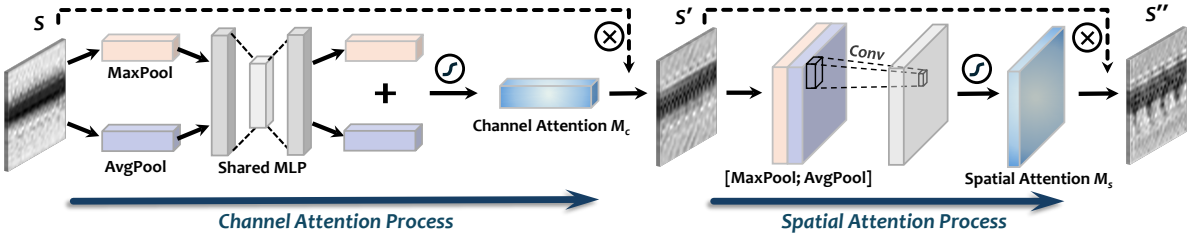
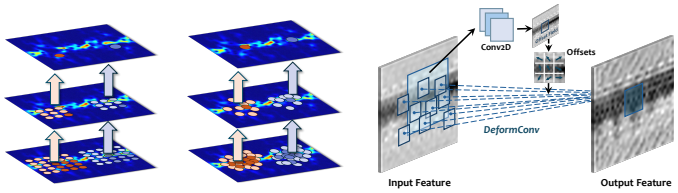


Fig. 7: Convolutional Attention used in Vec2RF Inversion Network.



(a) CNN. (b) DCN. (c) Offset Learning.

Fig. 8: Fixed vs. offset-adaptive receptive fields.

the spectrogram corresponding to key interference points or object signatures. This dual attention mechanism ensures that the inversion model reconstructs spectrograms that preserve key physical characteristics, suppress background noise, and improve the fidelity of the inverted RF data.

Deformable Convolution. While the convolutional attention block effectively enhances feature selection through channel and spatial recalibration, the reconstruction of RF spectrograms demands additional flexibility in capturing irregular and dynamic signal distortions, such as multipath propagation, reflections, and Doppler shifts. These phenomena often manifest as non-linear deformations in the time-frequency domain, making traditional convolutional operations, with their fixed receptive fields, insufficient to accurately model the complex physical characteristics of RF signals.

To address this, we integrate deformable convolution into the inversion module to improve the model's adaptability to dynamic and spatially variant RF features. As illustrated in Fig. 8, deformable convolution augments the standard convolution operation with learnable offsets, allowing its receptive field to flexibly deform according to the underlying spectrogram structure. This adaptive mechanism enables the network to better capture non-rigid signal components, which are crucial for modeling variations caused by multipath interference and environmental reflections [43].

2) Critic Module: The critic module evaluates the quality of the recovered RF data. It takes as input the concatenation of the prediction vector and the inverted RF data. The module consists of two convolutional layers with LeakyReLU activations to extract local features, followed by a linear layer with batch normalization and LeakyReLU for high-dimensional feature processing. The final linear layer outputs a critic score representing the Wasserstein distance [44] between the recovered and real data. This score measures the realism of the recovered RF data and provides feedback to guide the inversion module toward generating more authentic results.

D. Training Flow

We adopt WGAN-GP [44] as the framework for training the inverse module. Compared to traditional cGAN, WGAN-GP introduces Wasserstein distance and gradient penalty, effectively mitigating mode collapse and gradient vanishing issues, resulting in more stable training and higher-quality generation. In our RF data reconstruction task, WGAN-GP produces smoother and more detailed spectrograms, preserving weak signals and frequency peaks, significantly enhancing reconstruction accuracy.

1) Loss Design: Maintaining global energy distribution consistency while accurately recovering salient spectrogram features is essential for effective training. However, conventional pixel-wise losses (e.g., L1 or L2) impose uniform penalties on both background and signal regions, often blurring or suppressing high-energy Doppler traces that correspond to meaningful reflections. To overcome this limitation, we design a Spectrogram Alignment (SA) Loss combined with L1 loss to jointly preserve global structure and local signal fidelity. Unlike edge- or texture-based visual losses, SA Loss aligns the amplitude and location of energy peaks representing multipath and Doppler components, thereby enforcing the physical coherence of reconstructed RF spectrograms.

The core concept behind SA Loss is to extract and compare local peaks within 2D sliding windows, ensuring that high-energy areas in the inverted spectrogram align with those in the ground truth, both in terms of amplitude and spatial location. The function of the spectrogram alignment loss can be expressed as

$$L_{SA} = \frac{1}{K}\sum_{k=1}^{K} \underbrace{\left(\max(W_{inv}^k) - \max(W_{real}^k)\right)^2}_{\text{Magnitude Loss}} + \lambda_{\text{pos}} \cdot \underbrace{\left(p_{inv}^k - p_{real}^k\right)^2}_{\text{Position Loss}}, \tag{10}$$

where K is the number of sliding windows, W_{inv}^k and W_{real}^k are the inverted and real spectrogram patches in the k-th window, p_{inv}^k and p_{real}^k represent the peak positions within the window, and λ_{pos} is the position weight. By optimizing spectrogram positions and magnitudes within local windows, the inversion module ensures precise alignment of energy and location in target regions. This approach reduces background interference, guiding the inversion module to focus more effectively on salient target features. Meanwhile, we employ L1 loss as a

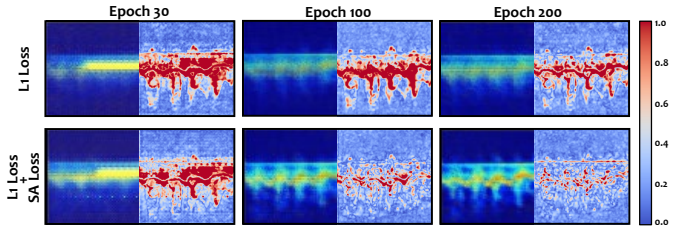


Fig. 9: Error Heatmap of RF Spectrogram Reconstruction: L1 vs. L1+SA. The left side shows the inverted data, while the right side presents the difference heatmap compared to the real data, where warmer colors indicate higher differences.

global constraint to maintain the consistency of the overall energy distribution across the RF spectrogram:

$$L_{L1} = \frac{1}{N}\sum_{i=1}^{N}|x^i - \hat{x}^i|, \tag{11}$$

where N is the total number of pixels, and x^i and $\hat{x}^i$ represent the i-th pixel in the inverted and real spectrograms.

The total generator loss is designed to balance **global consistency** with **local feature preservation** by combining L1 loss, SA loss, and adversarial loss:

$$L_I = \lambda_1 L_{L1} + \lambda_2 L_{SA} - \mathbb{E}_{\hat{x}}[C(\hat{x})], \tag{12}$$

where λ_1 and λ_2 control the contributions of the L1 and SA losses, respectively. Fig. 9 illustrates the comparison of RF spectrogram reconstruction using L1 loss and L1 + SA loss at different training epochs. The results demonstrate that the introduction of SA loss enhances the module's ability to align with the peak values of the RF spectrogram, significantly reducing reconstruction errors in the low-frequency region.

To ensure the inversion module continuously improves, the critic enforces a loss derived from the WGAN. It improves stability and mitigates mode collapse by minimizing the Wasserstein distance between real and inverted distributions:

$$L_{\mathrm{C}} = \underbrace{\mathbb{E}[D(\hat{x})] - \mathbb{E}[D(x)]}_{\text{Wasserstein Distance Loss}} + \lambda_{gp}\underbrace{\mathbb{E}\left[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2\right]}_{\text{Gradient Penalty}}, \tag{13}$$

where $D(\hat{x})$ and $D(x)$ represent the critic's output for the inverted and real spectrograms, respectively, and $\tilde{x}$ is a random interpolation between the real and inverted spectrograms. By penalizing deviations from the Lipschitz constraint, the critic enforces smooth gradients, promoting stable convergence and reducing training instability. This setup ensures that the inversion module continuously improves, producing spectrograms that closely resemble real data in terms of both overall structure and localized features.

2) Training Process: The training process is illustrated in Alg. 1. During each iteration, we first train the critic module multiple times (i.e., Line 3-10). This frequent training ensures that the critic module provides accurate feedback and enhances the stability of the entire training process. The critic's parameters θ_c are updated by minimizing the Wasserstein distance between the distributions of real and inverted RF data. A gradient penalty is applied to enforce the Lipschitz constraint, ensuring the critic remains effective in distinguishing real from inverted data and maintaining smooth gradients during optimization. After sufficient updates to the critic, the inversion module is trained by minimizing a composite loss function (i.e., Line 11-14). The inversion module's parameters θ_i are updated by minimizing the combined loss of the critic score on the recovered data, L1 loss and SA loss between the inverted and real data.

Our training process employs a joint training approach that combines the target model M, the inversion module I, and the critic module C. Meanwhile, the integration of adversarial loss, L1 loss, and SA loss ensures both global consistency and local feature preservation in the recovered RF data.

Algorithm 1: Vec2RF Training Process.

Input : $\mathbb{P}_r$, $\mathbb{P}_z$: distributions of real RF data and noise; α_c, α_i: learning rates of critic and inversion module; λ_{L1}, λ_{SA}, λ_{gp}: weights for L1 loss, SA loss, and gradient penalty; N_c: number of critic iterations per inversion update; N: total number of iterations

Output: θ_c: critic parameters; θ_i: inversion module parameters

1 Initialize θ_c and θ_i
2 **while** $t \leq N$ **do**
3 **for** $k = 1$ *to* N_c **do**
4 Samplea batch of real RF data $x \sim \mathbb{P}_r$ and noise vectors $z \sim \mathbb{P}_z$
5 Pass x through the target RF sensing model M to get prediction vectors $p = M(x)$
6 Reconstruct inverted RF data $\hat{x} = I(p, z)$
7 Interpolate between real and inversion RF data: $\tilde{x} = \alpha x + (1-\alpha)\hat{x}$ where $\alpha \sim U[0,1]$
8 Compute gradient penalty:
$$GP = \mathbb{E}_{\tilde{x}}\left[(\|\nabla_{\tilde{x}} C(\tilde{x})\|_2 - 1)^2\right]$$
9 Update critic module parameters θ_c:
$$\theta_c \leftarrow \theta_c - \alpha_c \nabla_{\theta_c}\left(\mathbb{E}_x[C(x)] - \mathbb{E}_{\hat{x}}[C(\hat{x})] + \lambda_{gp} GP\right)$$
10 **end**
11 Samplea batch of real RF data $x \sim \mathbb{P}_r$ and noise vectors $z \sim \mathbb{P}_z$
12 Reconstruct inverted RF data $\hat{x} = I(M(x), z)$
13 Compute spectrogram alignment loss L_{SA} and L1 loss L_{L1}.
14 Update inversion module parameters θ_i:
$$\theta_i \leftarrow \theta_i - \alpha_i \nabla_{\theta_i}\left(\lambda_1 L_{L1} + \lambda_2 L_{SA} - \mathbb{E}_{\hat{x}}[C(\hat{x})]\right)$$
15 **end**

V. Evaluation

A. Experimental Setup

1) Target RF Sensing Tasks and Datasets: To evaluate our proposed attack scheme, we employ datasets representing three distinct yet privacy-sensitive RF sensing tasks across different RF modalities.

- **T1) Gait Recognition.** Gait recognition is an identity verification task based on an individual's walking pattern. Due to its uniqueness and resistance to forgery, gait recognition is often regarded as a critical biometric identification method. However, if RF gait data is reversed, it may lead to leakage of user identity information and spoofing of the gait recognition system, thus realizing identity impersonation or bypassing authentication. We use the mmWave-based

TABLE II: Details of three basic target RF sensing models.

Architecture	Model Structure	Optimizer	Accuracy		
			T1	T2	T3
CNN	(Conv2D + BN[1] + Max Pooing + ReLU) × 3 Flatten+Dense+Dropout + ReLU Dense + Softmax	Adam	95.30%	93.73%	94.33%
RNN (LSTM)	Permute + Reshape (batch_size, 64, 64×3) (LSTM Bidirectional) × 2 (Dense + Dropout + ReLU) × 2 + Dense + Softmax	RMSprop	92.63%	91.42%	93.60%
Transformer	PatchEmbedding + Flatten + Dense TransformerBlock × 8 LayerNorm + GAP[2] + Dense + Softmax	AdamW	97.13%	93.33%	98.76%

[1]BN: Batch Normalization; [2]GAP: Global Average Pooling.

gait dataset MMRGait [45], which contains mmWave gait data from 121 volunteers collected using the TI AWR1843 radar. The dataset has eight walking paths per volunteer. To ensure the performance of the target model, we exclud paths walking along the mmWave radar normal direction.

- **T2) Gesture Recognition.** Gesture recognition enables contactless human-computer interaction by analyzing hand movements. If RF gesture data is reversed, attackers may forge user gestures to bypass system authentication or maliciously control devices. In addition, prolonged monitoring of gesture data could expose user behavior patterns and compromise privacy. We use the Wi-Fi-based gesture dataset Widar 3.0 dataset [46], which contains data from 22 types of gestures performed by 17 volunteers.
- **T3) Signature Recognition.** Signature recognition leverages users' unique handwriting characteristics for identity verification, widely applied in finance and security systems. Inversion of signature data could enable attackers to forge signatures, bypassing authentication and leading to identity impersonation or unauthorized access. We use the mmSign dataset [47], which contains mmWave signature data collected from 30 volunteers using the TI AWR1642 radar.

For each of these above tasks, we use half of the data to train the target RF sensing model and the remaining half of the data as an auxiliary dataset to train the Vec2RF inversion network. There is no intersection of categories between the data used to train the target RF sensing model and the data used to train the inversion network. For instance, we select classes 1- 60 of the mmWave gait dataset to train the target model and classes 61-120 to train the inversion network. Then we use classes 1-60 to test the performance of the trained inversion network.

2) Target RF Sensing Model: We evaluate our attack with three basic models and two advanced models. **1) Basic Model.** RFInv is tested on three representative tasks (i.e., gait, gesture, and signature recognition), each implemented with CNN, RNN, and Transformer architectures. All models share the same structure except for the output layer, which is adjusted to match the number of classes. The input size is $3 \times 64 \times 64$, trained with Cross-Entropy loss and a learning rate of 0.001. Their architectures and accuracies are summarized in Tab.II. **2) Advanced Model.** In addition to the basic models, we also evaluate the performance of our attack scheme on more advanced deep learning models. Specifically, we validate its effectiveness on the corresponding advanced deep learning models on three individual tasks. The experimental details are given in Sec. V-E.

3) Evaluation Metrics: We evaluate the proposed attack from two critical perspectives: the similarity between the inversion data and the original data, and the effectiveness of our proposed attack on the target system.

- To evaluate the similarity between the inversion and original data, we utilize the **Structural Similarity Index Measure (SSIM)** to quantify the similarity between the inversion data and the original data, calculated as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{14}$$

  where $\mu_{(\cdot)}$ is the mean of $(\cdot)$, $\sigma^2_{(\cdot)}$ represents the variance of $(\cdot)$, and σ_{xy} is the covariance between x and y. Constants c_1 and c_2 are introduced to avoid instability when the denominator is close to zero. We first convert the spectrogram into a grayscale map and then calculate the SSIM value.
- To assess the attack's effectiveness, we define the **Relative Attack Success Rate (RASR)** to evaluate the effectiveness of RFInv on the target model using inversion data, given that the recognition accuracy of the target RF sensing model is not 100%. The RASR can be expressed as follows:

$$RASR = P_{\text{inv}}/P_{\text{real}} \times 100\%, \tag{15}$$

  where P_{inv} and P_{real} represent the probabilities that the inversion data and real data are correctly recognized by the target RF sensing model, respectively. A RASR close to 100% means the inversion data is recognized nearly as accurately as the real data, indicating an effective attack. A lower RASR suggests the attack is less effective. RASR takes into account the baseline performance of the target model itself, which more accurately reflects the actual effectiveness of the proposed attack.

B. Overall Performance

We first evaluate the overall performance of our proposed attack across the aforementioned three different tasks without prediction vector truncation. For each task, we evaluate the effectiveness of our proposed attack scheme on target RF sensing models with different architectures, as listed in Tab. II. When the target model is CNN-based, the visualization results of the

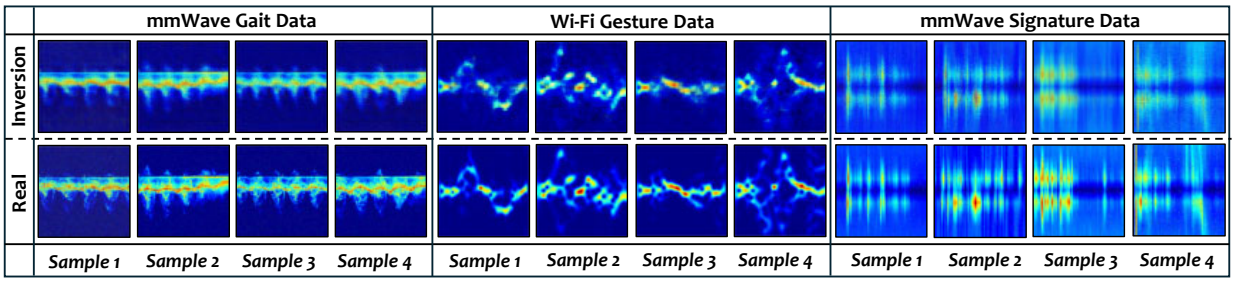
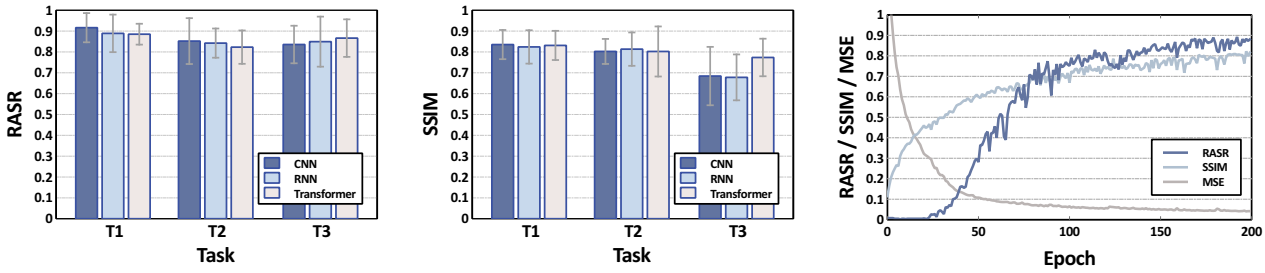


Fig. 10: Inversion results on different RF sensing tasks.



(a) Overall RASR. (b) Overall SSIM. (c) Training process.

Fig. 11: Overall Performance.

inversion are illustrated in Fig. 10. These examples clearly show that the inverted spectrograms preserve key semantic patterns, such as gait cycles (periodic torso/leg movements), gesture trajectories, and signature strokes, which are sufficient for downstream recognition. Such patterns are highly privacy-sensitive, as they can be exploited to infer user identity (via gait-based authentication), behavioral habits, or even health conditions. The performance of our proposed attack on three different target models is presented in Fig. 11a and Fig. 11b.

Our attack achieves average RASR values of 89.68%, 83.91%, and 85.04% for gait recognition, gesture recognition, and signature recognition tasks, respectively. The corresponding SSIM values are 0.83, 0.81, and 0.71. The inversion performance varies across different tasks. For the gait recognition task, the spectrogram primarily captures coarse-grained information, such as the large-scale movements of the torso and legs. The inversion network excels at recovering these critical spatio-temporal features, resulting in strong attack performance and high recovery accuracy. However, for the signature recognition task, the spectrogram reflects fine-grained hand movement speed variations and contains higher texture complexity. The inversion method faces challenges in reconstructing these high-frequency details, leading to a significantly lower SSIM. Despite this limitation, the inverted data still effectively captures the overall distribution characteristics, such as the key stripe patterns and feature structures. These preserved semantic features are sufficient to deceive the target RF sensing model, achieving RASR values above 80%, which demonstrates that the attacker could realistically impersonate user identities or bypass RF-based authentication systems.

To analyze the relationship between RASR and SSIM, we plot their changes during training for the gait recognition task. Meanwhile, we illustrate the variation of Mean Squared Error (MSE) across epochs, as shown in Fig. 11c. As training progresses, SSIM increases at a slower rate compared to RASR. By the time SSIM reaches approximately 0.75, RASR exceeds 80%. Simultaneously, the MSE value continues to decrease, eventually stabilizing around 0.04. This indicates that the inverted data retains all essential features, such as step speed and frequency, which the RF sensing model can effectively extract and recognize. Notably, our experiments utilize a resolution of 64 $\times$ 64, which constrains the amount of information encoded in each pixel. For detail-rich tasks such as signature recognition, while the inverted results closely approximate the overall distribution of the original data, minor deviations in details are amplified, causing a decrease in the SSIM value. Therefore, despite the lower SSIM, the inversion data successfully deceives the recognition system by incorporating the key features.

C. Effectiveness of the Designed Inversion Network

In this experiment, we evaluate the effectiveness of the designed inversion network. Specifically, we assess the impact of our Vec2RF network and the designed spectrogram loss function on the reconstruction results, respectively. The mmWave-based gait recognition task is used as a representative example.

1) Vec2RF Network Effectiveness.: We compare the performance of our designed Vec2RF network with that of traditional models. Specifically, we evaluate our approach against the traditional CGAN-based model and the WGAN-based model. The comparison results are presented in Tab.III and Tab.IV.

The RASR of Vec2RF network is improved by 21.69% and 10.48% compared to CGAN-based and WGAN-based inversion schemes, respectively. Similarly, the SSIM of Vec2RF shows an improvement of 16.46% and 9.62% over these traditional models. The training approach is based on WGAN-GP, which addresses the issues of gradient vanishing and pattern collapse by a gradient penalty, resulting in higher-quality samples. Meanwhile, Vec2RF incorporates L1 loss to

enhance stability and RF data similarity. The L1 loss reduces pixel-level variance, thereby improving the SSIM. In contrast, the CGAN model suffers from unstable training and poor sample quality, while the traditional WGAN model lacks both gradient penalty and L1 loss, leading to lower inversion performance. Therefore, the designed Vec2RF network achieves a more stable training process and higher quality inversion data, significantly improving both RASR and SSIM.

TABLE III: Performance comparison.

	CGAN		WGAN		Vec2RF	
	RASR	SSIM	RASR	SSIM	RASR	SSIM
CNN	76.30%	0.71	82.46%	0.77	91.63%	0.84
RNN	71.55%	0.70	79.70%	0.73	88.90%	0.82
Transformer	73.32%	0.72	81.36%	0.77	88.51%	0.83

2) Spectrogram Alignment Loss Effectiveness: We evaluate the impact of our designed loss function on the RF inversion results. Specifically, we compare three configurations: 1) using only L1 loss, 2) using only Spectrogram Alignment (SA) loss, and 3) using both L1 loss and SA loss in combination. The evaluation results are shown in Tab. IV.

Compared to using only L1 loss, our designed loss function can improve 11.12% on RASR and 9.83% on SSIM. And compared to using only SA loss, it can improve 25.91% on RASR and 19.73% on SSIM. The experimental results demonstrate that using only L1 loss or SA loss results in moderate performance in terms of RASR and SSIM. This can be attributed to the different roles each loss function plays: L1 loss provides a baseline reconstruction quality by minimizing global pixel-level errors but lacks the ability to recover fine-grained details in high-energy regions. On the other hand, SA loss focuses on enhancing local features in high-energy regions but struggles to maintain global structural consistency. The L1 + SA combined loss effectively balances global consistency and local feature optimization. Specifically, L1 loss ensures overall energy distribution consistency by minimizing pixel-level differences, thereby improving global structure reconstruction. Meanwhile, SA loss emphasizes local alignment in high-energy regions by optimizing the amplitude and position of salient features, enabling precise recovery of target motion details.

Furthermore, the results indicate that SA loss yields relatively smaller performance improvements compared to L1 loss. This is because L1 loss establishes a strong baseline performance by minimizing errors across the entire spectrogram. While SA loss significantly enhances alignment in high-energy regions, the low-energy background occupies a large portion of the RF spectrogram. Since global performance metrics like RASR and SSIM weigh all pixels equally, the overall contribution of SA loss to these metrics is more modest.

D. Impact of Auxiliary-Victim Dataset Similarity

To evaluate how the similarity between auxiliary and victim datasets affects attack performance, we conduct a hierarchical similarity experiment spanning four levels, ranging from ideal *in-domain* to highly divergent *cross-domain* settings. All target models are CNN-based. The similarity between the auxiliary and target datasets is categorized into four representative scenarios as follows:

- High Similarity (H.S.): Both the target and inversion models are trained on data from the same user categories within the mmWave gait recognition task, without overlapping samples. This represents an ideal in-domain condition.
- Moderate Similarity (M.S.): The target model is trained on half of the user categories in the mmWave gait dataset, while the remaining categories are used as the auxiliary dataset. This simulates a cross-user in-domain scenario.
- Cross-Source (C.S.): The two models share the same task and sensing modality but use data from different sources. We collected a small mmWave signature dataset (10 subjects) in our lab as victim data and used the public mmSign dataset as the auxiliary data. This introduces mild environmental and subject variations, representing a practical in-domain condition between moderate and low similarity.
- Low Similarity (L.S.): The target and inversion models are trained on different tasks. Specifically, the target model uses the mmWave gait dataset, while the inversion model is trained on the Wi-Fi gesture dataset, representing cross-domain condition.

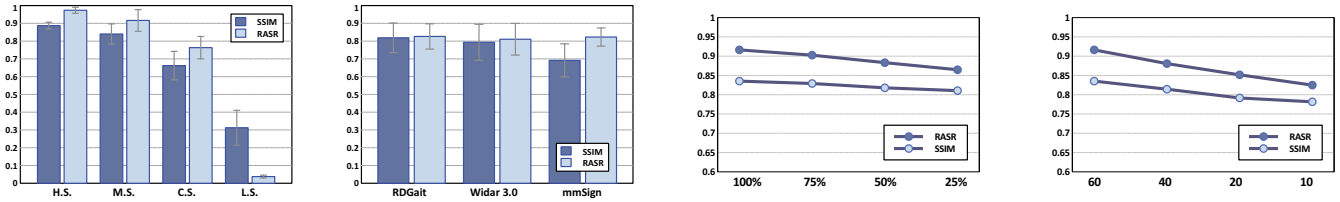
TABLE IV: Effectiveness of the designed loss.

	L1 Loss		SA Loss		L1 + SA Loss	
	RASR	SSIM	RASR	SSIM	RASR	SSIM
CNN	82.23%	0.76	73.22%	0.70	91.63%	0.84
RNN	78.61%	0.74	71.27%	0.69	88.90%	0.82
Transformer	81.30%	0.77	69.23%	0.68	88.51%	0.83

The evaluation results are shown in Fig. 12a. As the similarity between the auxiliary and victim datasets decreases, both the SSIM and the RASR consistently decline. In the high-similarity (H.S.) setting, Vec2RF reconstructs target data with high fidelity (SSIM=0.88) and achieves a RASR of 97.35%. Under the moderate-similarity (M.S.) condition, performance remains high (SSIM=0.84, RASR=91.6%). In the cross-source (C.S.) scenario, the inversion performance moderately declines (SSIM=0.66, RASR=76.32%) but remains stable overall, indicating Vec2RF's practical feasibility in real-world settings. When the similarity is low (L.S.), the inversion quality drops sharply, as heterogeneous data distributions prevent the network from learning aligned representations.

E. Effectiveness on Advanced Deep Learning Models

To further validate the robustness of RFInv, we evaluate its performance on three state-of-the-art deep learning models in RF sensing. For gait recognition, we adopt the RDGait [48] model, which integrates CNN and LSTM with attention mechanisms for feature embedding, with its multi-frame input adapted to our dataset. For gesture recognition, we use the Widar 3.0 [46] model combining CNN and GRU networks, modified to accept a single spectrogram as input. For signature recognition, we employ the mmSign [47] model based on attention mechanisms and patch embeddings, extending its binary output layer to support multi-class classification.



(a) Impact of auxiliary–victim dataset similarity. (b) Performance on advanced models. (c) Effect of the prediction truncation. (d) Effect of the prediction vector dimensionality.

Fig. 12: Experimental results.

As shown in Fig. 12b, RFInv remains effective against all three advanced models. It achieves an average SSIM of 0.82, 0.79, and 0.69, and a RASR of 82.63%, 81.06%, and 82.30% on RDGait, Widar 3.0, and mmSign, respectively. These results indicate that RFInv generalizes well to attention-based and hybrid CNN–RNN architectures, confirming its robustness and practicality across diverse advanced models.

F. Analysis of Prediction Vector Characteristics

We analyze two prediction-vector factors that affect attack success: truncation ratio (amount of retained values) and dimensionality (number of output classes), which together determine the information available for inversion.

1) Effect of Prediction Vector Truncation: In the mmWave-based gait recognition task, we truncate the prediction vectors by keeping the top 100%, 75%, 50%, and 25% of the values, respectively. We leverage the CNN-based model as the target model. The evaluation results are shown in Fig. 12c.

When the prediction vectors are not truncated, the average SSIM and RASR of the attack are 0.84 and 91.63%, respectively. However, when the prediction vectors are truncated to retain only the top 75%, 50%, and 25% of the values, the average SSIM drops to 0.83, 0.82, and 0.81, respectively. And the RASR drops to 90.26%, 88.32%, and 86.48%, respectively. Although the increase in the truncation ratio causes a slight decrease in the effectiveness of the attack, the decrease is relatively small, especially when only 25% of the data is retained, and the attack still manages to maintain a high inversion accuracy (SSIM close to 0.81) and attack success rate (RASR stays above 85%). This is because low-magnitude elements in the prediction vector contribute little to the model's final decision, whereas high-valued entries encode the most discriminative information. Retaining these dominant components preserves the key semantic structure, enabling the attack to maintain high reconstruction fidelity and success.

2) Effect of Prediction Vector Dimensionality: We evaluate the impact of the number of classes in the target RF sensing model on our proposed attack by training mmWave-based gait recognition models with 10, 20, 40, and 60 classification classes. The CNN-based model is employed as the target model, and the experimental results are presented in Fig. 12d.

RFInv achieves average RASR values of 91.63%, 88.07%, 85.16%, and 82.49% when the class number of the target model is 60, 40, 20, and 10, respectively. Similarly, the average SSIM values are 0.84, 0.81, 0.79, and 0.78 for the respective class numbers. These results indicate that the average RASR and SSIM of RFInv decrease as the number of classes decreases, indicating a degradation in the attack performance when fewer categories are present in the target model. This degradation mainly stems from three aspects. First, fewer output categories produce lower-dimensional prediction vectors, carrying less detailed information for inversion. Second, with fewer classes, the target model learns less discriminative features, weakening the cues that the inversion model can exploit. Third, reduced category diversity also removes redundant correlations among logits, making it harder to correct reconstruction errors. As a result, models with fewer classes provide less informative outputs, leading to lower inversion quality and attack success.

VI. Discussion

A. Challenges and Insights

1) Scope and Practical Implications: As stated in Sec. III, our threat model focuses on digital-domain inversion attacks rather than physical-layer OTA spoofing. High-fidelity OTA injection requires prior information such as precise synchronization and carrier/phase control, which are typically difficult in practical deployments. Therefore, OTA spoofing is treated as a high-cost, constrained scenario and is not assumed by default in this study. By contrast, a more realistic and lower-barrier threat stems from exposure of cloud/API intermediate representations. As RF sensing systems migrate to cloud-based architectures [49], devices commonly upload processed spectrograms or feature vectors to servers for recognition or authentication. If these intermediate representations can be accessed or estimated, an adversary can construct or tamper with inputs in the digital domain to mislead downstream classifiers without any physical signal transmission; our experiments are conducted under this assumption and verify its impact on authentication models. Moreover, spectrograms themselves encode semantically rich and privacy-sensitive information (e.g., gait periodicity, respiration/heartbeat modulation, and throat-vibration patterns). Access to such representations enables attribute inference, membership inference, and cross-modal misuse, leading to tangible privacy leakage.

2) Domain Shift and Generalization: While our experiments focus on in-domain scenarios, domain shift remains a fundamental challenge due to the sensitivity of wireless signals to multipath and geometry. Although Vec2RF's self-supervised feature alignment helps mitigate mild variations in SNR, distance, or orientation, large-scale cross-environment differences (e.g., across rooms or devices) are beyond the current scope.

To extend RFInv to cross-domain settings, future work can integrate adversarial domain adaptation [50], [51] to encourage the model to learn environment-invariant features (e.g., motion patterns) while suppressing domain-specific background noise. Furthermore, generative augmentation, utilizing physics-based ray-tracing [52] or diffusion synthesis [53], can be employed to synthesize training data under diverse spatial conditions, thereby expanding the auxiliary dataset to better cover unseen target distributions. These strategies will collectively bridge the gap between in-domain feasibility and robust cross-environment deployment.

3) Generalizability to Other RF Modalities: Vec2RF is specifically designed for spectrogram-based RF representations and thus operates within the transformed domain. Spectrograms provide rich time–frequency cues but inevitably lose certain low-level details of the raw signals during preprocessing. Reconstructing original IQ sequences would require inverse transformations (e.g., IDFT or IDWT), whose accuracy depends heavily on signal-processing parameters. Recent work such as RF-Diffusion [53] has shown that generative models can directly synthesize complex-valued IQ signals. In future work, we plan to explore end-to-end inversion by extending Vec2RF to raw-signal reconstruction using diffusion-based generators conditioned on its intermediate features.

4) Potential Privacy Consequences: Successful recovery of high-fidelity spectrograms enables concrete downstream attacks. Specifically, recovered gait or signature spectrograms serve as digital twins preserving unique biometric traits. Attackers can launch digital replay attacks by injecting these synthesized representations into authentication pipelines to impersonate legitimate users [47]. Inverted gesture trajectories further enable side-channel inference of sensitive inputs, such as deducing PINs from virtual keyboards [54]. Furthermore, recovered semantics allow for semantic injection, where attackers construct adversarial commands to remotely trigger smart home actions [19] (e.g., unlocking doors) without physical presence. Finally, recovered micro-Doppler features expose subtle physiological anomalies, such as Parkinsonian gait freezing [55]. This facilitates unconsented health profiling, allowing malicious entities to infer private medical conditions without user awareness, leading to potential discrimination.

B. Countermeasures

To counter inversion attacks against deep learning-empowered RF sensing systems, several countermeasures can be employed to enhance RF sensing data security.

First, implement differential privacy [56] in the RF sensing system to introduce noise during model training. Differential privacy can introduce noise during model training. This approach ensures data validity while protecting data privacy, thereby reducing the risk of input data being recovered through inversion attacks. For example, adding Gaussian noise to gradients during training can obscure the precise values of the input data, making it harder for attackers to invert the original RF data from the model's outputs. By balancing noise magnitude with model utility, this approach protects data privacy while preserving the accuracy required for RF sensing applications. **Second, apply fuzzification techniques [57] to the system's output to obscure the output probability distribution.** Methods such as adding small perturbations to the output values or using smoothing techniques can make it harder for attackers to obtain high-confidence prediction vectors. This approach directly reduces the effectiveness of inversion attacks by degrading the fidelity of the input-output mapping exploited by attackers. **Finally, enhance access control and permission management of RF sensing systems to restrict access to the RF sensing model.** Restricting access to the output of the deep learning model ensures that only authorized users can access sensitive data. Implementing role-based access control [58] and monitoring access logs can help mitigate the risk of reverse recovery attacks from the source. In addition, implementing query rate limits and logging all interactions [59], [60] with the RF sensing system can deter and detect potential malicious activity.

VII. Related Works

A. RF Sensing Security

RF sensing has been widely explored and applied in diverse domains, such as human identification [1], activity recognition [61], and vital signs monitoring [6]. However, with the rapid development of RF sensing systems, their inherent vulnerabilities have increasingly come to light.

Among these vulnerabilities, **adversarial attacks** have emerged as a significant threat, particularly for RF sensing systems powered by deep learning. Adversarial attacks exploit the vulnerabilities of sensing systems by introducing subtle and often imperceptible perturbations to input RF data, leading to erroneous decision outcomes. For instance, Yang et al. [62] were the first to investigate the vulnerability of wireless Doppler sensor-based human activity recognition (HAR) systems to adversarial attacks. Building upon this, Ambalkar *et al.* [16] studied the impact of adversarial attacks on deep learning-based Wi-Fi human activity recognition systems. Subsequently, Liu *et al.* [17] achieved both untargeted and targeted adversarial attacks on deep learning-based Wi-Fi HAR systems in the physical world. Li *et al.* [63] demonstrated a practical adversarial attack that is compatible with commercial Wi-Fi devices while having a minimal impact on the quality of Wi-Fi communications. Moreover, Xie *et al.* [18] proposed the first targeted adversarial attacks for mmWave-based HAR systems. Beyond HAR systems, RF-based gesture recognition systems are equally susceptible to adversarial threats. For example, WiAdv [19] is the first to explore the feasibility of physical adversarial attacks against Wi-Fi-based gesture recognition systems. By leveraging the dynamic multipath characteristics of Wi-Fi signals, WiAdv simulates gesture motion features to generate adversarial signals.

Meanwhile, **backdoor attacks** have also attracted considerable attention. These attacks involve embedding malicious behaviors into the RF sensing system during the training or development phase, such that the system behaves normally under standard conditions but exhibits compromised behaviors when specific triggers are activated. For example, Zhao *et al.* [21] investigated backdoor attacks on mmWave-based

localization systems, leveraging visible and invisible triggers to compromise system integrity.

While prior research has primarily focused on attacks targeting the systems themselves, the security of RF sensing data remains an underexplored area. Although recent studies [7], [64]–[67] have explored RF signals as a medium for side-channel recovery and eavesdropping on users' private data, no existing work has addressed data privacy leakage during the inference process of RF sensing models. In this paper, we uncover for the first time the data security threats to deep learning-empowered RF sensing systems.

B. Model Inversion Attack

Model inversion is an attack against machine learning models where the attacker infers sensitive information from the input data by analyzing the model's output.

The first model inversion attack was proposed in the context of genomic privacy [68], demonstrating that adversarial access to linear regression models used for personalized medicine could infer private genomic attributes of individuals. The follow-up work [69] utilized the confidence values of the predictions to infer the training category of neural networks by generating a representative sample of the target class. These previous optimization-based approaches treat the inversion task as a gradient optimization problem, seeking the optimal data for a given class. However, this approach is only effective for simple white-box networks and fails with more complex neural networks and black-box scenarios. To address this limitation, training-based methods were developed, which achieve the inversion of the input data by training another inversion model. For instance, Dosovitskiy *et al.* [70] proposed reconstructing the original image from the feature representation of a given layer (i.e., intermediate layer or predictive layer) by training a CNN-based inversion model. Similarly, Yang *et al.* [28] proposed to leverage prior background knowledge to train an inversion model to achieve input image inversion in adversary scenarios. These training-based methods offer a more robust solution for input data inversion, particularly when handling complex networks and black-box scenarios.

More recently, researchers have introduced generative paradigms for black-box model inversion. For instance, Ye *et al.* [37] proposed C2FMI, a coarse-to-fine framework that combines feature regression with gradient-free optimization to recover high-fidelity faces from limited black-box access. Liu *et al.* [38] further advanced this direction by proposing Prediction Alignment, which aligns model prediction vectors to guide a pretrained image generator, achieving realistic reconstructions without gradient information. Subsequent studies further improved generative inversion by exploiting richer feature representations and advanced generative architectures. Qiu *et al.* [71] enhanced inversion fidelity by aligning intermediate features within pretrained GANs, while Liu *et al.* [72] leveraged conditional diffusion models to perform label-only model inversion, demonstrating that realistic class-representative samples can be generated even without confidence scores. While these approaches achieve impressive results in the vision domain, they rely on natural image priors and focus on pixel-level visual recovery. In contrast, our work targets the inversion of RF spectrograms, where the goal is to reconstruct time–frequency structures that reflect physical signal semantics rather than visual appearance.

Building on these developments of model inversion in areas such as computer vision, we explore the first model inversion attack against deep learning-empowered RF sensing systems.

VIII. Conclusion

In this paper, we propose RFInv, the first model inversion attack targeting deep learning-empowered RF sensing systems, highlighting significant concerns about RF sensing data security. We achieve the inversion of the RF input data from the target model prediction vectors by designing a deep generative adversarial inversion network in black-box scenarios. RFInv employs a carefully designed deep generative inversion network that integrates attention mechanisms and deformable convolutions to model the complex temporal and spatial dynamics of RF spectrograms, preserving their physical consistency during reconstruction. The effectiveness of the proposed attack is validated by various types of RF sensing tasks and datasets with different modalities. To the best of our knowledge, this is the first work focused on the data security of RF sensing systems, aiming to draw attention to the importance of RF data security amidst the rapid development of RF sensing technology.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62422208, 62232010, 62572286, 62572273, 62302274, 62502274), Ministry of Industry and Information Technology of China (No. TC240A9ED-70), Shandong Science Fund (No. ZR2025LZH006, ZR2023QF113, ZR2024MF108, ZR2025QC1569), Research Project of Quancheng Laboratory, China (No. QCL20250106), Project of the Major Innovation Project of Key Laboratory of Computing Power Network and Information Security, Ministry of Education (No. 2024ZD012), Shandong Science Fund for Excellent Young Scholars (No. 2024HWYQ-021)

References

[1] X. Yang, J. Liu, Y. Chen, X. Guo, and Y. Xie, "MU-ID: Multi-user identification through gaits using millimeter wave radios," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2020, pp. 2589–2598.

[2] M. Han, L. Guo, J. Zhang, H. Ji, Z. Diao, and J. Sun, "WiID: Precise WiFi-based person identification via bio-electromagnetic information," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2022, pp. 1105–1112.

[3] S. Ding, Z. Chen, T. Zheng, and J. Luo, "RF-Net: A unified meta-learning framework for RF-enabled one-shot human activity recognition," in *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2020, pp. 517–530.

[4] H. Yang, M. Han, X. Li, D. Duan, T. Li, and W. Xu, "iRadar: Synthesizing millimeter-waves from wearable inertial inputs for human gesture sensing," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2025, pp. 1–10.

[5] Y. Cao, S. Zhang, F. Li, Z. Chen, and J. Luo, "hBP-Fi: Contactless blood pressure monitoring via deep-analyzed hemodynamics," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2024, pp. 1211–1220.

[6] C. Wang, L. Xie, W. Wang, Y. Chen, Y. Bu, and S. Lu, "Rf-ECG: Heart rate variability assessment based on cots RFID tag array," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–26, 2018.

[7] P. Hu, W. Li, R. Spolaor, and X. Cheng, "mmEcho: A mmWave-based acoustic eavesdropping method," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2023, pp. 1840–1856.

[8] M. Han, H. Yang, M. Jia, W. Xu, Y. Yang, Z. Huang, J. Luo, X. Cheng, and P. Hu, "Seeing the invisible: Recovering surveillance video with COTS mmWave radar," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 14 592–14 606, 2024.

[9] T. Zheng, Z. Chen, S. Ding, and J. Luo, "Enhancing RF sensing with deep learning: A layered approach," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 70–76, 2021.

[10] M. Han, H. Yang, W. Li, W. Xu, X. Cheng, P. Mohapatra, and P. Hu, "RF sensing security and malicious exploitation: A comprehensive survey," *arXiv preprint arXiv:2504.10969*, 2025.

[11] H. Yang, X. Li, J. Chen, M. Han, and W. Xu, "Poster Abstract: Uncovering mobile user gait patterns through contactless RF channels," in *Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2024, pp. 297–298.

[12] L. Zhao, R. Lyu, Q. Lin, A. Zhou, H. Zhang, H. Ma, J. Wang, C. Shao, and Y. Tang, "mmArrhythmia: Contactless arrhythmia detection via mmWave sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–25, 2024.

[13] N. Imran, J. Zhang, J. Ali, S. Hameed, M. Younas, M. J. Alenazi, F. Niaz *et al.*, "mm-HrtEMO: Non-invasive emotion recognition via heart rate using mm-Wave sensing in diverse scenarios," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2024.

[14] C. Li, M. Liu, and Z. Cao, "WiHF: Enable user identified gesture recognition with WiFi," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2020, pp. 586–595.

[15] M. Raja, V. Ghaderi, and S. Sigg, "WiBot! In-vehicle behaviour and gesture recognition using wireless network edge," in *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 376–387.

[16] H. Ambalkar, X. Wang, and S. Mao, "Adversarial human activity recognition using wi-fi csi," in *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2021, pp. 1–5.

[17] J. Liu, Y. He, C. Xiao, J. Han, L. Cheng, and K. Ren, "Physical-world attack towards wifi-based behavior recognition," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2022, pp. 400–409.

[18] Y. Xie, R. Jiang, X. Guo, Y. Wang, J. Cheng, and Y. Chen, "Universal targeted adversarial attacks against mmWave-based human activity recognition," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2023, pp. 1–10.

[19] Y. Zhou, H. Chen, C. Huang, and Q. Zhang, "WiADv: Practical and robust adversarial attack against WiFi-based gesture recognition system," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–25, 2022.

[20] A. Singha, Z. Bi, T. Li, Y. Chen, and Y. Zhang, "Securing contrastive mmwave-based human activity recognition against adversarial label flipping," in *Proceedings of the ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 2024, pp. 31–41.

[21] T. Zhao, X. Wang, and S. Mao, "Backdoor attacks against deep learning-based massive mimo localization," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 2023, pp. 2796–2801.

[22] R. R. Vennam, I. K. Jain, K. Bansal, J. Orozco, P. Shukla, A. Ranganathan, and D. Bharadia, "mmSpoof: Resilient spoofing of automotive millimeter-wave radars using reflect array," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2023, pp. 1807–1821.

[23] X. Chen, Z. Li, B. Chen, Y. Zhu, C. X. Lu, Z. Peng, F. Lin, W. Xu, K. Ren, and C. Qiao, "MetaWave: Attacking mmWave sensing with meta-material-enhanced tags," in *Proceedings of the Annual Network and Distributed System Security Symposium (NDSS)*, 2023, pp. 1–17.

[24] H. Fang, B. Chen, X. Wang, Z. Wang, and S.-T. Xia, "GIFD: A generative gradient inversion method with feature domain optimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4967–4976.

[25] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the ACM on Asia Conference on Computer and Communications Security (ASIACCS)*, 2017, pp. 506–519.

[26] A. W. Services, "Amazon rekognition," accessed: 2024-12-27. [Online]. Available: https://aws.amazon.com/rekognition

[27] G. Cloud, "Google cloud vision," accessed: 2024-12-27. [Online]. Available: https://cloud.google.com/vision

[28] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019, pp. 225–240.

[29] R. Xiao, J. Liu, J. Han, and K. Ren, "OneFi: One-shot recognition for unseen gesture via COTS WiFi," in *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2021, pp. 206–219.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[32] S. Savazzi, S. Sigg, M. Nicoli, S. Kianoush, F. Le Gall, H. Baqa, and D. Remon, "A cloud-IoT model for reconfigurable radio sensing: The Radio.Sense platform," in *Proceedings of the IEEE World Forum on Internet of Things (WF-IoT)*, 2018, pp. 179–185.

[33] S. Kianoush, M. Raja, S. Savazzi, and S. Sigg, "A cloud-IoT platform for passive radio sensing: Challenges and application case studies," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3624–3636, 2018.

[34] D. Green, C. McIrvin, R. Thaboun, C. Wemlinger, J. Risi, A. Jones, M. Toubeh, and W. Headley, "CLOUD-D RF: Cloud-based distributed radio frequency heterogeneous spectrum sensing," in *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2024, pp. 2395–2400.

[35] X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting explanations for model inversion attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 682–692.

[36] K.-C. Wang, Y. FU, K. Li, A. Khisti, R. Zemel, and A. Makhzani, "Variational model inversion attacks," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 9706–9719.

[37] Z. Ye, W. Luo, M. L. Naseem, X. Yang, Y. Shi, and Y. Jia, "C2FMI: Corse-to-fine black-box model inversion attack," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 3, pp. 1437–1450, 2023.

[38] Y. Liu, W. Zhang, D. Wu, Z. Lin, J. Gu, and W. Wang, "Prediction exposes your face: Black-box model inversion via prediction alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 288–306.

[39] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 253–261.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[42] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[43] H. Yang, M. Han, M. Jia, Z. Sun, P. Hu, Y. Zhang, T. Gu, and W. Xu, "XGait: Cross-modal translation via deep generative sensing for RF-based gait recognition," in *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2023, pp. 43–55.

[44] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[45] D. Lan, C. Xiaoyang, S. Yu, X. Shikun, and X. Meng, "MMRGait-1.0: A radar time-frequency spectrogram dataset for gait recognition under multi-view and multi-wearing conditions," *Journal of Radars*, vol. 12, no. 4, pp. 892–905, 2023.

[46] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2019, pp. 313–325.

[47] M. Han, H. Yang, T. Ni, D. Duan, M. Ruan, Y. Chen, J. Zhang, and W. Xu, "mmSign: mmWave-based few-shot online handwritten signature verification," *ACM Transactions on Sensor Networks*, vol. 20, no. 4, pp. 1–31, 2024.

[48] D. Wang, X. Zhang, K. Wang, L. Wang, X. Fan, and Y. Zhang, "Rdgait: A mmwave based gait user recognition system for complex indoor environments using single-chip radar," *Proceedings of the ACM on*

Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 8, no. 3, pp. 1–31, 2024.

[49] S. Kianoush, M. Raja, S. Savazzi, and S. Sigg, "A cloud-IoT platform for passive radio sensing: Challenges and application case studies," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3624–3636, 2018.

[50] Z. Zhou, F. Wang, and W. Gong, "i-Sample: Augment domain adversarial adaptation models for WiFi-based HAR," *ACM Transactions on Sensor Networks*, vol. 20, no. 2, pp. 1–20, 2024.

[51] Q. Feng, K. Cheng, and C. Duan, "mmWave radar-based unsupervised gesture recognition via image-aligned heterogeneous domain transfer," *IEEE Transactions on Mobile Computing*, pp. 1–17, 2025.

[52] X. Chen and X. Zhang, "RF genesis: Zero-shot generalization of mmWave sensing through simulation-based data synthesis and generative diffusion models," in *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2024, pp. 28–42.

[53] G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. X. Han, "RF-Diffusion: Radio signal generation via time-frequency diffusion," in *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2024, pp. 77–92.

[54] H. Wang, J. Hu, T. Zheng, J. Hu, Z. Chen, H. Jiang, Y. Zheng, and J. Luo, "MuKI-Fi: Multi-person keystroke inference with BFI-enabled Wi-Fi sensing," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9835–9850, 2024.

[55] W. Zhang, H. Dai, D. Xia, Y. Pan, Z. Li, W. Wang, Z. Li, L. Wang, and G. Chen, "mP-Gait: Fine-grained Parkinson's disease gait impairment assessment with robust feature analysis," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 3, pp. 1–31, 2024.

[56] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.

[57] S. Pattanayak and S. A. Ludwig, "Improving data privacy using fuzzy logic and autoencoder neural network," in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2019, pp. 1–6.

[58] A. M. Abdul, A. A. K. Mohammad, P. Venkat Reddy, P. Nuthakki, R. Kancharla, R. Joshi, and N. Kannaiya Raja, "Enhancing security of mobile cloud computing by trust-and role-based access control," *Scientific Programming*, vol. 2022, no. 1, p. 9995023, 2022.

[59] Y. Liu, M. R. Squires, C. R. Taylor, R. J. Walls, and C. A. Shue, "Account lockouts: Characterizing and preventing account denial-of-service attacks," in *Proceedings of the Security and Privacy in Communication Networks (SecureComm)*, 2019, pp. 26–46.

[60] K. A. Tarnowska and A. Patel, "Log-based malicious activity detection using machine and deep learning," *Malware Analysis Using Artificial Intelligence and Deep Learning*, pp. 581–604, 2021.

[61] T. Ni, Z. Sun, M. Han, Y. Xie, G. Lan, Z. Li, T. Gu, and W. Xu, "REHSense: Towards battery-free wireless sensing via radio frequency energy harvesting," in *Proceedings of the International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, 2024, pp. 211–220.

[62] Z. Yang, Y. Zhao, and W. Yan, "Adversarial vulnerability in doppler-based human activity recognition," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[63] C. Li, M. Xu, Y. Du, L. Liu, C. Shi, Y. Wang, H. Liu, and Y. Chen, "Practical adversarial attack on WiFi sensing through unnoticeable communication packet perturbation," in *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2024, pp. 373–387.

[64] L. Mei, R. Liu, Z. Yin, Q. Zhao, W. Jiang, S. Wang, S. Wang, K. Lu, and T. He, "mmSpyVR: Exploiting mmWave radar for penetrating obstacles to uncover privacy vulnerability of virtual reality," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–29, 2024.

[65] A. Adhikari and S. Sur, "Argosleep: Monitoring sleep posture from commodity millimeter-wave devices," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2023, pp. 1–10.

[66] Y. Yang, G. Wang, Z. An, G. Zhang, X. Cheng, and P. Hu, "RF-Parrot: Wireless eavesdropping on wired audio," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2024, pp. 701–710.

[67] S. Zhang, Q. Wang, M. Gan, Z. Cao, and H. Zeng, "RadSee: See your handwriting through walls using FMCW radar," in *Proceedings of the Annual Network and Distributed System Security Symposium (NDSS)*, 2025.

[68] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2014, pp. 17–32.

[69] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2015, pp. 1322–1333.

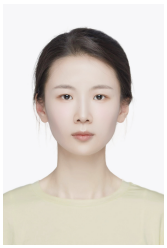
[70] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4829–4837.

[71] Y. Qiu, H. Fang, H. Yu, B. Chen, M. Qiu, and S.-T. Xia, "A closer look at GAN priors: Exploiting intermediate features for enhanced model inversion attacks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 109–126.

[72] R. Liu, D. Wang, Y. Ren, Z. Wang, K. Guo, Q. Qin, and X. Liu, "Unstoppable attack: Label-only model inversion via conditional diffusion model," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3958–3973, 2024.

Mingda Han is currently a Ph.D. student (since 2023) at School of Computer Science and Technology, Shandong University under the supervision of Prof. Pengfei Hu and Prof. Xiuzhen Cheng. Prior to this, he received his Bachelor's degree from the School of Information Science and Engineering, Shandong Normal University in 2016. His research interests include smart sensing and AIoT.

Huanqi Yang is currently a Ph.D. student (since 2021) at Department of Computer Science, City University of Hong Kong. He is supervised by Dr. Weitao Xu. Huanqi Yang received his Bachelor's degree (in 2021) from University of Electronic Science and Technology of China. His research interests lay in smart sensing, IoT security, IoT+AI, wireless network.



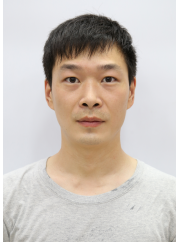
Yanni Yang received her Ph.D. degree in computer science from The Hong Kong Polytechnic University in 2021. Before that, she received the B.E. degree and M.Sc. degree from the Ocean University of China in Qingdao, in 2014 and 2017, respectively. She is currently an assistant professor in the School of Computing Science and Technology at Shandong University. She visited the Media Lab at MIT in 2019 as a visiting student. Her research interests include wireless human sensing, pervasive and mobile computing, and Internet of Things. She has published over 20 papers in top academic conferences and journals.

Guoming Zhang received his Ph.D. degree from the Department of Electrical Engineering of Zhejiang University, supervised by Prof. Wenyuan Xu and Donglian Qi. He received the Master degree in School of Mechanical Engineering of Beijing Institute of Technology, supervised by Prof. Jie Hu. His research interests include IoT security and acoustic communication. He won the best paper awards of ACM CCS 2017, Qshine 2019.

Yetong Cao received the BE degree from Shandong University, in 2017 and the PhD degree in computer science from the Beijing Institute of Technology, in 2023. She is a professor in the School of Computer Science and Technology, Shandong University. She also worked as a visiting student and postdoctoral research fellow at the College of Computing and Data Science at Nanyang Technological University from 2021 to 2024. Her current research interests include mobile computing and ubiquitous computing.

Weitao Xu is an Assistant Professor at the Department of Computer Science at City University of Hong Kong. Before that, he was a Postdoctoral Research Associate at the School of Computer Science and Engineering (CSE) at UNSW from June 2017 to August 2019. He obtained his PhD degree from the University of Queensland in 2017 (advised by Prof. Neil Bergmann and Dr. Wen Hu). His research areas include mobile computing, sensor network, and IoT. He is a member of IEEE.

Xiuzhen Cheng received her MS and Ph.D. degrees in computer science from the University of Minnesota – Twin Cities, in 2000 and 2002, respectively. She was a faculty member at the Department of Computer Science, The George Washington University, from 2002 to 2020. Currently, she is a professor of computer science at Shandong University, Qingdao, China. Her research focuses on blockchain computing, IoT Security, and privacy-aware computing. She is a Fellow of IEEE.



Pengfei Hu is a professor in School of Computer Science and Technology at Shandong University, China. He received Ph.D. in Computer Science from UC Davis. His research interests are in the areas of IoT security, AI security, mobile computing. He has published over 30 papers in premier conferences and journals on these topics, e.g. IEEE S&P, ACM CCS, IEEE INFOCOM, IEEE TMC, etc. He also served as TPC for numerous prestigious conferences and associate editors for IEEE TWC, TDSC, and IoTJ.