Frequency-Selective Adversarial Attack Against Deep Learning-Based Wireless Signal Classifiers

Da Ke, Xiang Wang, and Zhitao Huang

***Abstract*— Although Deep learning (DL) provides state-of-art results for most spectrum sensing tasks, it is vulnerable to adversarial examples. Based on this phenomenon, we consider a noncooperative communication scenario where an intruder tries to recognize the modulation type of the intercepted signal. Specifically, this paper aims to minimize the intruder's accuracy while guaranteeing that the intended receiver can still recover the underlying message with the highest reliability. This process is implemented by adding adversarial perturbations to the channel input symbols at the encoder. In image classification, the perturbation is limited to be imperceptible to a human observer by minimizing the ℓ_p norm, while in this work, we enriched the connotation of adversarial examples, and first proposed that the imperceptibility of adversarial examples in the field of wireless signals is the imperceptibility of filters. Based on this perspective, we optimized the model of adversarial examples and constrained the adversarial perturbation to a narrow frequency band so that filters cannot filter it out. We also define a new set of metrics to describe the imperceptibility of the wireless signal adversarial example. The simulation results demonstrate the viability of our approach in securing wireless communication against state-of-the-art DL-based intruders while minimizing communication performance reduction.**

***Index Terms*— Secure communication, deep learning, adversarial attacks, modulation classification.**

I. Introduction

ENSURING the security of wireless communication links is as important as improving their efficiency and reliability for military, commercial and civilian communication systems. The standard method of securing communications is to encrypt transmitted data. However, encryption may not always provide full security (e.g., side-channel attacks) or strong encryption due to complexity limitations (e.g., IoT devices). To further enhance security, encryption can also be complemented with other techniques that can even prevent an adversary from recovering encrypted bits.

An adversary implements its invasion on a wireless communication link in four steps [1]: 1) tuning the frequency of the transmitted signal; 2) detecting the presence or absence of a signal; 3) intercepting the signal by extracting signal features; 4) demodulating the signal using the extracted features and obtaining a binary data stream. Preventing any of these steps can significantly enhance the security of the communication link. While encryption focuses on protecting the demodulated bitstream, physical layer security [2] address the fourth step by minimizing the mutual information available to intruders. There has also been a strong interest in preventing the second step through secret communications [3]. In this work, we focus on the third step, which aims to prevent the adversary from detecting the modulation scheme used for communication.

Modulation recognition is the step between signal detection and demodulation in communication links, and it plays a significant role in data transmission and detection and jamming of unwanted signals in military communications and other applications [4]. Recently, deep learning (DL) has significantly contributed to modulation recognition. Indeed, DL-based methods extract features adaptively [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], surpassing the accuracy of conventional modulation classifiers utilizing likelihood function or feature-based representations [4].

The purpose of this paper is to prevent intruders employing state-of-the-art modulation detectors from successfully identifying the modulation scheme used. If an intruder is unable to recognize the modulation type, it is not possible to decode the underlying information or use modulation-dependent jamming attacks to block communications. In order to achieve this, the transmission signal needs to be modified. The main challenge here is to ensure that the intended receiver of the (modified) transmission signal still reliably receives the underlying information while preventing intruders from detecting the modulation scheme used. Otherwise, the cost of decreasing the accuracy of modulation-detecting intruders is that the bit error rate of the intended receiver may increase significantly. We assume that the intended receiver is blind to the modifications made by the transmitter, so the transmitter's goal is to make the smallest possible modifications to the transmitted signal that are sufficient to fool the intruder but do not exceed error-correcting capabilities of the intended receiver.

Introducing small changes in the modulation scheme that can fool the intruder is similar to adversarial attacks on classifiers, especially on deep neural networks (DNNs) [15], [16]. In the literature, adversarial attacks are considered mainly in the area of image classification, where they pose a security risk by exposing the vulnerability of the classifiers to very small changes in the inputs, which are imperceptible to humans but can lead to wrong decisions. In contrast, we utilize the similar approach here to protect communication links from intruders that employ DNNs methods for interception.

Manuscript received 26 April 2023; revised 23 October 2023 and 10 December 2023; accepted 10 December 2023. Date of publication 10 January 2024; date of current version 2 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 62271494. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Oliver Kosut. *(Corresponding author: Xiang Wang.)*

The authors are with the National University of Defense Technology, Kaifu, Changsha, Hunan 410073, China (e-mail: 1747884404@qq.com; christopherwx@163.com; huangzhitao@nudt.edu.cn).

Digital Object Identifier 10.1109/TIFS.2024.3352423

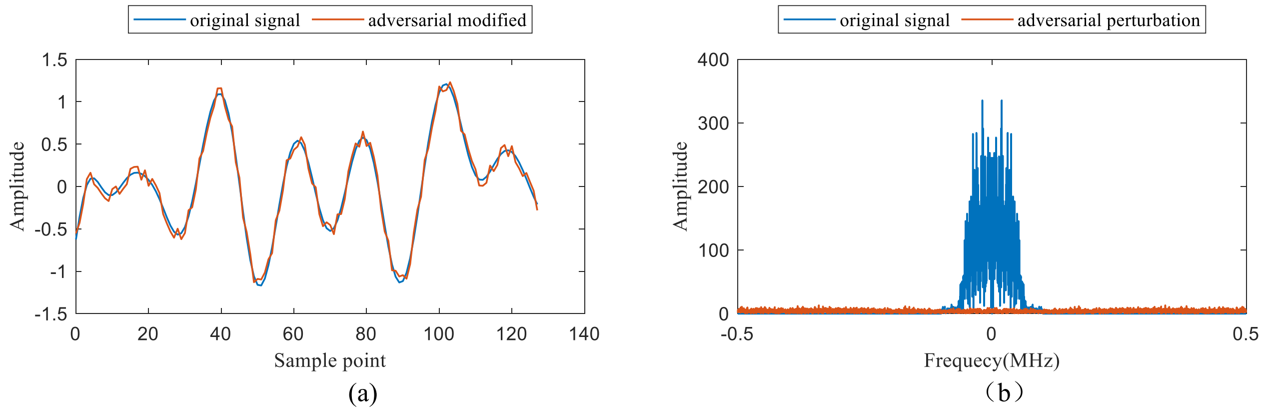

Fig. 1. (a) Time-domain waveform of a communication signal before and after being modified by an adversarial attack (we present only 128 points of the signal to enhance presentation) and (b) Spectrum of a communication signal and the adversarial perturbation added.

Existing literature suggests various methods exploiting adversarial attacks to defend a communication link against an intruder that employs DL classification methods for interception [17], [18], [19], [20], [21], [22], [23], [24], [25]. In [17], an adversarial attack for a DL-based modulation classifier has been proposed where the adversary assumes the availability of noisy symbols received at the modulation classifier for generating the adversarial attack, which makes it impractical and limited in scope. Similarly, [18] develops an autoencoder that the receiver uses to preprocess the modified signal. In this paper, the authors reveal that pre-training DL-based classifiers in the radio frequency (RF) domain using an autoencoder can mitigate the deceiving effect of adversarial examples. Moreover, [19] considers the impact of the adversarial examples on the RF domain and demonstrated that adversarial defense could improve the robustness of DL-based modulation classifiers. In [1], the authors suggested an adversarial attack method that reduces the modulation classification accuracy at the intruder while maintaining a low bit error rate (BER) at the legitimate receiver. In [20], the authors verified the effectiveness of various adversarial attacks by reconstructing the waveforms in a modulation classification scenario. Literature [21] investigates an adversarial attack scenario in a real wireless channel and proposes a maximum received perturbation power attack (MRPP). This method achieves the state-of-art attack performance against a spectrum-sensing deep learning model.

However, the adversarial examples generated by existing adversarial attack methods will introduce more high-frequency components due to the abrupt changes in the time domain. For example, Fig. 1(a) illustrates a comparison of the time-domain signal before and after being modified by Fast Gradient Sign Method (FGSM). The modified signal has more high-frequency perturbations compared to the original signal. Moreover, Fig. 1(b) illustrates the spectrum of the original signal and the adversarial perturbation, revealing that the energy of the original signal is concentrated only in a narrow frequency band. Opposing, the energy of the adversarial perturbation is distributed over the entire frequency band. In communication systems, both transmitters and receivers use narrowband filters to filter out interference or noise outside the signal band from the transmitted/received signals. This is helpful for communication systems to obtain clean signals. Nevertheless, it is challenging for most adversarial perturbations to access the intruder and prevent it from recognizing the modulation scheme.

To address this problem, a more realistic scenario must be considered. This paper presents a more realistic wireless attack built upon adversarial DL by accounting for filter effects while designing the algorithm for adversarial attacks. Motivated by the on-manifold adversarial attack [22], [23], we add a constraint in the frequency domain to the calculated adversarial example and thus restrict the generated adversarial perturbations in the signal's frequency band.

In summary, this paper's main contributions are as follows:

1. We consider a more realistic scenario of a wireless signal adversarial attack and establish a novel adversarial attack framework. Based on this framework we design 2 novel adversarial attack methods for the spectrum sensing task.

2. We define a new set of evaluation metrics to measure the "imperceptibility" of adversarial perturbations in the wireless signal scenario. The "imperceptibility" of the wireless signal field should not be easily filtered rather than just not be imperceptible to the human eye visual system. To the best of our knowledge, this work is groundbreaking.

3. Further analysis reveals that our method finds an on-manifold adversarial perturbation in the frequency subspace, which facilitates the study of the interpretability of the adversarial example.

4. Extensive experiments reveal that our approach works well in various settings compared to existing methods.

The remainder of this paper is organized as follows. Section II describes the problem model, section III describes the idealized adversarial example method and our improvement, section IV presents the experimental results, and Section IV discusses the findings and proposes future research directions.

II. Problem Analysis

We consider a wireless communication system that comprises a transmitter, a receiver, and an intruder as depicted in **Fig. 2** [24]. To simplify the problem, we assume all nodes have a single antenna operating on a Gaussian channel. The intruder classifies the intercepted signals using a deep neural network (DNN), aiming to determine the modulation type the transmitter uses. In the meantime, the transmitter transmits a signal with adversarial perturbation over the air to fool the intruder and impose it to make errors during the modulation classification.

For a transmitter, a binary information sequence $\mathbf{w} \in \{0,1\}^m$ is mapped into a sequence of channel symbols, $\mathbf{x} \in \mathrm{C}^n$, and then the modulated signal $\mathbf{x} = M_s(\mathbf{w})$ is obtained, where $s \in \mathrm{S}$ is the employed modulation scheme with S denoting the finite set of available modulation schemes, and $M_s : \{0,1\}^m \to \mathrm{C}^n$ denotes the whole modulation function. The modulated signal is transmitted into the physical world after undergoing frequency mixing, narrowband filtering, and power amplification in sequence. The signal $\mathbf{x}$ is sent over a noisy channel, which is assumed to be an additive white Gaussian noise (AWGN) channel for simplicity. The intercepted baseband signal $\mathbf{y}_1$ and $\mathbf{y}_2$, received by the intended receiver and the intruder, respectively, are expressed as:

$$\mathbf{y}_i = M_s(\mathbf{w}) + \mathbf{z}_i = \mathbf{x} + \mathbf{z}_i, \ \ i = 1, 2 \tag{1}$$

where $\mathbf{z}_i$ is the AWGN with zero mean and unit variance σ_i^2. As mentioned in Section I, both transmitters and receivers would filter the signal (dropping out the noise outside the signal's band), So the finally intercepted signal sequence of the intruder is expressed as:

$$\mathbf{y}_2 = H(\mathbf{x} + \mathbf{z}_2), \tag{2}$$

where $H(\bullet)$ is the filter used to drop out the noise. Such a filter includes but is not limited to band-pass and low-pass filters.

The intruder aims to recognize the transmitter's modulation type based on its intercepted noisy output $\mathbf{y}_2$. On the other hand, the transmitter wants to communicate without its modulation scheme being correctly detected by the intruder while keeping the transmitted signal as uncorrupted as possible.

Formally, the intruder aims to determine the transmitter's modulation type, which can be formulated as a classification problem where the label $s \in \mathrm{S}$ is the employed modulation scheme. The classifier assigns $\mathbf{y}_2$ to the predicted label:

$$\hat{s} = \arg\max_{s \in \mathrm{S}} f_{\boldsymbol{\theta}}(\mathbf{y}_2), \tag{3}$$

where $f_{\boldsymbol{\theta}} : \mathrm{C}^n \times \mathrm{S} \to \mathrm{R}$ is a score function parametrized by $\boldsymbol{\theta} \in \mathrm{R}^d$, which assigns a score to each possible class $s \in \mathrm{S}$. For the purpose of notation simplification, we denote the resulting class label by $\hat{s} = f_{\boldsymbol{\theta}}(\mathbf{y}_2)$. The goal of the intruder is to maximize the probability $\Pr(s = \hat{s})$ of correctly detecting the modulation scheme, which is considered as the intruder's success probability. State-of-the-art modulation classification schemes consider $f_{\boldsymbol{\theta}}$ is a DNN classifier [5], [6], [7], [8], [9], $\boldsymbol{\theta}$ denotes the neural network weights, and $f_{\boldsymbol{\theta}}(\mathbf{y})$ are the so-called logit values for the class labels $s \in \mathrm{S}$.

III. Methodology

A. Adversarial Attack in an Idealized Scenario

This paper modifies the encoding processes M_s so that, given a modulation type $s \in \mathrm{S}$, the new encoding method M_s'

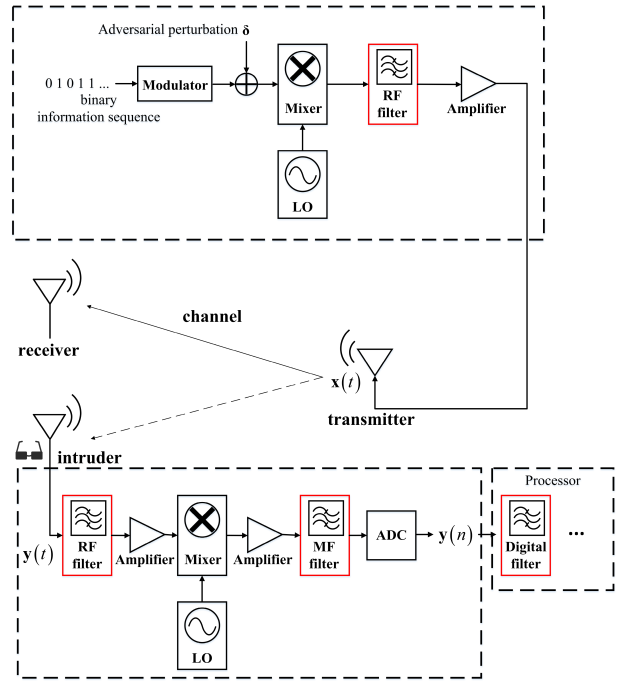

Fig. 2. System model where a transmitted signal is intruded.

ensures that the intruder's success probability becomes smaller and the receiver's signal is not modified substantially. This idea is motivated by adversarial attacks on image classification [15], where modifying images makes the modification imperceptible to a human observer, but advanced image classifiers perform significantly poor. By applying the same idea to the investigated problem, we aim to find defensive modulation schemes $M_s'(\mathbf{w}) \approx M_s(\mathbf{w})$, but the intruder misclassifies the new intercepted signal $\mathbf{y}_2' = M_s'(\mathbf{w}) + \mathbf{z}_2$ with higher probability.

Following the idea of adversarial attacks on image classifiers directly, [10] proposes an idealized yet impractical adversarial attack mechanism that modifies a correctly classified channel output sequence $\mathbf{y}_2$ (i.e., for which $s = f_{\boldsymbol{\theta}}(\mathbf{y}_2)$) with a perturbation $\boldsymbol{\delta} \in \mathrm{C}^n$ such that $f_{\boldsymbol{\theta}}(\mathbf{y}_2 + \boldsymbol{\delta}) \neq f_{\boldsymbol{\theta}}(\mathbf{y}_2)$, the true label while imposing the restriction $\|\boldsymbol{\delta}\|_2 \leqslant \epsilon$ for some small positive constant ϵ. For a DNN-based classifier $f_{\boldsymbol{\theta}}$, the problem above can be formulated as a constrained optimization problem:

$$\max L(\boldsymbol{\theta}, \mathbf{y}_2 + \boldsymbol{\delta}, s) \ \ s.t. \ \|\boldsymbol{\delta}\|_p \leqslant \epsilon, \tag{4}$$

where $L(\bullet)$ is a loss function often used to train the classifier f_θ. $L(\bullet)$ is typically defined as a cross-entropy function in classification problems. There are several methods to solve approximately, i.e., when $p = \infty$ (4) can be solved using the fast gradient sign method (FGSM) [15], [26] while considering a one-step adversarial attack:

$$\boldsymbol{\delta} = \epsilon sign\left(\nabla_{\mathbf{y}} L(\boldsymbol{\theta}, \mathbf{y}, s)\right), \tag{5}$$

where ∇ denotes the gradient operator. Moreover, the projected gradient descent (PGD) [27] attack can iteratively achieve state-of-the-art attack performance. Starting from $\mathbf{y}^0 = \mathbf{y}$, at each iteration t it calculates:

$$\mathbf{y}^t = \prod\nolimits_{\mathbf{B}_\epsilon(\mathbf{y})}\left(\mathbf{y}^{t-1} + \beta sign\left(\nabla_{\mathbf{y}} L\left(\boldsymbol{\theta}, \mathbf{y}^{t-1}, s\right)\right)\right), \tag{6}$$

where β denotes the step size, $sign(\bullet)$ denotes the sign operation, $\prod_{\mathbf{B}_\epsilon(\mathbf{y})}(\bullet)$ is the Euclidean projection operator to the ℓ_2-ball $\mathbf{B}_\epsilon(\mathbf{y})$ of radiusϵ centered at $\mathbf{y}$. The attack is typically run for a specified number of steps, which depends on the computational resources. In practice $\mathbf{y}$ is more likely to be a successful adversarial example for larger values of t. Note that this formulation assumes we can access the intruder's logit function $f_{\boldsymbol{\theta}}$. These methods are called white-box attacks. If $f_{\boldsymbol{\theta}}$ is unknown, one can create adversarial examples against another classifier $f'_{\boldsymbol{\theta}}$, hoping it will also work against the targeted model $f_{\boldsymbol{\theta}}$. Such methods are called black-box attacks. In this paper, we only consider white-box attacks against intruders.

B. Frequency-selective Adversarial Attack

As mentioned above, directly applying an idealized adversarial attack to our problem is practically infeasible as the intruder can filter the intercepted signal $\mathbf{y}_2$, prohibiting most adversarial perturbation power from accessing the intruder. Thus, the newly received signal model is expressed as:

$$\mathbf{y}'_2 = H(\mathbf{x} + \mathbf{z}_2 + \boldsymbol{\delta}), \tag{7}$$

Accordingly, the new modulation scheme aims to find defensive modulation schemes that allow the intruder to misclassify the newly received signal $\mathbf{y}'_2 = H(\mathbf{x} + \mathbf{z}_2 + \boldsymbol{\delta})$ as much as possible. Our scenario involves a new connotation regarding the constraint that modifications are imperceptible to human observers. Specifically, we require the perturbation power $\boldsymbol{\delta}$ to be as small as possible and not easily detected by the filter. This ensures that the adversarial attack will not affect the intended receiver and will access successfully on the intruder.

Following the improved idea of adversarial attacks, we propose a more realistic mechanism. The optimization problem (4) can be improved as:

$$\max_{\boldsymbol{\delta}} L(\boldsymbol{\theta}, H(\mathbf{y}_2 + \boldsymbol{\delta}), s) \;\; s.t. \;\; \|\boldsymbol{\delta}\|_p \leqslant \epsilon, \tag{8}$$

where $H(\bullet)$ is a frequency filter. With this improvement, we expect to restrict the search space in the adversarial direction to the subspace of the signal.

Next, we solve (8) using the approximation technique. Specifically, we assume that $H(\bullet)$ is a linear transformation (later, we will implement $H(\bullet)$ using the linear method) so $H(\mathbf{y}_2 + \boldsymbol{\delta})$ can be expanded as $H(\mathbf{y}_2) + H(\boldsymbol{\delta}) = \tilde{\mathbf{y}} + \tilde{\boldsymbol{\delta}}$. Then (8) can be rewritten as [1]:

$$\max_{\boldsymbol{\delta}} L\left(\tilde{\mathbf{y}} + \tilde{\boldsymbol{\delta}}\right) \;\; s.t. \;\; \|\boldsymbol{\delta}\|_p \leqslant \epsilon \quad . \tag{9}$$

In practice, $\tilde{\mathbf{y}}$ can be easily obtained by filtering $\mathbf{y}$. So, we can directly approximate the loss function by its first order Taylor expansion at point $\tilde{\mathbf{y}}$. The problem in (9) then becomes:

$$\max_{\delta} L(\tilde{\mathbf{y}}) + \nabla_{\tilde{y}} L(\tilde{\mathbf{y}})^{\mathrm{T}} \tilde{\boldsymbol{\delta}} \text{ s.t. } \|\boldsymbol{\delta}\|_p \leq \epsilon. \tag{10}$$

[1] $L(\boldsymbol{\theta}, \mathbf{x}, s)$ might be shorthanded as $L(\mathbf{x})$ or L in the following text.

This problem is trivially linear and hence convex w.r.t. $\boldsymbol{\delta}$. We obtain a closed-form solution using the Lagrangian multiplier method (see the Appendix), yielding:

$$\boldsymbol{\delta} = \epsilon sign\left(J_H \cdot \nabla_{\mathbf{y}} L\right)\left(\frac{\left|J_H \cdot \nabla_{\mathbf{y}} L\right|}{\left\|J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right\|_{p^*}}\right)^{\frac{1}{p-1}}, \tag{11}$$

where p^* is the dual of p, i.e., $\frac{1}{p^*} + \frac{1}{p} = 1$, and J_H is the Jacobian of $H(\bullet)$. When $p = \infty$, (11) reduces to:

$$\boldsymbol{\delta} = \epsilon sign\left(J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right), \tag{12}$$

which is similar to (5).

Discrete Fourier transform (DFT). DFT decomposes a signal into complex index components and represents a natural signal in the frequency space. More precisely, given a discrete signal $\mathbf{y} \in \mathrm{C}^n$, the $\mathbf{Y} = \mathrm{DFT}(\mathbf{y})$ is:

$$\mathbf{Y}_k = \sum_{i=0}^{n-1} \mathbf{y}_i e^{-j(2\pi/n)ki}, \;\; 0 \leqslant k \leqslant n-1, \tag{13}$$

where $\phi(i) = e^{-j(2\pi/N)ki}, 0 \leqslant k \leqslant n-1$ are a set of primary functions. DFT is invertible, with the inverse $\mathbf{y} = \mathrm{IDFT}(\mathbf{Y})$,

$$\mathbf{y}_i = \frac{1}{n} \sum_{k=0}^{n} \mathbf{Y}_k e^{j(2\pi/n)ki} \;\; , \;\; 0 \leqslant i \leqslant n-1. \tag{14}$$

To calculate $\boldsymbol{\delta}$ in (8), we employ a filter $H(\bullet)$ in the DFT process involving a set of weights $\mathbf{h} \in \mathrm{R}^n$:

$$h_i = \begin{cases} 1 & n_l \leqslant i \leqslant n_h \\ 0 & otherwise \end{cases} \quad i = 1, 2, \ldots, n, \tag{15}$$

where n_l and n_h are related to the sampling rate Fs of the signal and the number of DFT points N_{FFT}. Specifically, let f_l and f_h be the minimum and maximum frequencies of the signal, respectively. Then n_l and n_h are $\frac{f_l}{(Fs/2)} \cdot N_{FFT}$ and $\frac{f_h}{(Fs/2)} \cdot N_{FFT}$. Next, we can implement

$$H(\mathbf{y}) = \mathrm{IDFT}(\mathbf{h} \cdot \mathrm{DFT}(\mathbf{y})). \tag{16}$$

Since DFT is a linear transformation, IDFT is also a linear transformation. So $H(\bullet)$ is a linear transformation proving that our assumptions in holds. Then, we compute J_H.

First, through the chain rule, we have:

$$J_H = J_{\mathrm{IDFT}} \cdot \mathbf{h} \cdot J_{\mathrm{DFT}}, \tag{17}$$

where J_{DFT} and J_{IDFT} are the Jacobian of DFT and IDFT respectively. According to (13), the DFT operation can be written in the form of matrix multiplication:

$$\mathbf{Y} = \mathbf{F} \cdot \mathbf{y}$$
$$= \begin{bmatrix} e^{-j\frac{2\pi}{n}\times 0\times 0} & e^{-j\frac{2\pi}{n}\times 0\times 1} & \cdots & e^{-j\frac{2\pi}{n}\times 0\times(n-1)} \\ e^{-j\frac{2\pi}{n}\times 1\times 0} & e^{-j\frac{2\pi}{n}\times 1\times 1} & \cdots & e^{-j\frac{2\pi}{n}\times 1\times(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-j\frac{2\pi}{n}\times(n-1)\times 0} & e^{-j\frac{2\pi}{n}\times(n-1)\times 1} & \cdots & e^{-j\frac{2\pi}{n}\times(n-1)\times(n-1)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{n-1} \end{bmatrix}. \tag{18}$$

Matrix $\mathbf{F}$ denotes the matrix form of the Fourier transform and it is a symmetric matrix, so

$$J_{\mathrm{DFT}} = \mathbf{F}^{\mathrm{T}} = \mathbf{F}. \tag{19}$$

Similarly, we can get $J_{\mathrm{IDFT}} = \mathbf{F}_I$, where $\mathbf{F}_I$ is the matrix of IDFT. In summary, $J_H(\boldsymbol{\delta}) = H(\boldsymbol{\delta})$ and can be rewritten as:

$$\boldsymbol{\delta} = \epsilon sign\left(H\left(\nabla_{\tilde{\mathbf{y}}} L\right)\right)\left(\frac{\left|H\left(\nabla_{\tilde{\mathbf{y}}} L\right)\right|}{\left\|H\left(\nabla_{\tilde{\mathbf{y}}} L\right)\right\|_{p^*}}\right)^{\frac{1}{p-1}}. \tag{20}$$

The physical meaning of (20) can be expressed as filtering the gradient to obtain the in-band subspace of the gradient. It can be observed that the proposed algorithm in this paper is achieved by constraining the frequency of the gradient vector during its computation. Therefore, the proposed algorithm can be easily embedded into existing gradient-based adversarial attack frameworks, such as PGD and C&W. Accordingly, we refer to them as Frequency Selective PGD(FS-PGD) and Frequency Selective C&W(FS-C&W) algorithms.

The value of p has been discussed in [26] and will not be repeated in this paper. Thus, we directly set $p = \infty$ and design a multi-step iterative attack, as presented in Algorithm 1.

Algorithm 1 Frequency-selective adversarial perturbation

Input: Original signal example $\mathbf{y}$; ground-truth label s; loss function L; The filter coefficients $\mathbf{h}$.
Input: The perturbation size ϵ and the iteration N.
Output: An adversarial example $\mathbf{y}^*$, where $\|\mathbf{y}^* - \mathbf{y}\|_\infty \leqslant \epsilon$.

```
y*_0 = h · y   β = ε/N
for iter = 0 : N − 1:
    Calculate the gradient ∇_{h·y} L;
    Compute h · ∇_{h·y} L;
    Update y*_{iter+1} = y*_{iter} + β sign(h · ∇_{h·y} L)
end for
return y* = y*_N
```

C. Imperceptibility

In the realm of adversarial examples, imperceptibility is a crucial metric for evaluating their performance. In particular, in the domain of wireless signals, the imperceptibility of adversarial examples should be manifested as being undetectable by filters. This attribute is of paramount importance in assessing the efficacy of adversarial attacks on wireless communication systems, where the ability to evade detection is a critical factor in their success.

For approximating the "imperceptibility", existing work [26]would like to use ℓ_p norms to quantify average variations of the basic structure information between the original and adversarial examples. Several works on the topic of wireless signal adversarial examples have established the perturbation-signal power ratio (PSR) [1], [17] as a metric for quantifying the imperceptibility of adversarial examples:

$$\mathrm{PSR} = 10\lg\left(\frac{P_{\boldsymbol{\delta}}}{P_{\mathbf{y}}}\right), \tag{21}$$

where $P_{\boldsymbol{\delta}}$ and $P_{\mathbf{y}}$ represent the power of the perturbation and the signal, respectively. This metric describes the relative energy of the perturbation to that of the signal.

From a frequency domain perspective, both the ℓ_p norm and PSR metrics describe the relative energy of the perturbation and signal across the entire frequency band. However, when a signal contaminated with adversarial perturbation passes through a filter, the perturbation and signal components outside the filter bandwidth become irrelevant. In this context, novel evaluation metrics are required to measure the relative energy of the perturbation and signal within the filter bandwidth, while also assessing the imperceptibility of the adversarial perturbation to the filter.

Based on this perspective, we define in-band perturbation-to-signal power ratio (IB-PSR) to quantify average variations of the basic structure information between the original and adversarial examples. The IB-PSR is defined as:

$$\text{IB- PSR} = 10\lg\left(\frac{P_{H(b)}}{P_y}\right). \tag{22}$$

Correspondingly, we introduce the concept of filter loss Δ PSR to measure imperceptibility in our study. The Δ PSR is the difference between PSR and IB-PSR.

IV. Experiments

In this part, we perform extensive experiments to evaluate our proposed FS-PGD and FS-C&W in attacking DL-based modulation classifier.

A. Experimental Setup

1) Datasets: We assume that the binary source data is generated independently and uniformly at random. Ten standard baseband modulation schemes are considered: 2ASK, 2FSK, 8PSK, BPSK, QPSK, OQPSK, pi/4-QPSK, 16QAM, 32QAM, 64QAM. The square root cosine filter is used for pulse shaping of the modulated data with a roll-off factor of 0.3, a code rate of 0.1 MHz, and sampling rates of 1MHz. In particular, the frequency offset of the 2FSK modulated signal is 0.75 times the code rate. The modulated data is sent through an AWGN channel with a signal-to-noise ratio (SNR) varying between −12 dB and 20 dB, with a small Doppler bias.

The intruder has to estimate the modulation scheme after receiving 4096 complex I/Q (in-phase /quadrature) channel symbols. All previous work used 128-point complex I/Q [17], [18], [19], [20], [21], which we believe is unreasonable. This is because the number of data points is too short to effectively traverse all the code element states of the modulated signal. For example, 64QAM theoretically has 64 states, and the theoretical minimum sampling multiple is 2 times (which is far from enough in practice), and thus use 10 times sampling. Then at least 640 I/Q are theoretically required to traverse all the code element states. Considering the balance of algorithm performance and computational resources, we chose 4096 I/Q as the input to the model.

We generated 2720 blocks of data for each modulation type, each containing 4096 I/Q sampling data. We divided the data into a training, validation, and test set according to a 7:2:1

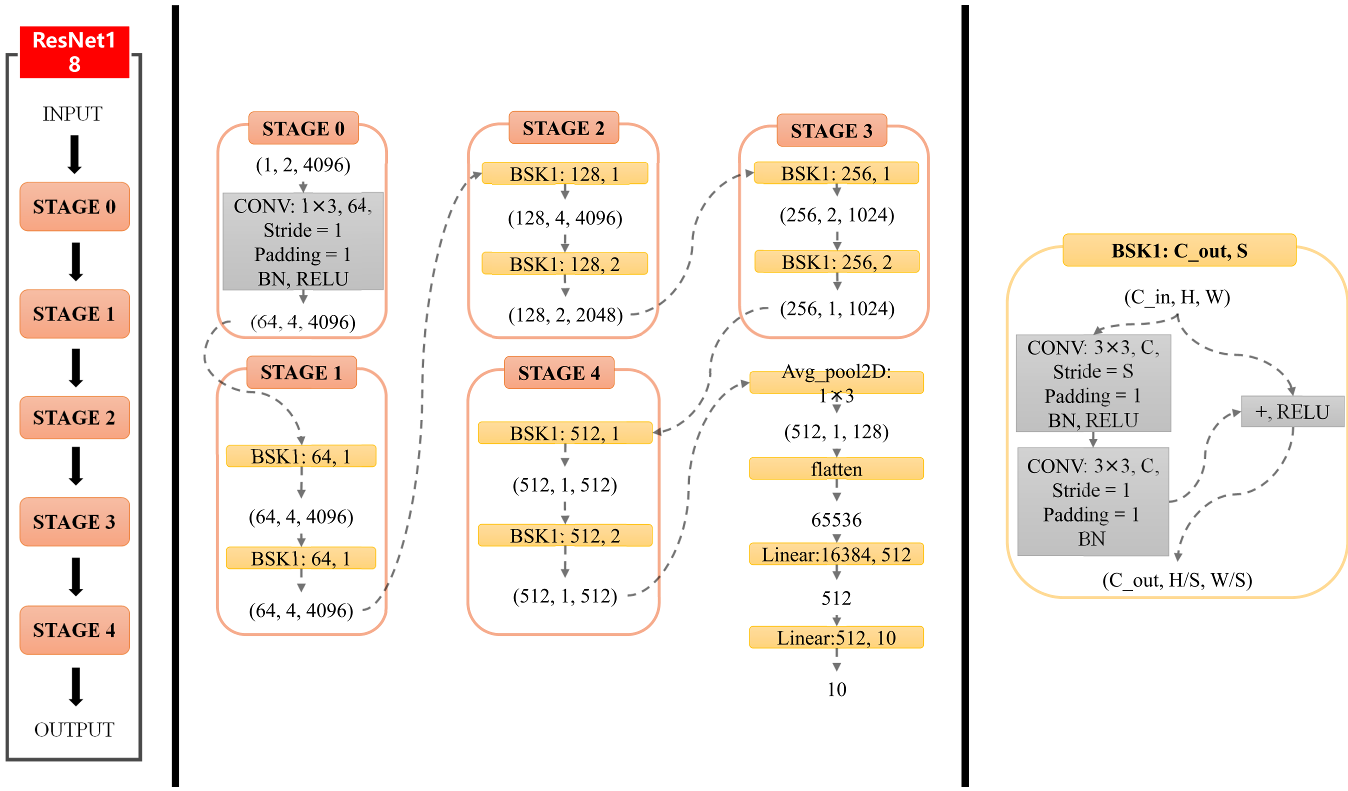

Fig. 3. The structure of fine-tuned Resnet18.

TABLE I
RESULTS OF THE FR BY SEVEN ATTACK APPROACHES WITH PSR= −20dB

Attack	**FR (Before filtering)**	**FR (After filtering)**	**IB-PSR**	**Δ FR**	**Δ PSR**
WGN	3.86%	1.20%	-27dB	-2.66%	-7dB
FGSM	**100%**	42.20%	-29.15dB	-57.8%	-9.15dB
PGD	**100%**	43.55%	-27.39dB	-56.45%	-7.39dB
C&W	**100%**	47.32%	-27.67dB	-52.68%	-7.67dB
MRPP	94.23%	88.66%	-21.01dB	-5.57%	-1.01dB
FS-PGD	99.98%	**98.18%**	**-20.42dB**	**-1.8%**	**-0.42dB**
FS-C&W	99.96%	**98.41%**	**-20.42dB**	**-1.55%**	**-0.42dB**

ratio. We fine-tuned the Resnet18 network as a classifier to fit our data dimensionality, and the detailed network structure is illustrated in **Fig. 3**.

2) Implementation Details: All experiments were performed on an NVIDIA GeForce GTX 3090. Each of the attack methods was implemented in PyTorch. We use Adam to optimize the target model and the ReLU as the activation function in all layers. In the experiments, we utilized the categorical-cross entropy loss functions. The training epoch was set to 200, and we used the early-stopping strategy with a patience of 20.

3) Evaluation Metrics: The performance evaluation relied on the fooling rate (FR), which is defined as follows:

$$\mathrm{FR} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}\left(f\left(\mathbf{y}_i + \boldsymbol{\delta}\right) \neq f\left(\mathbf{y}_i\right)\right)$$
$$where \;\; f\left(\mathbf{y}_i\right) = true \;\; label, \tag{23}$$

where $\mathbf{I}(\bullet)$ is the indicator function, which means that when the event occurs, it takes the value of 1, and if it does not occur, it takes the value of 0. The implication of FR is that the attack is considered successful if the classifier misclassifies a sample that has been correctly identified by adding an adversarial perturbation.

We illustrate the performance of our methods through the following experiments:

- Compare the white-box attack performance of our methods with that of white gaussian noise(WGN), FGSM [15], [20], PGD [20], [27], C&W [28] and MRPP [21]. While FGSM, PGD and C&W methods were first used for image recognition tasks, literature [20] shows that the above methods also perform well in spectrum sensing tasks. MRPP, on the other hand, is an adversarial attack method designed specifically for spectrum sensing tasks, which solves the problem of poor performance of tiny

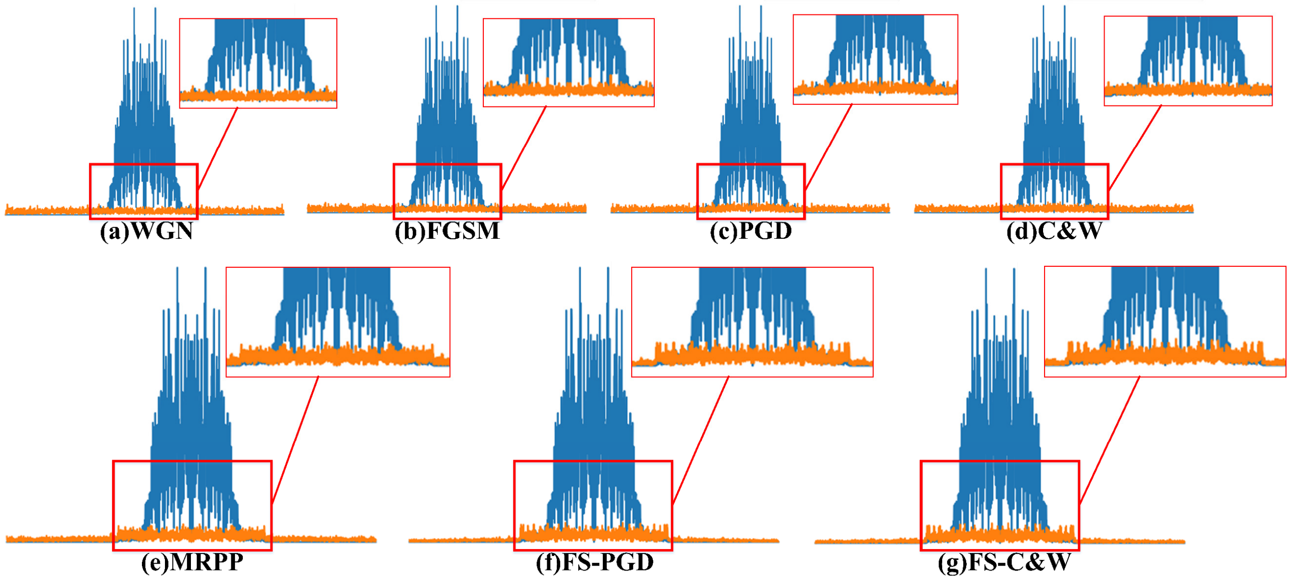

Fig. 4. A comparison of the spectrum of a 16QAM modulated signal with a SNR of 20dB under 7 different adversarial attacks. (The blue line represents the spectrum of the original signal, while the orange line represents the spectrum of the adversarial perturbation.)

adversarial perturbations in the real physical world, and is currently the state-of-the-art method for spectrum sensing tasks. The power of adversarial perturbation was uniformly set to PSR=−20dB, and the iteration number was set to 10 for iterative algorithms such as PGD and C&W. We will also visualize the distribution of the perturbations generated by the above method in the frequency domain. As mentioned above, the frequency offset of the 2FSK signal is 0.75 times the code rate, so the actual bandwidth occupied by the 2FSK signal is 1.75 times the code rate. Therefore, to avoid destroying the original signal information, the bandwidth for all experiments is set to at least 0.2MHz.

- To evaluate the robustness of our algorithm to adversarial defense models, we compare its performance with other white-box attacks such as WGN, FGSM, PGD, C&W and MRPP under adversarial defense conditions.
- Test our methods' robustness when the bandwidth and types of filters are mismatched (i.e., the bandwidth of the constrained perturbation does not match the filter bandwidth used by the actual receiver). This trial evaluates robustness of our methods against filter mismatch.

B. Attack Performance

In order to evaluate the attack performance of our proposed FS-PGD and FS-C&W adversarial attack methods, we compare them with WGN as well as four state-of-the-art adversarial attack methods including FGSM, PGD, C&W and MRPP. This section evaluates the adversarial power and imperceptibility of the examples generated by different approaches in a white-box scenario, where the knowledge of the target system is fully accessible. Tab. I reports the performance of 7 attack approaches over 5 different metrics. Without considering filtering, FGSM, PGD and C&W achieve FR of 100%, and the FR of MRPP is 94.23%. Our proposed FS-PGD and FS-C&W methods also achieve FR close to the optimal comparison algorithms without the use of filter in the classifier, reaching 99.98% and 99.96%, respectively. When the effect of filtering exists, our method has an obvious advantage presenting the smallest filtering loss ΔPSR (−0.42dB) and the highest FR (98.18% and 98.41%). In contrast, the comparison algorithms only realize FRs of 1.20%, 42.20%, 43.55%, 47.32%, 88.66%, respectively. In addition to the two methods proposed in this paper. The MRPP is better adapted to the real physical world in the presence of filters than existing adversarial attack methods.

This infers that, compared to the comparison algorithms, our attack implements an adversarial perturbation that avoids being dropped out by the filter as much as possible, thus achieving a more effective adversarial attack for intruders. This is numerically demonstrated as our method has the lowest ΔPSR with −0.42dB. Such experimental results also show that our proposed methods are more adaptable to the filter effects at the transmitter/receiver than the state-of-the-art adversarial attack method (MRPP) for spectrum sensing tasks.

Further, we visualize the spectrum and time-domain waveform of the signal before and after the attack with a 16QAM modulated signal with 20dB SNR, as illustrated in **Fig. 4** and **Fig. 5**. **Fig. 4 (b)**∼(d) demonstrate that the spectrum of the adversarial perturbations generated by the three methods FGSM, PGD and C&W are closer to that of the WGN (**Fig. 4(a)**) and is characterized by a uniform distribution of the energy over the entire frequency band. However, according to **Fig. 4(e)**∼(g), the spectrum of the adversarial perturbations generated by WRPP and our proposed methods are closer to the signal itself, characterized by the energy of the perturbation being within the band of the signal. In contrast, our proposed methods produce a more concentrated perturbation energy, and this analysis also demonstrated by the fact that our method has a smaller ΔPSR.

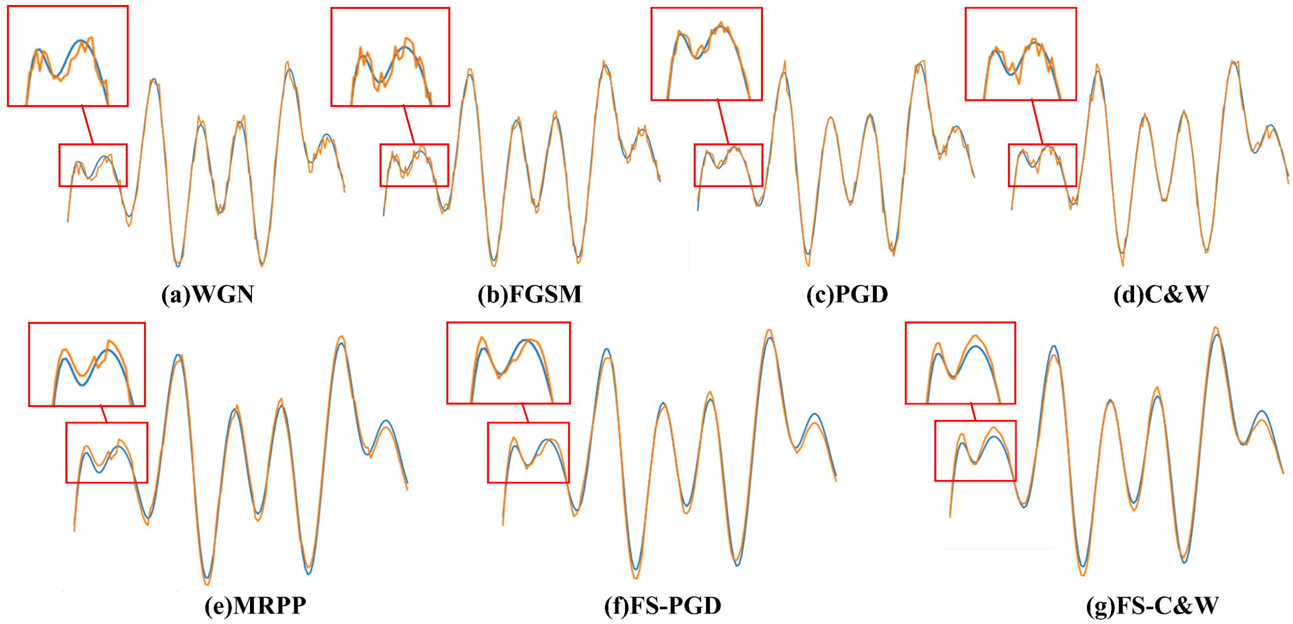

Fig. 5. A comparison of the waveform of a 16QAM modulated signal with a SNR of 20dB under 7 different adversarial attacks. (The blue line represents the waveform of the original signal, while the orange line represents the waveform of the adversarial example.)

From the perspective of the time domain, **Fig. 5 (b)**~(d) demonstrate that the waveforms of adversarial examples generated by FGSM, PGD and C&W have more violent perturbation corresponds to more high-frequency components in the frequency domain than the original signal, which are easily dropped by filters and abrupt to the human visual system. **Fig. 5 (e)**~(g) show that the time-domain waveforms of MRPP, FS-PGD and FS-C&W are smoother, i.e., predominantly low-frequency components. More low-frequency components mean that the adversarial attack is less likely blocked by the filter, and smoother waveform means that it is less noticeable to the human visual system. Therefore, our adversarial attack methods achieve imperceptibility to both the filter and the human visual system and an appealing attack performance.

C. Attack Adversarial Training Model

To illustrate further the attack performance of our adversarial attack methods against the adversarial defense model, we retrain the model using adversarial training and attack the model with our methods. The corresponding results are reported in **Tab. II**, highlighting that when an adversarial attack retrains the model, it has a certain defense ability against various adversarial attack methods, and the FR before filtering decreases to different degrees. The best performance before filtering is the C&W attack, with FR of 96.56%. Our methods are close to the optimal performance with an FR of 95.53% and 93.88% before filtering, still maintaining the ability to adapt to the filter and demonstrating the highest FR after filtering, attaining 91.43% and 90.7%.

Interestingly, the adversarial examples generated by the competitor algorithms (FGSM, PGD, C&W and MRPP) against the robust model did not present a catastrophic performance reduction after filtering. Their FR decreased only by −13.16%, −11.95%, −11.16% and −5.96%, respectively. Moreover, the filtering losses ΔPSR infer a −1.78dB, −1.27dB, −1.36dB and −0.82dB performance reduction, which is much smaller than attacking the original model. Besides, we visualize the spectrum of adversarial perturbations generated by the four algorithms attacking the robust model. **Fig. 6** depicts that the energy of the adversarial perturbations generated by the four attack methods are concentrated in the frequency band of signal.

In response to this phenomenon, Tsipras et al. [29] tries to give an explanation. Tsipras separated the adversarial examples created with small epsilon, clean-trained the network, and created one using large epsilons and robust networks. While the former looks like noise, the latter creates examples that resemble the target class. Shamir et al. [23] used on-manifold and off-manifold adversarial examples to demonstrate a very similar phenomenon. In their view, the on-manifold perturbation does what the human would expect, while the off-manifold perturbation is hard to interpret. Our experimental results corroborate the above two views from the spectrum's perspective, revealing that the perturbations generated against the robust model are closer to the target class, i.e., on-manifold perturbation. In the scenario investigated in this paper, we reveal that the energy of the perturbation against the robustness model is more concentrated in the signal's frequency band, while the proposed method can naturally find the on-manifold perturbation, both for clean and robust models. Further, since our method sets a strong constraint on the frequency of the adversarial perturbations, our method has better performance even for robust models.

D. Robustness to Filter Bandwidth Mismatch

The most critical improvement made by the suggested methods are to consider the effect of the filter $H(\bullet)$. The above

TABLE II

RESULTS OF THE FR BY SEVEN ATTACK APPROACHES AGAINST THE ADVERSARIAL TRAINING MODEL WITH PSR= −20dB

Attack	FR (Before filtering)	FR (After filtering)	IB-PSR	Δ FR	Δ PSR
WGN	1.84%	0.82%	-26.98dB	-1.02%	-6.98dB
FGSM	91.10%	77.49%	-21.78dB	-13.16%	-1.78dB
PGD	**96.19%**	84.24%	-21.27dB	-11.95%	-1.27dB
C&W	**96.56%**	85.40%	-21.36dB	-11.16%	-1.36dB
MRPP	87.76%	81.80%	-20.82dB	-5.96%	-0.82dB
FS-PGD	95.53%	**91.43%**	**-20.47dB**	-4.10%	**-0.47dB**
FS-C&W	93.88%	**90.70%**	**-20.48dB**	-3.18%	**-0.48dB**

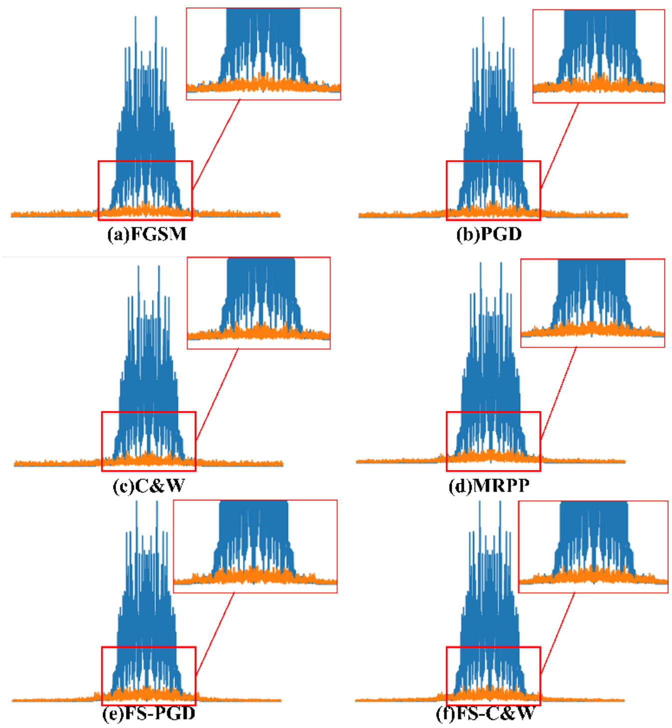

Fig. 6. Spectrum of the adversarial perturbations and original signals attacked by 6 different adversarial attack methods on an adversarial training model. (The blue line represents the spectrum of the original signal, while the orange line represents the spectrum of the adversarial perturbation.)

experiments set the constrained perturbation and the receiver of the intruder to the same kind of filter and bandwidth. It is clear that this is impossible in practical applications, as the transmitter cannot know the intruder's filter type and bandwidth. Hence, we designed the following experiment to evaluate performance of our methods in case of filter mismatch. We compared the adaptability performance of our algorithms under four different filter implementation methods used by the intruder, namely FFT filter, Chebyshev Type I filter, Chebyshev Type II filter, and Butterworth filter. For the transmitter of the adversarial example, the code rate of the signal is known. Therefore, we fix the bandwidth of the constrained perturbation constant to 0.2MHz and then vary the filtering bandwidth of the receiver from 0.2MHz to 0.4MHz to evaluate our method's performance.

TABLE III

THE FR OF FS-PGD UNDER DIFFERENT FILTER SETTING

	0.2MHz	0.3MHz	0.4MHz
FFT	98.18%	97.68%	97.95%
Chebyshev I	95.27%	92.67%	94.67%
Chebyshev II	95.70%	95.72%	95.56%
Butterworth	95.75%	95.62%	95.56%

We only presented the performance of the FS-PGD algorithm, and the performance of the FS-C&W follows a similar pattern. Tab. III shows the FR of our algorithm under different filter configuration conditions. It can be observed that the performance of our algorithm remains unchanged regardless of whether there is bandwidth mismatch or implementation mismatch in the filter. This indicates that our algorithms have strong robustness to changes in both the parameters of the filter and the type of filter.

V. Discussion and Future Work

This paper considers a more realistic scenario of a wireless signal adversarial attack, where the wireless receiver filters the signal, resulting in adversarial perturbations that cannot access the target model. Hence, we propose a novel framework to accommodate this scenario, aiming to perturb signals by attacking their frequency subspace. The experimental results demonstrate that such an approach to attacking wireless signals works well in various settings. Further analysis demonstrates that the essence of our approach is to find the on-manifold adversarial perturbation.

We also define a new set of evaluation metrics to measure the "imperceptibility" of adversarial perturbations in the wireless signal domain. To our knowledge, this work is groundbreaking, enriching the connotation of adversarial samples.

The conclusion about the on-manifold adversarial perturbation is only illustrated experimentally, lacking theoretical support. Thus, an in-depth study poses a future research direction. Besides, the receiver is only part of the physical world factor, and the wireless channel also affects the performance of the adversarial attack. Hence, we will study the adversarial attack under the influence of a wireless channel to make our adversarial attack scenario closer to the real world.

Appendix
Solving δ

$$\max_{\tilde{\delta}} L(\tilde{\mathbf{y}}) + \nabla_{\tilde{y}} L(\tilde{\mathbf{y}})^{\mathrm{T}} \tilde{\boldsymbol{\delta}} \text{ s.t. } \|\boldsymbol{\delta}\|_p \leq \epsilon.$$

Since $L\left(\tilde{\mathbf{y}}\right)$ is independent of $\boldsymbol{\delta}$. Our problem is

$$\max_{\boldsymbol{\delta}} \nabla_{\tilde{\mathbf{y}}} L^T \tilde{\delta} \ s.t. \ \|\boldsymbol{\delta}\|_p \leqslant \epsilon \tag{24}$$

Clearly, the optimal $\boldsymbol{\delta}$ would have a norm of ϵ, otherwise, we can normalize $\boldsymbol{\delta}$ to get a greater loss. Therefore, we are set to solve

$$\max_{\boldsymbol{\delta}} \nabla_{\tilde{\mathbf{y}}} L^T \tilde{\delta} \ s.t. \ \|\boldsymbol{\delta}\|_p = \epsilon \tag{25}$$

This could be solved by standard Lagrangian multiplier method, where $f\left(\boldsymbol{\delta}\right) = \nabla_{\tilde{\mathbf{y}}} L^T \tilde{\delta}$, and $g\left(\boldsymbol{\delta}\right) \equiv \|\boldsymbol{\delta}\|_p = \epsilon$. Set $\nabla f\left(\boldsymbol{\delta}\right) = \lambda \nabla g\left(\boldsymbol{\delta}\right)$, we have

$$\nabla f\left(\boldsymbol{\delta}\right) = \lambda \nabla g\left(\boldsymbol{\delta}\right) \tag{26}$$

$$J_H \cdot \nabla_{\tilde{\mathbf{y}}} L = \lambda \frac{\boldsymbol{\delta}^{p-1}}{p\left(\sum_i \boldsymbol{\delta}_i^p\right)^{1-\frac{1}{p}}} \tag{27}$$

$$J_H \cdot \nabla_{\tilde{\mathbf{y}}} L = \frac{\lambda}{p}\left(\frac{\boldsymbol{\delta}}{\epsilon}\right)^{p-1} \tag{28}$$

$$\left(J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right)^{\frac{p}{p-1}} = \left(\frac{\lambda}{p}\right)^{\frac{p}{p-1}} \left(\frac{\boldsymbol{\delta}}{\epsilon}\right)^p \tag{29}$$

Sum over two sides

$$\sum \left(J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right)^{\frac{p}{p-1}} = \sum \left(\frac{\lambda}{p}\right)^{\frac{p}{p-1}} \left(\frac{\boldsymbol{\delta}}{\epsilon}\right)^p \tag{30}$$

$$\left\| J_H \cdot \nabla_{\tilde{\mathbf{y}}} L \right\|_{p^*}^{p^*} = \left(\frac{\lambda}{p}\right)^{p^*} * 1 \tag{31}$$

$$\frac{\lambda}{p} = \left\| J_H \cdot \nabla_{\tilde{\mathbf{y}}} L \right\|_{p^*} \tag{32}$$

Combine (28) and (32), it is easy to see

$$\boldsymbol{\delta} = \epsilon sign\left(J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right) \left(\frac{\left|J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right|}{\left\|J_H \cdot \nabla_{\tilde{\mathbf{y}}} L\right\|_{p^*}}\right)^{\frac{1}{p-1}} \tag{33}$$

References

[1] M. Z. Hameed, A. György, and D. Gündüz, "The best defense is a good offense: Adversarial attacks to avoid modulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1074–1087, 2021, doi: 10.1109/TIFS.2020.3025441.

[2] D. Gunduz, D. Richard Brown, and H. Vincent Poor, "Secret communication with feedback," in *Proc. Int. Symp. Inf. Theory Its Appl.*, Dec. 2008, pp. 1–6.

[3] B. A. Bash, D. Goeckel, and D. Towsley, "Square root law for communication with low probability of detection on AWGN channels," in *IEEE Int. Symp. Inf. Theory Proc.*, Jul. 2012, pp. 448–452.

[4] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, no. 2, pp. 137–156, Apr. 2007.

[5] G. J. Mendis, J. Wei, and A. Madanayake, "Deep learning-based automated modulation classification for cognitive radio," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, pp. 1–6, Dec. 2016.

[6] T. O'shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Oct. 2017.

[7] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Mar. 2017, pp. 1–6.

[8] J.-K. Kim, B.-D. Kim, D.-W. Yoon, and J.-W. Choi, "Deep neural network-based automatic modulation classification technique," *J. Korean Inst. Inf. Technol.*, vol. 14, no. 12, p. 107, Dec. 2016.

[9] X. Liu, D. Yang, and A. E. Gamal, "Deep neural network architectures for modulation classification," in *Proc. 51st Asilomar Conf. Signals, Syst. Comput.*, Oct. 2017, pp. 915–919.

[10] Y. Zhao, X. Wang, and Z. Huang, "Concentrate on hardware imperfection via aligning reconstructed states," *IEEE Commun. Lett.*, vol. 26, no. 12, pp. 2934–2938, Dec. 2022, doi: 10.1109/LCOMM.2022.3204170.

[11] Y Zhao, X Wang, Z Lin, and Z Huang, "Multi-classifier fusion for open-set specific emitter identification," *Remote Sensing*, vol. 14, no. 9, p. 2226, May 2022.

[12] L. Sun, D. Ke, X. Wang, Z. Huang, and K. Huang, "Robustness of deep learning-based specific emitter identification under adversarial attacks," *Remote Sensing*, vol. 14, no. 19, p. 4996, Oct. 2022.

[13] Y. Zhao, X. Wang, L. Sun, and Z. Huang, "A novel framework for extracting moment-based fingerprint features in specific emitter identification," *EURASIP J. Adv. Signal Process.*, vol. 2023, no. 1, pp. 1–17, 2023.

[14] Y. Zhao, X. Wang, L. Sun, and Z. Huang, "A novel signal representation in SEI: Manifold," *J. Franklin Inst.*, vol. 360, no. 7, pp. 5292–5318, May 2023.

[15] J. Bruna et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2014, pp. 2–11.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, pp. 1–11, 2015.

[17] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 213–216, Feb. 2019.

[18] S. Kokalj-Filipovic, R. Miller, N. Chang, and C. L. Lau, "Mitigation of adversarial examples in RF deep classifiers utilizing autoencoder pre-training," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2019, pp. 1–6.

[19] D. Ke, Z. Huang, X. Wang, and L. Sun, "Application of adversarial examples in communication modulation classification," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 877–882.

[20] Y. Lin, H. Zhao, X. Ma, Y. Tu, and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Trans. Rel.*, vol. 70, no. 1, pp. 389–401, Mar. 2021.

[21] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3868–3880, Jun. 2022.

[22] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[23] A. Shamir, O. Melamed, and O. BenShmuel, "The dimpled manifold model of adversarial examples in machine learning," 2021, *arXiv:2106.10151*.

[24] M. Shi, Y. Huang, and G. Wang, "Receiver distortion normalization method for specific emitter identification," in *Proc. IEEE 9th Joint Int. Inf. Technol. Artif. Intell. Conf.*, vol. 9, Dec. 2020, pp. 1728–1732.

[25] D. Ke, X. Wang, K. Huang, H. Wang, and Z. Huang, "Minimum power adversarial attacks in communication signal modulation classification with deep learning," *Cognit. Comput.*, vol. 15, no. 2, pp. 580–589, Mar. 2023.

[26] C. Lyu, K. Huang, and H.-N. Liang, "A unified gradient regularization family for adversarial examples," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 301–309.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2018, pp. 1–10.

[28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57, doi: 10.1109/SP.2017.49.

[29] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. 7th Int. Conf. Learn. Represent.*, 2019, pp. 1–24.

Da Ke received the B.S. degree in electronic engineering and the M.S. degree in information and communication engineering from the National University of Defense Technology in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in electronic science and technology. His research interests include signal processing and deep learning.

Xiang Wang was born in 1985. He received the bachelor's degree in electronic science and engineering from the National University of Defense Technology, Changsha, Hunan, China, in 2007, where he is currently pursuing the Ph.D. degree. His research interests include blind signal processing in radar and communication applications.

Zhitao Huang was born in 1976. He received the B.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, Hunan, China, in 1998 and 2003, respectively. He is currently a Professor with the College of Electronic Science and Engineering, National University of Defense Technology. His research interests include radar and communication signal processing, and array signal processing.