remote sensing

MDPI

Article

Robustness of Deep Learning-Based Specific Emitter Identification under Adversarial Attacks

Liting Sun [1,†], Da Ke [1,†], Xiang Wang [1,*], Zhitao Huang [1] and Kaizhu Huang [2]

[1] College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China
[2] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
* Correspondence: xwang@nudt.edu.cn
† These authors contributed equally to this work.

Abstract: Deep learning (DL)-based specific emitter identification (SEI) technique can automatically extract radio frequency (RF) fingerprint features in RF signals to distinguish between legal and illegal devices and enhance the security of wireless network. However, deep neural network (DNN) can easily be fooled by adversarial examples or perturbations of the input data. If a malicious device emits signals containing a specially designed adversarial samples, will the DL-based SEI still work stably to correctly identify the malicious device? To the best of our knowledge, this research is still blank, let alone the corresponding defense methods. Therefore, this paper designs two scenarios of attack and defense and proposes the corresponding implementation methods to specializes in the robustness of DL-based SEI under adversarial attacks. On this basis, detailed experiments are carried out based on the real-world data and simulation data. The attack scenario is that the malicious device adds an adversarial perturbation signal specially designed to the original signal, misleading the original system to make a misjudgment. Experiments based on three different attack generation methods show that DL-based SEI is very vulnerability. Even if the intensity is very low, without affecting the probability density distribution of the original signal, the performance can be reduced to about 50%, and at −22 dB it is completely invalid. In the defense scenario, the adversarial training (AT) of DL-based SEI is added, which can significantly improve the system's performance under adversarial attacks, with ≥60% improvement in the recognition rate compared to the network without AT. Further, AT has a more robust effect on white noise. This study fills the relevant gaps and provides guidance for future research. In the future research, the impact of adversarial attacks must be considered, and it is necessary to add adversarial training in the training process.

Keywords: specific emitter identification; adversarial attack; wireless network security; adversarial training; radio frequency fingerprint; Internet of Things

check for updates

Citation: Sun, L.; Ke, D.; Wang, X.; Huang, Z.; Huang, K. Robustness of Deep Learning-Based Specific Emitter Identification under Adversarial Attacks. *Remote Sens.* **2022**, *14*, 4996. https://doi.org/10.3390/rs14194996

Academic Editors: Prem Prakash Jayaraman, Federico Montori, Charith Perera and Felip Marti

Received: 25 July 2022
Accepted: 29 September 2022
Published: 7 October 2022

1. Introduction

Specific Emitter Identification (SEI), also known as Radio Frequency Fingerprinting (RFF), is a technique to identify the source of emission by measuring external characteristics of the radio frequency (RF) signal and obtaining the emitter-specific information [1–3]. The technique does not rely on the content of the signal transmission; however, it extracts physical-layer information specific to the device hardware as an individual label. This information is Radio frequency (RF) fingerprints, which are unintentionally generated by the imperfect characteristics of hardware and cannot be avoided or forged [4,5]. Therefore, SEI becomes a promising lightweight non-key authentication solution [6,7]. It can be used in many scenarios, especially for the security of Internet of Things (IoT) [8,9]. Because, IoT applications are highly resource-constrained, hence they can be not fully effective to utilize even many modern cryptographic methods for security purposes, let alone the traditional cryptography solutions [10,11]. And the size and flexibility of IoT also bring other unique security concerns [12].

The RF fingerprint of the emitter arises from the differences caused by the non-ideal nature of the hardware, which is extremely subtle within the tolerance of the system's normal operation [13]. RF fingerprint has five main characteristics: universality, stability, uniqueness, measurability, and independence [14]. Therefore, the core of SEI lies in the acquisition of RF fingerprint information [1].

Up to now, there are two major feature extraction methods: (1) traditional human-predefined feature method, (2) automatic extraction method based on neural networks. The traditional human-predefined feature methods can be divided the transient-features [15–17] and the state state-features [2,18,19]. However, with the rapid development of the IoT technology, the number of devices and the amount of data have increased dramatically. Therefore, the traditional methods have become increasingly difficult to meet the requirements of large-scale data with high real-time [12,20].

Recent studies have shown that deep learning (DL)-based SEI methods can achieve end-to-end automatic feature extraction with satisfying recognition results [5,8,13,21,22]. Despite their remarkable performance, the model robustness of deep neural networks (DNN) emerges as one of the greatest challenges in safety-critical applications (e.g., self-driving and healthcare). As indicated in many researches, DNNs can easily be fooled by adversarial examples or perturbations of the input data. Adding a well-designed adversarial perturbation to the input signal makes the neural network much less effective or even fail to work at all in countering the attack [23].

An adversarial attack was first proposed in image recognition and computer vision fields [24]. In the signal processing field, the research is primarily focused on automatic modulation classification (AMC). The vulnerability of DL-based modulation classification technique to adversarial examples was analyzed in [25].

SEI is specially designed to enhance the network security for adversarial problems, such as DoS attack [26], injection attack [27], spoofing attack [28], etc. [29]. Moreover, compared with modulation differences, SEI deals with more subtle fingerprint differences, even requiring high robustness to identify malicious devices. However, no studies relevant to the adversarial attacks have been conducted on SEI.

In SEI, the adversarial examples may be a new type of attack. If an unregistered malicious device contains these specially designed adversarial perturbations on the transmitted signal, will the performance of the DL-based SEI system be affected, and can the illegal device be identified normally? Therefore, to make DL-based SEI techniques work stably and reliably, the following questions must be answered:

(1) DL-based SEI is used to identify hardware differences between radiation sources, is it robust? Will it be affected by adversarial examples?
(2) Does it work properly after being affected against these attacks? What is the form of adversarial examples against SEI? What are the characteristics of the attack signal?
(3) Is there any way to improve system performance against these attacks? Can the recognition performance of the system be fully recovered?

To the best of our knowledge, although there are many studies on DL-based SEI, they have never focused on the these issues. Therefore, in this paper, we design two scenarios to conduct systematical research on the above problems.

1.1. Contribution

To tackle the above problems, this paper focuses on the robustness of DL-based SEI, especially the performance under adversarial attacks. The main contribution of this work can be summarized as follows:

(1) The security and robustness of DL-based SEI under adversarial attacks are studied for the first time, and the in-depth system analysis answers the above questions, which has certain guiding significance for the practical application of this technology in the future.

(2) The concept of adversarial attack and defensive training is introduced into the SEI problem, two new scenarios are designed on the basis of the original SEI, the specific implementation methods are given, and rigorous experiments are carried out on the real-world and simulated datasets. Scenarios include ① attack scenario: malicious devices use adversarial attacks to fool DL-based SEI, and ② defense scenario: adversarial training is performed to improve system performance to correctly identify illegal devices.

(3) In the attack scene, the adversarial perturbations and system loss are investigated based on three adversarial example generation methods. The waveform characteristics of adversarial examples, degree of performance degradation, and influences on different emitters are analyzed and studied. It is discovered that DL-based SEI is very vulnerable to adversarial attacks. Adversarial perturbation, even with a quite low energy, can make the DL-based SEI fail, with much higher destructiveness than the white noise of the same strength. The adversarial perturbations at the strength of −25 dB on the real-world data can reduce the recognition performance from 99% to below 10%.

(4) In the defense scenario, a corresponding adversarial training (AT) method inspired by the normalized training in [30], which deals with adversarial attacks, is proposed to enhance the robustness of DL-based SEI. Through AT, the highest improvement of the performance can be more than 60%. Facing the attacks at the strength of −32 dB, performance of DL-based SEI recovers from 55.29% to above 85.59%. AT also improves the robustness of the system to white noise.

(5) In addition, it is also found that different datasets are affected by attacks differently, and the improvement effect after AT is also different; there are also differences between individual emitters. Moreover, there is a certain threshold for the improvement effect of AT.

The research found that the problem of SEI is originally fine-grained identification, so it is greatly affected by adversarial attacks, which makes the original system unable to identify malicious devices and threatens network security. The system performance can be improved after AT. In the future research and application, attention should be paid to this kind of attack, and AT should be carried out.

1.2. Related Work

1.2.1. DL-Based SEI

At present, SEI based on deep learning has become a research hotspot. Different from data such as images and texts, RF signals are often stored as In-phase/Quardrature (I/Q) data, with the dimension of $2 * N$, where N represents the sequence length along the time direction. The input of DL-based SEI is generally I/Q data or preprocessed data. For the latter, it is common to use the Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), and Recurrence Plots (RP) methods to convert the signal to image form [31] as input.

In terms of network structure, the CNN model is the most widely used and earliest network structure in SEI [32,33]. Wong, Riyaz et al. [34] confirmed that the convolutional neural network (CNN) can be used to directly process I/Q data to achieve RFF, showing the advantages of end to end. The residual design can well alleviate the problem of network degradation, and reduce the risk of gradient disappearance [35]. In 2019, Pan et al. [36] used Residual Network (ResNet) to achieve emitter classification. In 2021, Zhang et al. [37] designed a RFFResNet model with reduced parameters based on the ResNet model.

Therefore, this paper uses the I/Q data as the input of the neural network and conducts experiments on the ResNet.

1.2.2. Adversarial Attack

Szegedy et al. [38] first revealed the vulnerability of deep learning models to adversarial perturbations by solving for the following optimization problem:

$$\min \|\rho\|_2 \qquad \text{s.}t.\mathcal{M}(\mathbf{I}+\rho) = \bar{\ell}; \mathbf{I}+\rho \in [0,1]^m,$$

where $\mathcal{M}(\cdot)$ is a deep learning model, $\mathbf{I}$ is the input of model and ρ is the adversarial perturbation. $\bar{\ell}$ is the ground truth label.

This is a hard problem. Szegedy computed a approximate solution by Limited Memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. However, solving this approximate problem is computationally prohibitive. This inspired the Fast Gradient Sign Method (FGSM) [24]. FGSM is a gradient-based single-step method for computing adversarial perturbations, focusing on the "efficiency" rather than achieving a high fooling rate. Building on the basis of FGSM, iterative FGSM, the Basic Iterative Method (BIM) [39] was proposed. It is an influential contribution that introduced the Physical World attacks. The Projected Gradient Descent (PGD) [40] attack is widely considered as one of the most powerful attacks in the literature. It is regarded the iterative FGSM as the ℓ_∞-bound PGD. The research on adversarial vulnerability also started a parallel line of research on the defense of deep learning adversarial attacks. Defensive distillation [41] was a prominent technique that promised an effective solution to the problem, by building on the insights of knowledge distillation in deep networks. However Carlini & Wagner [42] developed a set of attacks that computes norm-restricted additive perturbations that completely break defensive distillation, referred as "C&W" in the following.

In this work, FGSM, PGD and C&W are all utilized to generate adversarial perturbations for SEI.

1.2.3. Adversarial Attack in Communication Signal Processing

In [25], 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, the research is primarily focused on automatic modulation classification (AMC). The vulnerability of DL-based AMC to adversarial examples was analyzed in [43]. A similar method has been proposed recently in [44], where modifications are employed by the transmitter to evade a DNN-based jammer, and the receiver uses another DNN (an autoencoder) to preprocess the received signal and filter out the modifications. Lin et al. [45] found that the attack effect of the iterative methods to AMC was better than that of the one-step method, inspiring researchers to further promote the convolutional neural network reliability against adversarial attacks. Ref [41] considered the impact of the defensive perturbations on the bit error rate (BER) at the legitimate receiver and proposed a novel defense mechanism that modifies the channel input symbols at the transmitter in order to reduce the modulation-classification accuracy at the intruder while maintaining a low BER at the legitimate receiver.

However, no relevant studies of adversarial attacks on SEI have been conducted yet. Since SEI focuses on subtle features between different emitters, the inter-class distances are small. This leads to SEI being more susceptible to weak adversarial perturbations. The security of DL-based SEI should receive more attention. In this paper, we first verify the effectiveness and feasibility of adversarial attack in the SEI. Then, we give an alternative method, adversarial training, to deal with the adversarial attack in the SEI.

1.2.4. Adversarial Training

Adversarial training is a set of techniques to improve the robustness of classifiers [30,46]. It feeds adversarial data instead of clean data into the DNNs. Methods for improvement extend the conventional adversarial training by injecting the adversarial perturbation to hidden layers to boost the robustness of latent space [42,47,48]. All these methods generate adversarial examples by maximizing the loss function of the label information.

1.3. Organization

The remainder of the paper is organized as follows. Section 2 introduces the model of SEI and the transmitter distortion models. Section 3 describes three adversarial attacks and the method of adversarial training. In Section 4, we detail the proposed three scenarios for

SEI and the implement scheme of adversarial attacks and AT-SEI. Section 5 presents and analyzes the experimental results based on different datasets. Finally, Section 6 sets out the conclusion.

2. Problem Formulation

RFF arise from non-ideal distortion of the hardware device, i.e., deviations caused by device imperfections or not operating in an ideal state. Such deviations are inevitably incidental to the actual transmitted signal in the form of unintentional modulation (UM) [18,49].

The signal with UM received at the receiver side can be expressed as

$$r(t) = (A(t) + \Delta A(t))e^{j(2\pi f_0 t + \phi(t) + \Delta\phi(t) + \theta_0)} + \nu(t), \tag{1}$$

where $A(t)$ and $\phi(t)$ represent the intentional modulation (IM) on amplitude and phase, respectively; $\Delta A(t)$ is the UM on amplitude; $\Delta\phi(t)$ is the UM on phase; f_0 is the frequency offset, θ_0 is the initial phase; $\nu(t)$ is the additive noise.

UM information is emitter-specific, representing the device difference between emitters, i.e., RF fingerprint. Usually, the energy of UM is far less than the IM [50]. Compared to IM, UM is subtle, difficult to extract, and susceptible to perturbation [51].

SEI aims to extract the exact representation of subtle UMs from IM and environmental influences. In this study, UMs are automatically extracted using neural networks based on the received time series $r(t)$, similar to [8,21].

2.1. Transmitter Distortion Model

A typical transmitter structure is shown in the Figure 1 [52]. Usually, the devices in the figure all lead to the generation of UM, that is, they are closely related to the RFF generation of the emitter, including filter, mixer, power amplifier, and local oscillator. In this paper, they are classified and modeled based on distortion effects, namely filter distortion (FD), IQ modulation error (IQE), power amplifier nonlinear distortion (PAD), carrier leakage and spurious tone (CST).

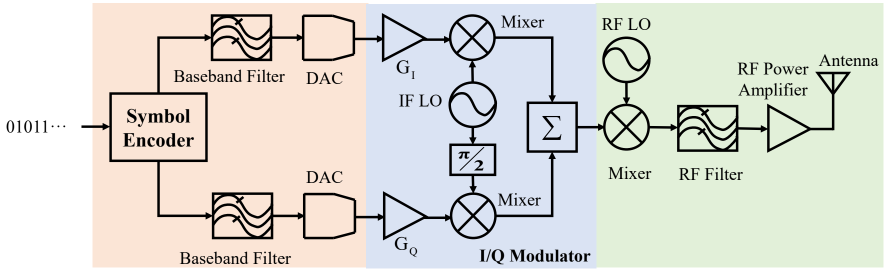


Figure 1. Structure of the typical transmitter.

2.1.1. Filter Distortion (FD)

The ideal bandpass filter should behave as follows throughout the transmitting band: the amplitude of the frequency response and the group delay are constant. However, the actual analog filter's frequency response amplitude may be skewed or rippled within the band, and its group delay varies with frequency. The filter distortions are mainly manifested as: (1) skew and ripple in amplitude and frequency response; (2) fluctuations in group delay [53].

These distortions of the filter are closely related to the characteristics of its internal circuit components, and thus the filter distortion characteristics may vary from transmitter to transmitter, thus constituting a transmitter fingerprint.

Although there are multiple filters in the transmitter structure, they can be modeled approximately as a whole-path filter from the point of view of their impact on the complex signal.

Let $H(f)$ be the frequency response of an ideal transmitter bandpass filter, the whole path distortion filter of the mth emitter can be expressed as

$$G_m(f) = H(f)A_m(f)e^{j\phi_m(f)}, \tag{2}$$

where $A_m(f)$ and $\phi_m(f)$ denotes the distortions of amplitude and phase respectively. The following discussion is all about the mth emitter. For brevity, the symbol m is omitted when there is no ambiguity.

Further, the distortions can be expressed by the Fourier series, given as [52]

$$\begin{aligned} A(f) &= a_0 + \sum_{k=1}^{\infty} a_k \cos(2\pi kf/T_A) \\ \phi(f) &= 2\pi b_0 f + \sum_{k=1}^{\infty} b_k \sin\big(2\pi kf/T_\phi\big). \end{aligned} \tag{3}$$

where a_0, b_0, T_A, and T_B are the distortion parameters for the current emitter. The symbol m has been omitted.

Taking the second-order approximation, we can obtain [54]

$$\begin{aligned} A(f) &= a_0 + a_k \cos(2\pi kf/T_A) \\ \phi(f) &= 2\pi b_0 f + b_k \sin(2\pi kf/T_B). \end{aligned} \tag{4}$$

Therefore, by equating the filter to baseband, the transmitter equivalent baseband signal for the entire path is expressed as

$$s^{\mathrm{FD}}(t) = \sum_{k=-\infty}^{\infty} S_k g(t - kT), \tag{5}$$

where S_k is the transmit complex modulation symbol and $g(t) = \int_{-\infty}^{\infty} G(f)e^{j2\pi ft}df$.

The decomposition of $g(t)$ yields that the filter distortion leads to parasitic modulation on the carrier, similar to the summation of multiple paths of delay, which is spread in time, causing delay expansion and thus introducing trace amounts of inter-code crosstalk.

2.1.2. I/Q Quadrature Modulation Errors (IQE)

I/Q modulation errors are specifically manifested as gain mismatches in the two IQ paths and quadrature errors, i.e., not strictly 90°.

Suppose $s_I(t)$ and $s_Q(t)$ are the baseband waveforms of I/Q respectively, then the ideal baseband signal is

$$s_0(t) = s_I(t) + js_Q(t). \tag{6}$$

And the baseband signal carrying the I/Q modulator distortion can be expressed as

$$\begin{aligned} s^{\mathrm{IQE}}(t) &= \rho(s_I(t) + c_I)\cos(2\pi ft + \frac{\xi}{2}) \\ &\quad + (s_Q(t) + c_Q)\sin(2\pi ft - \frac{\xi}{2}), \end{aligned} \tag{7}$$

where $\rho = \frac{G_I}{G_Q}$ is the gain mismatch, G_I and G_Q are the amplitude gains of the two channels, respectively, and ξ are the phase deviations of the quadrature errors, c_I and c_Q are the DC bias components generated by the two mixers, respectively. Typically, c_I and c_Q can be treated as constants.

2.1.3. Carrier Leakage and Spurious Tone (CST)

Oscillators and various other active devices (such as mixers and amplifiers) generate various harmonic signals. If the frequency of the parasitic harmonic signal lies within the signal passband range, it will not be filtered and will be emitted along with the normal radiation source modulated signal, which is spurious tone. If the frequency of the parasitic harmonics is equal to the carrier frequency, it is called carrier leakage.

Due to differences in devices, the frequency and amplitude of the parasitic harmonics will likely differ, thus constituting an RFF.

This distortion can be modeled as

$$s^{\mathrm{CST}}(t) = \left(x(t) + \varsigma^{\mathrm{CL}}\right)\cos(2\pi f t) + A^{\mathrm{ST}}\cos\left(2\pi\left(f^{\mathrm{ST}} + f\right)t\right), \tag{8}$$

where A^{ST} is the amplitude of the parasitic harmonics, f^{ST} is the frequency of the parasitic harmonics, and ς^{CL} is the carrier leakage.

2.1.4. Power Amplifier Nonlinear Distortion (PAD)

The amplifier is designed to amplify the input signal power linearly, but due to the constraints of the non-ideality of the device, the amplification effect of the amplifier on the input signal shows nonlinear characteristics. Effectively, the amplifier nonlinearity will produce the amplitude modulation-amplitude modulation (AM-AM) compression effect and the amplitude modulation-phase modulation (AM-PM) conversion effect [55].

The nonlinear characteristic of the power amplifier is usually described by Taylor series [2,54]. The output signal can be modeled as

$$s^{\mathrm{PAD}}(t) = \sum_{i=1}^{L} b_{m,i}(x(t))^{i}, \tag{9}$$

where $x(t)$ is the input signal of power amplifier, L is the order and $b_{m,i}$ denotes the coefficient of the Taylor polynomial.

3. Adversarial Attack and Adversarial Training

3.1. Adversarial Attack

In recent years, it has been found that DNNs can be easily deceived or attacked by a special perturbation called adversarial examples, which are often imperceptible, generated using adversarial learning, and have some generalization, i.e., they can successfully attack different DL models. Specifically, in pattern classification, when the identified samples, are slightly perturbed, it is very likely that the neural network will be misclassified even if it is well trained [23]. As DL has been widely used in practice, especially in some fields with high security requirements, the robustness of artificial intelligence (AI) algorithmic models against adversarial examples becomes particularly important.

Szegedy et al. first demonstrated the existence of small perturbations to the input samples, such that the perturbed samples could fool DNNs into misclassification [38].

3.1.1. Definition

Let the input signal be x, and the most likely class of output be l. The purpose of training DNNs is to obtain the optimal parameters. The formula is as follows:

$$\arg\min_{\theta} L(f(\theta, x), l), \tag{10}$$

where f is the selected network model, and L is the loss function.

Generally, imperceptibly tiny perturbations of a given sample do not normally change the underlying class.

The most basic optimization problem for adversarial examples can be defined as

$$x_{\mathrm{adv}} = \arg\max_{x'} L(x, \theta). \quad s.t. \parallel x' - x \parallel_p \leq \epsilon. \tag{11}$$

The goal is to find a worst sample x_{adv} in a very small neighborhood of radius ϵ of the original sample x such that the loss of the classifier is maximized.

Several methods have been proposed to generate adversarial examples.

3.1.2. Type of Attack

These generation methods can be classified into white-box attacks and black-box attacks based on whether they can access the target network parameters [56].

Under white-box attacks, the attacker is usually assumed to know all the information of the target model, such as the model structure, model parameters, training strategy, and even training data.

In contrast, under a black-box attack, the attacker is usually assumed to know only the inputs of the target model but not the internal parameters. The fact that black-box attacks are set up is more in line with the actual reality, because it is generally difficult to obtain the internal parameters of the target model. But this does not make black-box attacks more meaningful than white-box attacks. In fact, testing the performance of a model under a white-box attack can reflect the performance of the model in the worst case.

Therefore, we choose three relatively typical methods of generating adversarial examples under black-box attacks to test the performance of DL-based SEI under the worst-case adversarial sample attacks.

3.1.3. Adversarial Example Generation

To explore the attack performance in SEI, we select three most popular methods: fast gradient sign (FGSM) [24], projected gradient descent (PGD) [46] and C&W [41] methods.

FGSM

FGSM was developed to efficiently compute an adversarial perturbation for a given input signal by solving the following problem,

$$\rho = \varepsilon \cdot sign(\nabla L(\theta, x, l)), \tag{12}$$

where $\nabla L(\cdot,\cdot,\cdot)$ computes the gradient of the lost function around the current value of the model parameters θ w.r.t. x; $sign(\cdot)$ denotes the sign function; ε is the small scalar value that restricts the norm of the perturbation.

The pseudo code of FGSM can be summarized in Algorithm 1.

```
**Algorithm 1** FGSM.
Input:   Original signal example $x$; ground-truth label $l$;
         Loss function $L$ of classifier; perturbation size $\varepsilon$.
Output: Adversarial example $x_{adv}$
    1. Calculate the gradient $\nabla_x L(\theta, x, l)$
    2. Acquire $\rho$ by applying the gradient method as
           $\rho = \varepsilon \cdot sign(\nabla_x L(\theta, x, l))$
    3. Applying the perturbation to the original sample as:
           $x_{adv} = x + \varepsilon \cdot sign(\nabla_x L(\theta, x, l))$
    4. Return $x_{adv}$
```

PGD

PGD is an intuitive extension of FGSM, also known as Basic Iterative Method (BIM). It is to iteratively take multiple small steps while adjusting the direction after each step and limit the overall perturbation to the input space. PGD is iteratively calculated as follows [46]:

$$x_\rho^{i+1} = Clip_\varepsilon\{x_\rho^i + \lambda sign(\nabla L(\theta, x_\rho^i, l))\}, \tag{13}$$

where x_ρ^i is the perturbed signal at the ith iteration; $Clip_\varepsilon\{\cdot\}$ clips (the values of the pixels of) the signal in its argument at ε and λ determines the step size ($0 < \lambda < \varepsilon$).

The pseudo code of PGD is given in Algorithm 2.

Algorithm 2 PGD.

```
Input:  Original signal example $x$; ground-truth label $l$;
        Loss function $L$ of classifier; perturbation size $\varepsilon$;
        Iteration number $N$.
Output: Adversarial example $x_{adv}$
    1. Initialize $x_\rho^0 \leftarrow x$
    2. for $i = 0$ to $N - 1$ do
    3.     Calculate the gradient $\nabla_x L(\theta, x, l)$
    4.     Update $x_\rho^{n+1}$ by applying the gradient method, and use Clip$_{x,\lambda}\{\cdot\}$ to clip $x_\rho^{i+1}$ as
               $x_\rho^{i+1} = Clip_{x,\lambda}\left\{x_\rho^{i+1} + \lambda \cdot sign\left(\nabla_x L\left(x_\rho^i, l\right)\right)\right\}$
    5.     Applying the perturbation to the original sample as:
    6. end for
    7. Return $x_{adv} = x_\rho^{i+1}$
```

C&W

A set of adversarial attacks was introduced by Carlini and Wagner [41], that is C&W method. These attacks make the perturbations quasi-imperceptible by restricting their l_2, l_∞ , and l_0 norms. The C&W method can achieve attacks with different confidence levels.

$$L_{\mathrm{C\&W}}(x', t) = \max\left(\max_{i \neq t}\{Z(x')_{(i)}\} - Z(x')_{(t)'} - \kappa\right), \tag{14}$$

where $Z(x')_{(i)'}$ represents the ith component of the classifier output, t is the target labels, κ is the parameter reflexing the minimum expected confidence bound of the adversarial example.

There are many ways to implement C&W. In this paper, we implement C&W attacks using the C&W-loss within the PGD framework [57].

The pseudo code of C&W is shown in Algorithm 3.

Algorithm 3 C&W.

```
Input:  Original signal example $x$; ground-truth label $l$;
        C&W loss function $L_{C\&W}$ of classifier; Perturbation size $\varepsilon$;
        Iteration number $N$.
Output: Adversarial example $x_{adv}$
    1. Initialize $x_\rho^0 \leftarrow x$
    2. for $i = 0$ to $N - 1$ do
    3.     Calculate the gradient $\nabla_x L_{\mathrm{C\&W}}(\theta, x, l)$
    4.     Update $x_\rho^{n+1}$ by applying the gradient method, and use Clip$_{x,\lambda}\{\}$ to clip $x_\rho^{i+1}$ as
               $x_\rho^{i+1} = Clip_{x,\lambda}\left\{x_\rho^{i+1} + \lambda \cdot sign\left(\nabla_x L_{\mathrm{C\&W}}\left(x_\rho^i, l\right)\right)\right\}$
    5.     Applying the perturbation to the original sample as:
    6. end for
    7. Return $x_{adv} = x_\rho^{i+1}$
```

3.2. Adversarial Training

The aim of AT is to train the model with both the natural samples and adversarial examples to alleviate the influence caused by adversarial attacks [58].

The new training objective can be given as

$$\alpha L(x, y) + (1 - \alpha)(L(x', y), \tag{15}$$

where $L(x, y)$ denotes the classification loss, x' represents the corresponding adversarial example of natural example x, and α is the weight, usually set to 0.5.

The optimization problem of AT with the l_p norm constraint can be formulated as follows:

$$\begin{aligned} &\min_{\theta} \max_{\rho} D[q(\mathbf{y}|\mathbf{x}), p_m(\mathbf{y}|\mathbf{x},\theta)] + D[q(\mathbf{y}|\mathbf{x}), p_m(\mathbf{y}|\mathbf{x}+\rho,\theta)] \\ &\qquad s.t. \qquad \|\rho\|_p \leq \sigma, \end{aligned} \tag{16}$$

where $q(\mathbf{y}|\mathbf{x})$ defines the true posterior distribution; ρ is the perturbation added to the input within a small range; and σ is the small positive value. The nonnegative function $D[\cdot]$ is used to measure the divergence between the true posterior and model distribution. The objective of AT is to fit the true distribution with the model distribution on natural and adversarial examples.

3.2.1. Training Method

To implement AT, the inner maximization problem must first be solved to obtain the worst perturbation. The first Taylor expansion of $D[\cdot]$ is used as the approximation to simplify the optimization problem for the original objective function. With the l_p constraint, we can approximate the worst perturbation as follows:

$$\rho = \sigma sign(\nabla L)\left(\frac{|\nabla L|}{\|\nabla L\|_{p^*}}\right)^{\frac{1}{p-1}}, \tag{17}$$

where L is the loss function of DNN and p^* is the dual of p.

When $p = \infty$, this method can be degraded to FGSM [24]. Consequently, the worst perturbation becomes Equation (12). Details of the proof can be found in [30]. After computing the adversarial perturbation, the model distribution $p_m(\mathbf{y}|\mathbf{x},\theta)$ is trained to approximate the true distribution $q(\mathbf{y}|\mathbf{x})$ on natural and adversarial examples. It has been shown that AT can achieve better generalization and robustness than traditional training methods of DNNs [30].

4. Scenario Description and Implementation

The robustness of DL-based SEI under adversarial attacks is the main focus in this work. On the basis of the original SEI, i.e., Natural SEI, we have added two new scenarios, namely attack and defense. The schematic diagram of the scenes is shown in Figure 2, where the black devices represent legal devices and red one for illegal device. The lines in the same color represent the signals of the corresponding devices. And the green line represents a situation where a malicious rogue device conducts an adversarial attack.

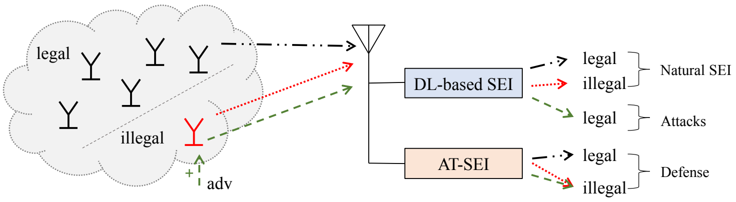


Figure 2. Scheme of DL-based SEI in three scenarios.

The goal of SEI is to stably identify emitters, especially distinguish legitimate and malicious devices, so as to ensure network security.

The three scenarios are described in detail below.

4.1. Natural SEI

Much of the current literature on DL-based SEI pays attention to improving the correct recognition rate of the system in the case of no adversarial attack [22,59]. This case is called as “Natural SEI” in this work.

As shown in Figure 2, there are registered legitimate devices (in black) and unregistered malicious device (in red) in the environment. In Natural SEI, the original DL-based SEI can correctly identify these devices and classify them as legal and illegal.

4.1.1. Model

The implementation framework of DL-based SEI is shown in Figure 3. There are three main steps. (1) At first, the actual received signal should be preprocessed. In this work, the input of the network is time-domain waveform of the IQ raw signal, so the preprocessing includes signal detection, filtering noise reduction, and energy normalization. (2) Input the preprocessed signal into the DNN, where the UM-related fingerprint feature extraction and classification identification are automatically completed. In the figure, $y(t)$ represents the pre-processed results of $r(t)$. (3) Finally, output the emitter ID of the specific emitter corresponding to the received signal.

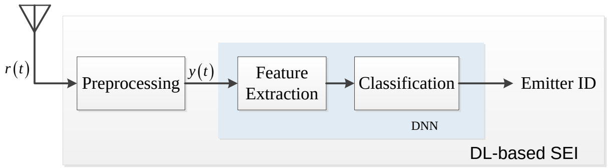


Figure 3. Implementation Framework of DL-based SEI.

4.1.2. Implementation

Improving the neural network structure to improve the correct recognition rate is not the focus of this paper, instead, it focuses on the impact assessment of adversarial attacks and AT methods on classical neural networks. Therefore, the experiments are based on the classic ResNet18, the network settings refer to [60]. The details of the network setup are shown in Figure 4. As shown , the activation function is chosen as ReLu, and N denotes the number of sample points of the signal samples. Other necessary signal processing procedures are according to the DL-based SEI given in Figure 3.

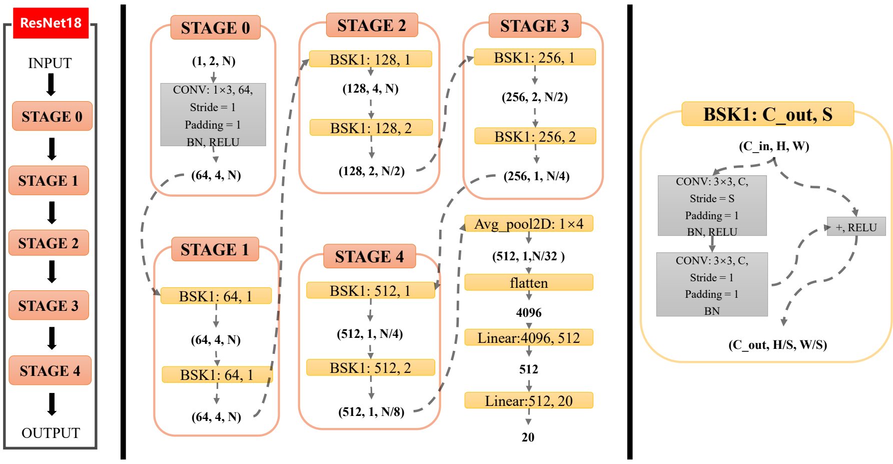


Figure 4. Deep learning network structure based on ResNet18 of SEI. *N* represents the data length.

All experiments in this paper are based on this network structure. Without loss of generality, the DNN is still used to denote the selected ResNet18 architecture in this following.

4.2. Attack: SEI under Adversarial Attacks

This is a new scenario proposed in this paper. This kind of attack s not an attack method in the traditional sense, such as a spoofing attack, nor does it change the hardware of the emitter. The specific process is to generate a very subtle adversarial perturbations according to the characteristics of the neural network, and then attach the adversarial perturbations to the signal to be transmitted. Adding this perturbation signal will cause the original DL-based SEI system to misjudge the red illegal device as a legal device. This completes the attack process.

The green line associated with DL-based SEI in Figure 2 represents the scenario where a malicious device (red) emits a signal with a specially designed perturbation that makes the original DL-based SEI system misjudgment.

4.2.1. Model

When faced with an adversarial attack, the signal can be expressed as

$$\tilde{r}(t) = r(t) + \delta(t), \tag{18}$$

where $\delta(t)$ denotes the subtle perturbation introduced by the adversarial attack (short as "adv" in the following), $r(t)$ represents the received signal without adversarial attacks in Equation (1), and $\tilde{r}(t)$ is referred to as the adversarial example in the following.

To express the strength of the adversarial attack, we define the perturbation-to-signal ratio (PSR) as an evaluation metric, denoted as the energy ratio of the perturbation $\delta(t)$ to the received signal $r(t)$,

$$\text{PSR} = 10\log\left(\frac{\int_0^T \delta^2(t)dt}{\int_0^T r^2(t)dt}\right). \tag{19}$$

4.2.2. Implementation

In this work, three kinds of adversarial perturbations are directly generated according to Section 3.1.3. Then add the digital adversarial perturbations before DAC, shown in Figure 5. In the experiment, the adversarial perturbations are directly added to the received signals. The signal with adversarial perturbation is aforementioned adversarial example $\tilde{r}(t)$. Finally, input the adversarial examples to the trained DNN model to test the change of the recognition performance for DL-based SEI under adversarial attacks.

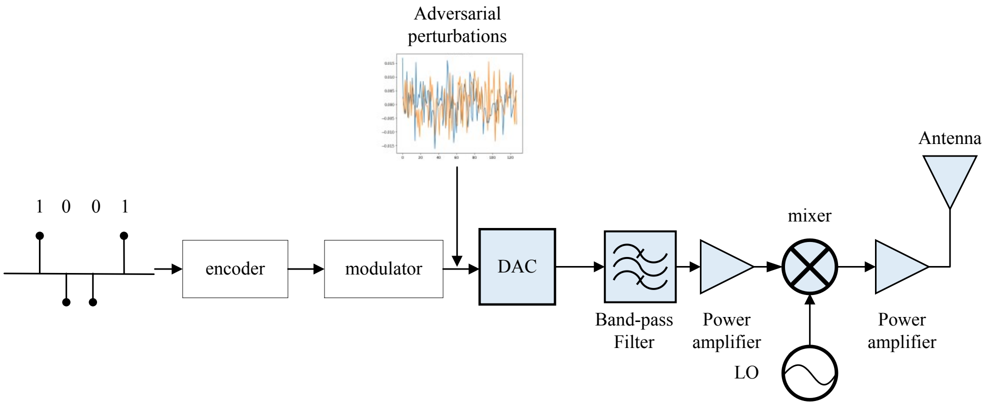


Figure 5. Scheme of the addition of subtle adversarial perturbation at the emitter.

4.3. Defense: AT-SEI under Adversarial Attacks

This is another scenario proposed in this work, which is to defend against the above-mentioned attacks. The specific process is to add an adversarial training link in the training procedure of DL-based SEI to improve the stability of the system against adversarial attacks.

In Figure 2, AT-SEI represents the DL-based SEI system after AT, and the green line associated with AT-SEI corresponds to this scenario. After AT, AT-SEI can correctly identify illegal devices. That is, defense means (AT) can ensure that the DNN model works properly to identify the correct emitter identity (ID).

Implementation

We choose the single-step gradient attack FGSM as the way to generate adversarial examples, and add adversarial examples to the training set to retrain in this work. The training schematic is shown in Figure 6. The training method is introduced in Section 3.2.1. And the other related settings can be kept consistent with the previous two scenarios.

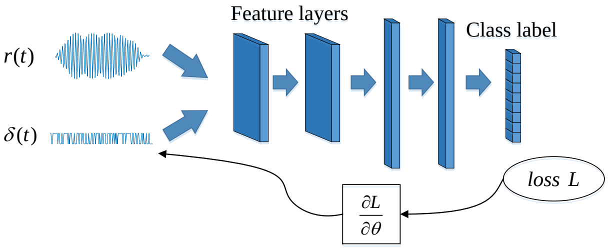


Figure 6. Diagram of AT for DL-based SEI.

In addition, the generation settings of adversarial attacks are also kept the same as Section 4.2.

To sum up, scheme of DL-based SEI including all three different scenarios is shown in Figure 7.

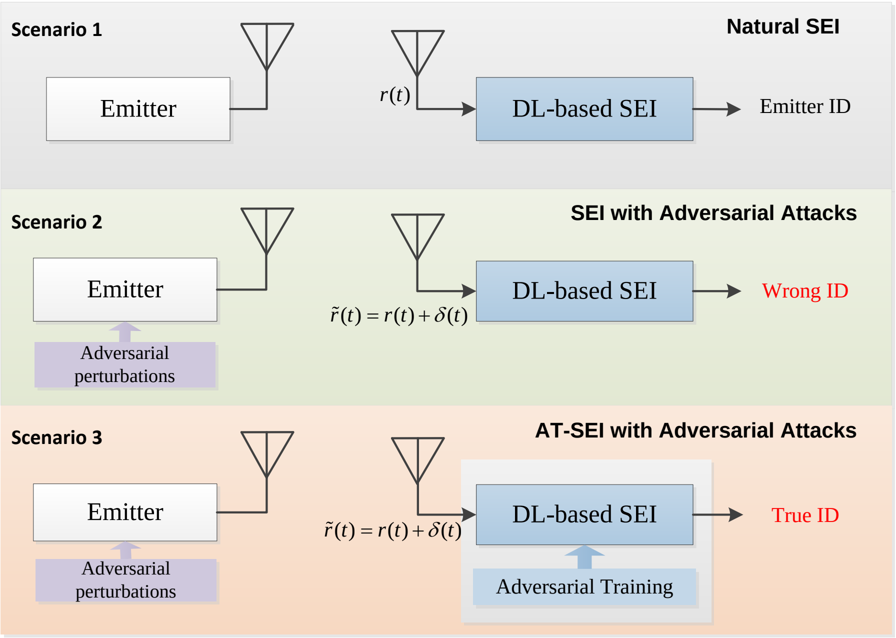

Scenario 1
Natural SEI
Emitter
$r(t)$
DL-based SEI
Emitter ID
Scenario 2
SEI with Adversarial Attacks
Emitter
Adversarial perturbations
$\tilde{r}(t) = r(t) + \delta(t)$
DL-based SEI
Wrong ID
Scenario 3
AT-SEI with Adversarial Attacks
Emitter
Adversarial perturbations
$\tilde{r}(t) = r(t) + \delta(t)$
DL-based SEI
Adversarial Training
True ID

Figure 7. Scheme of DL-based SEI in three different scenarios.

5. Experimental Results

In this section, we first introduce the situation of data and related configurations. There are two types of datasets, radar data (Dataset 1) and communication data (Dataset 2), each containing 20 emitters to be identified. Experiments are then conducted primarily on Dataset 1 for adversarial attack and defense training, i.e., AT, and repeated on Dataset 2 for validation. Both sets of experiments use the same DNN structure (ResNet18) and are trained independently.

All experiments are performed on an NVIDIA GeForce GTX 3090 using on GPU per run. Each of the four attack methods is implemented using pytorch1.10 and cuda11.3. We use Adam with learning rate = 1×10^{-5} to optimize the target model, and the ReLU activation function was used in all layers, and the categorical-crossentropy loss functions are tried in the experiments. We divide the whole dataset into the training set, validation set and test set according to the ratio of 7:2:1. The training set is used to adjust the model parameters. After each training epoch, the model is validated on the validation set to get the validation loss. Besides, we adopt an early-stopping training strategy, i.e., we stop training when the validation loss is no longer decreasing after several epochs of training, and the model with the lowest validation loss is used as the optimal model. Finally, we test the optimal model on the test set and get all the experimental results. The three datasets are completely independent, so the results on the test set can guarantee that the model is not obtained in an overfitting situation.

5.1. Dataset and Configure

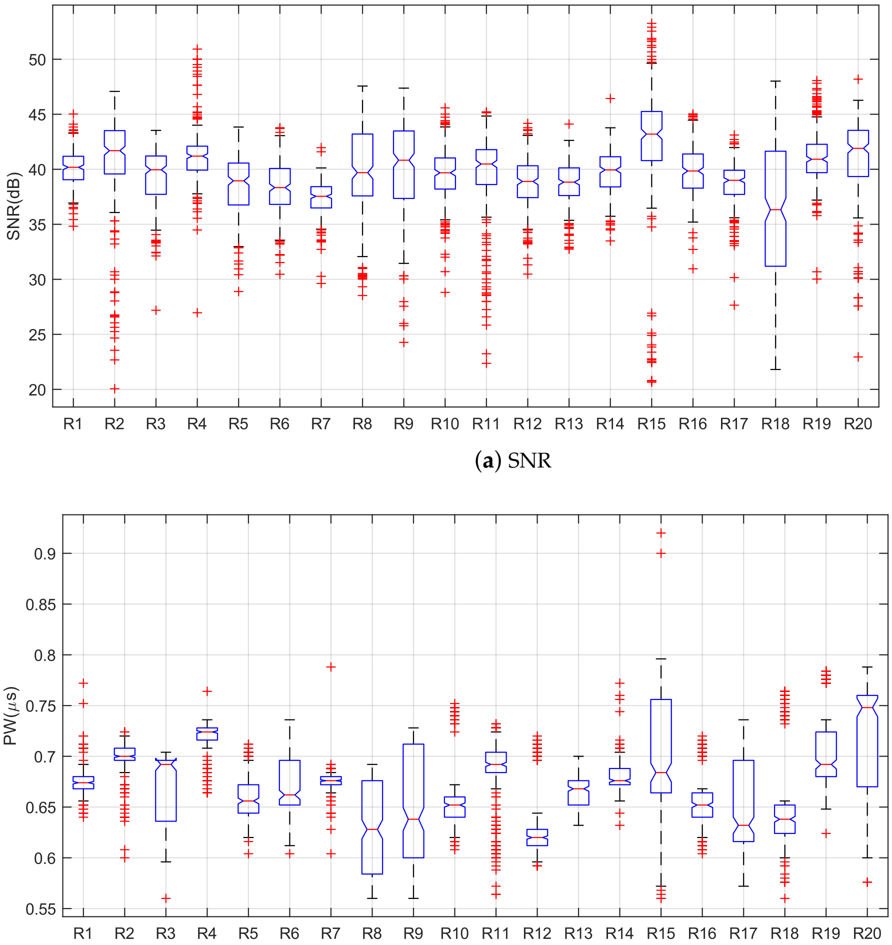
5.1.1. Dataset 1

As for the radar signals, they are emitted by 20 secondary radars of 20 different civil aircrafts, labeled as R1 to R20. There are more than 240 pulse samples of each emitter, and the detailed pulse numbers are listed in Table 1. And there are 6267 pulses in total and 264 points for each sample. The RF and intermediate frequency (IF) are 1090 and 60 MHz, respectively. The sampling frequency is $f_s = 250$ MHz.

Table 1. Samples of Each Emitter for Dataset 1.

Emitter ID	Sample	Emitter ID	Sample
R1	334	R11	321
R2	398	R12	302
R3	417	R13	280
R4	398	R14	308
R5	262	R15	303
R6	300	R16	292
R7	304	R17	300
R8	241	R18	301
R9	305	R19	304
R10	289	R20	306

The signal-to-noise ratio (SNR) and the pulse width (PW) distribution of all samples are shown in Figure 8. It can be seen the SNR of most pulses is above 30 dB. And the majority of samples range in pulse width from 0.6 μs to 0.7 μs.



(**a**) SNR

(**b**) Pulse Width

Figure 8. SNR and pulse width distribution of real-world signals of the 20 radars in Dataset 1.

5.1.2. Dataset 2

According to the device distortion models of Section 2.1, 20 different emitters are simulated, which are represented by T1-T20, and the specific parameter settings are shown in Table 2.

Signals of Quadrature Phase Shift Keying (QPSK) modulation are generated. The shaped filter is a square root rising cosine-shaped filter with a roll-off factor set to 0.35. The sampling rate (f_s) is 20 MHz, the code rate (f_z) is 1 MHz, and the carrier frequency (f_0) is 5 MHz. There are 3000 samples per emitter and 60,000 samples in total. Each sample has 1000 points, containing 50 random codes. Here, 80% samples of each emitter are randomly selected for training, and the rest are used as the test set. These parameters are summarized in Table 3.

Table 2. Distortion parameters of simulated communication transmitters for Dataset 2.

Emitter ID	**FD**		**IQE**		**CST**			**PAD**		
	(a_0, a_1, T_A)	(b_0, b_1, T_B)	ρ	ξ	A^{ST}	f^{ST}	$\varsigma^{\mathrm{CL}}(10^{-3})$	d_1	d_2	d_3
T1	(1, 0.06, 4)	(1, 0.0309, 4)	0.9996	−0.016	0.0072	0.0197	$1.1 + 9.8i$	1	0.03	0.1
T2	(1, 0.085, 4)	(1, 0.0308, 4)	0.9997	−0.014	0.0074	0.0193	$1.2 + 9.6i$	1	0.05	0.09
T3	(1, 0.073, 4)	(1, 0.0307, 4)	0.9998	−0.012	0.0076	0.0189	$1.3 + 9.4i$	1	0.07	0.08
T4	(1, 0.04, 4)	(1, 0.0306, 4)	0.9999	−0.010	0.0078	0.0185	$1.4 + 9.2i$	1	0.09	0.07
T5	(1, 0.06, 4)	(1, 0.0305, 4)	1.0000	−0.008	0.0080	0.0181	$1.5 + 9.0i$	1	0.11	0.06
T6	(1, 0.085, 4)	(1, 0.0304, 4)	1.0001	−0.006	0.0082	0.0177	$1.6 + 8.8i$	1	0.13	0.05
T7	(1, 0.073, 4)	(1, 0.0303, 4)	1.0002	−0.004	0.0084	0.0173	$1.7 + 8.6i$	1	0.15	0.04
T8	(1, 0.04, 4)	(1, 0.0302, 4)	1.0003	−0.002	0.0086	0.0169	$1.8 + 8.4i$	1	0.17	0.03
T9	(1, 0.03, 4)	(1, 0.0301, 4)	1.0004	0.000	0.0088	0.0165	$1.9 + 8.2i$	1	0.19	0.02
T10	(1, 0.03, 4)	(1, 0.0300, 4)	1.0005	0.002	0.0090	0.0161	$2.0 + 8.0i$	1	0.21	0.01
T11	(1, 0.03, 4)	(1, 0.0299, 4)	1.0006	0.004	0.0092	0.0157	$2.1 + 7.8i$	1	0.25	0
T12	(1, 0.03, 4)	(1, 0.0298, 4)	1.0007	0.006	0.0094	0.0153	$2.2 + 7.6i$	1	0.3	−0.01
T13	(1, 0.03, 4)	(1, 0.0297, 4)	1.0008	0.008	0.0096	0.0149	$2.3 + 7.4i$	1	0.35	−0.02
T14	(1, 0.03, 4)	(1, 0.0296, 4)	1.0009	0.010	0.0098	0.0145	$2.4 + 7.2i$	1	0.4	−0.03
T15	(1, 0.03, 4)	(1, 0.0295, 4)	1.0010	0.012	0.0100	0.0141	$2.5 + 7.0i$	1	0.45	0.1
T16	(1, 0.03, 4)	(1, 0.0294, 4)	1.0011	0.014	0.0102	0.0137	$2.6 + 6.8i$	1	0.5	0.2
T17	(1, 0.03, 4)	(1, 0.0293, 4)	1.0012	0.016	0.0104	0.0133	$2.7 + 6.6i$	1	0.55	0.3
T18	(1, 0.03, 4)	(1, 0.0292, 4)	1.0013	0.018	0.0106	0.0129	$2.8 + 6.4i$	1	0.6	0.4
T19	(1, 0.03, 4)	(1, 0.0291, 4)	1.0014	0.020	0.0108	0.0125	$2.9 + 6.2i$	1	0.65	0.5
T20	(1, 0.03, 4)	(1, 0.0290, 4)	1.0015	0.022	0.0110	0.0121	$3.0 + 6.0i$	1	0.7	0.6

Table 3. Parameters of communication signals in Dataset 2.

Parameter	**Value**	**Parameter**	**Value**
Sampling Rate	20 MHz	Modulation	QPSK
Symbol Rate	5 MHz	Roll-off factor	0.35
Carrier Frequency	1 MHz	Codes/sample	50
Emitter ID	T1–T20	Sample/emitter	3000

5.1.3. Configure

Recognition Rate

In this work, the correct recognition rate is used as the evaluation index. And the reliability is evaluated by using the change of the recognition rate in different scenarios. The overall recognition rate is defined as P:

$$P = \sum_{i=1}^{I} N_i^c / \sum_{i=1}^{I} N_i^t, \tag{20}$$

where N_i^c denotes the number of correctly identified samples of the ith emitter, and N_i^t represents the number of tested samples. The recognition rate of the ith emitter is given as

$$P_i = N_i^c / N_i^t. \tag{21}$$

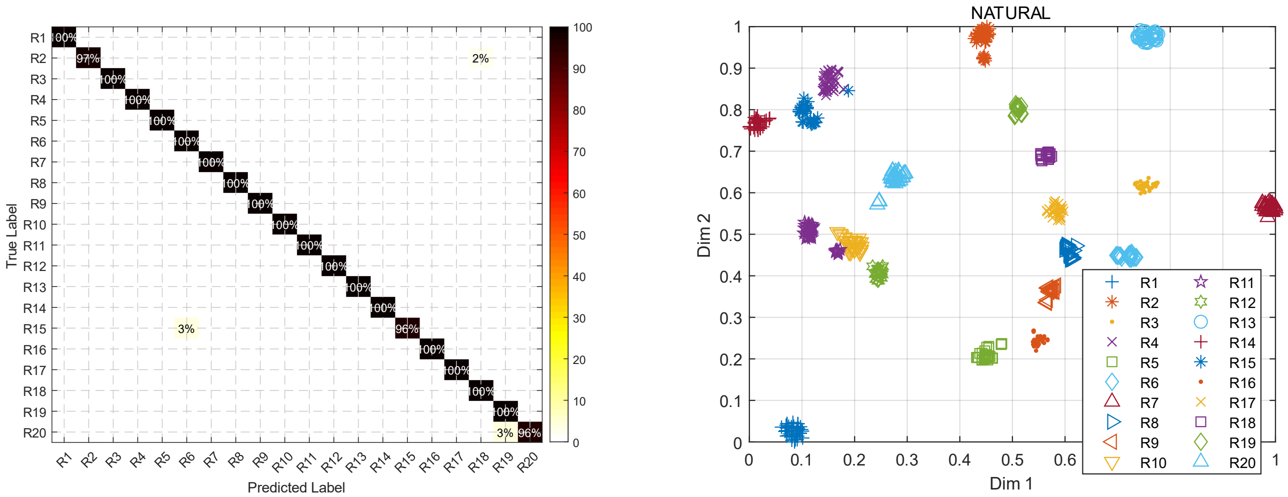
Parameter

The learning rate for Dataset 1 is 5×10^{-5} and that for Dataset 2 is 1×10^{-6}. The batch size for Dataset 1 is 32, and 16 for Dataset 2. The patience of early stopping is set to 20. The noise in this work is addition white gaussian noise (AWGN).

5.2. Experiment on Dataset 1

5.2.1. Natural SEI

Figure 9 shows the confusion matrix of ResNet18 after training for all targets on the validation set, with an average recognition rate of 99.52%, and the distribution of extracted features after T-sne dimensionality reduction [61]. The classification boundaries of features of different emitter are obvious and relatively dispersed in the feature space, which can achieve accurate identification, proving that the neural network is well trained and can achieve the SEI task without perturbation.



(**a**) Confusion matrix (**b**) Low-dimensional feature distribution

Figure 9. Identification performance and the feature distribution of directly trained DNN for Dataset 1. The average identification accuracy is $P = 99.522\%$.

Floating Point Operations Per Second (FLOPS) is often used to measure the time complexity in neural network. The FLOPS of the whole neural network is $O\left(\sum\limits_{l=1}^{D} M_{l-1}^2 K_l^2 C_{l-1} C_l\right)$, where D is the depth of CNN, M and K is the length of feature map and kernel of L-th layer, C is the number of channel. In our model, the FLOPS is 1.429Gflops, and the parameters size is 13.27522 M.

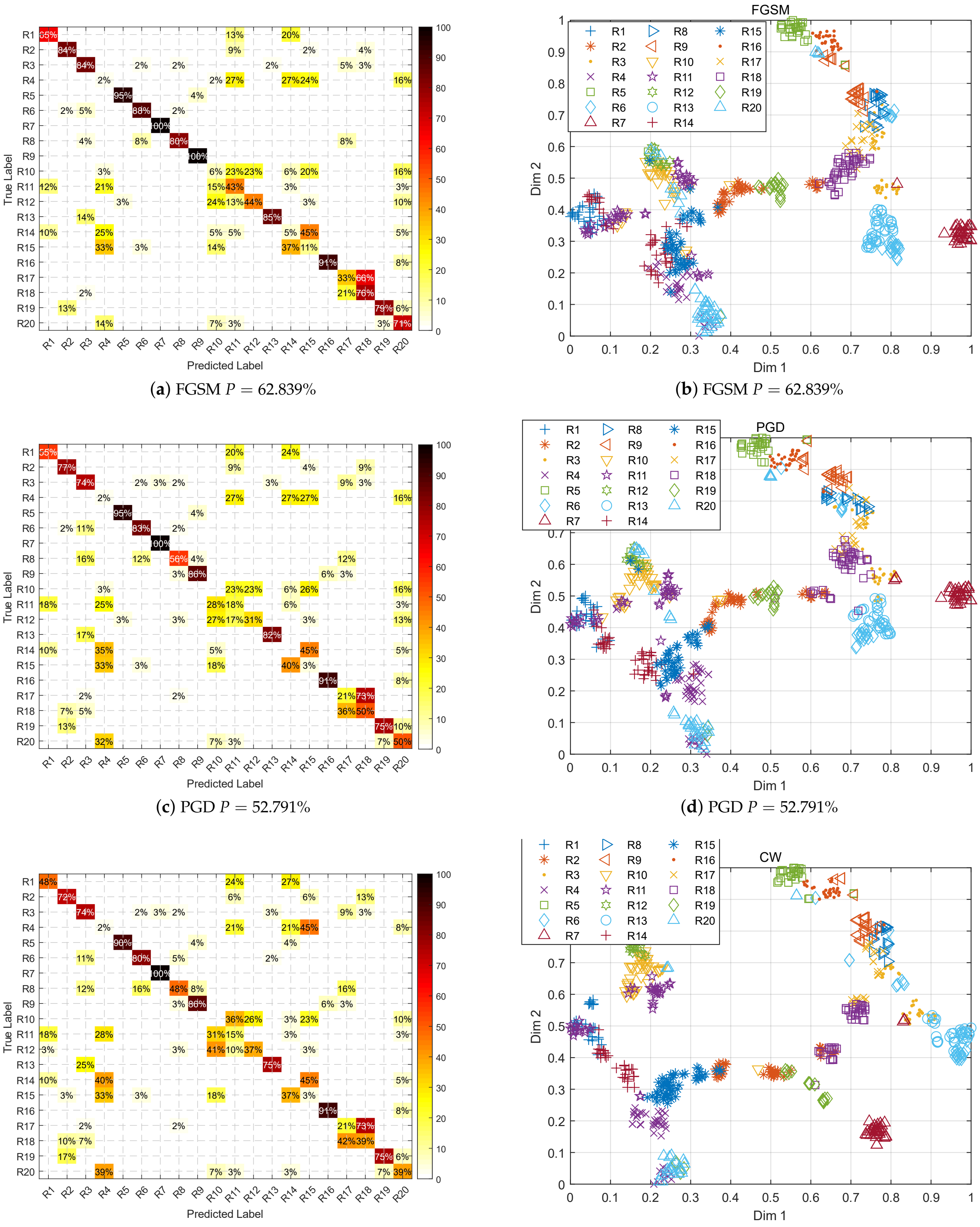
5.2.2. Adversarial Attacks

We test the recognition performance of directly trained DNNs when facing three types of adversarial attacks.

Overall Performance

The specific recognition of each emitter is shown in the confusion matrices in Figure 10. The recognition performance decreases by more than 40% when the adversarial sample strength is at −32 dB (for Dataset 1, the case with PSR = −32 dB is chosen as the focus of the analysis, because it is relatively representative (not the best nor the worst, as shown in Figure 11 below)). The average correct recognition rate of the targets decreases to 61.72%, 51.20%, and 51.36% under FGSM, PGD, and C&W attacks, respectively, indicating that the system is no longer working properly.

After adding an adversarial attack under PSR = −32 dB, different emitters are misclassified, which is manifested in the confusion matrix as the appearance of more samples for elements beyond the diagonal, and different emitters are affected by the adversarial attack to different degrees. The degree of impact of the three different attacks on different targets also varies. Under the C&W attack, the performance dropped the most, from 99.52% to 50.239%, with a drop of nearly 50%.



(**a**) FGSM $P = 62.839\%$

(**b**) FGSM $P = 62.839\%$

(**c**) PGD $P = 52.791\%$

(**d**) PGD $P = 52.791\%$

(**e**) C&W $P = 50.239\%$

(**f**) C&W $P = 50.239\%$

Figure 10. Identification performance and the feature distribution after attacks of PSR = −32 dB for Dataset 1. The figures of the first column represent the confusion matrices, and the second column is for low-dimensional feature distribution.

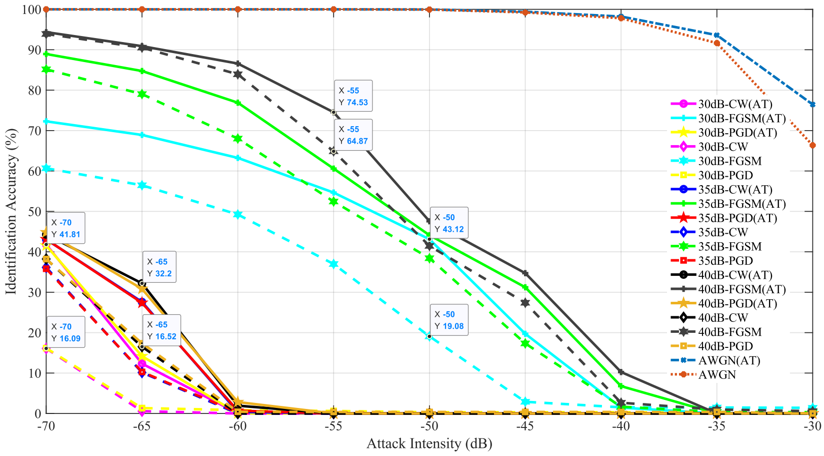


Figure 11. Recognition effect of the system with and without AT for different attack strengths, different attack types and different noise levels for Dataset 2.

Compared with Figure 9, the feature boundaries of different emitters in Figure 10 become blurred and mixed in distribution, making it difficult to distinguish accurately and proving the fragility of the DL-based SEI.

Adding adversarial examples causes the distribution of features to shift, and the feature shifts are not consistent. For example, in the case of no attack (Figure 9), the features of R19 are concentrated in the middle upper part of the feature plane without overlapping with other emitters, while after FGSM and PGD attacks the features are concentrated in the middle of the whole feature plane and have overlapping with R14. And the features of R19 after C&W attack are distributed in the lower middle of the whole feature plane and scattered into three parts, which are not concentrated.

Emitter

The recognition results of each emitter under different types of attacks are shown in Figure 12. In the absence of attacks, the correct recognition rate of each emitter can reach almost 100%; in the face of attacks, the recognition effect decreases significantly. However, the impact on different targets varies, with R7 being almost unaffected (almost 100%), while all test samples of R4, R10, R14, and R15 are misidentified. This is due to the fact that the differences between emitters are inherently subtle and unevenly distributed in the feature space, and thus are affected to different degrees by the adversarial attacks. Although there are several relatively robust targets under the three attacks, the overall recognition rate is only 50–60%, and the system is actually not working at all.

Waveform

As aforementioned, identification accuracy decrease dramatically at PSR = −32 dB, the corresponding waveforms of the three attack methods are shown in Figure 13. Original waveforms and waveforms with the added adversarial attack of the three attack methods (waveform + adv) are shown on the left column. The second column of subplots shows added adversarial perturbations (adv). The probability density function (PDF) of the waveforms with and without adversarial perturbation are given in the third column. The PDF is fitted by the frequency distribution histogram of the signal [54].

It can be seen from the figures:

- The overall signal waveform after adding the attack is similar to the original signal. The difference is so subtle that the human eye can-not detect the difference; however, these perturbation can make the recognition performance of SEI significantly worse.
- The waveform of the perturbation is not significantly regular. The normalized energy amplitude does not exceed 0.02, which is very low compared with the original signal.

- The PDF of the waveform did not change significantly before and after being attacked.

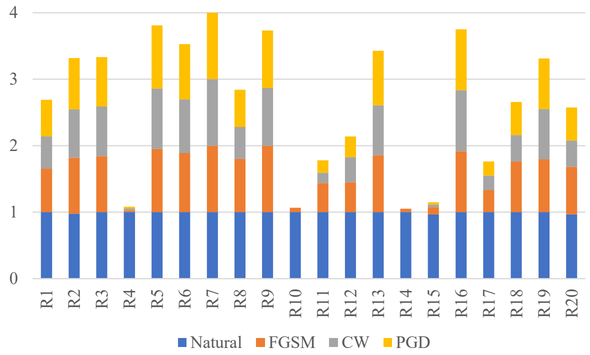


Figure 12. Recognition results of each emitter under different types of attacks at a strength of −32 dB for Dataset 1. The horizontal axis indicates different emitters, and the vertical axis indicates the corresponding recognition rate values. Different color blocks indicate the recognition effect of different cases, the length of the blocks is proportional to the size of the correct recognition rate. The longer the length, the higher the recognition rate. The values range from 0–1 (100%). The total range is 0–4 for all four cases.

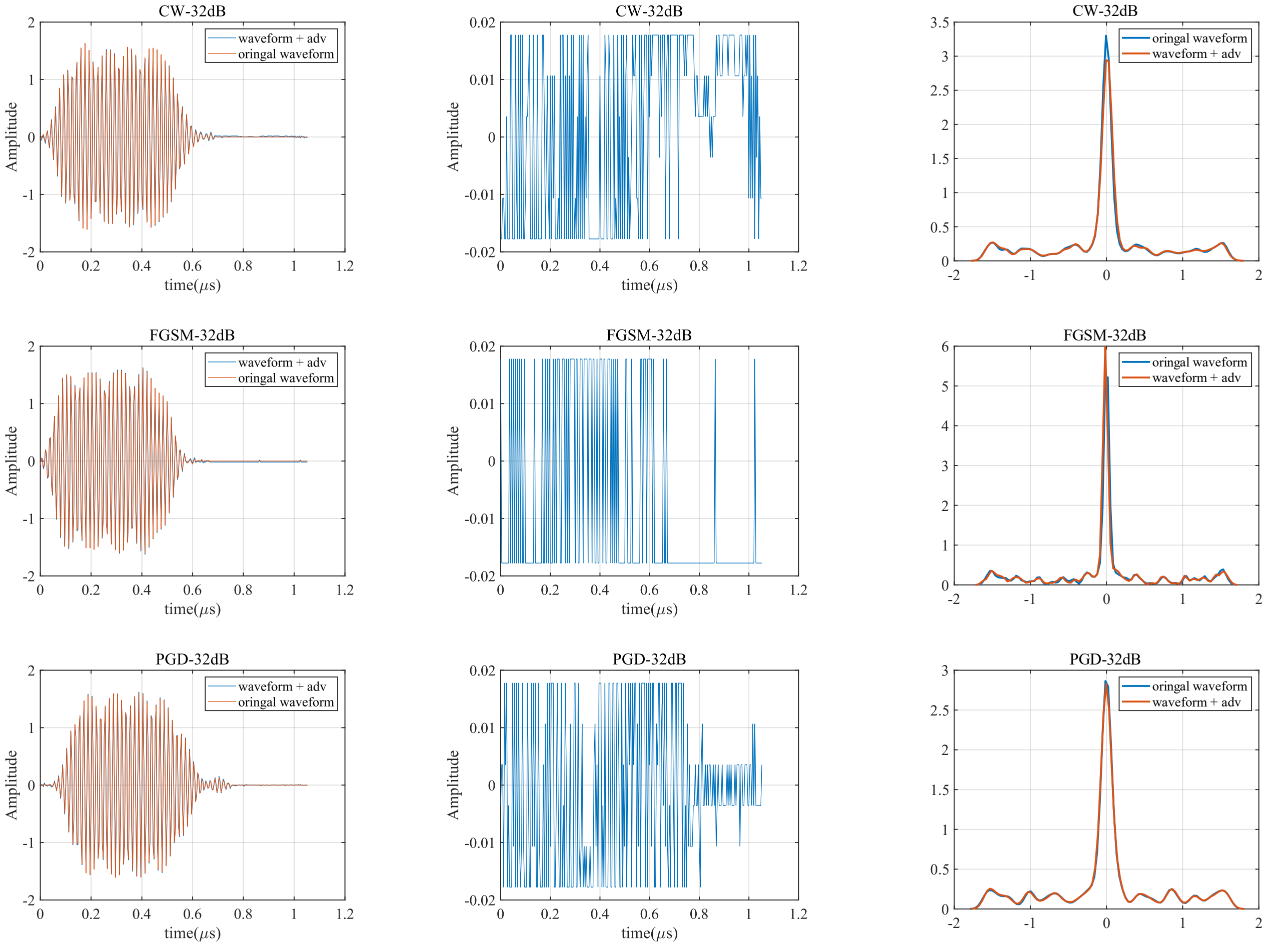


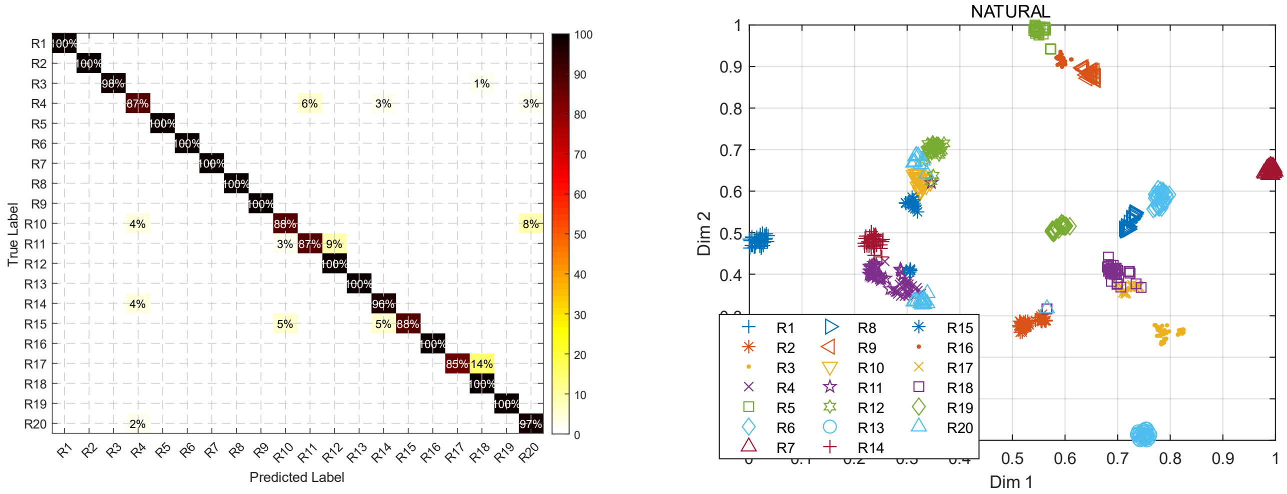
Figure 13. Waveforms of the three attack methods at PSR = −32 dB. The original waveforms and waveforms with the added adversarial attack of the three attack methods (waveform + adv) are shown in the left column. The second column of subplots shows the waveforms of the added adversarial examples. The third column shows the probability density function (PDF) of the waveforms.

Experiments show that the waveform with adv is not significantly different from the original waveform, and only a very low energy perturbation needs to be added to disable the system compared to the original signal.

In practice, if these subtle adversarial perturbations are artificially added to the malicious device signal, they are difficult to detect and can prevent a specific SEI system from working properly, seriously affecting network security, so targeted robustness enhancement is essential.

5.2.3. Performance after AT

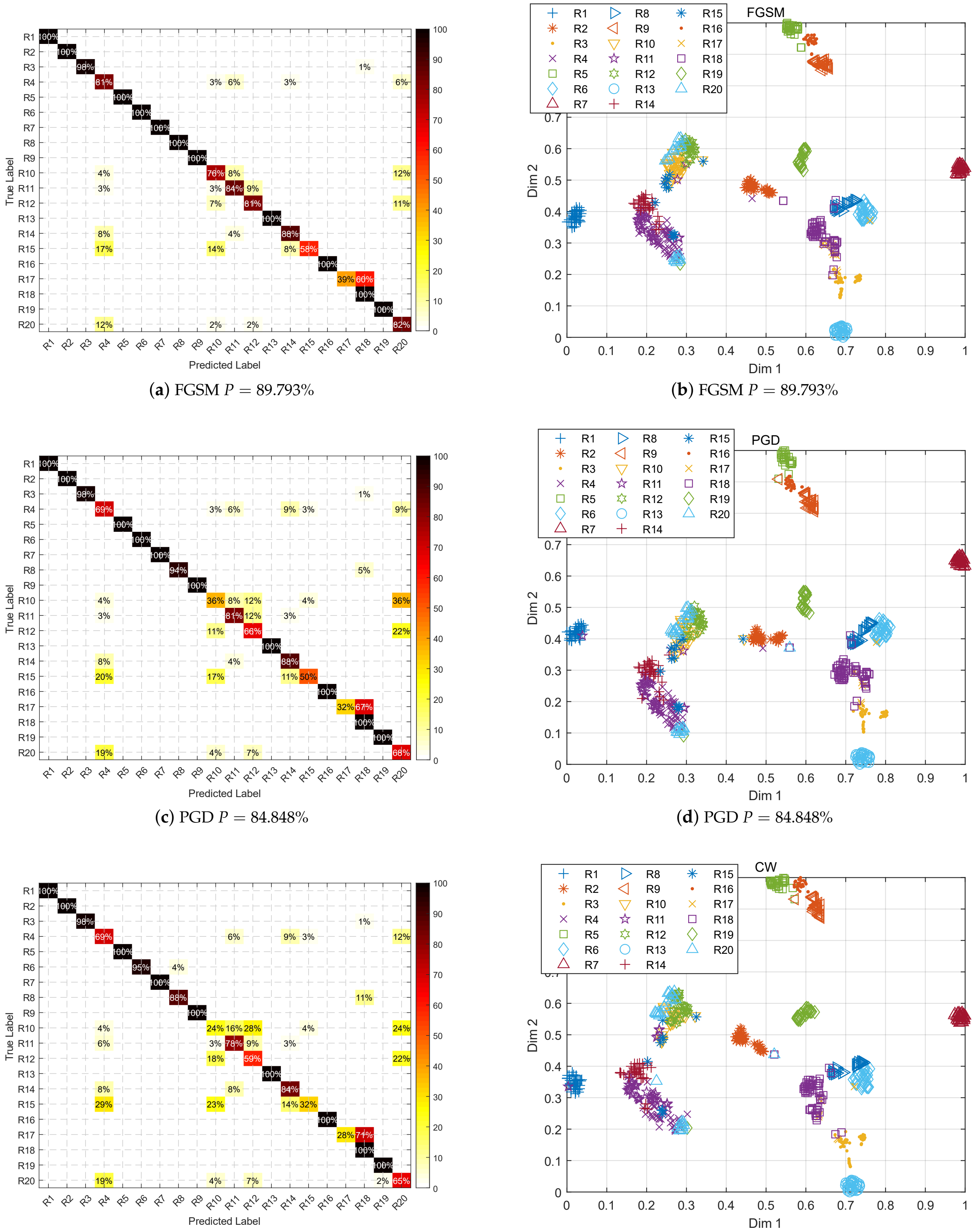
To improve the robustness against adversarial attacks, the same structure of ResNet18 network as above is trained with AT. In the absence of an adversarial attack, the average recognition rate of the system on the test set is 96.491%, and the confusion matrix of 20 emitters is shown in the first subfigure of Figure 14. The second subfigure in Figure 14 visualizes the feature distribution for different emitters.



(**a**) Confusion matrix (**b**) Low-dimensional feature distribution

Figure 14. Identification performance and the feature distribution of the DNN after AT for Dataset 1. The average identification accuracy is $P = 96.491\%$.

Faced with adversarial attacks of the same intensity as that of Section 5.2.2, the corresponding confusion matrices of DNN after AT are shown in Figure 15, with an average recognition rate of 89.79%, 82.14%, and 84.85% for the three cases, respectively. Under the three adversarial attacks, the corresponding confusion matrices are shown in Figure 15. Although the performance is slightly reduced compared to the result of AT without attacks in Figure 14, the improvement is obvious for each emitter compared with Figure 10. According to corresponding feature distributions after attacks, shown in the second column of Figure 15, the effectiveness of AT is pretty obvious compared with Figure 10. From the perspective of the confusion matrix, when there are adversarial examples after AT, most of the values are concentrated on the diagonal line, indicating that they can basically be correctly identified; From the scatter plot, the scatter distribution of each emitter is more concentrated.



(**a**) FGSM $P = 89.793\%$

(**b**) FGSM $P = 89.793\%$

(**c**) PGD $P = 84.848\%$

(**d**) PGD $P = 84.848\%$

(**e**) C&W $P = 82.137\%$

(**f**) C&W $P = 82.137\%$

Figure 15. Identification performance and the feature distribution with attacks of PSR = −32 dB after AT for Dataset 1. The figures of the first column represents the confusion matrices, and the second column is for low-dimensional feature distribution.

Table 4 summarizes the comparison of DNN with or without AT. Although the AT performance slightly decreases in the absence of an attack (−3.03%), it is still >96% and does not affect normal recognition. However, the performance improvement in the adversarial case is dramatic. When PSR = −32 dB, the average accuracy under three attacks before and after AT are 55.29% and 85.59%, respectively, with an average improvement of 30.30%. The system recognition rate under all attacks are maintained at more than 82%; the basic recognition function is recovered.

Table 4. Performance comparison under PSR = −32 dB for Dataset 1(%).

Method	No AT	AT	Improvement
Natural	99.52	96.49	−3.03
FGSM	62.84	89.79	26.95
C&W	50.24	82.14	31.90
PGD	52.79	84.85	32.06
Average (attacks)	55.29	85.59	30.30

Performance of Each Emitter

The statistics of each emitter device with and without AT under attacks at PSR = −32 dB are shown in Table 5, where "Average" indicates the average value in the three attack cases. It can be seen that the overall law of the effect of attack and AT is consistent for all emitters. That is, the recognition effect decreases when facing the attack, and improves significantly after AT.

Table 5. Identification Performance of Each Emitter under Attacks with the Intensity of PSR = −32 dB (%).

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Natural	100.00	97.67	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
FGSM	65.52	84.09	84.31	2.70	95.24	88.89	100.00	80.00	100.00	6.67
C&W	48.28	72.73	74.51	2.70	90.48	80.56	100.00	48.00	86.67	0.00
PGD	55.17	77.27	74.51	2.70	95.24	83.33	100.00	56.00	86.67	0.00
Average	67.24	82.94	83.33	27.03	95.24	88.19	100.00	71.00	93.33	26.67
Natural (AT)	100.00	100.00	98.18	87.88	100.00	100.00	100.00	100.00	100.00	88.00
FGSM (AT)	100.00	100.00	98.18	81.82	100.00	100.00	100.00	100.00	100.00	76.00
C&W (AT)	100.00	100.00	98.18	69.70	100.00	95.65	100.00	88.89	100.00	24.00
PGD (AT)	100.00	100.00	98.18	69.70	100.00	100.00	100.00	94.44	100.00	36.00
Average (AT)	100.00	100.00	98.18	77.27	100.00	98.91	100.00	95.83	100.00	56.00
	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20
Natural	100.00	100.00	100.00	100.00	96.43	100.00	100.00	100.00	100.00	96.77
FGSM	43.75	44.83	85.71	5.00	11.11	91.67	33.33	76.32	79.31	71.43
C&W	15.63	37.93	75.00	0.00	3.70	91.67	21.43	39.47	75.86	39.29
PGD	18.75	31.03	82.14	0.00	3.70	91.67	21.43	50.00	75.86	50.00
Average	44.53	53.45	85.71	26.25	28.74	93.75	44.05	66.45	82.76	64.37
Natural (AT)	87.88	100.00	100.00	96.00	88.24	100.00	85.71	100.00	100.00	97.56
FGSM (AT)	84.85	81.48	100.00	88.00	58.82	100.00	39.29	100.00	100.00	82.93
C&W (AT)	78.79	59.26	100.00	84.00	32.35	100.00	28.57	100.00	100.00	65.85
PGD (AT)	81.82	66.67	100.00	88.00	50.00	100.00	32.14	100.00	100.00	68.29
Average (AT)	83.33	76.85	100.00	89.00	57.35	100.00	46.43	100.00	100.00	78.66

On the other hand, we can see that different attacks have different effects on different emitter devices. For example, for R10, the recognition result is 100% after direct training, however, the recognition rate drops rapidly to single digits in the presence of attacks. After AT, although the direct recognition rate dropped to 88%, the performance improved significantly with the addition of adversarial examples, and the average correct rate recovered from 26.67% to 56.00% and 76% on FGSM. And after AT, nine emitters can reach $P_i = 100\%, i \in \{1, 2, 5, 7, 9, 13, 16, 18, 19\}$ under all three attacks.

The performance of different devices with or without AT is visually displayed in Figure 16. It can be concluded that the recognition rate of some devices will drop slightly after adversarial training, but the recognition performance improves significantly when facing adversarial attacks.

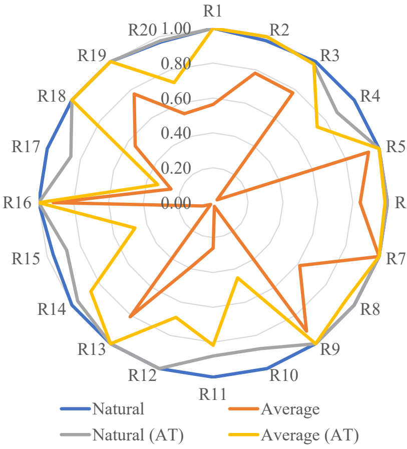


Figure 16. Identification performance for each emitter after AT under three attack scenarios with PSR = −32 dB for Dataset 1. "Average" denotes the average identification accuracy of original DNN under three attacks. "Average (AT)" denotes the average identification accuracy of the DNN after AT under three attacks.

5.2.4. Intensity Analysis

Figure 17 shows the recognition performance of DNN with and without AT in the presence of three attacks with PSR ranging from −40 to 0 dB for Dataset 1. For comparison, the performance under the influence of the additive white noise of the corresponding strength is also tested. To be consistent with the PSR, the ratio of noise-to-signal ratio (NSR) is used here to express the AWGN intensity,

$$\mathrm{NSR} = \frac{1}{\mathrm{SNR}} = 10 \log \left[\frac{\frac{1}{T}\int_0^T s^2(t)dt}{\frac{1}{T}\int_0^T v^2(t)dt} \right]. \tag{22}$$

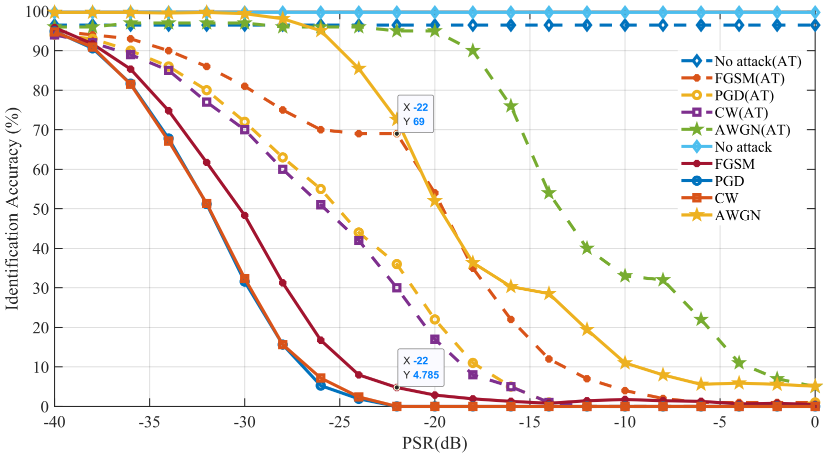


Figure 17. Recognition effect of the system with and without AT for different attack strengths, different attack types and noise levels for Dataset 1.

The solid lines in the figure represent no adversarial training, and the dashed line represents the system after AT. Overall, the dashed lines are generally above the solid ones, proving the effectiveness of AT. The following conclusions can be obtained from the figure:

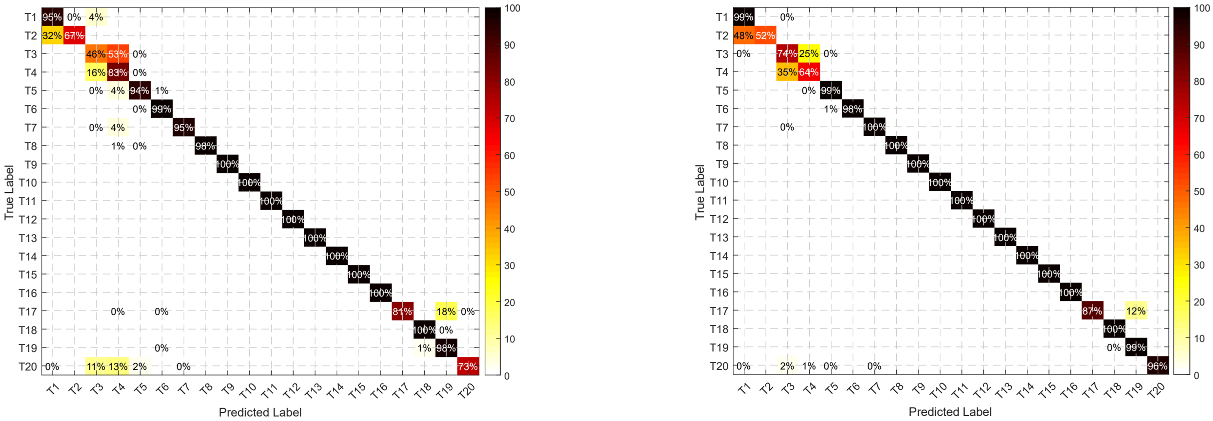
1. The directly trained DNN can achieve a highly accurate recognition rate without attack. However, in the presence of an adversarial attack, the performance degrades rapidly and drops below 10% at −25 dB, demonstrating the vulnerability of the DNN-based SEI.
2. It can be seen that the result curve of the adversarial attack is at the bottom of the figure. All three attack methods are effective for DNN-based SEI, and minor perturbations can cause significant performance degradation, for their result lines are on the bottom of the figure. Among them, C&W and PGD attacks are more effective (because the implement of C&W is based on PGD framework in Section 3.1.3), their performance are similar, and FGSM has a slightly lesser effect on the performance. It is worth noting that the line of "AWGN" is higher than the lines of adversarial attacks, it means that the adversarial examples are more destructive to DNN than the white noise of the same strength.
3. After AT, the performance of direct recognition is slightly reduced compared to that without AT but can still reach 96.49%. However, the DNN's performance is significantly improved compared to that of direct training when facing adversarial attacks. In particular, under the FGSM attack, the accurate recognition rate of 69% at −22 dB is still achieved, as marked in the figure. Compared with the case without AT, there is an improvement of more than 60%. Moreover, the line of "AWGN(AT)" is obviously higher that "AWGN", it means that AT can improve the neural network's robustness to additive Gaussian white noise.
4. The system fails even after AT when a well-designed adversarial perturbation with PSR > -10 dB is added. It shows that the trained DNN for SEI is concerned with the fine-grained variability at that level on Dataset 1. Additionally, the RF fingerprints between emitters are very subtle. Thus, it reaches a performance boundary when faced with an adversarial attack above that strength, losing the recognition effect.

5.3. Validation on Dataset 2

Based on Dataset 2, the above experiments on Dataset 1 are all repeated at SNR = 30, 35, 40 dB, respectively. The experimental results are shown in Figure 11. As can be seen from the figure, similar conclusions can be obtained on Dataset 2 as on Dataset 1.

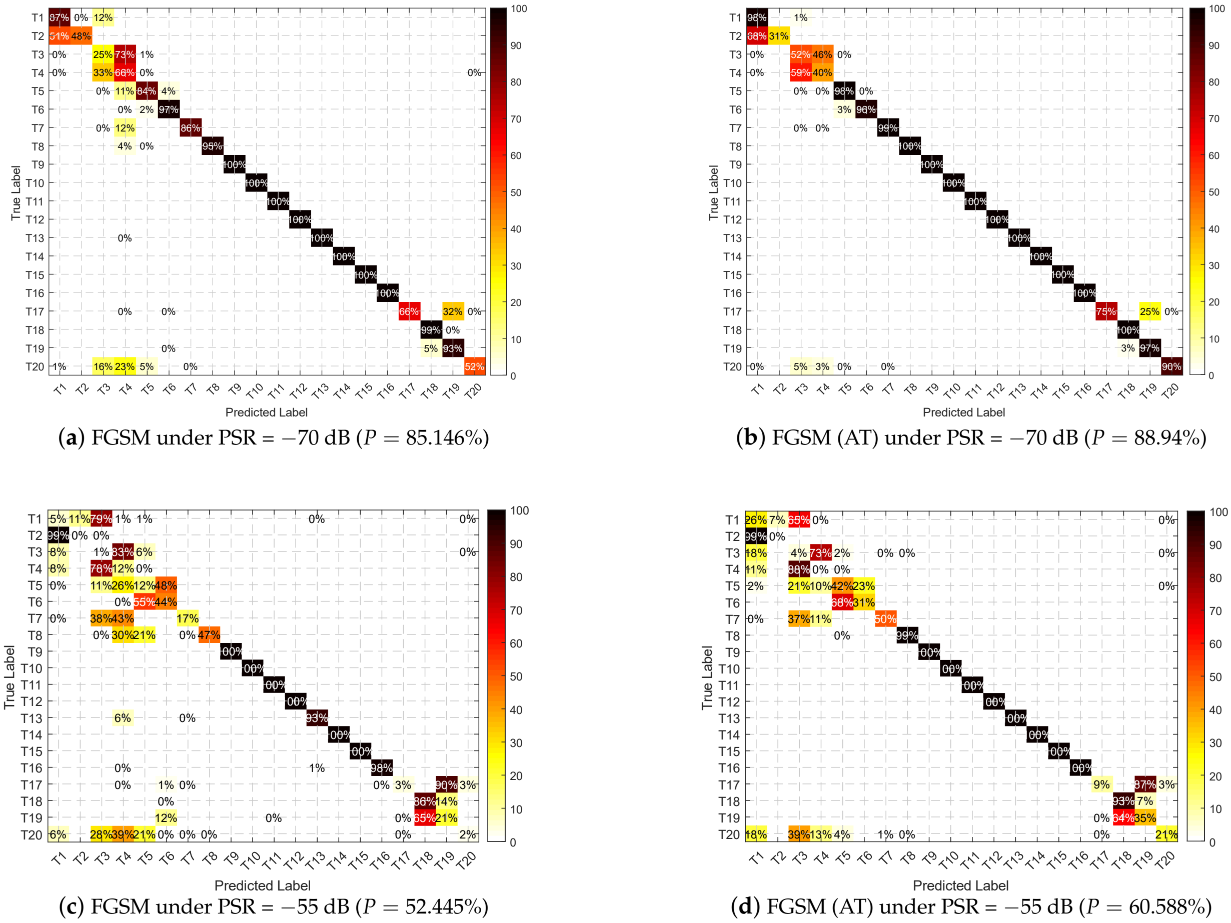
1. The performance of DNN, which has originally $P > 90\%$, decreases rapidly when encountering adversarial attacks. Compared to C&W and PGD, FGSM has a smaller impact. Also, unlike Dataset 1, Dataset 2 is more sensitive to counter perturbations and requires a lower attack intensity to compromise the system. It is more vulnerable because the differences between transmitters are set to be small, as shown in the Table 2.
2. The adversarial perturbation is more destructive compared to AWGN of the same intensity. The perturbations are carefully designed to be able to affect the DL-based SEI with less energy.
3. AT can improve the robustness to adversarial examples. The enhancement of AT can also reach 25.72% (PSR = −70 dB), 24.04% (PSR = −50 dB). Also, AT enhances the robustness of the system to AWGN. However, the enhancement effect here is lower compared to Dataset 1.
4. The AT fails when the strength of perturbations reaches PSR = −35 dB. The differences between the emitters in Dataset 2 are more subtle, so that the performance boundary is at PSR = −35 dB.

The confusion matrices of the different methods are given for SNR = 35 dB at intensities of −70 dB and −55 dB from Figures 18–21. The closer the values in the confusion matrix are to the diagonal, the better the identification is, since the adjacent emitters are set closer in simulation parameters as given in Table 2. At lower attack intensities, the system performance of the three methods is obviously improved after adding AT. When the attack intensity increases (−55 dB), in which although the C&W and PGD correct recognition rates are not significantly improved, the clutter of elements beyond the diagonal decreases as seen in the figures.



(**a**) Natural with $P = 91.652\%$ (**b**) Natural (AT) with $P = 93.604\%$

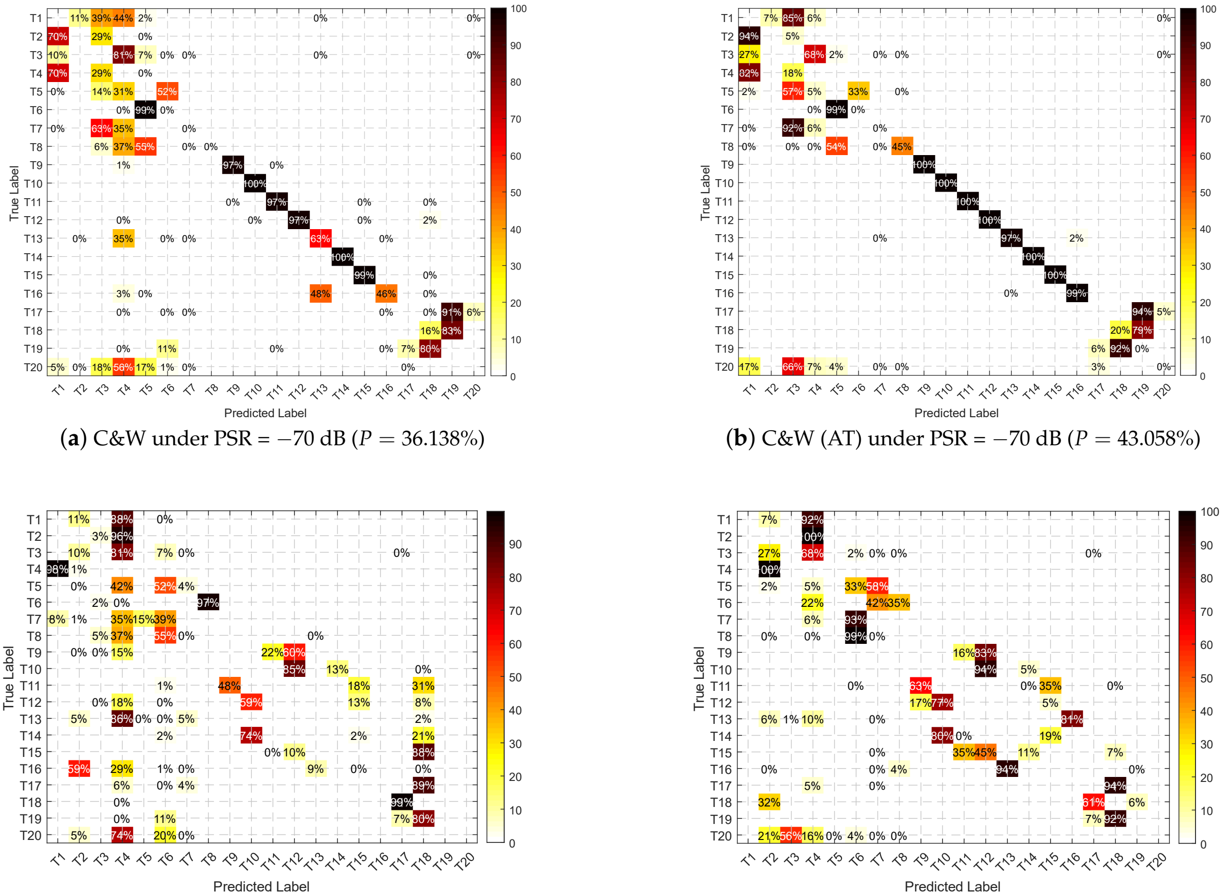
Figure 18. Recognition effect of the system with and without AT when SNR = 35 dB for Dataset 2.



(**a**) FGSM under PSR = −70 dB ($P = 85.146\%$) (**b**) FGSM (AT) under PSR = −70 dB ($P = 88.94\%$)

(**c**) FGSM under PSR = −55 dB ($P = 52.445\%$) (**d**) FGSM (AT) under PSR = −55 dB ($P = 60.588\%$)

Figure 19. Recognition effect of the system with and without AT under FGSM attack when SNR = 35 dB for Dataset 2.



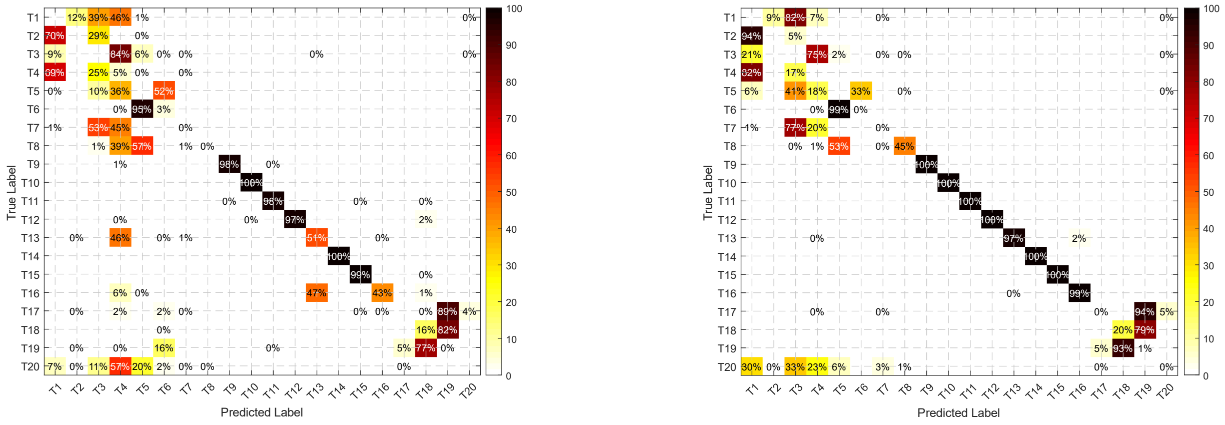
(**a**) C&W under PSR = −70 dB (P = 36.138%)

(**b**) C&W (AT) under PSR = −70 dB (P = 43.058%)

(**c**) C&W under PSR = −55 dB (P = 0%)

(**d**) C&W (AT) under PSR = −55 dB (P = 0%)

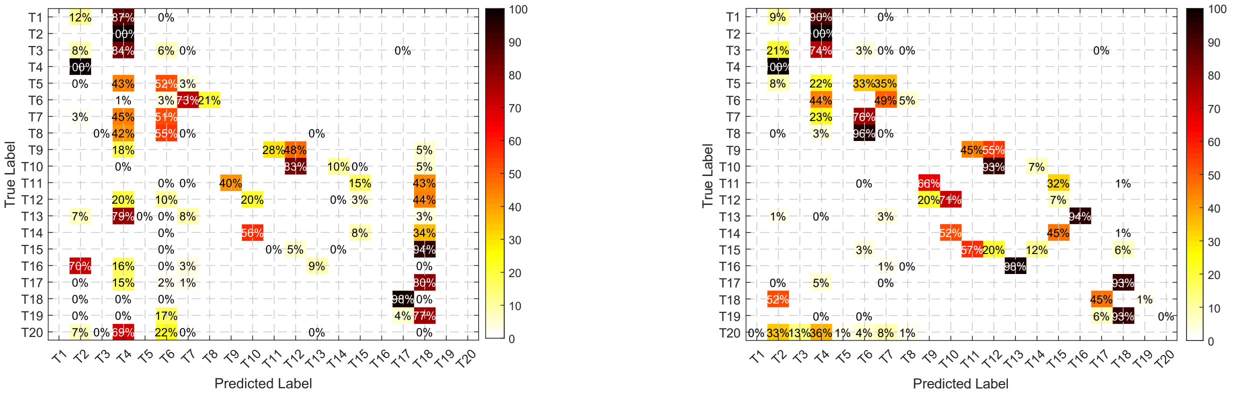
Figure 20. Recognition effect of the system with and without AT under C&W attack when SNR = 35 dB for Dataset 2.



(**a**) PGD under PSR = −70 dB (P = 35.816%)

(**b**) PGD (AT) under PSR = −70 dB (P = 43.11%)

Figure 21. *Cont.*



(**c**) PGD under PSR = −55 dB ($P = 0\%$) (**d**) PGD (AT) under PSR = −55 dB ($P = 0\%$)

Figure 21. Recognition effect of the system with and without AT under PGD attack when SNR = 35 dB for Dataset 2.

5.4. Analysis

From the above experimental results and analysis, the following points can be seen.

In the attack scenario, in terms of the overall recognition rate, the recognition performance of the system will drop significantly after being attacked. Taking Dataset 1 as an example, when PSR is −32 dB, the maximum performance drop of the system reaches 50%. With the increase of the attack intensity, at PSR = −22 dB, compared with the noise system of the same intensity, the reduction rate exceeds 60%, and the recognition performance directly drops to about 0%, which completely fails.

In terms of details, ① the impact of different attack methods is different, and the effect of C&W and PGD attacks is more obvious; ② After the attack, the distribution of the feature methods extracted by the neural network changes; ③ Different radiation sources are affected differently.

As the characteristics of the adversarial sample, the adversarial perturbation is of very low energy and irregular, and it has little effect on the probability density of the original signal.

For the defence scenario, after AT, when attacked at the same intensity (−32 dB), the performance of AT-SEI is improved by an average of 30%, and the average recognition rate can be restored to 85.59%.

For the communication signal Dataset 2, there are similar laws to Dataset 1, but different datasets are affected to different extents. Using a lower strength attack can disable the DL-based SEI system of the communication signal.

In addition, there is a certain threshold for improving the performance of AT, which is about −10 dB for Dataset 1 and −35 dB for Dataset 2.

To sum up, in practical applications, if a malicious device superimposes a very small perturbation waveform on the original waveform, without affecting the signal PDF, the existing DL-based SEI will fail, which cannot correct distinguish between malicious and legitimate devices. After AT, the robustness in this regard can be greatly improved, and the performance can be recovered to a certain extent. Therefore, in the practical application of DL-based SEI, the impact of adversarial attacks must be considered, and AT must be carried out.

6. Conclusions

In this paper, adversarial attack and corresponding defense training are introduced into SEI for the first time, which answers the relevant questions about the robustness of DL-based SEI under adversarial attacks and fills the relevant gaps. Aiming at the SEI problem, two practical scenarios of attack and defense are designed to carry out research, and specific implementation methods are given.

It is found that DL-based SEI, similar to the field of image recognition, is also sensitive to adversarial attacks. Experiments on different datasets show that after being attacked, the system will be misjudged or even invalid, so that malicious devices cannot be identified and the network security is facing serious threats. For example, only attack at the strength of −32 dB on the real-world dataset can reduce identification performance of the system to less than 50%, and the destructive power is far greater than the noise of the same intensity. At the same time, this paper gives a solution as adversarial training to improve the robustness of the system under adversarial attacks, and at −32 dB the system can be recovered to more than 85%.

Therefore, the following conclusions can be drawn. The reliability of DNN-based SEI systems under strong adversarial conditions is crucial. In future DL-based SEI related research and practical applications, in order to stabilize the work, the attack of adversarial samples must be considered, and operations such as adversarial training must be performed to improve the robustness in this scenario.

In addition, the following are also found in this work, which can be the focus of future research. (1) Different types of data are affected to different degrees, among which the simulated communication signals are more easily affected, and further research should be conducted on the relationship between signal types and attacks; Different emitters even from the same dataset are affected by different degrees, and more in-depth research is needed. (2) There is still a performance boundary after adversarial training. It is necessary to design a more effective defense method combined with the radio frequency fingerprint mechanism to further improve the performance, especially for communication signals with subtle device-specific differences. (3) More signal types, more neural network structures, and more attack methods should be involved in future experiments, especially the more realistic black-box attacks.

Author Contributions: All authors collaborated to conduct this study. L.S., formal analysis, manuscript writing, and editing; D.K., supervision, project management, and editing; X.W., Z.H. and K.H., review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Youth Science and Technology Innovation Award of National University of Defense Technology, grant number 18/19-QNCXJ.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Talbot, I.K.; Duley, R.P.; Hyatt, H.M. Specific Emitter Identification and Verification. *Technol. Rev. J.* **2003**, *113*, 113–133.
2. Zhang, J.; Wang, F.; Dobre, O.A.; Zhong, Z. Specific Emitter Identification via Hilbert–Huang Transform in Single-Hop and Relaying Scenarios. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1192–1205. [CrossRef]
3. Man, P.; Ding, C.; Ren, W.; Xu, G. A Specific Emitter Identification Algorithm under Zero Sample Condition Based on Metric Learning. *Remote Sens.* **2021**, *13*, 4919. [CrossRef]
4. Peng, L.; Zhang, J.; Liu, M.; Hu, A. Deep Learning Based RF Fingerprint Identification Using Differential Constellation Trace Figure. *IEEE Trans. Veh. Technol.* **2020**, *69*, 1091–1095. [CrossRef]
5. Xu, Q.; Zheng, R.; Saad, W.; Han, Z. Device Fingerprinting in Wireless Networks: Challenges and Opportunities. *IEEE Commun. Surv. Tutorials* **2016**, *18*, 94–104. [CrossRef]
6. Wang, W.; Sun, Z.; Piao, S.; Zhu, B.; Ren, K. Wireless Physical-Layer Identification: Modeling and Validation. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 2091–2106. [CrossRef]
7. Gok, G.; Alp, Y.K.; Arikan, O. A New Method for Specific Emitter Identification With Results on Real Radar Measurements. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3335–3346. [CrossRef]
8. Sankhe, K.; Belgiovine, M.; Zhou, F.; Angioloni, L.; Restuccia, F.; D'Oro, S.; Melodia, T.; Ioannidis, S.; Chowdhury, K. No Radio Left Behind: Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 165–178. [CrossRef]
9. Sun, L.; Wang, X.; Huang, Z.; Li, B. Radio Frequency Fingerprint Extraction based on Feature Inhomogeneity. *IEEE Internet Things J.* **2022**, *9*, 17292–17308. [CrossRef]
10. Nguyen, D.D.N.; Sood, K.; Nosouhi, M.R.; Xiang, Y.; Gao, L.; Chi, L. RF Fingerprinting based IoT Node Authentication using Mahalanobis Distance Correlation Theory. *IEEE Netw. Lett.* **2022**, *4*, 78–81. [CrossRef]

11. Gope, P.; Sikdar, B.; Millwood, O. A scalable protocol level approach to prevent machine learning attacks on PUF-based authentication mechanisms for Internet-of-Medical-Things. *IEEE Trans. Ind. Informat.* **2021**, *18*, 1971–1980. [CrossRef]
12. McGinthy, J.M.; Wong, L.J.; Michaels, A.J. Groundwork for Neural Network-Based Specific Emitter Identification Authentication for IoT. *IEEE Internet Things J.* **2019**, *6*, 6429–6440. [CrossRef]
13. Shen, G.; Zhang, J.; Marshall, A.; Peng, L.; Wang, X. Radio Frequency Fingerprint Identification for LoRa Using Deep Learning. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2604–2616. [CrossRef]
14. SUN Liting, HUANG Zhitao, WANG.Xiang, WANG.Fenghua, and LI. Baoguo. Radio Frequency Fingerprint Extraction in Specific Emitter Identification. *J. Radars* **2020**, *9*.
15. Guo, S.; Akhtar, S.; Mella, A. A Method for Radar Model Identification Using Time-Domain Transient Signals. *IEEE Trans. Aerosp. Electron. Syst.* **2021**, *57*, 3132–3149. [CrossRef]
16. Ureten, Oktay, Serinken, Nur. Bayesian detection of radio transmitter turn-on transients. In Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), Antalya, Turkey, 20–23 June 1999.
17. Zhao, C.; Huang, L.; Hu, L.; Yan, Y. Transient fingerprint feature extraction for WLAN cards based on polynomial fitting. In Proceedings of the 2011 6th International Conference on Computer Science & Education (ICCSE), Singapore, 3–5 August 2011.
18. Ru, X.; Liu, Z.; Huang, Z.T.; Jiang, W.L. Evaluation of unintentional modulation for pulse compression signals based on spectrum asymmetry. *IET Radar Sonar Navig.* **2017**, *11*, 656–663. [CrossRef]
19. Sun, L.; Wang, X.; Yang, A.; Huang, Z. Radio Frequency Fingerprint Extraction Based on Multi-Dimension Approximate Entropy. *IEEE Signal Process. Lett.* **2020**, *27*, 471–475. [CrossRef]
20. Rajendran, S.; Sun, Z.; Lin, F.; Ren, K. Injecting Reliable Radio Frequency Fingerprints Using Metasurface for The Internet of Things. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 1896–1911. [CrossRef]
21. Youssef, K.; Bouchard, L.; Haigh, K.; Silovsky, J.; Thapa, B.; Valk, C.V. Machine Learning Approach to RF Transmitter Identification. *IEEE J. Radio Freq. Identif.* **2018**, *2*, 197–205. [CrossRef]
22. Du, M.; He, X.; Cai, X.; Bi, D. Balanced Neural Architecture Search and Its Application in Specific Emitter Identification. *IEEE Trans. Signal Process.* **2021**, *69*, 5051–5065. [CrossRef]
23. Huang, X.; Kroening, D.; Ruan, W.; Sharp, J.; Sun, Y.; Thamo, E.; Wu, M.; Yi, X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.* **2020**, *37*, 100270. [CrossRef]
24. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *Comput. Sci.* **2014**.
25. Ke, D.; Huang, Z.; Wang, X.; Sun, L. Application of Adversarial Examples in Communication Modulation Classification. In Proceedings of the 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 8–11 November 2019; pp. 877–882. [CrossRef]
26. Raymond, D.R.; Midkiff, S.F. Denial-of-service in wireless sensor networks: Attacks and defenses. *IEEE Pervasive Comput.* **2008**, *7*, 74–81. [CrossRef]
27. Ohigashi, T.; Morii, M. A practical message falsification attack on WPA. *Proc. JWIS* **2009**, *54*, 66.
28. Kannhavong, B.; Nakayama, H.; Nemoto, Y.; Kato, N.; Jamalipour, A. A survey of routing attacks in mobile ad hoc networks. *IEEE Wirel. Commun.* **2007**, *14*, 85–91. [CrossRef]
29. Balakrishnan, S.; Gupta, S.; Bhuyan, A.; Wang, P.; Koutsonikolas, D.; Sun, Z. Physical layer identification based on spatial–temporal beam features for millimeter-wave wireless networks. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 1831–1845. [CrossRef]
30. Lyu, C.; Huang, K.; Liang, H.N. A Unified Gradient Regularization Family for Adversarial Examples. In Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM), Atlantic City, NJ, USA, 14–17 November 2015.
31. Baldini, G.; Gentile, C.; Giuliani, R.; Steri, G. Comparison of techniques for radiometric identification based on deep convolutional neural networks. *Electron. Lett.* **2019**, *55*, 90–92. [CrossRef]
32. Wong, L.J.; Headley, W.C.; Andrews, S.; Gerdes, R.M.; Michaels, A.J. Clustering learned CNN features from raw I/Q data for emitter identification. In Proceedings of the MILCOM 2018—2018 IEEE Military Communications Conference (MILCOM), Los Angeles, CA, USA, 29–31 October 2018; pp. 26–33.
33. Wong, L.J.; Headley, W.C.; Michaels, A.J. Specific emitter identification using convolutional neural network-based IQ imbalance estimators. *IEEE Access* **2019**, *7*, 33544–33555. [CrossRef]
34. Riyaz, S.; Sankhe, K.; Ioannidis, S.; Chowdhury, K. Deep learning convolutional neural networks for radio identification. *IEEE Commun. Mag.* **2018**, *56*, 146–152. [CrossRef]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 630–645.
36. Pan, Y.; Yang, S.; Peng, H.; Li, T.; Wang, W. Specific emitter identification based on deep residual networks. *IEEE Access* **2019**, *7*, 54425–54434. [CrossRef]
37. Zhang, T.; Ren, P.; Ren, Z. Deep Radio Fingerprint ResNet for Reliable Lightweight Device Identification. In Proceedings of the 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), Norman, OK, USA, 27–30 September 2021; pp. 1–6.
38. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *Comput. Sci.* **2013**.
39. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial Examples in the Physical World. 2016. Available online: https://arxiv.org/abs/1607.02533 (accessed on 24 July 2022).

40. Papernot, N.; Mcdaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016.
41. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017.
42. Sadeghi, M.; Larsson, E.G. Adversarial Attacks on Deep-Learning Based Radio Signal Classification. *IEEE Wirel. Commun. Lett.* **2018**, *8*, 213–216. [CrossRef]
43. Kokalj-Filipovic, S.; Miller, R.; Chang, N.; Lau, C.L. Mitigation of Adversarial Examples in RF Deep Classifiers Utilizing AutoEncoder Pre-training. In Proceedings of the 2019 International Conference on Military Communications and Information Systems (ICMCIS), Budva, Montenegro, 14–15 May 2019.
44. Lin, Y.; Zhao, H.; Ma, X.; Tu, Y.; Wang, M. Adversarial Attacks in Modulation Recognition With Convolutional Neural Networks. *IEEE Trans. Reliab.* **2021**, *70*, 389–401. [CrossRef]
45. Hameed, M.Z.; Gyorgy, A.; Gunduz, D. The Best Defense Is a Good Offense: Adversarial Attacks to Avoid Modulation Detection. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1074–1087. [CrossRef]
46. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2017**, arXiv:1706.06083.
47. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial Examples Are Not Bugs, They Are Features. In Proceedings of the NeurIPS Conference, Vancouver, BC, Canada, 8–14 December 2019.
48. Liu, A.; Liu, X.; Zhang, C.; Yu, H.; Liu, Q.; He, J. Training Robust Deep Neural Networks via Adversarial Noise Propagation. *IEEE Trans. Image Process.***2019**, *30*, 5769–5781. [CrossRef]
49. Sun, L.; Wang, X.; Huang, Z. Unintentional modulation microstructure enlargement. *J. Syst. Eng. Electron.* **2022**, *33*, 522–533. [CrossRef]
50. Sun, L.; Wang, X.; Huang, Z. Unintentional modulation evaluation in time domain and frequency domain. *Chin. J. Aeronaut.* **2021**, *35*, 376–389. [CrossRef]
51. Sun, L.; Wang, X.; Zhao, Y.; Huang, Z.; Du, C. Intrinsic Low-Dimensional Nonlinear Manifold Structure of Radio Frequency Signals. *IEEE Commun. Lett.* **2022**. [CrossRef]
52. Huang. Yuanling.; Zheng. H.; Theoretical performance analysis of radio frequency fingerprinting under receiver distortions. *Wirel. Commun. Mob. Comput.* **2015**, *15*, 823–833. [CrossRef]
53. Yiwei, P.; Sihan, Y.; Hua, P.; Tianyun, L.; Wenya, W. Specific emitter identification using signal trajectory image. *J. Electron. Inf. Technol.* **2020**, *42*, 941–949.
54. He, B.; Wang, F. Cooperative Specific Emitter Identification via Multiple Distorted Receivers. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3791–3806. [CrossRef]
55. Huang, Y.; Zheng, H. Radio frequency fingerprinting based on the constellation errors. In Proceedings of the 2012 18th Asia-Pacific Conference on Communications (APCC), Jeju, Korea, 15–17 October 2012.
56. Naveed Akhtar; Ajmal Mian; Navid Kardan;Mubarak Shah Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* **2021**, *9*, 155161–155196. [CrossRef]
57. Zhang, H.; Wang, J. Defense against adversarial attacks using feature scattering-based adversarial training. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1831–1841.
58. Manoj, B.R.; Sadeghi, M.; Larsson, E.G. Adversarial Attacks on Deep Learning Based Power Allocation in a Massive MIMO Network. In Proceedings of the ICC 2021—IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021; pp. 1–6. [CrossRef]
59. Wang, Y.; Gui, G.; Gacanin, H.; Ohtsuki, T.; Dobre, O.A.; Poor, H.V. An Efficient Specific Emitter Identification Method Based on Complex-Valued Neural Networks and Network Compression. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2305–2317. [CrossRef]
60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
61. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **2014**, *15*, 3221–3245.