Channel-Robust Class-Universal Spectrum-Focused Frequency Adversarial Attacks on Modulated Classification Models

Sicheng Zhang, *Student Member, IEEE*, Jiangzhi Fu, *Member, IEEE*, Jiarun Yu, *Student Member, IEEE*, Huaitao Xu, *Student Member, IEEE*, Haoran Zha, *Student Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, and Yun Lin, *Senior Member, IEEE*

***Abstract*—With the improvement of basic designs and the evolution of key algorithms, artificial intelligence (AI) has been considered by both industry and academia as the most promising solution for many electromagnetic space problems, such as automatic modulation classification (AMC). However, the fact that AI-based AMC models are vulnerable to adversarial examples mystifies the optimism. Adversarial attacks help researchers to reexamine AI-based AMC models and promote safe applications. In this paper, we study the frequency leakage and glitch problems caused by high frequency components in the adversarial perturbations of existing attack algorithms. We propose a Spectrum-focused Frequency Adversarial Attack (SFAA) algorithm to suppress the high frequency components to alleviate such problems. Next, we leverage meta-learning to improve the transferability of the proposed algorithm for black-box attacks. We also train a Channel-robust Class-universal Spectrum-focused Frequency Adversarial Attack (CrCu-SFAA) generative model using the generative adversarial network framework. Finally, extensive experiments using qualitative and quantitative indicators demonstrate that the proposed algorithm achieves an improved attack performance, and our proposed approach of reducing out-of-band high frequency components of the adversarial perturbations improves the concealment and adversarial signal quality.**

***Index Terms*—Automatic modulation classification (AMC), frequency adversarial attack, spectrum focus, channel-robustness, class universal.**

I. Introduction

BENEFITING from the rapid development of big data, high-performance computing equipment and other basics, the new paradigm of data-intensive scientific discovery, spawned by the information technology, have been widely studied and applied [1]. Due to the openness, easy access to data, and abstract features, the electromagnetic space becomes rich in interesting problems for researchers to explore. Artificial Intelligence (AI) based on deep learning has been considered the most promising solution to various problems in the electromagnetic space, especially, Automatic Modulation Classification (AMC) [2]. AMC is a key technology in many wireless applications such as cognitive radio, spectrum sensing, and spectrum management [3]. Various data preprocessing methods and model architectures have been proposed to continuously improve the state-of-the-art of the recognition rates [4], [5], [6], [7], [8]. Research under different conditions expands the applicability of AMC models in various practical scenarios [9], [10], [11], [12], [13]. However, adversarial examples have been shown to fool the AI-based computer vision recognition models [14]. Adversarial examples, resulting from adding carefully crafted imperceptible perturbations to the input data, can cause the AI model to output wrong results with high confidence. Subsequently, the same problem was found in various fields, including the electromagnetic space [15], [16]. The study of electromagnetic adversarial attacks and defense methods will promote efficient, safe, and credible AI applications in the electromagnetic space.

There has been increasing interest in adversarial attacks against AI-based AMC models. The authors in [17] imposed gradient-based adversarial attack methods, including the Fast Gradient Sign Method (FGSM) [18], the Projected Gradient Descent (PGD) [19], the Basic Iterative Method (BIM) [20], and the Momentum Iterative Method (MIM) [21], on AMC models and investigated the different degrees of degradation of the model classification performance by single-step and iterative attacks. The authors in [22] applied the optimization-based Carlini-Wagner (CW) [23] attack to AMC and individual recognition tasks, and proposed a defense mechanism based on the autoencoder. Taking advantage of the openness of the electromagnetic space, the authors in [24] proposed a multi-antenna attack against the AMC model that significantly reduces the accuracy of the classifier under different channel variances and correlations between antennas. The authors in [25] conducted various attacks and discovered an inverse relationship between signal confidence and attack success rate. The authors in [26] leveraged remap and

Manuscript received 12 November 2023; accepted 20 March 2024. Date of publication 26 March 2024; date of current version 8 August 2024. This work is supported by the National Natural Science Foundation of China (No. 62201172), the National Key Research and Development Program of China (2022YFE0136800). This work is also supported by Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin, China. The associate editor coordinating the review of this article and approving it for publication was Y. Gao. *(Corresponding author: Jiangzhi Fu.)*

Sicheng Zhang, Jiangzhi Fu, Jiarun Yu, Huaitao Xu, Haoran Zha, and Yun Lin are with the College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: fujiangzhi@hrbeu.edu.cn).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

Digital Object Identifier 10.1109/TCCN.2024.3382126

regularization functions to enforce undetectability constraints on perturbations to generate high-secretion universal adversarial perturbations. By considering the channel propagation effect, the authors in [27] proposed channel-aware adversarial attacks against the AMC model, and further introduced broadcast adversarial attacks by jointly considering all channel effects. Considering the impact of adversarial perturbation on the Bit Error Rate (BER), the authors in [28] proposed the BER-aware adversarial attack, and achieved a better trade-off between the classification accuracy of the intruder and the BER at the cooperative receiver. The authors in [29] proposed detection tolerant black-box adversarial-attacks, which greatly reduced the number of queries to the target AMC model and effectively improved the transferability. Currently, there is no report on the effectiveness of adversarial examples for the AMC systems based on traditional feature. The inexplicability of deep learning models leads to the existence of adversarial examples.

Efforts to defend against adversarial attacks also continue. The authors in [30] proposed the use of adversarial training to improve the robustness of AMC models, and concluded that robustness is essential for ensuring that AMC models learn features relevant to the task. The authors in [31] proposed to use the peak-to-average-power ratio statistical features of the signal to detect adversarial examples to mitigate the attack on the AMC model. The authors in [32] designed a binary modulation classification defense network that combines the benefits of low storage complexity, low computational complexity, and gradient masking to developed a lightweight defense method against white-box gradient attacks. The authors in [33] proposed a two-fold defense mechanism, which consists of correcting misclassifications on mild attacks and detecting adversarial examples on stronger attacks. The authors in [34] proposed a wireless receiver architecture consisting of both time and frequency domain feature-based AMC models, which improves the defense against black-box attacks. A comprehensive review of this topic is presented and summarized in [35].

However, existing adversarial attack methods all focus on the signal waveform in the time domain, while ignoring the impact of adversarial perturbations on the signal spectrum. The high frequency components in the data are easily captured by the AI model and can affect the prediction at a lower cost. It is easy to form an attack capability in the high frequency region for the unconstrained adversarial attack methods [36]. It has been shown that the perturbations generated by typical adversarial attack methods contain a large number of high frequency components. Time and frequency domain comparisons of the original signal and the adversarial signals of the FGSM, PGD, and Universal Adversarial Perturbations (UAP) [37] are presented in Fig. 1. These high frequency components cause frequency leakage, making the adversarial signal easier to be detected in the frequency domain. In addition, they also cause a large number of glitches in the signal, which degrades the signal quality.

To address these issues, we propose a Spectrum-focused Frequency Adversarial Attack (SFAA) method, which mitigates frequency leakage by suppressing out-of-band perturbations, to improve concealment and reduce glitches for

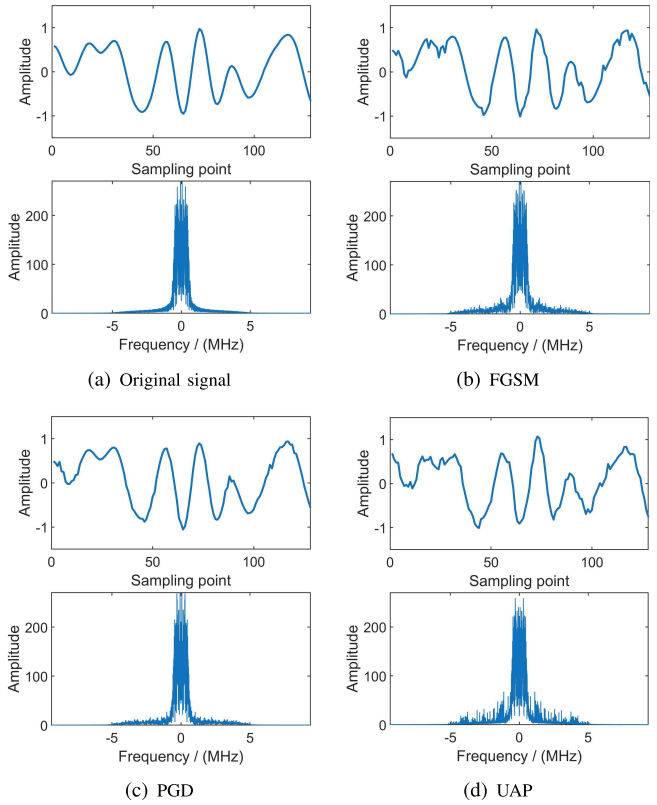


Fig. 1. The original and adversarial signals in time and frequency domains. The original signals and adversarial examples, which have high similarity in the time domain, exhibit obvious differences in the frequency spectrum. The adversarial attack leads to a frequency leakage outside of the signal bandwidth.

adversarial signal quality. Meanwhile, we leverage meta-learning to perform black-box and white-box attacks simulated in turn to improve the transferability of the algorithm in black-box attacks (termed Meta-SFAA). Further, considering the real-world adversarial scenario in the electromagnetic space, where adversarial perturbations are added at the transmitting end, and must reach a certain generation rate, we draw on the design of Generative Adversarial Network (GAN) [38] to carry out a Channel-robust Class-universal Spectrum-focused Frequency Adversarial Attack (CrCu-SFAA). This method allows for the generation of a batch of class universal adversarial perturbation libraries that can be superimposed in real time during signal transmission. Extensive experiments using qualitative and quantitative indicators demonstrate that the proposed algorithms achieve a stronger attack performance, reduce the out-of-band high frequency components of the adversarial perturbations, and improve concealment and the adversarial signal quality.

In summary, our contributions are given as follows:

- The frequency leakage and glitch problems caused by the high frequency components of adversarial perturbations generated by typical attack methods are discovered. We propose the SFAA algorithm to effectively alleviate the problem and improve attack effectiveness.
- We simulate and alternately perform black-box and white-box attacks using meta-learning, which enhances the transferability of the SFAA algorithm to black-box attacks.

- Based on the GAN framework, we propose the CrCu-SFAA algorithm, which allows pre-generation of adversarial perturbations that are robust to channel effects.

The remainder of this paper is organized as follows. The characteristics of electromagnetic space adversarial attacks and defense are analyzed in Section II. The system model and evaluation indicators for the proposed algorithms are introduced in Section III. The detailed procedures of the proposed SFAA, Meta-SFAA and CrCu-SFAA algorithms are introduced in Section IV. Extensive experiments with qualitative and quantitative indicators are conducted to evaluate the proposed algorithms in Section V. A summary of this paper and potential future work are given in Section VI.

II. Electromagnetic Adversarial Characteristics

In the adversarial attack and defense task in the electromagnetic space, imperceptible perturbations are introduced to electromagnetic signals by the adversarial program to trick AMC models in target devices into producing wrong predictions with high probability. AMC models may be located in an adaptive modulation receiver or in an eavesdropper. In these two different scenarios, it could lead to the disruption of normal communication or the blocking of eavesdropping. Here, we will consider these scenarios comprehensively to illustrate the electromagnetic adversarial attributes. In Section III, we will provide specific eavesdropping scenarios. Compared with image processing in computer vision and speech signal processing in natural language processing, adversarial attack and defense in the electromagnetic field has its unique domain characteristics, as shown in Fig. 2. Understanding these unique characteristics can help to develop effective adversarial attack and defense methods.

A. Channel Dynamics

The signals in the electromagnetic space environment are undoubtedly dense and diverse, and the transmission channel is highly dynamic. For the adversarial attack, the adversarial perturbations superimposed on the signal are usually carefully designed and calculated. The complex and dynamic channel environment brings new challenges to the design and assurance of the effect of adversarial perturbations [26]. On the other hand, noise and interferences in the channel also provide concealment for adversarial attacks, in that it is usually difficult to distinguish between noise interferences and adversarial perturbations. For the adversarial defender, the dynamics of the channel can provide a natural means of defense and protection. Moreover, once the attacker designs a channel-robust adversarial attack method, the defender will then lose such protection.

B. Difference Between Channels

In the electromagnetic space, the channel response between any two pairs of transmitter and receiver show great degrees of difference. Even the channel effects between the same pair of transmitter and receiver can vary over time. The difference in channel effects is also a double-edged sword. On one hand, this provides attackers with more flexible attack methods to achieve a certain strategy, such as only attacking some specific targets [24]. On the other hand, the difference in channel effects has widened the gap between adversarial attack simulation and actual deployment. As shown in Fig. 2, an adversarial signal generated based on the information of channel 2 will be less effective for channel 1.

C. Indirectness of Interaction

In an ideal adversarial attack task, the attacker can directly access the AI model of the target device as well as the output corresponding to the modified input. However, in the actual adversarial attack process, it will be very difficult for an attacker to access the target devices. The attacker needs to repeatedly send test data and observe the behavior of the target model to analyze the input-output relationship, characteristics, and performance of the target model [39]. Then it can modify the test data according to the obtained information, and repeat this procedure. After acquiring sufficient information concerning the target AI model, it then establishes a substitute model locally and carries out development of adversarial attack algorithms. Notably, the number of such test is often limited.

D. Openness of the Electromagnetic Space

The electromagnetic space is an open space. Any radio device has the ability to transmit and receive electromagnetic signals into and from the electromagnetic space. The openness of the electromagnetic space provides flexible and varied means for adversarial attack and defense [27]. For example, an attacker could broadcast adversarial perturbations in an area to affect targeted receivers.

E. Constraints of Communications Networks

Compared with computer vision which relies on the human eye for evaluation, radio systems are used as an important tool for verification and evaluation in the electromagnetic field. There are clear and standard evaluation indicators available, including bandwidth, power spectrum, perturbation-to-noise ratio, and more. In addition, compared with the l_0 norm, which reflects the number of changing pixels, the l_2 and l_∞ norms, which reflect the changing power and the maximum changing value, respectively, are more suitable for measuring the electromagnetic signal against perturbations [40].

III. System Models and Evaluation Indicators

In this section, we first introduce the adversarial scenarios. Further, the system models of the adversarial attacks in the transmitter and receiver are refined, respectively. Finally, the corresponding evaluation indicators are proposed around the system models.

A. Research Scenarios and System Models

Consider a scenario with a transmitter (Alice), a receiver (Bob), and an eavesdropper (Eve). Alice maintains a cooperative communication relationship with Bob. Eve wants to intercept their signals and classify their modulation type for

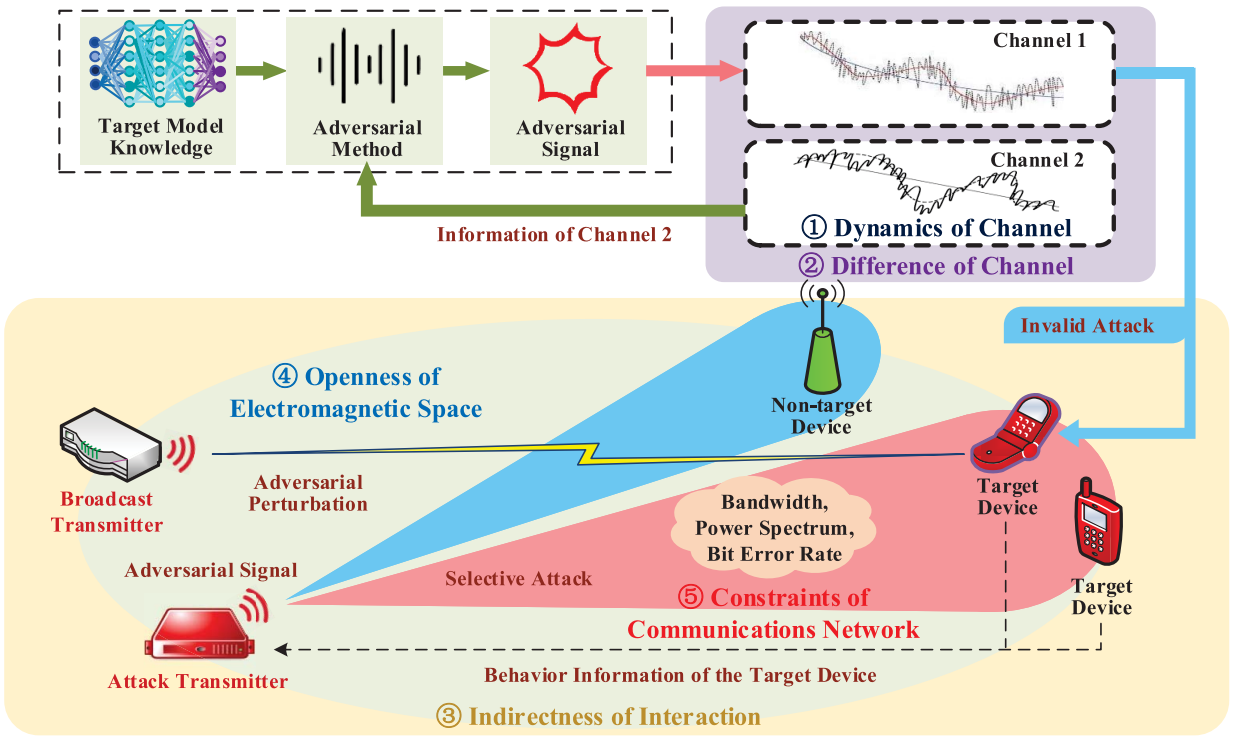


Fig. 2. Overview of Characteristics of adversarial attack and defense in the electromagnetic field.

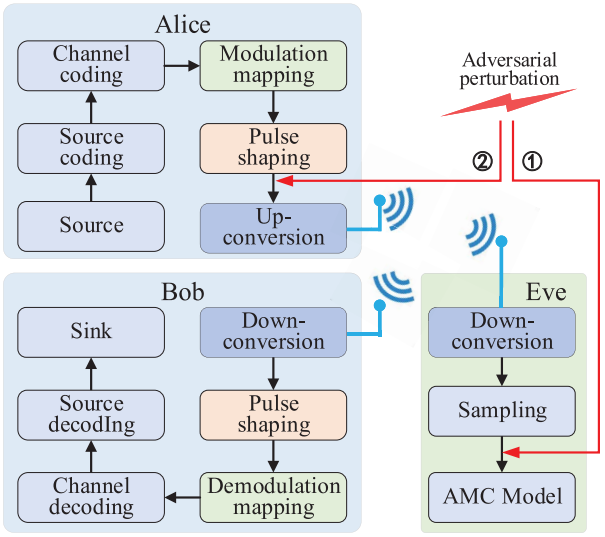


Fig. 3. Research scenarios and corresponding adversarial attack positions.

further analysis. To protect the communication content, Alice can add imperceptible adversarial perturbations to the signal to reduce Eve's classification accuracy, as shown in Fig. 3.

We first investigate the ideal case, that is, by injecting Eve with a virus program. This virus program can directly access the information of the AMC model and add adversarial perturbations to the input data without going through the channel. Spectrum leakage and time-domain glitches are both caused by the high frequency components in the adversarial perturbations. To address these two issues, we will try to suppress the out-of-band energy of the perturbations relative to the original signal as much as possible. Therefore, the system model for adversarial attacks at the receiving end can be defined as

$$\min_{\delta}\left(\frac{\mathrm{eng}\left(\delta_x^{ob}\right)}{\mathrm{eng}(\delta_x)}\right), \mathrm{s.t.}\, f(x+\delta_x)\neq y \,\mathrm{and}\, x+\delta_x\sim D_{\mathbf{x}}, \tag{1}$$

where x is the original signal data example and y is the class label; $D_{\mathbf{x}}$ is the distribution of original signal data; δ_x is the adversarial perturbation of example x and δ_x^{ob} is the out-of-band component of δ_x; $f(\cdot)$ is the AMC model; and $\mathrm{eng}(\cdot)$ is the energy calculation function, its calculation method is as

$$\mathrm{eng}(x)=\sum_{i=1}^{N} x_i^2, \tag{2}$$

where N is the length of the original signal example; The purpose of this system model is to generate perturbations with as little out-of-band energy as possible, but without affecting the data distribution, and to cause the model to produce erroneous outputs.

Then we study how to overcome the influence of the dynamics of the channel, to add adversarial perturbations at the transmitter, and to affect the target model in Eve through the channel.

$$\min_{\delta}\left(\frac{\mathrm{eng}\left(\delta_y^{ob}\right)}{\mathrm{eng}\left(\delta_y\right)}\right)$$
$$\mathrm{s.t.}\, f\left(h\left(x_y+\delta_y\right)\right)\neq y \,\mathrm{for}\, x_y \,\mathrm{and}\, x_y+\delta_y\sim D_{\mathbf{x}_y}, \tag{3}$$

where x_y is the original signal data example of class y; δ_y is the class universal adversarial perturbation of class y and δ_y^{ob} is

the out-of-band component of δ_y; $h(\cdot)$ is channel model. The purpose of this system model is to generate channel-robust and class-universal adversarial perturbations based on the previous model.

B. Evaluation Indicators

Qualitative and quantitative indicators are designed for comprehensive analysis and evaluation. Qualitative indicators, including perturbation statistical distribution and perturbation spectrum distribution, help visualize the distribution of adversarial perturbations and the degree of in-band convergence of the spectrum. Quantitative indicators, including Out-of-band Energy Ratio (OBER), Fitting Difference (FD) [40], and Bit Error Rate (BER) provide more precise results.

1) Out-of-Band Energy Ratio: Reducing the out-of-band energy can effectively reduce spectrum leakage and time-domain glitches. We define the energy of the adversarial perturbation out-of-band of the original signal as Out-of-band Energy (OBE), and the rest as In-band Energy (IBE). So we design an indicator as (4) to measure the ratio of OBE relative to the total energy of the adversarial perturbation.

$$OBER = 10\log_{10}\left(\frac{\sum_{B\cdot N/f_s\le i<N-(B\cdot N/f_s-1)}^{N} s^2(i)}{\sum_{i}^{N} s^2(i)}\right), \quad (4)$$

where B is the set focus bandwidth, which is generally the original signal's bandwidth; and f_s is the sampling rate.

2) Fitting Difference: The infinity norm can only represent the maximum difference of a sample point in a segment of the signal, but not the overall quality. Therefore, we use FD [40] as an evaluation indicator for the signal quality in the time domain, given by

$$FD = \frac{\sum_{j=1}^{N}\left(x_j - x_j'\right)^2}{\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^2}, \quad (5)$$

where x_j is the j-th sampling point of the original signal; x_j' is the j-th sampling point of the adversarial signal; and $\bar{x}$ is the average value of the original signal.

3) Bit Error Rate: Different identification methods need to be used for different types of data. For example, image data needs to be identified by the human eye, and speech needs to be identified by the human ear. The most direct way to identify the degree of damage to a signal is to compare the BER changes before and after the perturbation is added.

IV. Adversarial Attack Algorithm

In this section, we will introduce the proposed algorithms step by step around two system models. We first introduce the spectrum focused frequency adversarial attack, which is designed for the spectrum leakage and time-domain glitches problem, under a white-box scenario at the receiver-end. To enhance the black-box transferability of the attack method, we further introduce a version based on meta-learning. Finally, to extend this method to more realistically significant transmitter-end attack scenarios, we introduce the channel-robust class-universal spectrum-focused frequency adversarial attack.

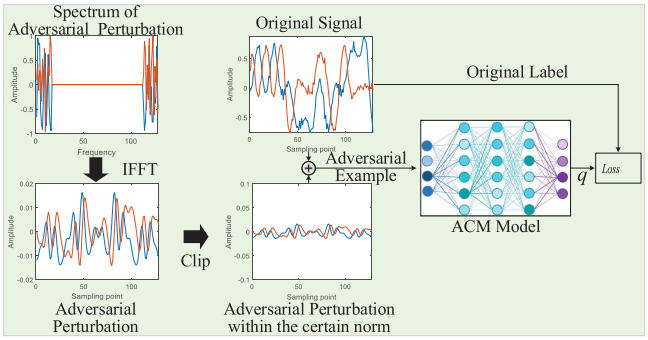


Fig. 4. Overview of the proposed SFAA algorithm.

A. Spectrum Focused Frequency Adversarial Attack

Spectrum leakage and time-domain glitches are caused by the high frequency components in the adversarial perturbations. Reducing the out-of-band high frequency components relative to the original signal is an intuitive way to address these issues. The first system model is defined in the ideal adversarial attack scenario, that is, the adversarial attacker can directly access and modify the input of the target model. Therefore, this paper proposes to carry out SFAA on the received data, and its overview is shown in Fig. 4.

Before entering the main algorithm, we analyze the composition of the spectrum data. The Fast Fourier Transform (FFT) formula for transforming from the time domain to the frequency domain is given by

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, \quad (6)$$

where $x[n]$ is the n-th complex point of the signal example; and $X[k]$ is the k-th frequency domain sample. For the positive frequency part of the signal, the frequency samples corresponding to the signal bandwidth are first N_s points of $X[k]$, which can be calculated by

$$N_s = B \cdot N/f_s, \quad (7)$$

and the frequency sampling points corresponding to the negative frequency part are the last $N_s - 1$ points of $X[k]$.

First, we initialize the random spectrum data s_0 at the focus location. Therefore, s_0 can be expressed as

$$s_0[i] = \begin{cases} 0, & N_s \le i < N - (N_s - 1) \\ \text{random}, & \text{otherwise.} \end{cases} \quad (8)$$

Next, use the Inverse Fast Fourier Transform (IFFT) to obtain the corresponding time domain adversarial perturbations of s_0 as

$$\delta[n] = \frac{1}{N}\sum_{n=0}^{N-1} s_0[k] e^{j2\pi kn/N}, \quad (9)$$

where δ is the adversarial perturbation. In order to ensure that the infinite norm is within a certain range, we truncate the adversarial perturbations with the $\mathrm{Clip}(\cdot)$ operation as in (10). Note that the $\mathrm{Clip}(\cdot)$ function may introduce some

Algorithm 1: The SFAA Algorithm

Input : The AMC model $f(\cdot)$; The initialized spectrum data s_0; The original data x; The correct label y; The update step size α; The epoch times M.

Output: The final perturbation spectrum s_M; The final adversarial example x'_M.

Randomly initialize s_0;
for *epoch* $m = 0$ *to* $M-1$ **do**
  Take the IFFT for s_m and use the truncation function to get the adversarial perturbation δ_m;
  Superimpose adversarial perturbation δ_m to get adversarial example x'_m;
  Feed the adversarial example x'_m into the model to get confidence list q;
  According to the loss function (12), calculate the loss value between q and y;
  Get the gradient of s_m by backpropagation;
  Get the s_{m+1} by (13) according to α;
end for
return s_M *and* x'_M;

high frequency components, but these components are usually very small.

$$\delta_0 = \mathrm{Clip}(\delta), \tag{10}$$

where δ_0 is the adversarial perturbation corresponding to s_0. The $\mathrm{Clip}(\cdot)$ function is defined as

$$\mathrm{Clip}(a) = \begin{cases} a, & a > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Furthermore, the adversarial perturbation is fed into the AMC model along with the original signal, and the confidence list q is obtained. A loss function is designed between the correct label y and q to punish the corresponding confidence as in (12).

$$Loss(q, y) = -\log_{10}\big(1 - q_y + \varepsilon\big), \tag{12}$$

where q_y is the confidence corresponding to the correct label; $\varepsilon = 1e^{-6}$ is used to prevent $1 - q_y$ from being 0. The update rule for s are given in (13).

$$s_{m+1} = s_m + \alpha \cdot \mathrm{sign}(\nabla_{s_m} Loss(q, y)), \tag{13}$$

where α is the update step size; $\nabla_{s_m} Loss(\cdot, \cdot)$ represents the differential operation of the $Loss(\cdot, \cdot)$ function with respect to s_m. m is the current number of updates. Using (10) and (13) to iteratively update s_0 for M times, the final perturbation spectrum and adversarial perturbation can be thus obtained. The pseudo code is provided in Algorithm 1.

B. Spectrum-Focused Frequency Adversarial Attack Based on Meta Learning

As with adversarial attack algorithms in the time domain, adversarial examples generated using the SFAA algorithm in a white-box attack scenario do not tend to have a strong transferability to black-box scenarios. It is difficult to obtain the satisfactory attack effect by directly applying the adversarial examples used in the white-box attack scenario to a black-box one because of the differences in model decision boundaries, model initialization, connection structure, training optimizers, etc.

To improve the transferability of adversarial examples generated using the SFAA algorithm to black-box scenarios, we propose a Meta-SFAA algorithm based on the concept of meta-gradient adversarial attack [41]. As shown in Fig. 5, the algorithm consists of multiple tasks including meta-train and meta-test phases. During each task, the meta-train phase simulates a white-box attack, while meta-test phase simulates a black-box attack. The tasks are performed sequentially, with meta-train and meta-test executed in turn. By adaptively reducing the gradient difference between the white-box and black-box scenarios, the Meta-SFAA algorithm enhances the transferability of adversarial examples to the black-box attack.

Specifically, we first build a model collection containing C different models. Note that the model collection should not contain the model for testing black-box attacks. Then, $R + 1$ models are randomly selected from the model collection to form a task, and a total of T tasks are generated. In each task, R models are used for meta-train and the other one is used for meta-test.

During the meta-train phase, the average logits output of the R models is computed as the final output. The final confidence list $q_{t,h}$ can be obtained further by the $\mathrm{softmax}(\cdot)$ function as:

$$q_{t,h} = \mathrm{softmax}\left(\sum_r \mathrm{logits}\Big(r, x'_{t,h}\Big)/R\right), \tag{14}$$

where $x'_{t,h}$ is the input data for the h-th meta-train in the t-th task; the value ranges of t and h are both in $\mathbb{N}$(natural number); and $\mathrm{logits}(r, \cdot)$ is the output logits of the r-th model. Then, one-step update is done by computing the gradient using the loss function (15).

$$s_{t,h+1} = s_{t,h} + \alpha \cdot \mathrm{sign}\big(\nabla_{s_{t,h}} Loss\big(q_{t,h}, y\big)\big). \tag{15}$$

After performing meta-train phase H times, we perform meta-test phase for one time. This phase is similar to the original SFAA algorithm, with the exception that the model used is randomly selected from the collection. Therefore, the process of meta-test is not repeated. The adversarial example x_t and adversarial perturbation spectrum s_t generated by each task are then passed on to the next task until all tasks are completed. The final x_T generated by the Meta-SFAA algorithm is expected to have strong transferability to black-box scenarios.

The main idea of meta-learning is to get common knowledge and skills in multiple tasks, and apply these knowledge and skills to new tasks. During the meta-train process of each task, the average gradient is used to update the adversarial perturbation, which makes the adversarial perturbation capable of attacking multiple networks. However, such adversarial attack is often weak. Based on this, a single model used in the meta-test process is used to optimize the adversarial perturbation to make it more targeted and enhance the adversarial attack capability. Repeatedly performing multiple tasks, the resulting adversarial perturbation learns attack capability that

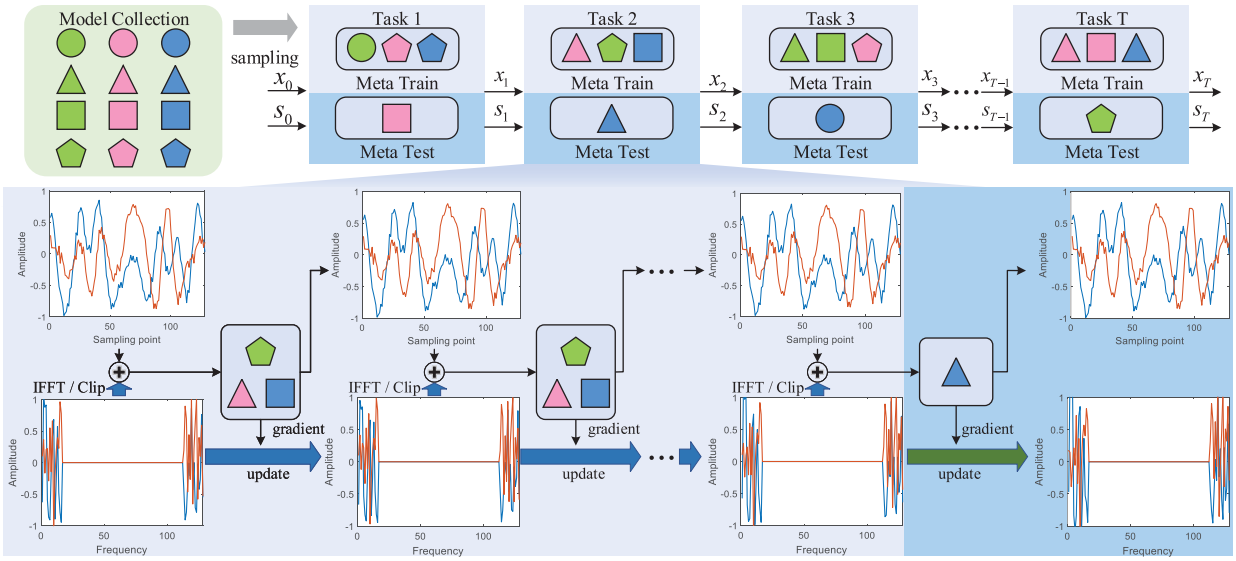


Fig. 5. Overview of the proposed Meta-SFAA algorithm.

is transferable across multiple models. This also enhances the adversarial attack transferability to new models.

C. Channel-Robust Class-Universal Spectrum-Focused Frequency Adversarial Attack

In real-world electromagnetic adversarial attack scenario, the attacker does not have the right to directly access and modify the input data of the target model as defined in the second system model. However, traditional adversarial attack perturbations, which are carefully crafted to manipulate the input data, often fail to achieve an excellent attack effectiveness when faced with random noise in the wireless channel. Additionally, the high timeliness requirements of communication tasks often require the generation rate of adversarial communication signals should not be limited by the generation and superposition of adversarial perturbations. To address these challenges, we propose the CrCu-SFAA algorithm.

We draw on the design of GAN [38] to build the framework of the CrCu-SFAA algorithm, as shown in Fig. 6. The input of the generative (G) model is a batch of Gaussian white noise $\{z\}_{\mathrm{B}}$ and class labels $\{y\}_{\mathrm{B}}$. The deconvolutional layers progressively scale up the input to a tensor with the same size as the original signal examples. This tensor is seen as an adversarial perturbation in the time domain and is then transformed into the frequency domain by the FFT algorithm. In the frequency domain, the high frequency components of the adversarial perturbation are zeroed. The final adversarial perturbation will be obtained after the IFFT algorithm and the $\mathrm{Clip}(\cdot)$ function. The adversarial perturbation is fed into the generalized channel along with the original signal examples. The generalized channel refers to all links from the superimposed adversarial perturbation in Alice to the sampling in Eve. The adversarial signal examples $\{r'\}_{\mathrm{B}}$ sampled by Eve are the input to the AMC model. The parameters and structure of the AMC model are fixed, but backpropagation is supported. The training of the generative model is performed under the supervision of loss function $Loss = (1-\rho)Loss_f + \rho Loss_d$ with the tuning factor $\rho \in [0,1]$. The $Loss_f$ is defined as

$$Loss_f = -\log_{10}(1-q_y), \tag{16}$$

where q_y is the confidence corresponding to class label y in the confidence list. The goal of $Loss_f$ is to reduce q_y. $Loss_d$ is defined as

$$Loss_d = \frac{\left(\left|\sum_j^L x_i(j)\right| + \left|\sum_j^L x_q(j)\right|\right)}{2L}, \tag{17}$$

where x_i and x_q are the in-phase and quadrature components in the adversarial perturbations corresponding to the signal, respectively. The goal of $Loss_f$ is to make the average absolute value of the generated perturbations as close to 0 as possible to reduce the impact on the signal energy distribution. The G model trained under the supervision of the loss function $Loss$ can generate class universal adversarial perturbations that are robust to the channel.

V. Experimentation and Evaluation

In this section, we conduct extensive experiments to qualitatively and quantitatively analyze the performance of the proposed algorithms.

A. Datasets

Two datasets are used in this paper to validate the proposed algorithms. The first is the publicly available, actual collected RML2016.10a [42] dataset, which contains 3 analog modulations, AM-DSB, AM-SSB, WBFM and 8 digital modulations, 8PSK, BPSK, CPFSK, GFSK, PAM4, 16QAM, 64QAM, QPSK. This dataset accounts for channel effects such as frequency offset and sample rate offset, and is commonly used in the AMC research. The second is a simulated dataset from a communication system built using MATLAB. This dataset includes the 8 digital modulations as in the RML2016.10a

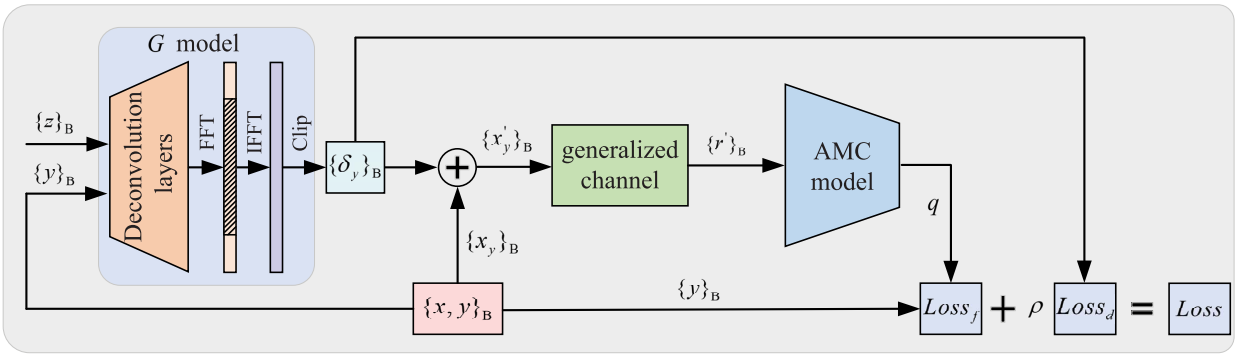


Fig. 6. Overview of the proposed CrCu-SFAA algorithm.

TABLE I
DATASET PARAMETERS

Item	Value
Symbol rate	1MHz
Baseband sampling rate	8MHz
Carrier frequency	300Mhz
Direct path frequency offset	2Hz
Indirect path time delay	0.3us
Indirect path maximum frequency shift	2Hz
Average path gain	-5.9dB
Rician factor	10
Signal-to-Noise Ratio (SNR)	-20dB: 2dB: 18dB
Example size	2×128
Example number	1,000/type/SNR
Train: Valid: Test	8:1:1

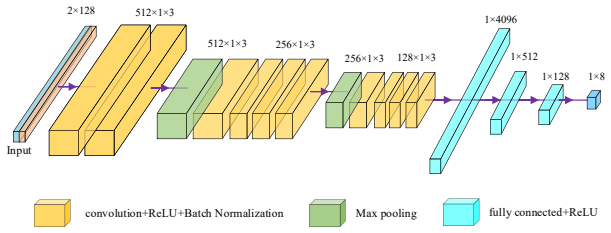


Fig. 7. Network structure diagram of the DeepModNet model.

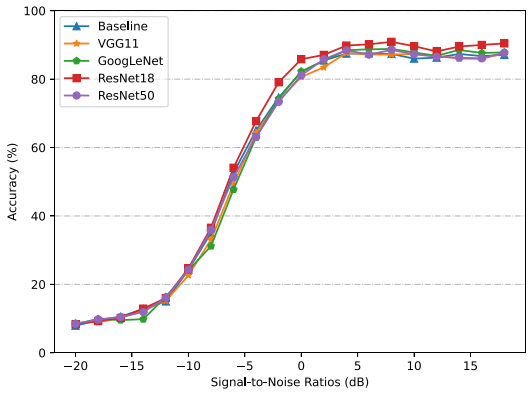


Fig. 8. The classification accuracy of the models without attack.

TABLE II
THE CLASSIFICATION ACCURACY UNDER DIFFERENT NUMBERS OF ITERATION STEPS

SFAA	Iteration steps	40	60	80	100	120
	Accuracy (%)	36.19	24.36	18.70	15.86	14.19
PGD	Iteration steps	5	10	15	20	30
	Accuracy (%)	29.28	20.40	18.65	18.15	17.74
UAP	Iteration steps	5	10	15	20	30
	Accuracy (%)	42.91	42.56	41.28	41.73	41.02

dataset, and the channel model is a Rician channel with Additive White Gaussian Noise (AWGN). The parameters for this dataset are shown in Table I.

B. Comparison of Classification Performance

In this paper, we build a model to classify modulated signals and use it as a target model for adversarial attacks. We call it DeepModNet. The network structure of DeepModNet model is shown in Fig. 7.

To evaluate the performance of the DeepModNet model, we compare it to several classical models on the RML2016.10a dataset, including ResNet-18, ResNet-50, VGG-11, and GoogLeNet [43], [44], [45], [46]. The classification accuracy of each model at different SNRs is shown in Fig. 8.

As can be seen from Fig. 8, the DeepModNet model can achieve an almost comparable classification performance to the other classical models. The classification accuracy increases with SNR and gradually stabilizes around 4dB. Therefore, we will conduct adversarial attacks on the DeepModNet model in the subsequent experiments.

C. Comparison of White-Box Attacks on the Receiver

The RML2016.10a dataset and DeepModNet model are used in a receiver-side white-box attack. Since the number of samples per symbol in the RML2016.10a dataset is 8, we set N_s to 16 according to (7). Before comparing the attack performance of the SFAA algorithm with others, we first experimentally study the effect of the number of iterations on the performance and determine the optimal hyperparameters. The perturbation strength is set to 0.1, the iteration step size is set to 0.02, and the number of iteration steps is set varied. Under the attack of the SFAA, PGD, and UAP algorithm with different numbers of iteration steps, the classification accuracy of the target model on the RML2016.10a dataset is shown in Table II. To obtain better attack performance and avoid serious overfitting, we set the number of iteration steps of the SFAA to 100, PGD and UAP to 10, in the following experiments.

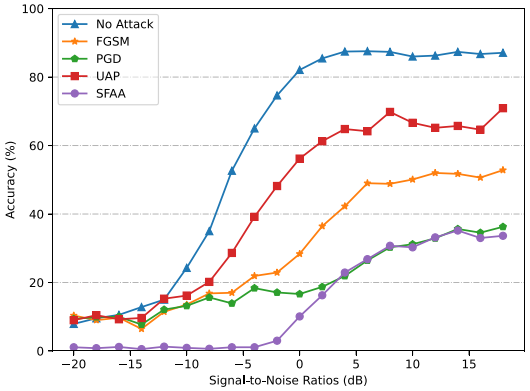


Fig. 9. The classification accuracy of the DeepModNet model under various white-box attack algorithms at the receiver-side.

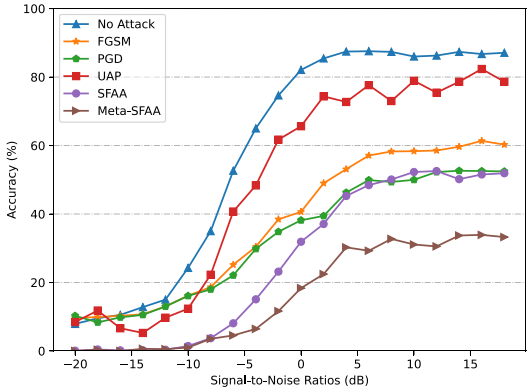


Fig. 10. The classification accuracy of the DeepModNet model under various black-box attack algorithms at the receiver-side.

In the experimental setup, the perturbation strength for all four algorithms is set to 0.1. The iteration step size and the number of iteration steps for the PGD and UAP algorithms are 0.02 and 10, respectively. The classification accuracies under no attack and the four attack algorithms on the RML2016.10a dataset are shown in Fig. 9.

In Fig. 9, the four attack algorithms cause the accuracy to decrease to varying degrees. The accuracy drop at low SNR is smaller than that at high SNR, which is because the strong noise can obscure various types of data, making it difficult for the attacks to be effective. The UAP algorithm generates a single adversarial perturbation for the entire dataset, rather than design one for each example, resulting in a relatively weak attack effectiveness. The iterative algorithm can update multiple times to find a more optimal point that reduces the target confidence, making its attack effectiveness stronger than the single-step attack FGSM. In contrast, SFAA can completely fool the target model at low SNR. When the SNR is higher than 10dB, the attack performance of SFAA is similar to that of PGD. Analysis of the above attack results shows that the SFAA algorithm has the strongest attack performance.

D. Comparison of Black-Box Attacks on the Receiver

Compared to the white-box attack, the black-box attack is more realistic as the attacker has no information about the structure and parameters of the target AMC model, only the input-output relationship. Next, we experimentally evaluate the black-box attack performance of the SFAA algorithm among the five models using the RML2016.10a dataset. The data used are the data of all 11 types of signals and 20 SNRs in the RML2016.10a dataset. The classification accuracy under the attack of SFAA algorithm across different networks is shown in Table III.

In Table III, the rows indicate the models for which the adversarial perturbations are computed, and the columns indicate the models for which the classification accuracy is reported. The diagonal lines indicate the accuracy under the white-box attack, while the other positions are the accuracy under the black-box attack. Among them, the DeepModNet model appears to be more difficult to attack, while the ResNet-18 model has the weakest capability to defeat the adversarial attack. However, the effectiveness of the white-box attack using the proposed SFAA algorithm is significantly higher than that of black-box attack on each model.

To address this issue, we propose to leverage meta-learning to improve attack transferability of the SFAA algorithm. In the Meta-SFAA algorithm, we use three different optimization algorithms, including RMSprop, Adam, and Adamax [43], to train the DeepModNet, ResNet-18, ResNet-50, and VGG-11 models, to obtain the model collection. The black-box target model is GoogLeNet. Four models are randomly selected for each task, three of which are used for meta-train for 50 times and one is used for meta-test. 15 tasks like this are created. The classification accuracies of the black-box target AMC model under these attack algorithms are shown in Fig. 10.

Comparing Fig. 9 and Fig. 10, it can be seen that the attack performance of all four algorithms in the black-box attack case is weakened. Among them, the performance of the UAP algorithm is reduced the least, because adversarial perturbations that are universal to dataset are also more transferable in black-box attacks. The attack performance of the FGSM and the PGD have a similar weakening degree, with the latter is still stronger than the former. The SFAA algorithm has the most degraded performance, as it achieves a very impressive attack performance in the white-box attack but tends to overfit, leading to the maximum decay in the black-box attack. Finally, the attack performance of Meta-SFAA is greatly improved, surpassing the PGD algorithm. Meta-learning combines the gradient directions of multiple models and simulates the process of alternating between white-box and black-box attacks, enabling SFAA to achieve excellent transferability in black-box attacks.

E. Comparison of Adversarial Attacks on the Transmitter

Next, we consider the scenario of an adversarial attack on the transmitter side. The simulation system is used in this section to allow the signal to pass through the generalized channel model. We set the same parameters as in Section V-C for the FGSM, PGD, and UAP algorithms. For the CrCu-SFAA algorithm, we set the same perturbation strength. The optimizer used to update the G model is Adam. A dynamic learning rate scheme is used, with the learning rate starting at

TABLE III
CLASSIFICATION ACCURACIES UNDER THE SFAA ATTACK ACROSS DIFFERENT MODELS

		Test Model				
		DeepModNet	VGG-11	GoogLeNet	ResNet-18	ResNet-50
Surrogate Model	DeepModNet	**13.98%**	39.14%	39.46%	34.56%	39.78%
	VGG-11	37.41%	**10.08%**	40.60%	32.14%	38.26%
	GoogLeNet	32.10%	38.01%	**6.42%**	32.66%	38.08%
	ResNet-18	37.10%	35.39%	38.17%	**5.13%**	35.44%
	ResNet-50	32.38%	32.46%	35.07%	26.20%	**6.54%**

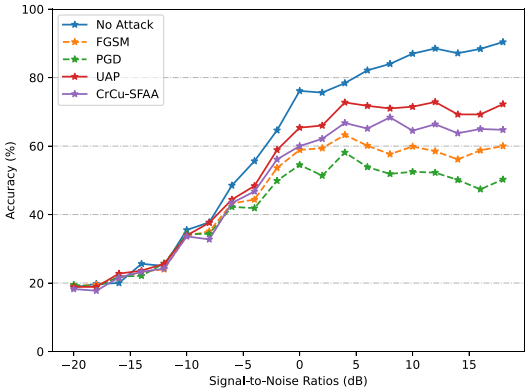


Fig. 11. The classification accuracy of the DeepModNet model under various attack algorithms at the transmitter-side.

1e-2 and decaying by half every two epochs. In each batch, various types of data are randomly selected for training to prevent the model from skewing certain types of data. The tuning factor ρ in the loss function is set to 0.5. In the testing phase, random noise and labels are fed into the G model to generate a batch of adversarial perturbation libraries, which can be superimposed on the transmitted signal in real time. Under various attack methods carried out at the transmitter, the classification accuracy of the target AMC model is shown in Fig. 11.

In Fig. 11, at low SNR, the channel effects overwhelm the adversarial perturbations. When the SNR increases to about -8dB, the different attack algorithms show various attack performances. Among them, the UAP and CrCu-SFAA algorithms do not generate perturbations specific to individual examples. However, the FGSM and PGD algorithms carefully craft adversarial perturbations for each example. The performance of the two example-specific adversarial attacks is better than the two example-universal ones. The attack performance of the proposed CrCu-SFAA algorithm is higher than that of the UAP algorithm, demonstrating its effectiveness in conducting channel-robust class-universal adversarial attacks.

F. Comparison of Computational Complexity

Next, we compare the computational complexity of the proposed SFAA, Meta-SFAA and CrCu-SFAA with existing algorithms, FGSM, PGD, and UAP. Whether pre-training is required, batchability, universality, the number of forward and backward propagation required in real-time generation are used as evaluation indicators, as shown in Table IV.

Among the above algorithms, only CrCu-SFAA is based on the generative model, so pre-training is required. Once trained, the model can generate class-universal adversarial perturbations, thus requiring only one forward. FGSM requires only one update step, thus requiring one forward-and-backward. The Meta-SFAA varies with the number of tasks T and the number of meta-training H. They are set to 10 and 15 in this paper. The SFAA, PGD, and UAP vary according to the setting of the number of iteration steps. As discussed above, these are set to 100, 10, and 10, respectively. Only the UAP algorithm cannot generate adversarial perturbations in batches, but the perturbations he generates are common to the entire data set.

G. Property Evaluation of Adversarial Perturbation

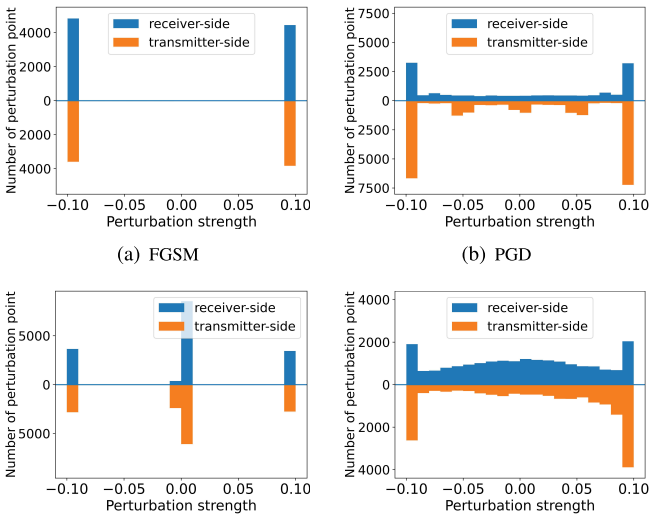
Next, we evaluate the properties of the adversarial perturbations of the proposed algorithm from the following qualitative and quantitative indicators, and compare it with several existing algorithms.

1) Perturbation Statistical Distribution: The L_∞ norm can only evaluate the largest magnitude in perturbations. The perturbation statistical distribution, as a qualitative indicator, can capture the true distribution position of the perturbation amplitudes. If the distribution is concentrated in several positions, it means that there are more glitches in the perturbations, and vice versa. If the absolute value of the distribution position is larger, it means that the power consumption of the disturbance is larger, under the same L_∞ constraint. For comparative analysis, we plot the perturbation statistical distributions of the four attack algorithms at the transmitter and receiver in pairs as shown in Fig. 12. The perturbations corresponding to 100 randomly selected samples are counted in each algorithm.

Each of these four algorithms has similar statistical distribution characteristics at the transmitter and receiver. In Fig. 12(a), the perturbations of all sample points produced by the FGSM algorithm touches the norm boundary because of its single-step update. This reflects that the exploration of the FGSM algorithm in the feasible space is simple and crude. Such perturbations will introduce a lot of glitches on the signal waveform. However, the PGD algorithm saves unnecessary perturbations due to its iterative updates as in Fig. 12(b), and a large number of perturbations are updated to the interior of the region. In the UAP algorithm as shown in Fig. 12(c), the magnitudes of the perturbation are very small or at the boundary. These perturbations will also introduce a lot of glitches. In contrast, for the SFAA and CrCu-SFAA algorithms shown in Fig. 12(d), the frequency band is limited so that the perturbation does not change drastically. The perturbations of the SFAA and CrCu-SFAA algorithm are more widely

TABLE IV
COMPARISON OF COMPUTATIONAL COMPLEXITY OF DIFFERENT ALGORITHMS.

	Pre-training	Batchability	Universality	Forward(real-time)	Backward(real-time)
SFAA	✗	✓	✗	100	100
Meta-SFAA	✗	✓	✗	15×(50+1)	15×(50+1)
CrCu-SFAA	✓	✓	Class-universal	1	0
FGSM	✗	✓	✗	1	1
PGD	✗	✓	✗	10	10
UAP	✗	✗	Universal	10	10



(a) FGSM (b) PGD

(c) UAP (d) SFAA/CrCu-SFAA

Fig. 12. The perturbation statistical distribution of attack algorithms.

distributed within the perturbation boundary, which makes perturbations appear softer and less glitches.

2) Perturbation Spectrum Distribution: The spectrum distribution of perturbations is an intuitive qualitative method to observe the degree of aggaregation of the perturbations in the spectrum. We plot the perturbation spectrum distribution of these algorithms at the transmitter and receiver in Fig. 13.

In Fig. 13, there are a lot of energy outside the frequency band (1MHz) where the signal is located, for the FGSM, PGD, and UAP algorithms at the transmitter and receiver. Moreover, universal adversarial perturbations including the UAP and CrCu-SFAA algorithms all show dense peaks in the spectrum. Importantly, the perturbations of the SFAA and CrCu-SFAA algorithms are obviously more focused, which concentrates the attack energy in the original signal spectrum, thus obtaining a significant attack performance and excellent frequency concealment.

3) Fitting Difference: The FD indicator reflects the degree of the difference between the adversarial examples and the original signal examples. We randomly select 100 signals from the datasets at the transmitter and receiver, and generate adversarial examples using the four algorithms. The average FD for each algorithm at the transmitter and receiver is calculated and shown in Fig. 14.

In Fig. 14, the left column of each group is from the receiver attack, and the right column is from the transmitter attack. In the receiver attack, the FD of the FGSM algorithm is the largest, and that of the PGD and UAP algorithms are

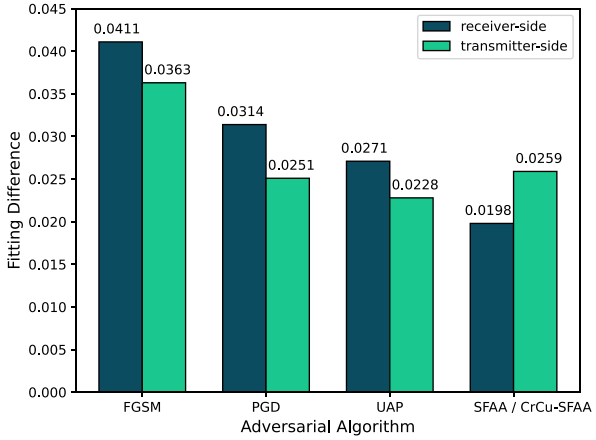


Fig. 14. The Fitting difference of attack algorithms.

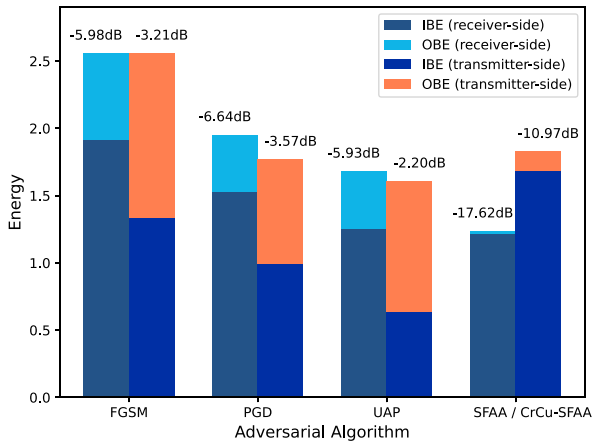
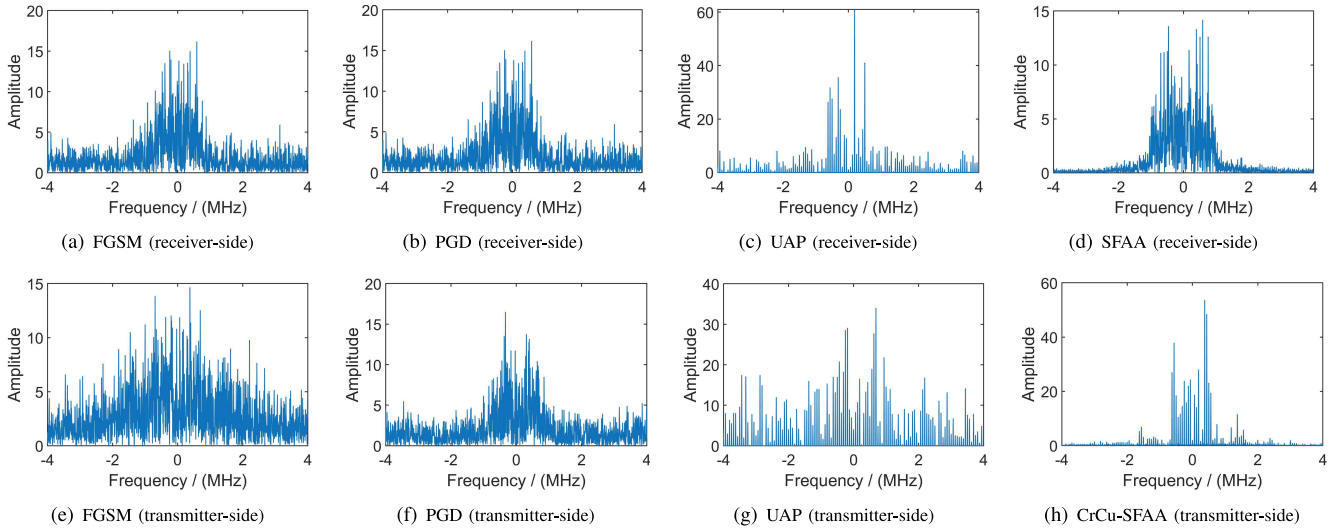


Fig. 15. The out-of-band energy ratio of attack algorithms.

smaller. The FD of the SFAA algorithm is the smallest, which is close to a half of that of the FGSM algorithm. While in the transmitter attack, the FD of the FGSM algorithm is also the largest. The FD of PGD, UAP, and CrCu-SFAA algorithms are very similar and significantly lower than that of the FGSM algorithm. The perturbations of the SFAA and CrCu-SFAA algorithms are concentrated in the low frequency region and are more flexible, so there is minimal extra waste and difference in the waveforms.

4) Out-of-Band Energy Ratio: In order to quantitatively analyze the spectrum concentration of the adversarial attack algorithm in the frequency domain, we calculate the average IBE, OBE, and OBER for the perturbations of the above 100 examples at receiver and transmitter, as shown in Fig. 15.



(a) FGSM (receiver-side) (b) PGD (receiver-side) (c) UAP (receiver-side) (d) SFAA (receiver-side)

(e) FGSM (transmitter-side) (f) PGD (transmitter-side) (g) UAP (transmitter-side) (h) CrCu-SFAA (transmitter-side)

Fig. 13. The perturbation spectrum distributions of the attack algorithms.

In Fig. 15, each energy column consists of IBE and OBE, and OBER as marked above the energy column. The left column of each group is from the receiver attack, and the right column is from the transmitter attack. Comparing Fig. 14 and Fig. 15, it can be seen that the overall trend of FD of adversarial perturbations for each algorithm is similar to the total energy. Further observation, in the receiver attack, is that the perturbations of the FGSM, PGD, and UAP algorithms all have obvious frequency leakages, and the OBER reaches over -6.64dB. However, the OBER of the SFAA algorithm is only -17.62dB. In comparison, there is a similar trend in transmitter attack. The OBER of the CrCu-SFAA algorithm is only -10.97dB, and that of others is over -3.57dB. The minimum frequency leakage is achieved by the proposed the SFAA and CrCu-SFAA algorithms.

5) Bit Error Rate: The measurement of semantic information of different types of data requires different methods. For example, image data needs to be identified by the human eye, and speech needs to be identified by the human ear. In our scenario, in order to prevent Eve from identifying the modulation class and further obtaining the communication content, Alice adds adversarial perturbations in the signal to attack the AMC model of Eve. However, there is a premise that the added adversarial perturbations cannot affect the original semantic information of the signal in cooperative communication. The most direct way to identify the degree of damage to a signal is to compare the BER changes before and after the perturbation is added.

In Fig. 16, we can see that the impact of the adversarial perturbations of the four attack algorithms added at the transmitter on the BER is very similar and very small. Channel noise plays a major role and masks the impact of perturbations to some extent. On the contrary, in a receiver attack, the impact of the adversarial disturbances of the four algorithms on the BER cannot be ignored. Moreover, the BER under the FGSM, PGD, and UAP attack algorithms are relatively similar. The proposed SFAA algorithm has significantly less impact on the BER than the other algorithms. In a receiver attack, the high frequency component in the disturbance may be the main factor affecting the BER. The proposed SFAA algorithm suppresses out-of-band high frequency components of the adversarial perturbations, thereby significantly reducing the impact on BER.

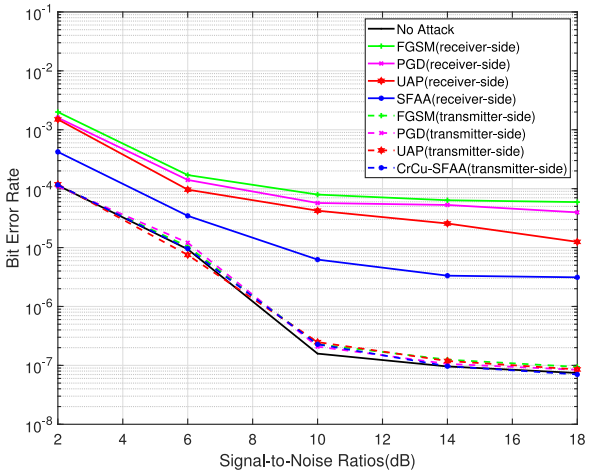


Fig. 16. The bit error rate under various attack algorithms.

VI. Conclusion

In this paper, We studied the redundant high frequency components of existing adversarial attack methods for AI-based AMC models, and analyzed the resulted frequency leakage and glitch problems. To address these issues, we proposed the SFAA algorithm, which concentrates the adversarial perturbation energy in the signal band by suppressing the OBE. Further, we proposed the Meta-SFAA algorithm to enhance the transferability to black-box attacks, using meta-learning to generate adversarial perturbations under multiple models. Considering the requirement for transmitter attacks and perturbation generation rate, we proposed the CrCu-SFAA algorithm, inspired by the GAN framework, which generates channel-robust class-universal perturbations. Qualitative

and quantitative indicators were used to demonstrate that the proposed algorithms can achieve a stronger attack performance, mitigate the frequency leakage, and improve the adversarial signal quality. However, the generative model is trained under the channel model rather than using explicit channel information, which is a passive training method. Future work should be conducted on how to train a perturbation generative model based on the channel information to actively generate the desired adversarial perturbations. Overall, a new perspective has been provided in this paper for adversarial attacks in the electromagnetic field, which provided a critical research value.

REFERENCES

[1] K. M. Tolle, D. S. W. Tansley, and A. J. G. Hey, "The fourth paradigm: Data-intensive scientific discovery [point of view]," *Proc. IEEE*, vol. 99, no. 8, pp. 1334–1337, Aug. 2011.

[2] T. Huynh-The et al., "Automatic modulation classification: A deep architecture survey," *IEEE Access*, vol. 9, pp. 142950–142971, 2021, doi: 10.1109/ACCESS.2021.3120419.

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

[4] S. Peng et al., "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 718–727, Mar. 2019.

[5] Y. Lin, Y. Tu, Z. Dou, L. Chen, and S. Mao, "Contour stella image and deep learning for signal recognition in the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 34–46, Mar. 2021.

[6] Z. Ke and H. Vikalo, "Real-time radio technology and modulation classification via an LSTM auto-encoder," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 370–382, Jan. 2022.

[7] P. Qi, X. Zhou, S. Zheng, and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 21–33, Mar. 2021.

[8] Y. Tu, Y. Lin, J. Wang, and J.-U. Kim, "Semi-supervised learning with generative adversarial networks on digital signal modulation classification," *Comput. Mater. Continua*, vol. 55, no. 2, pp. 243–254, 2018.

[9] Q. Zhou, R. Zhang, J. Mu, H. Zhang, F. Zhang, and X. Jing, "AMCRN: Few-shot learning for automatic modulation classification," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 542–546, Mar. 2022, doi: 10.1109/LCOMM.2021.3135688.

[10] Y. Dong, X. Jiang, H. Zhou, Y. Lin, and Q. Shi, "SR2CNN: Zero-shot learning for signal recognition," *IEEE Trans. Signal Process.*, vol. 69, pp. 2316–2329, Mar. 2021.

[11] M. Wang, Y. Lin, Q. Tian, and G. Si, "Transfer learning promotes 6G wireless communications: Recent advances and future challenges," *IEEE Trans. Rel.*, vol. 70, no. 2, pp. 790–807, Jun. 2021.

[12] S. Zhang, Y. Lin, Y. Tu, and S. Mao, "Electromagnetic signal modulation recognition technology based on lightweight deep neural network," *J. Commun.*, vol. 41, no. 11, pp. 12–21, Nov. 2020.

[13] L. Zhang, W. Xiang, and X. Tang, "An efficient bit-detecting protocol for continuous tag recognition in mobile RFID systems," *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 503–516, Mar. 2018.

[14] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[15] 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.

[16] Z. Bao, Y. Lin, S. Zhang, Z. Li, and S. Mao, "Threat of adversarial attacks on DL-based IoT device identification," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9012–9024, Jun. 2022.

[17] 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.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[20] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Artif. Intell. Safety Secur.*, 2018, pp. 99–112.

[21] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 9185–9193.

[22] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted adversarial examples against RF deep classifiers," in *Proc. ACM Workshop Wireless Secur. Mach. Learn. (WiseML)*, New York, NY, USA, 2019, pp. 6–11.

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[24] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, and S. Ulukus, "Adversarial attacks with multiple antennas against deep learning-based modulation classifiers," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1–6.

[25] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in DNN-based modulation recognition," in *Proc. INFOCOM IEEE Conf. Comput. Commun.*, 2020, pp. 2469–2478.

[26] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against DNN-based wireless communication systems," in *Proc. ACM SIGSAC CCS*, New York, NY, USA, 2021, pp. 126–140.

[27] 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.

[28] 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.

[29] P. Qi, T. Jiang, L. Wang, X. Yuan, and Z. Li, "Detection tolerant black-box adversarial attack against automatic modulation classification with deep learning," *IEEE Trans. Rel.*, vol. 71, no. 2, pp. 674–686, Jun. 2022.

[30] J. Maroto, G. Bovet, and P. Frossard, "SafeAMC: Adversarial training for robust modulation recognition models," 2021, *arXiv:2105.13746*.

[31] S. Kokalj-Filipovic, R. Miller, and G. Vanhoy, "Adversarial examples in RF deep learning: Detection and physical robustness," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Ottawa, ON, Canada, 2019, pp. 1–5.

[32] S. Zhang, Y. Lin, Z. Bao, and J. Fu, "A lightweight modulation classification network resisting white box gradient attacks," *Secur. Commun. Netw.*, vol. 2021, Oct. 2021, Art. no. 8921485.

[33] R. Sahay, D. J. Love, and C. G. Brinton, "Robust automatic modulation classification in the presence of adversarial attacks," in *Proc. Annu. Conf. Inform. Sci. Syst. (CISS)*, 2021, pp. 1–6, doi: 10.1109/CISS50987.2021.9400326.

[34] R. Sahay, C. G. Brinton, and D. J. Love, "A deep ensemble-based wireless receiver architecture for mitigating adversarial attacks in automatic modulation classification," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 1, pp. 71–85, Mar. 2022, doi: 10.1109/TCCN.2021.3114154.

[35] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using RF data: A review," *IEEE Commun. Surveys Tutor.*, vol. 25, no. 1, pp. 77–100, 4th Quart., 2023.

[36] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8681–8691.

[37] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 1765–1773.

[38] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[39] M. Usama, J. Qadir, and A. Al-Fuqaha, "Black-box adversarial ML attack on modulation classification," 2019, *arXiv:1908.00635*.

[40] H. Zhao, Y. Lin, S. Gao, and S. Yu, "Evaluating and improving adversarial attacks on DNN-based modulation recognition," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, 2020, pp. 1–5.

[41] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, and S. Shan, "Meta gradient adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 7748–7757.

[42] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Proc. Eng. Appl. Artif. Intell.*, Cham, Switzerland, 2016, pp. 213–226.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[46] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 1–9.

Sicheng Zhang (Student Member, IEEE) received the B.S. degree from the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China in 2019, where he is currently pursuing the Ph.D. degree. His research interests include communication technology, signal processing, artificial intelligence, and adversarial threat in electromagnetic space and lightweight deep learning model.

Jiangzhi Fu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Harbin Engineering University, Harbin, China, in 2000, 2005, and 2010, respectively. From 2018 to 2019, he was a Visiting Scholar with Utah State University, Logan, UT, USA. He is currently an Instructor with Harbin Engineering University. His current research interests include communication technology, signal processing, communication anti-interference, cognitive radio, software-defined radio, and D2D communication.

Jiarun Yu (Student Member, IEEE) received the B.S. degree in electronic information engineering from the Wuhan University of Science and Technology, Wuhan, China, in 2021. He is currently pursuing the M.S. degree with the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China. His research interests include wireless communication, deep learning, and AI security.

Huaitao Xu (Student Member, IEEE) received the B.S. degree from the Wenzheng College of Soochow University, Soochow, China, in 2020. He is currently pursuing the M.S. degree with the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China. His research interests include digital communication technology, and interfered speech quality assessment.

Haoran Zha (Student Member, IEEE) received the B.S. degree from the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China, in 2019, where he holds a doctoral position. His current research interests include signal processing, machine learning, and data analysis.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA. He is a Professor and an Earle C. Williams Eminent Scholar, and the Director of the Wireless Engineering Research and Education Center, Auburn University. His research interests include wireless networks and multimedia communications. He is a co-recipient of the 2021 Best Paper Award of Elsevier/KeAi Digital Communications and Networks, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems, and several best conference paper/demo awards. He is the Editor-in-Chief of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is a Distinguished Lecturer of the IEEE Communications Society and IEEE Council of RFID.

Yun Lin (Senior Member, IEEE) received the B.S. degree from Dalian Maritime University, Dalian, China, in 2003, the M.S. degree from the Harbin Institute of Technology, Harbin, China, in 2005, and the Ph.D. degree from Harbin Engineering University, Harbin, in 2010. He was a Research Scholar with Wright State University, USA, from 2014 to 2015. He is currently a Full Professor with the College of Information and Communication Engineering, Harbin Engineering University. He had published more than 200 international peer-reviewed journal/conference papers, such as the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, IEEE TRANSACTIONS ON RELIABILITY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, INFOCOM, GLOBECOM, ICC, VTC, and ICNC. His current research interests include machine learning and data analytics over wireless networks, signal processing and analysis, cognitive radio and software defined radio, artificial intelligence, and pattern recognition. He is a recipient of IEEE Outstanding Service Award of Trustcom 2021 and the IEEE Outstanding Track Chair Award of MASS 2021. He has gotten the Best Paper of ICCC 2023, Mobimedia 2022, ADHIP 2021, and CSPS 2018. He is serving as the Editor-in-Chief for *EAI Endorsed Transactions on Mobile Communications and Applications*, an Editor for the IEEE TRANSACTIONS ON RELIABILITY, IEEE INTERNET OF THINGS JOURNAL, *Digital Communications and Networks*, *Wireless Network*, *KSII Transactions on Internet and Information Systems*, and *International Journal of Performability Engineering*. He serves as the GC2022 Co-Chair of Mobile and Wireless Networking Symposium, the General Vice Chair of VTC-2021 Fall, the General Chair of ADHIP 2020, ADHIP 2023, and Mobimedia 2022, the TPC Chair of MOBIMEDIA 2020, ICEICT 2019, and ADHIP 2017, and a TPC Member of GLOBECOM, ICC, ICNC, ICCC, WCSP, and VTC. He had successfully organized several international workshops and symposia with top-ranked IEEE conferences, including INFOCOM, GLOBECOM, DSP, and ICNC.